Evaluation of Computational and Experimental Parameters in RNA Bisulfite Sequencing Analysis and Applications in Brain Development Studies

Zachary Johnson

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

In

Genetics, Bioinformatics, and Computational Biology

Hehuang Xie, Chair
Deborah Good
Liqing Zhang
Alicia Pickrell

August 8, 2022
Blacksburg, Virginia

Keywords: RNA cytosine-5 methylation, RNA bisulfite sequencing, brain development, neuronal stem cells, mitochondria

Copyright© 2022, Zachary Johnson
Evaluation of Computational and Experimental Parameters in RNA Bisulfite Sequencing Analysis and Applications in Brain Development Studies

Zachary Johnson

ABSTRACT

Epitranscriptomics, the study of RNA modifications, has become a hotspot of research over the last decade. Over 170 unique modifications have been discovered with a widespread occurrence in a diverse range of RNAs. 5-methylcytosine, m$^5$C, is an evolutionarily conserved and reversible modification that regulates the stability and export of tRNAs, rRNAs, and mRNAs. m$^5$C has recently been implicated in many biological phenomena including tumorigenesis, embryonic cell expansion and differentiation, brain development, and neuronal functions. While we are just beginning to understand the functions of m$^5$C, a gold standard of m$^5$C detection has yet to be established due to the low signal-to-noise presence of m$^5$C. In this work, we utilize RNA bisulfite sequencing as a transcriptome-wide approach to understand the computational and chemical parameters needed to optimize m$^5$C discovery in the mitochondria and the developing brain.

In Chapter 1, we systematically evaluate four preparation conditions of bisulfite sequencing to identify potential presence of m$^5$C-mRNAs localized to the mitochondria in neuronal stem cells. In tandem, we utilize unique molecular identifiers and a consortium of control template transcripts to evaluate sources of false positive m$^5$C sites that may emerge from sequencing errors, PCR amplification, and the inadequate bisulfite conversion of transcripts. While improvements to mitochondrial transcript bisulfite conversion and false positive filtering were observed, no mitochondrial mRNAs were identified to be methylated, indicating no or very few methylated cytosines in mitochondrial mRNAs and the need for improved chemical methods to detect mitochondrial m$^5$C-mRNAs if any.

In Chapter 2, we employ the computational approaches established in Chapter 1 to survey the m$^5$C landscape of the developing mammalian brain. We discover a general increase in unique m$^5$C sites in mouse whole brain tissue when compared to neuronal cell cultures. Of these sites, we found the post-natal day 0 and 17 brain time points to undergo significant methylation level changes in comparison to the 6-week-old brain. These differentially methylated sites were
significantly enriched for brain development, synaptic development, and transcriptional control gene network pathways.

In Chapter 3, we expand on our findings in Chapter 2 to understand the impact of m\(^5\)C reader FMRP and m\(^5\)C eraser TET1 loss in the mouse post-natal day 17 brain. Among a set of m\(^5\)C sites identified in wildtype or knockout samples, few were differentially methylated after protein ablation, suggesting m\(^5\)C may rely on compensatory enzymes. Using FMRP-RNA pulldown assays to validate FMRP binding positions, we identified *Ralbp1* to be hypermethylated and overexpressed in *Fmr1*-KO brain tissues. RalBP1 is a binding protein responsible for the endocytosis of AMPA receptors, a process critical for neuronal long term depression and brain development.
Evaluation of Computational and Experimental Parameters in RNA Bisulfite Sequencing Analysis and Applications in Brain Development Studies

Zachary Johnson

GENERAL AUDIENCE ABSTRACT

Ribonucleic acid (RNA) is the product of deoxyribonucleic acid (DNA) transcription and the precursor to protein translation. Chemical modifications can be made to the bases of DNA, known as epigenetic modifications, to elicit new functions and responses to the environment. Epitranscriptomics refers to the study of RNA modifications that also serve unique roles and functions depending on the type of modification made. Here, we study the presence of 5-methylcytosine, a methyl group added to the cytosine (C) base of RNA. This modification is found throughout all branches of life and is known to promote the stability and export of many RNA types.

Recently, studies have utilized many techniques including RNA bisulfite sequencing to find links between the presence of m\(^5\)C-RNAs and cancer progression, stem cell development, and brain development. RNA bisulfite sequencing uses chemical applications to convert non-methylated “C”s to the RNA base “U”, while retaining a “C” signature on methylated “C”s. However, due to the extremely low presence of RNA-m\(^5\)C in comparison to DNA-m\(^5\)C, sources of noise make it difficult to identify a true m\(^5\)C signal. Because of this discrepancy, established analytical methods based on DNA biology may not be suitable for RNA analysis.

To address shortcomings in current detection methods of RNA-m\(^5\)C, we performed systematic analysis of 1) different preparation methods for improved m\(^5\)C detection methods and 2) computational approaches for the filtering of false positive m\(^5\)C sites, as described in Chapter 1. To achieve these goals, we expanded the breadth of analytical methods by including unique molecular identifiers and expanding the set of control RNA sequences to better grasp how false positive sites might be introduced into non-methylated sequences. While noticeable improvements were made to control RNA sequence false positive detection, we found that most mitochondrial RNAs did not carry the same m\(^5\)C signatures as RNAs from other sources. Because of this difference, we could not conclude that mitochondrial mRNAs were methylated. Therefore, we
suggest that future studies may need to develop better or alternative methods for the detection of mitochondrial RNA-\(\text{m}^5\text{C}\).

In **Chapters 2 and 3**, we utilize the computational methods developed in **Chapter 1** to understand how m\(\text{5C}\) levels change throughout the development of a mouse’s brain. By investigating the m\(\text{5C}\) profiles of mouse newborn, young child, and juvenile brains, we found significant changes in m\(\text{5C}\) levels specific to certain RNAs. These RNAs are associated with neuronal growth, development, and maturation, which may have implications for m\(\text{5C}\)’s role in cognitive development, intellectual disabilities, and neurodegenerative disorders. To discover if these RNAs could be affected by the absence of m\(\text{5C}\)-specific proteins, we created mice deficient in a protein m\(\text{5C}\) reader, FMRP, and an m\(\text{5C}\) eraser protein, TET1. Interestingly, we did not find a significant difference in mice deficient in the proteins, indicating m\(\text{5C}\) may rely on multiple proteins to serve redundant functions. However, one RNA, *Ralbp1*, was found to be significantly methylated in FMRP deficient models. This RNA is essential for developmental changes in the brain as well as neuronal growth and could be an interesting target for future research.
DEDICATION

To my parents, brothers Brooks and Michael, and partner Natalie.
ACKNOWLEDGMENTS

I would like to acknowledge my advisor, Dr. David Xie, and the opportunity he has given me to succeed in his lab. Collectively, we have published two high quality manuscripts, submitted two more, and have laid the groundwork for two more still. Special thanks to my lab mate Dr. Xiguang Xu for initiating this lab’s journey into RNA methylation, for generating all data I have analyzed in this publication and others, and for the thoughtful discussion he has provided during my time in this lab. In addition, thank you to my current and former lab mates Dr. Christina Pacholec, Dr. Cody Swilley, Yu Lin, Alex Murray, Dr. Amy Smith, Dr. Jaiyi Fan, Razan Alajoleen, Natalie Melville, and Xiaoran Wei as well as the numerous undergraduate students for their help, comradery, and friendship during our shared successes and adversities in this lab. Thank you to my entire committee: Dr. Deborah Good, Dr. Alicia Pickrell, and Dr. Liqing Zhang for their great feedback, discussion, and guidance throughout my academic career at Virginia Tech. I hope that we will cross paths again in the future.

I would also like to praise the incredible support provided by the Graduate School and the Genetics, Bioinformatics, and Computational Biology family, especially Dennie Munson, who I cannot express enough gratitude for. Dennie has helped me navigate through many treacherous positions I have found myself in, and I can easily say that I would not have made it this far without her. Further, thank you to the GBCB department head, Dr. Liwu Li, and the Graduate School Ombudsman, Bryan Hanson, for their invaluable discussions and insights throughout my time at Virginia Tech.

Lastly, I would like to thank my friends and family for their unending support to help me find success in this program: my parents Doug and Harriet Johnson, and my brothers Brooks Johnson and Michael Palmieri. Thank you to my close friends Jeff, Dan, Jonas, Missy, and many others for their empathetic support and patience during my rants. My deepest gratitude goes to my partner Natalie Wickenkamp for her ceaseless support and dedication through both of our graduate school careers. Without her and our four-legged children this journey would not have been possible.
ATTRIBUTIONS

Several colleagues assisted in the design, data generation, and writing of each chapter presented in this dissertation. Brief descriptions of their contributions are listed here:

Hehuang Xie, Ph.D., is the Principal Investigator of the Xie lab, and helped to design, interpret, and write each chapter presented here.

Xiguang Xu, Ph.D., is currently a post-doctoral research in the Xie lab. Dr. Xu generated the data used and helped to design and write each chapter presented here.

Christina Pacholec, DVM., is a Ph.D. student of the Xie lab and helped to interpret findings and edit Chapter 1.

Yu Lin is a Ph.D. student of the Xie lab and assisted in the validation of data analysis and data interpretations in Chapter 2.
# Table of Contents

ABSTRACT ........................................................................................................................................ ii

GENERAL AUDIENCE ABSTRACT ............................................................................................... iii

DEDICATION ......................................................................................................................................... iv

ACKNOWLEDGMENTS ...................................................................................................................... v

ATTRIBUTIONS .................................................................................................................................... vi

Table of Contents .............................................................................................................................. vii

Chapter 1 - Systematic Evaluation of Parameters in RNA Bisulfite Sequencing Data

Generation and Analysis .................................................................................................................. 1

1.1 Abstract ........................................................................................................................................ 2

1.2 Introduction .................................................................................................................................... 3

1.3 Materials and Methods ................................................................................................................. 4

1.4 Results ......................................................................................................................................... 7

1.5 Discussion ..................................................................................................................................... 22

1.6 References ................................................................................................................................... 24

Chapter 2 - Dynamics of RNA-m$^5$C Modification during Brain Development ......................... 28

2.1 Abstract ........................................................................................................................................ 29

2.2 Introduction .................................................................................................................................... 30

2.3 Materials and Methods ................................................................................................................. 31

2.4 Results ......................................................................................................................................... 35

2.5 Discussion ..................................................................................................................................... 48

2.6 References ................................................................................................................................... 50

Chapter 3 - Transcript specific targets of m$^5$C methylation in Fmr1-KO and Tet1-KO models .................................................................................................................................. 54

3.1 Abstract ........................................................................................................................................ 55

3.2 Introduction .................................................................................................................................... 56

3.3 Materials and Methods ................................................................................................................. 57

3.4 Results ......................................................................................................................................... 60

3.5 Discussion ..................................................................................................................................... 68

3.6 References ................................................................................................................................... 69

Chapter 4 - Conclusions and Future Directions .............................................................................. 72
4.1 Future Directions .................................................................................................................. 73
4.2 References............................................................................................................................ 74
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>RNA library constructed in this study</td>
</tr>
<tr>
<td>1-2</td>
<td>Quality score analysis of p-m^5C bases</td>
</tr>
<tr>
<td>1-3</td>
<td>PCR deduplication and bisulfite conversion of ERCC spike-in transcripts</td>
</tr>
<tr>
<td>1-4</td>
<td>Effects of m^5C filtering steps on bisulfite sequencing data analysis</td>
</tr>
<tr>
<td>1-5</td>
<td>Interrogation of bisulfite preparation conditions used in mitochondrial RNA BS-seq libraries</td>
</tr>
<tr>
<td>1-6</td>
<td>Characterization of m^5C sites among cellular compartments of mouse NSCs</td>
</tr>
<tr>
<td>2-1</td>
<td>Experimental design</td>
</tr>
<tr>
<td>2-2</td>
<td>Reproducibility of replicates among RNA BS-seq samples</td>
</tr>
<tr>
<td>2-3</td>
<td>Characteristics of high-confidence m^5C sites identified in NSCs, neuron, and brain samples</td>
</tr>
<tr>
<td>2-4</td>
<td>Methylation dynamics throughout brain development</td>
</tr>
<tr>
<td>2-5</td>
<td>Differentially expressed transcripts among NSCs, neurons, and brain tissues</td>
</tr>
<tr>
<td>2-6</td>
<td>Temporal and spatial expression of differentially methylated transcripts</td>
</tr>
<tr>
<td>3-1</td>
<td>Experimental design</td>
</tr>
<tr>
<td>3-2</td>
<td>WT, Fmr1-KO, and Tet1-KO RNA BS-seq replicate reproducibility</td>
</tr>
<tr>
<td>3-3</td>
<td>Figure 3-3: Methylation level changes in Fmr1-KO and Tet1-KO models</td>
</tr>
<tr>
<td>3-4</td>
<td>Differentially methylated and expressed transcripts in Fmr1-KO and Tet1-KO models</td>
</tr>
</tbody>
</table>
List of Tables

All tables are included in the Supplementary Materials file.
### List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALYREF</td>
<td>Aly/REF export factor</td>
</tr>
<tr>
<td>AMPA</td>
<td>α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor</td>
</tr>
<tr>
<td>BP</td>
<td>Biological process</td>
</tr>
<tr>
<td>BS</td>
<td>Bisulfite sequencing</td>
</tr>
<tr>
<td>C</td>
<td>Cytosine</td>
</tr>
<tr>
<td>CC</td>
<td>Cellular component</td>
</tr>
<tr>
<td>CDS</td>
<td>Coding sequence</td>
</tr>
<tr>
<td>CLIP-seq</td>
<td>Crosslinking immunoprecipitation sequencing</td>
</tr>
<tr>
<td>CPM</td>
<td>Counts per million</td>
</tr>
<tr>
<td>DAPI</td>
<td>4',6-diamidino-2-phenylindole</td>
</tr>
<tr>
<td>DAVID</td>
<td>Database for Annotation, Visualization, and Integrated Discovery</td>
</tr>
<tr>
<td>DMEM</td>
<td>Dulbecco's Modified Eagle Medium</td>
</tr>
<tr>
<td>DMS</td>
<td>Differentially methylated site</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>DSB</td>
<td>Double-strand break</td>
</tr>
<tr>
<td>EGF</td>
<td>Epidermal Growth Factor</td>
</tr>
<tr>
<td>ERCC</td>
<td>External RNA Controls Consortium</td>
</tr>
<tr>
<td>ESR1</td>
<td>Estrogen receptor α</td>
</tr>
<tr>
<td>FDR</td>
<td>False Discover Rate</td>
</tr>
<tr>
<td>FMRP</td>
<td>Fragile X messenger ribonucleic protein</td>
</tr>
<tr>
<td>FXS</td>
<td>Fragile X syndrome</td>
</tr>
<tr>
<td>GFAP</td>
<td>Glial fibrillary acidic protein</td>
</tr>
<tr>
<td>GO</td>
<td>Gene Ontology</td>
</tr>
<tr>
<td>HR</td>
<td>Homologous recombination</td>
</tr>
<tr>
<td>IACUC</td>
<td>Institute Animal Care and Use Committee</td>
</tr>
<tr>
<td>IP</td>
<td>Immunoprecipitation</td>
</tr>
<tr>
<td>KEGG</td>
<td>Kyoto Encyclopedia of Genes and Genomes</td>
</tr>
<tr>
<td>KO</td>
<td>Knockout</td>
</tr>
<tr>
<td>LC-MS</td>
<td>Liquid chromatography mass spec</td>
</tr>
<tr>
<td>LTD</td>
<td>Long term depression</td>
</tr>
<tr>
<td>m5C</td>
<td>5-methylcytosine</td>
</tr>
<tr>
<td>m6A</td>
<td>N6-methyladenosine</td>
</tr>
<tr>
<td>MF</td>
<td>Molecular function</td>
</tr>
<tr>
<td>MT</td>
<td>Mitochondria</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>NGS</td>
<td>Next-generation sequencing</td>
</tr>
<tr>
<td>NIH</td>
<td>National institute for Health</td>
</tr>
<tr>
<td>NSC</td>
<td>Neuronal Stem Cell</td>
</tr>
<tr>
<td>NSUN</td>
<td>NOP2/Sun RNA methyltransferase enzyme family</td>
</tr>
<tr>
<td>PBS</td>
<td>Phosphate buffered serum</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>RAPGEF4</td>
<td>Rap guanine nucleotide exchange factor 4</td>
</tr>
<tr>
<td>RIP-seq</td>
<td>RNA immunoprecipitation sequencing</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribonucleic acid</td>
</tr>
<tr>
<td>RNA BS-seq</td>
<td>RNA bisulfite sequencing</td>
</tr>
<tr>
<td>RNA-seq</td>
<td>RNA sequencing</td>
</tr>
<tr>
<td>RT</td>
<td>Room temperature</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>RT-qPCR</td>
<td>Real-time quantitative polymerase chain reaction</td>
</tr>
<tr>
<td>scRNA-seq</td>
<td>Single-cell RNA sequencing</td>
</tr>
<tr>
<td>SVZ</td>
<td>Subventricular Zone</td>
</tr>
<tr>
<td>TET1</td>
<td>Tet methylcytosine dioxygenase 1</td>
</tr>
<tr>
<td>TIS</td>
<td>Transcription initiation site</td>
</tr>
<tr>
<td>TPM</td>
<td>Transcripts per million</td>
</tr>
<tr>
<td>TRDMT1</td>
<td>TRNA Aspartic Acid Methyltransferase 1</td>
</tr>
<tr>
<td>tSNE</td>
<td>t-distributed stochastic neighbor embedding</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcription start site</td>
</tr>
<tr>
<td>U</td>
<td>Uracil</td>
</tr>
<tr>
<td>UCSC</td>
<td>University of California Santa Cruz</td>
</tr>
<tr>
<td>UMI</td>
<td>Unique Molecular Identifier</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
<tr>
<td>WBP2</td>
<td>WW domain binding protein 2</td>
</tr>
<tr>
<td>WT</td>
<td>Wildtype</td>
</tr>
<tr>
<td>YBX1</td>
<td>Y-box binding protein 1</td>
</tr>
</tbody>
</table>
Chapter 1 - Systematic Evaluation of Parameters in RNA Bisulfite Sequencing 
Data Generation and Analysis

Zachary Johnson\textsuperscript{1, 2$}, Xiguang Xu\textsuperscript{1, 3$}, Christina Pacholec\textsuperscript{1, 3}, Hehuang Xie\textsuperscript{1, 2, 3, 4, 5}$

\textsuperscript{1}Epigenomics and Computational Biology Lab, Fralin Life Sciences Institute, Virginia Tech, Blacksburg, VA 24061, USA
\textsuperscript{2}Genetics, Bioinformatics and Computational Biology Program, Virginia Tech, Blacksburg, VA 24061, USA
\textsuperscript{3}Department of Biomedical Sciences and Pathobiology, Virginia-Maryland College of Veterinary Medicine; Virginia Tech, Blacksburg, VA 24061, USA
\textsuperscript{4}Translational Biology, Medicine and Health Program, Virginia Tech, Blacksburg, VA 24061, USA
\textsuperscript{5}School of Neuroscience, Virginia Tech, Blacksburg, VA 24061, USA

$These authors contributed equally to this work.

* Correspondence: Hehuang Xie, davidxie@vt.edu

Status: Published

https://doi.org/10.1093/nargab/lqac045
1.1 Abstract

The presence of 5-methylcytosine (m$^5$C) in RNA molecules has been known for decades and its importance in regulating RNA metabolism has gradually become appreciated. Despite recent advances made in the functional and mechanistic understanding of RNA m$^5$C modifications, the detection and quantification of methylated RNA remains a challenge. In this study, we isolated the mitochondria of neuronal stem cells to compare four library construction procedures for RNA bisulfite sequencing and implemented an analytical pipeline to assess the key parameters in the process of m$^5$C calling. We found that RNA fragmentation after bisulfite conversion increased the yield significantly, and an additional high temperature treatment improved bisulfite conversion efficiency especially in reads mapped to the mitochondria. Using Unique Molecular Identifiers (UMIs), we observed that PCR favors the amplification of unmethylated templates. The low sequencing quality of bisulfite-converted bases is a major contributor to the methylation artifacts.

In addition, we found that mitochondrial transcripts are frequently resistant to bisulfite conversion. Taken together, this study reveals the various sources of artifacts in RNA bisulfite sequencing data and provides an improved experimental procedure together with analytical methodology.

Keywords: RNA cytosine methylation, RNA bisulfite sequencing, RNA-seq, Mitochondrial transcripts, Mouse neural stem cells
1.2 Introduction

Post-transcriptional modification of RNA molecules plays a fundamental role in the regulation of RNA function and metabolism [1-4]. Among the more than 170 types of RNA modifications that have been identified [5], RNA 5-methylcytosine (m^5C) is one of the most well-known and widely present in transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), and messenger RNAs (mRNAs) [6-8]. RNA m^5C modification in tRNAs, mediated by DNA methyltransferase 2 (DMNT2) and members of the NOP2/Sun RNA methyltransferase enzyme family (NSUN) [6, 9-11], promote tRNA stability and protein synthesis [9]. RNA m^5C modification in rRNAs, introduced by NSUN5, serves as a conserved mechanism in rRNA-mediated translational regulation [12]. Compared to tRNAs and rRNAs, mRNAs carry few m^5C modifications, the functions of which have been better understood in recent years [8, 13-16]. Specifically, the m^5C modification promotes the export of mRNAs from the nucleus to the cytoplasm via the RNA binding protein ALYREF [16], stabilizes mRNAs by facilitating the binding of the m^5C reader protein YBX1 [17-19], and modulates mRNA translation efficiency [20]. Moreover, mRNA m^5C modification is involved in diverse physiological and pathological conditions including facilitating the maternal-to-zygotic transition in early embryos of zebrafish [18], promoting ovarian germ line stem cell development in drosophila [19], and driving the pathogenesis of bladder cancer in humans [17].

Along with advances in high-throughput sequencing, RNA bisulfite sequencing (RNA BS-seq) was developed and widely used for the identification of RNA m^5C modification at single nucleotide resolution [8, 14, 16, 21, 22]. Despite the successful confirmation of m^5C sites in tRNAs [9, 23, 24] and rRNAs [7, 12] using RNA BS-seq, it remains a challenge to obtain reproducible sets of m^5C in mRNAs, even among biological replicates. Currently, a wide range of m^5C sites in the mammalian transcriptome has been reported, ranging from less than 100 to more than 10,000 sites per transcriptome [8, 14-16, 22]. Such a large variation in the number of m^5C sites determined in mRNAs is speculated to be associated with differences in experimental versus computational approaches including inefficient bisulfite conversion, sequencing data-quality controls, methylation calling, and methylation filtering strategies (Table S1) [14-16, 22, 25, 26]. From an experimental aspect, several versions of RNA BS-seq library construction protocols have been published [14-16, 22]. The primary differences in these protocols lie in the timing of RNA fragmentation, the temperature and duration of thermal conditions during bisulfite conversion, and
the usage of ACT or regular random hexamers for first strand cDNA synthesis. Despite an elegant toolkit meRanTk [25] implemented to provide accurate sequence mapping, methylation calling, and high-confidence filtering; the pipelines used to process RNA BS-seq data vary across different research groups.

Recent studies utilizing high-stringency bisulfite conditions, a “C-cutoff” of RNA BS-seq reads, and other statistical techniques have identified hundreds of high-confidence m^5C sites in mRNAs in mouse and human tissues [20, 22, 27]. mRNAs carrying high-confidence m^5C sites were found to be enriched in the mitochondrial gene pathway [22]. Research groups focusing on non-coding RNAs have identified the m^5C modification of mitochondrial tRNAs and one rRNA [11, 28-30], indicating the presence of NSUN2 [29, 31], NSUN3 [32, 33], and NSUN4 [30, 34-36] activity within the mitochondrial complex. Some studies identified methylated mRNAs originating from the mitochondrial genome [14, 16, 22], however, other studies were unable to support this finding [15, 37].

Despite the promising results obtained in recent RNA BS-seq studies, it remains a challenge to select an ideal experimental protocol for library construction and appropriate parameters in the data processing procedure to accurately identify m^5C sites. In this study, we compared four different protocols for RNA BS-seq library construction. RNA samples isolated from the mitochondria of mouse neural stem cells (NSCs) was used as starting materials. The mitochondrial transcriptome was chosen to test preparation protocols to minimize artifacts from multi-mapping and to validate any sites identified in previous studies [11, 29-33, 38]. To provide a robust technical analysis of mitochondrial RNA BS-seq data, Unique Molecular Identifiers (UMI) were introduced to estimate the error rates resulting from PCR and sequencing steps [39], and a stringent analytical pipeline was implemented to assess key parameters in m^5C calling.

### 1.3 Materials and Methods

**Mouse neural stem cell isolation and culture**

Adult mouse neural stem cells (NSCs) were isolated from the subventricular zone (SVZ) of the lateral ventricles as described previously [40]. NSCs were seeded on poly-Ornithine and laminin-coated plates and cultured in DMEM/F12 medium supplemented with 2% B27 supplement, 2
mmol/L L-glutamine, 1x penicillin-streptomycin, 20 ng/ml epidermal growth factor (EGF, PeproTech), 20 ng/ml basic fibroblast growth factor (bFGF, PeproTech).

**Mitochondrial BS-seq library construction**

Mitochondria were isolated from NSCs using a mitochondrial isolation kit (Abcam, ab110171) following the manufacturer's instructions. RNA was extracted from the isolated mitochondria and subjected to DNase digestion. One round of poly(A) selection was performed to enrich mitochondrial molecules. ERCC RNA mixes (Thermo) and unmethylated Xef mRNA were spiked into the samples as external RNA controls. The mitochondrial BS-seq libraries were constructed using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (NEB, E7760S) and bisulfite treatment was performed using the EZ RNA methylation kit (Zymo Research) under four different conditions: A) bisulfite conversion using three cycles of 70°C for 10 min and 64°C for 45 min. After bisulfite conversion, RNA fragmentation and priming was performed by incubation at 94 °C for 8 minutes in 1st strand reaction buffer and 6 bp random primers for 1st strand cDNA synthesis; B) RNA fragmentation was performed before bisulfite conversion by incubation at 90 °C for 50s in 1x RNA fragmentation buffer and quenched by adding 1x stop buffer, then purified by Zymo Research RNA clean and concentrator-5 kit. Then, bisulfite conversion was performed using three cycles of 70°C for 10 min and 64°C for 45 min. After bisulfite conversion, RNA fragmentation was omitted and priming was performed by incubation at 65 °C for 5 minutes in 1st strand reaction buffer and random primers for 1st strand cDNA synthesis; C) RNA fragmentation was performed first, then bisulfite conversion was performed using three cycles of 95°C for 1 min, 70°C for 10 min and 64°C for 45 min. After bisulfite conversion, RNA fragmentation was omitted and priming was performed by incubation at 65 °C for 5 minutes in 1st strand reaction buffer and random primers for 1st strand cDNA synthesis; D) RNA fragmentation was performed first, then bisulfite conversion was performed using three cycles of 95°C for 1 min, 70°C for 10 min, and 64°C for 45 min. After bisulfite conversion, RNA fragmentation was omitted, and priming was performed by incubation at 65 °C for 5 minutes in 1st strand reaction buffer and 6 bp ACT random hexamer primers for 1st strand cDNA synthesis.

**Methylation calling and post-call filtering of BS-seq reads**

5
Raw reads were processed using fastp v0.20 [26] using the parameters (-Q -l 50 --trim_poly_x --poly_x_min_len 10). We then removed low-quality reads and trimmed read ends using the parameters (-q 25 -5 -3 -M 25 -f 6 -t 6). Clean reads were then mapped to the mm10 genome using meRanGh of the meRanTk package [25]. Methylation calling was performed using meRanCall. A p-m^5C site was defined as any C->T variants (or G->A variants in the complementary strand) compared to the converted reference genome. All p-m^5C sites with quality above Q30 and at least 10x (C+T) coverage were called using the parameters (-mBQ 30 -sc 10 -cr 1 -mr 0.00001 -mcov 10). To achieve high-confidence in methylation calling, a “standard filter” was applied to each site: 1) at least 3 variants (i) to be called at a position; 2) the (C+T) coverage (j) to be 20 or greater; 3) the methylation level, defined as i/j, to be at least 0.1. Bisulfite converted reads with multiple cytosines identified were considered as incomplete conversion artifacts [15, 20, 24]. To determine the threshold of cytosines (C-cutoff) identified in a read, we calculated the Gini coefficient following the previously described procedure [22]. After C-cutoff filtering, the p-m^5C sites with “signal/noise” ratios greater than 0.9 [20, 22] and FDR adjusted p-values less than 0.05 [14, 25] were retained. Lastly, RNAfold of the ViennaRNA v2.2.9 software (--maxBPspan 150, -T 70, --MEA 0.1) was used to predict conversion-resistant regions [41]. p-m^5C sites located in these regions were removed. To ensure high-confidence in methylation calling among biological replicates, a methylated site must pass all the filtering steps described above in at least one sample and was present in at least one other replicate after the C-cutoff. Sites were annotated using a custom script and the Ensembl mm10 v79 GTF.

**UMI deduplication and analysis**

UMIs of mitochondrial libraries were grouped and deduplicated using umi-tools [42]. Concordance and discordance rates of ERCC sequences were analyzed using a custom python script. A UMI-group was considered discordant if reads reported different nucleotides at a given variant position.

**RNA-seq library analysis**

RNA-seq libraries were filtered using the same parameters applied to BS-seq libraries and mapped to the reference genome using meRanGh. The expression values for each gene were calculated using featureCounts of the Subread package suite v2.0.0 using default parameters.
**Statistical Analysis**

Statistical analyses were performed using SciPy v1.7 and R v4.1.1. Fisher Exact test was used to determine differentially methylated sites among mitochondrial replicates. Wilcoxon rank-sum was used to compare methylation levels of shared m\(^5\)C sites among RNA BS-seq libraries.

**Availability of data and materials**

Data generated in this study have been submitted to the NCBI Gene Expression Omnibus under accession number GSE190614. Analyses in this study was performed using the R v4.1.1, and Python 3.9.4 packages Biopython v1.78, matplotlib v3.3.4, Seaborn v0.11, and Pysam v0.16. The software package developed in this study is available in GitHub repository (https://github.com/zaustinj33/SysAnalysisRNABS).

1.4 Results

1.4.1 Experimental design and construction of RNA bisulfite sequencing libraries

Considering the RNA bisulfite treatment conditions used in previous studies, we isolated mitochondria from mouse NSC culture and constructed RNA BS-seq libraries with four different conditions (Figure 1-1) to examine the impacts of: 1) the order of RNA fragmentation and bisulfite treatment; 2) the inclusion of a heat denaturation step during bisulfite treatment; and 3) the use of random hexamers containing all four nucleotides vs ACT-only primers for first strand cDNA synthesis. NSCs were chosen in this study since previous reports identified RNA m\(^5\)C methylation plays a critical role in stem cell differentiation [24, 43]. Western blot and RT-qPCR were performed to confirm the successful enrichment of mitochondrial isolation (Figure 1-S1A & 1-S1B). For each condition, RNA-seq and RNA BS-seq libraries were constructed for two biological replicates and sequenced on the HiSeq 4000 platform in 150 bp paired end mode. RNA BS-seq libraries were constructed using four different procedures, which we named MT-A/B/C/D. In these libraries, adaptors carrying UMIs were used to remove PCR duplicates and assess the errors generated during PCR amplification. External RNA Controls Consortium (ERCC) consisting of pre-formulated blends of 92 transcripts were spiked in as unmethylated controls to estimate the bisulfite conversion rate. In addition to the eight RNA BS-seq libraries constructed in this study,
we included an external RNA BS-seq dataset, Huang libraries, generated from mouse muscle tissues [22]. Throughout this study, putative methylated sites (C in mRNA strands or G in the complementary cDNA strands) were denoted as “p-m5C”. We aimed to assess the effects of each analytical step in the pipeline for p-m5C identification and determine the potential sources of p-m5C artifacts.

Figure 1-1. RNA library constructed in this study. Fragmentation timing, bisulfite conversion conditions, and primers used in RNA BS-seq libraries are described as individual conditions. “ACT” denotes the use of ACT primers rather than random primers.

1.4.2 Read pre-processing and the influence of sequencing quality filter on methylation calling

Read pre-processing and cleaning are essential steps in most NGS analyses. These steps are especially critical in RNA BS-seq data processing, as sequencing artifacts heavily influence
downstream analysis due to the extremely low m$^5$C signal. In this study, raw reads were processed using fastp [26] to identify low-quality reads and called bases. First, non-overlapping pair-end reads and reads with lengths shorter than 50 bp after adapter trimming were discarded. For each subsequent step, this criterion was maintained. Second, reads were subjected to polyX trimming with a threshold of a 10-base nucleotide repeat. Two quality filters were applied to remove reads with: 1) an average score less than Q25, and 2) more than 40% of the bases with a Phred33 score less than Q25. Last, we trimmed 6 bp from the 5’ and 3’ ends of both the forward and reverse reads. This was performed to reduce the influence of methylation bias resulting from any residual bases derived from the hexamer primers used in first-strand cDNA synthesis (Figure 1-S2).

We evaluated the sequencing quality of the four types of nucleotides (A, T, C, G) at each step of read pre-processing. For libraries generated in this study, the average Phred score of cytosine in unprocessed reads was 3 points lower than those of the other three kinds of nucleotides, and 6 points lower in the RNA BS-seq dataset generated with Huang libraries (Figure 1-2A). The overall low sequencing quality of Cs in forward reads and Gs in reverse reads is presumably due to the composition of nucleotides in the RNA BS-seq libraries being unbalanced during sequencing. In addition, a significant drop in the Phred score of cytosine occurred starting from the 70th base position, with this trend diminishing after sequence trimming (Figure 1-2A & 1-2B). Such a phenomenon was observed in RNA BS-seq libraries, but not in the regular RNA-seq libraries (Figure 1-S3). Despite the stringent filters employed to remove low quality reads and/or bases in the pre-processing steps, the average Phred scores of p-m$^5$C sites in our BS-seq libraries was 2 points lower than other nucleotides, and 7 points lower in Huang libraries. (Figure 1-2C). In Huang libraries, the quality of p-m$^5$C sites in clean reads was 1 point lower on average than in raw reads due to removal of high-quality p-m$^5$Cs within adaptors. Therefore, additional removal of those p-m$^5$C sites with low quality scores is necessary to minimize false-positive methylation calls resulting from sequencing errors. For this reason, we included an additional Q30 cutoff filter for all p-m$^5$C sites.
Figure 1-2. Quality score analysis of p-m^5C bases. (A-B) Mean Phred score per base sequence of adenosine (blue), thymine (yellow), guanine (green), and cytosine (red) in MT-A replicate 1 and Huang replicate 2 datasets before (A) and after (B) all cleaning steps. (C) Mean base-level Phred scores at each step of the sequence cleaning pipeline (shades of red). p–m^5C sites are “Cs” in Read 1 and “Gs” in Read 2. The p-m^5C bases are depicted in the top figure, and the average of all nucleotides are represented in the bottom figure.

1.4.3 Estimation of the influence of bisulfite conversion rate and PCR error on methylation calling

Using the built-in mapper functions of the meRanTk toolkit [25], all clean reads from both the RNA BS-seq and RNA-seq libraries were mapped to the mm10 reference genome (meRanGh) and transcriptome (meRanT). Sequence reads derived from RNA-seq show a higher percentage of uniquely mapped reads compared to those derived from RNA BS-seq. On average, the percentages of uniquely mapped reads using meRanGh are 52.4% and 44.4% higher than those using meRanT for RNA BS-seq and RNA-seq, respectively. We also examined the mapping efficiencies of the aggregated approach to recover multi-mapped and unmapped reads using meRanGh or meRanT alone. meRanGh alone was able to provide unique mapping rates similar to the combination of
meRanGh and meRanT (Figure 1-S4A). More than 50% of mapped reads were mapped to exonic regions in all analyzed samples (Figure 1-S4B).

Using UMI adaptors and the mapping coordinates, uniquely mapped reads in this study’s libraries were grouped using the ‘group’ command of the umi-tools package [42]. Reads that were mapped to the same genomic coordinate and contained an identical UMI-ID were considered to be PCR amplicons. These PCR amplicons may contain small sequence variations due to PCR error, and so the most prevalent sequence was retained for methylation calling. In this study, all bisulfite converted libraries were subjected to PCR amplification to obtain enough DNA suitable for Illumina sequencing. We found that the cDNA yields of the MT-B/C/D libraries were much lower than that of the MT-A libraries. Thus, 20 cycles of PCR were performed to amplify MT-B/C/D libraries while only 16 cycles were needed for MT-A libraries. Such a difference in the number of PCR cycling across libraries was manifested by UMI-based PCR deduplication. More specifically, less than 20.0% of uniquely mapped reads in MT-A libraries were derived from PCR amplicons. Compared to those of MT-A libraries, PCR duplication rates for MT-B/C/D libraries increased by an average of 57.9% (Table 1-S2). This indicated that in all four conditions tested for RNA BS-seq library construction, RNA fragmentation after bisulfite sequencing (MT-A libraries) is the best in terms of cDNA yield and reducing the need for additional PCR cycles.

Besides PCR deduplication, the UMI-IDs also allowed for the examination of PCR errors within a UMI-group. We focused on reads mapped to ERCC references to determine PCR or sequencing error, which was reported as the discordance rate at each nucleotide position within a UMI group. As mentioned, MT-B/C/D libraries exhibited higher percentages of PCR amplicons than those of MT-A libraries. Consequently, the read depths of UMI groups identified in MT-B/C/D libraries were found to be much larger than those of MT-A libraries (Figure 1-3A). The increased read depth within a UMI group led to a higher probability of a PCR and/or sequencing error. Indeed, compared with MT-A libraries, discordance ratios were found to be higher in MT-B/C/D libraries (Figure 1-3B). Interestingly, discordance ratios were similar for three types of nucleotides (cytosine, guanine, and thymine), but the discordance ratios of adenine were two to six times higher (Figure 1-3B). This is likely due to the high proportion of adenine in mRNA molecules, i.e., shorter poly-A tails not removed by the polyX filter. Importantly, for libraries generated under all four conditions, the discordant rates of p-m^5C sites ranged from 0.1% for MT-A libraries to 0.7% for MT-D libraries. This indicates that PCR and sequencing errors at p-m^5C
sites are very low, even with 20 cycles of PCR amplification in the RNA BS-seq procedure. We further examined the nucleotide ratios at each discordant p-m^5C site and found that C/T was the discordance type most frequently observed (Figure 1-3C). In addition, MT-A libraries had the highest C/T ratio while MT-D libraries had the lowest C/T ratio. This suggests an increase in PCR amplicons enriched for reads carrying thymine but not cytosine.

Since ERCC references were unmethylated spike-in controls, they were ideal for the estimation of bisulfite conversion rate. In other words, any p-m^5C site in reads mapped to ERCC should be an artifact. Over 90% of ERCC reads in all libraries were found to be free of p-m^5C sites. After PCR deduplication, the proportion of reads without m^5C artifacts decreased by 1% for MT-A libraries, but 3% to 6% for MT-B/C/D libraries (Figure 1-3D). As a result, ERCC conversion rates for libraries MT-B/C/D decreased after deduplication (Figure 1-3E). Further examination of the ERCC reads carrying methylation artifacts revealed that the majority of these reads only carried one cytosine while some ERCC reads contained more than twenty cytosines. This suggests that, for some RNA molecules such as ERCC 00002/00096/000130 (Figure 1-S5A), bisulfite conversion reactions may not take place properly due to RNA secondary structure [13, 44]. In addition, PCR deduplication increased the percentages of reads carrying more than twenty cytosines, particularly for MT-B/C/D libraries (Figure 1-3F). This result is consistent with the observation that PCR amplification favors reads with fewer cytosines (Figure 1-3D). Regardless of PCR deduplication, bisulfite conversion rates for two MT-A libraries were higher than 99.7%. However, incomplete bisulfite conversion was observed in the MT-B and MT-D libraries.

Since each ERCC reference was provided with a known concentration, we further examined the influence of the bisulfite sequencing procedure on the abundance of transcripts. For the regular RNA-seq libraries, the read coverages of the ninety-two ERCC references were highly correlated with the concentrations provided by the manufacturer. Similar trends were observed in RNA BS-seq libraries except for two ERCC molecules: ERCC-00004 (7,500 attomoles/ul) and ERCC-00096 (15,000 attomoles/ul). The read coverages of these two ERCCs were significantly below the expected concentrations in all mitochondrial BS-seq libraries (Figure 1-S5B).

To determine the transcriptome-wide effect of the bisulfite sequencing procedure, the expression levels of all mapped transcripts were determined using featureCounts [45]. For RNA BS-seq libraries, the CPM (counts per million) values were determined with and without UMI-deduplication. After deduplication, the MT-B/C/D libraries reported at least a log two-fold
reduction in CPM for 10.2%-12.7% of genes, while less than 1% of genes experienced a change in expression level in MT-A libraries (Figure 1-S6A). We further examined the effect of bisulfite conversion on gene expression values by comparing CPM values of bisulfite converted libraries to non-converted RNA-seq libraries. MT-A RNA BS-seq libraries reported the highest correlation to the RNA-seq control with a Spearman correlation of 0.99, and only 2.6% of transcripts with changes greater than two-fold. In contrast, MT-B/C/D libraries reported over 60% of transcripts with a greater than log two-fold change (Figure 1-S6B). This result suggests that for the majority of genes, expression profiles remain comparable to regular RNA-seq if bisulfite sequencing libraries are constructed using the MT-A condition with 16 cycles of PCR amplification.
Figure 1-3. PCR deduplication and bisulfite conversion of ERCC spike-in transcripts. (A) UMI-duplicate coverage of ERCC transcripts containing p-mC artifacts. UMI-duplicates are defined as reads containing identical UMI barcodes and mapping coordinates. (B) Discordance ratio of each nucleotide (p-mC, T, G, A) among duplicated ERCC-mapped reads as defined above. Discordance rates were measured as the proportion of UMI-groups containing discordant positions over total UMI-groups. Read-positions with the same UMI barcode were considered concordant if the read-position shared complete similarity with all other read-positions in its group. The number of discordant reads for each nucleotide (A, T, C, and G) were calculated if any dissimilarity at a position was observed. The
concordance-discordance ratio is displayed above the chart. (C) Nucleotide frequencies of discordant positions containing p-m5C bases within ERCC-mapped reads. (D) Percent of reads without any p-m5C sites in ERCC reads before and after UMI-deduplication. (E) p-m5C content of ERCC-mapped reads before and after UMI-deduplication. RNA-seq and RNA BS-seq libraries are shown. p-m5C content was quantified and binned accordingly. (F) Conversion rates of individual ERCC IDs before and after UMI-deduplication. Dashed lines represent the mean conversion rate of the library. Mean conversion rates are displayed above each violin plot.

1.4.4 Multi-level filter for highly confident methylation callings

After determination of p-m5C sites in uniquely mapped reads, multi-level filters with various strategies were widely used to achieve highly confident methylation callings (Table 1-S1). For each library, p-m5C sites with at least 10X read coverage were compiled as a starting set. We followed a multi-step filtering procedure (Figure 1-4A) to evaluate the influence of each filtering step on the number of methylation calling (Table 1-S3). The first step was a “Standard filter,” which filtered sites based on the read depth and the frequencies of p-m5C observed for a given p-m5C site. Approximately 30-40% of the p-m5C sites identified in the MT-B/C/D libraries exhibited shallow read depths of less than 20, which may have been due to the loss of coverage from deduplication. In contrast, over 95% of p-m5C sites in MT-A libraries exhibited read depths over 20 (Figure 1-4B). The majority of p-m5C sites, ranging from 62%-85% across libraries, were filtered when the frequencies of p-m5C observed was less than three at a given site (Figure 1-4C).

To determine an appropriate C-cutoff for each library, the Gini coefficient was employed to assess the distribution of incomplete bisulfite conversion events. The C-cutoff is the threshold of cytosine identified in a bisulfite sequencing read that is considered as an incomplete conversion artifact. The Gini coefficient was calculated with the number of sites per gene (Figure 1-S7A) and the number of unique genes (Figure 1-S7B) for each sample. The Gini coefficient decreases when the p-m5C sites are evenly distributed across genes. By increasing C-cutoff stringency, reads which bear the largest proportion of Cs are removed resulting in a smaller Gini coefficient (Figure 1-S7C). For all RNA BS-seq libraries, the majority of reads carrying any candidate methylated base only had one p-m5C site identified (Figure 1-S8). Following a Gini coefficient threshold of 0.15, recommended previously [22], the C-cutoffs of MT-A and MT-B RNA BS-seq datasets were determined to be between 3 and 5. Interestingly, for MT-C/D libraries, the Gini coefficient was below 0.15 when the C-cutoff was set as 15 and 8, respectively (Figure 1-4D). For a given position, the frequencies of p-m5C observed before and after the C-cutoff were used to calculate the signal/noise ratio. p-m5C sites with a signal/noise ratio less than 0.9 were removed due to the
high proportion of poorly converted reads mapped to those sites (Table 1-S3, Figure 1-S9A & 1-S9B).

To delineate the filter effect on methylation calling, 100% was used as the initial number of p-m\textsuperscript{5}C sites for each library. Combining the “standard filter” with the C-cutoff filter resulted in the removal of more than 98% of p-m\textsuperscript{5}C sites in all RNA BS-seq libraries (Figure 1-4E). All p-m\textsuperscript{5}C sites identified in unmethylated ERCC reference transcripts were not able to pass the thresholds of these two filters (Figure 1-S10). Therefore, the combination of “standard filter” with a C-cutoff filter was sufficient to minimize the chance of false-positive methylation callings. Furthermore, RNA secondary structure was predicted using the ViennaRNA package as previously reported [41] with hundreds of p-m\textsuperscript{5}C sites found in regions predicted to be resistant to bisulfite conversion. Finally, the Benjamin-Hochberg procedure of false discovery rate (FDR) correction removed 43%-95% of the remaining p-m\textsuperscript{5}C sites in the MT-B/C/D libraries but did not remove any in libraries with bisulfite conversion rates higher than 99.9%. For datasets generated in this study, the p-m\textsuperscript{5}C sites were retained for downstream analysis if they passed all filters in another technical replicate.
Figure 1-4. Effects of \( m^5C \) filtering steps on bisulfite sequencing data analysis. (A) The summarized \( m^5C \) filtering pipeline after methylation calling. Methylation sites were called using meRanCall. (B) Binned (C+T) coverage values of \( m^5C \) sites in each bisulfite sequencing library. RNA BS-seq library replicates were merged into single libraries for \( m^5C \) calling purposes (C) Binned \( p-m^5C \) coverage per site identified. (D) Gini coefficient was measured after iterative \( C \)-cutoffs were performed for each library. A \( C \)-cutoff describes the criteria for retaining reads with multiple \( p-m^5C \) sites. The Gini coefficients of all bisulfite-converted libraries are displayed as a heatmap. (E) The proportion of remaining \( p-m^5C \) sites after each step of \( p-m^5C \) filtering for bisulfite sequencing libraries. The original reported sites are determined using meRanCall with a 10x (C+T) coverage filter. Blue lines represent RNA BS-seq libraries generated in this study, red-orange lines represent the Huang datasets used.

1.4.5 RNA bisulfite sequencing analysis of mitochondrial mRNAs

A previous study reported high methylation levels of mitochondria-related genes in heart and muscle tissues [22]. The methylation of mitochondrial tRNAs and rRNAs has also been identified [11, 28-33]. However, the methylation of mitochondrial mRNAs remains largely unexplored. Sequencing reads mapped to the mitochondrial genome were visualized on the University of California Santa Cruz (UCSC) genome browser using Huang RNA BS-seq replicate
2 and MT-A as representatives (Figure 1-5C). Abundant aggregation of mapped reads centered on the coding regions of the mitochondrial chromosome were observed for both kinds of libraries. Successful enrichment of mitochondrial mRNA was demonstrated by the RNA-seq that was performed. Using the meRanT mapping tool, 45.9% of reads were mapped to the mitochondrial transcriptome and the remaining reads were mapped to nuclear transcriptomes or spike-in controls (Figure 1-5D). Such an enrichment for mitochondrial transcripts was more prominent in MT-A libraries than MT-B/C/D. This is likely due to performing RNA fragmentation after bisulfite conversion, resulting in a larger proportion of short RNA fragments and a greater loss of RNA template. In particularly, the median length of mt-mRNAs is much shorter than that of mRNAs derived from nuclear genome. However, despite higher proportions of mitochondrial mapped reads in MT-A, only 32.4% of those reads passed the C-cutoff filter, compared to 65.8% in MT-B and 81.1% of reads in MT-C (Figure 1-5E). Thus, the procedure used for bisulfite library construction can lead to a distorted proportion of mitochondrial mRNAs in the entire RNA population. Compared to regular RNA-seq libraries, the mapping rate of the mitochondrial genome was reduced approximately six to ten times in bisulfite sequencing libraries. Thus, an enrichment procedure is recommended for mitochondrial epitranscriptome studies.

Libraries constructed with enriched mitochondrial transcripts allowed us to compare epitranscriptomes derived from mitochondria and nuclear genomes. More than 95% of reads mapped to non-mitochondrial reads contained no p-m\textsuperscript{5}C sites, while the proportion of mitochondrial mapped reads without any p-m\textsuperscript{5}C varied from 28% (MT-A) to 79% (MT-C) (Figure 1-5F). While the number of p-m\textsuperscript{5}C was low in the majority of reads mapped to the nuclear genome, a substantial portion of reads mapped to the mitochondrial genome carried more than twenty p-m\textsuperscript{5}C. The non-converted reads bearing more than twenty p-m\textsuperscript{5}C were found to be enriched in mitochondrial coding regions (Figure 1-S11). This suggests that those reads did not result from mitochondria genomic DNA contamination but rather were derived from transcripts resistant to bisulfite conversion, presumably due to intramolecular RNA secondary structure. The percentage of non-converted reads was lowest in libraries generated with the MT-C condition (Figure 1-5F). This suggests that RNA fragmentation after bisulfite conversion in combination with a high temperature bisulfite conversion step may be the most suitable for generation of RNA BS-seq data.

Libraries generated in this study were constructed with an equal aliquot from the same pool of RNAs, which allowed us to examine the influence of the four experimental procedures on
methylation data generation. For pair-wise comparisons, we identified the p-\(\text{m}^{\text{5}}\text{C}\) sites shared in libraries generated with two different conditions (Figure 1-S12A). The methylation level correlations were found to have a Spearman coefficient above 0.75 (Figure 1-S12B). We further performed differential methylation analysis and identified 7, 1, and 0 differentially methylated sites (DMSs) in the pair-wise comparisons of MT-A vs MT-B, MT-B vs MT-C, and MT-C vs MT-D, respectively. All DMSs were removed from the list of high-confidence sites. The use of random primers during 1st strand cDNA synthesis has commonly been used in RNA-bisulfite studies, while ACT primers have been suggested to avoid reverse transcription of inefficiently deaminated RNA templates [16, 46]. In this study, we did not observe a significant advantage of using ACT primers.
Figure 1-5. Interrogation of bisulfite preparation conditions used in mitochondrial RNA BS-seq libraries. (A) Read pile-up visualization of a representative Huang and Mitochondrial RNA BS-seq library MT-A using the UCSC genome browser. Peaks are scaled according to max peak height for each library. GENCODE gene annotations are displayed below. (B) Proportion of reads mapped to the mitochondrial chromosome using meRanGh (genome...
We further compared the methylation profiles of RNAs obtained with four different conditions (Table 1-S4). Highly confident m$^5$C sites were defined as Ensembl-annotated p-m$^5$C sites which passed all filtering criteria in at least one replicate and contained at least one m$^5$C count and 10x read coverage after the C-cutoff in another replicate. Using the above criteria, 77 and 684 sites were identified to be m$^5$C sites with high confidence in this study and the Huang dataset respectively (Table 1-S5). Library MT-C reported a m$^5$C site per mapped read rate comparable to Huang and MT-A libraries despite containing ~40 million fewer reads (83.2% fewer) (Figure 1-6A & 1-6B). Replicates from the Huang study reported 61.0% of high-quality sites present in at least two replicates, and 37.7% of sites were present in all four replicates (Figure 1-S13A & 1-S13B). Of high-confidence sites identified in this study’s libraries, 59.7% were also identified in Huang libraries, suggesting some m$^5$C sites may be unique to NSCs (Figure 1-6C).

As reported previously [20, 22, 47], a “GGG” motif was identified downstream of the m$^5$C sites of high confidence (Figure 1-6D). Interestingly, we found that the m-bias filter was able to remove a strong 5’GGG motif upstream of p-m$^5$C sites (Figure 1-S14), which was suggested to be an indication of false positive sites [47]. The difference in methylation levels of high-confidence m$^5$C sites was insignificant (Wilcoxon rank-sum, p > 0.05) (Figure 1-6E). Analysis of m$^5$C distribution on mRNA transcripts was calculated as previously reported [14, 20, 22]. Analysis revealed sites biased to the 5’UTR of mRNAs, with the lowest density in the 3’UTR (Figure 1-6E). Our characterization of high-confidence m$^5$C sites in this study reveals features consistent with previously established reports [20, 22, 47], namely the down-stream “GGG” motif and the enrichment near the transcription starting sites of mRNA transcripts. In mitochondria 12S ribosomal RNA (MT 911, mt-Rnr1), one heavily methylated p-m$^5$C site was found to have a methylation level above 80% in all four conditions and in the published Huang dataset [22]. However, we were not able to consistently identify any p-m$^5$C sites with high confidence on mitochondrial mRNAs (Table 1-S6).
Figure 1-6. Characterization of m^5C sites among cellular compartments of mouse NSCs. (A) Average mapped read count for bisulfite-converted libraries. Bars represent standard deviation. (B) Methylation calling efficiencies of m^5C sites per 10^7 mapped reads. Libraries constructed in this study and Huang replicates are shown in blue and red, respectively. (C) Overlap of high-confidence m^5C sites among libraries generated in this study and Huang muscle tissue datasets. (D) Sequence logo surrounding the high-confidence m^5C sites. (E) Methylation level of the high-confidence m^5C sites. Significance was tested using Wilcoxon Rank-Sum. (F) Distribution of m^5C sites across binned mRNA transcripts. 5'UTR and 3' UTR positions are indicated by dashed lines at bins 5 and 18, respectively.

1.5 Discussion

Substantial differences in the prevalence and magnitude of mRNA methylation reported call into question whether the best practice of RNA BS-seq data generation and analysis has been achieved [22, 47]. In this study, we examined the impact of key parameters in both experimental and computational procedures on the detection of RNA cytosine methylation.

Using the established RNA bisulfite analysis pipeline, RNA BS-seq data was analyzed in a systematic fashion. We observed that the procedure for bisulfite library construction reduced the proportion of sequence reads mapped to the mitochondrial genome. Compared with transcripts derived from the nuclear genome, the overall bisulfite conversion rate of mitochondrial transcripts was poor. More specifically, after bisulfite conversion, a substantial percentage of mitochondrial
transcripts had over twenty cytosines. This may be due to the intramolecular secondary RNA structure within mitochondrial transcripts. Although no m\textsuperscript{5}C sites on mitochondrial mRNAs could be determined with high confidence for mouse neural stem cells, we confirmed a highly methylated cytosine on mitochondrial rRNA as previously reported [22, 30]. However, our approach is limited to poly-A selected mitochondrial transcripts which may not cover the entirety of the mitochondrial transcriptome, leaving the possibility of methylated mitochondrial transcripts without a poly-A tail. In addition, some association has been made between truncated mitochondrial transcripts and poly-adenylation, suggesting poly-A selection may be biased towards degraded RNA [48].

Previous studies have conflicting viewpoints regarding performing RNA fragmentation before or after bisulfite conversion [16, 22]. We found that RNA fragmentation performed after bisulfite conversion (condition MT-A) significantly improved the yield of the cDNA library, compared with MT-B/C/D conditions. Utilizing UMIIs, we observed that the PCR error rate positively correlates with the number of PCR cycles and PCR favors unmethylated templates. Such a bias in PCR amplification of sequences carrying thymidine vs cytosine may lead to an underestimation of the methylation level. In addition, the inclusion of a high-temperature treatment helps to reduce the proportion of unconverted reads originating from the mitochondrial genome. Altogether, our study recommends the following procedures for RNA bisulfite sequencing study: 1) perform RNA fragmentation after bisulfite conversion; 2) include a high-temperature denaturation step in bisulfite treatment cycling; and 3) include a UMI-deduplication strategy for low-input RNA samples or amplify the library with a low number (less than 16) of PCR cycling.

One important characteristic of RNA BS-seq data is the low Phred scores of p-m\textsuperscript{5}C sites. The stringent filters employed to remove low quality reads and/or bases in the pre-processing steps help but cannot fully compensate the difference in sequencing quality between the p-m\textsuperscript{5}C sites and the other three kinds of nucleotides. A previous study indicated that an upstream “GGG” motif was frequently associated with false positive sites [47]. We found that the m-bias filter was able to remove sites with such a motif. In addition, all false positive p-m\textsuperscript{5}C sites in the ERCC reference controls were removed when the C-cutoff filter was applied together with the “Standard filter”. Therefore, our study supports the following parameters/steps in methylation calling: 1) an additional quality filter with Q30 as a cutoff for all p-m\textsuperscript{5}C sites; 2) a stringent m-bias correction; and 3) a combination of a “Standard filter” with the C-cutoff filter. In summary, our study conducted a systematic evaluation of parameters used in RNA bisulfite sequencing and may shed
new light on RNA methylation data generation and analysis. Further improvement may be achieved with improved characterization of false-positive sites [47], alternative deamination techniques [49], and advance computational modeling for m\(^5\)C calling [50].

Acknowledgements
This work was supported by NIH grant ES031521, NS094574, MH120498, NSF1922428, the Center for One Health Research at the Virginia-Maryland College of Veterinary Medicine and The Edward Via College of Osteopathic Medicine, and the Fralin Life Sciences Institute faculty development fund for H.X. We recognize The Center for Engineered Health, the Virginia-Maryland College of Veterinary Medicine at Virginia Tech, and VT’s Open Access Subvention Fund. We thank Dr. Janet Webster for English language editing.

Contributions
X. X. and H. X. conceived and designed the study; X. X. cultured mouse NSCs, performed mitochondrial RNA isolation, and constructed libraries for RNA-seq and RNA bisulfite sequencing; Z. J. performed the bioinformatic analyses; C. P., X. X., and Z.J. contributed to data interpretation and presentation; H. X., X. X., and Z. J. wrote the manuscript. All authors discussed the results and edited the manuscript.

Competing financial interests
The authors declare no competing financial interests.

1.6 References


Chapter 2 - Dynamics of RNA-m\textsuperscript{5}C Modification during Brain Development

Zachary Johnson\textsuperscript{1,2,8}, Xiguang Xu\textsuperscript{1,3,4,8}, Yu Lin\textsuperscript{1,4}, Hehuang Xie\textsuperscript{1,2,3,4,5,6*}

1. Epigenomics and Computational Biology Lab, Fralin Life Sciences Institute, Virginia Tech, Blacksburg, VA 24061, USA
2. Genetics, Bioinformatics and Computational Biology program, Virginia Tech, Blacksburg, VA 24061, USA
3. Department of Biological Sciences, College of Science, Virginia Tech, Blacksburg, VA 24061, USA
4. Department of Biomedical Sciences and Pathobiology, Virginia-Maryland College of Veterinary Medicine, Virginia Tech, Blacksburg, VA 24061, USA
5. Translational Biology, Medicine, and Health program, Virginia Tech, Blacksburg, VA 24061, USA
6. School of Neuroscience, Virginia Tech, Blacksburg, VA 24061, USA

\textsuperscript{8}These authors contributed equally to this work.

*Corresponding author: Email: davidxie@vt.edu

Status: Submitted
2.1 Abstract

Post-transcriptional RNA modifications have been recognized as key regulators of neuronal differentiation and synapse development in the mammalian brain. While distinct sets of 5-methylcytosine (m$^5$C) modified mRNAs have been detected in neuronal stem cells and brain tissues, no study has been performed to characterize methylated mRNA profiles in the developing brain. Here, together with regular RNA-seq, we performed transcriptome-wide bisulfite sequencing to compare RNA cytosine methylation patterns in neural stem cells (NSCs), cortical neuron cultures, and brain tissues at three postnatal stages. Among 501 m$^5$C sites identified, approximately 6% are consistently methylated across all five conditions. Compared to m$^5$C sites identified in NSCs, 96% of them were hypermethylated in neurons and enriched for genes involved in positive transcriptional regulation and axon extension. In addition, brains at the early postnatal stage demonstrated substantial changes in both RNA methylation and gene expression of RNA methylation readers, writers, and erasers. Furthermore, differentially methylated transcripts were significantly enriched for genes regulating synaptic plasticity. Altogether, this study provides a brain epitranscriptomic dataset as a rich new resource and lays the foundation for further investigations into the role of RNA methylation during brain development.

Keywords: Brain development, neuron, neural stem cell, RNA methylation, RNA bisulfite sequencing, RNA-seq
2.2 Introduction

Over the past decade, epitranscriptomics has emerged as a new field to study the post-transcriptional modifications of RNA bases on a transcriptome-wide scale [1]. Among the over 170 kinds of RNA modifications, cytosine methylation (m\(^5\)C) has gradually been recognized as an important form regulating RNA metabolism. The m\(^5\)C modification on tRNAs regulates tRNA stability and protein translation [2], while cytosine methylation on rRNAs coordinates mitochondrial assembly and translation activity [3, 4]. Despite their lower abundance, maternally derived m\(^5\)C-mRNAs are essential for embryonic development during the maternal-to-zygotic transition [5] and ablation of maternal m\(^5\)C-mRNAs results in developmental delay and embryonic defects [6]. Cytosine methylation on mammalian mRNAs is primarily catalyzed by NOP2/Sun RNA Methyltransferase 2 (NSUN2) [7, 8] and is removed via m\(^5\)C oxidation by the Ten Eleven Translocation enzymes [9, 10]. At DNA damage sites, the RNA methyltransferase, tRNA Aspartic Acid Methyltransferase 1 (TRDMT1), is recruited to introduce m\(^5\)C to mRNA in the form of DNA:RNA hybrids [11]. The methylated RNA in such hybrids was recognized by the m\(^5\)C readers RAD51/52. Another m\(^5\)C binding protein, Aly/REF Export Factor (ALYREF), mediates the exportation of methylated mRNAs from the nucleus to the cytoplasm [7]. mRNAs carrying m\(^5\)C can also be recognized by the Y-box binding protein 1 (YBX1), which recruits the poly(A) binding protein cytoplasmic 1a (PABPC1A) to achieve stabilization of methylated mRNAs [12].

Accumulating literature emphasizes the important roles of RNA post-transcriptional modifications in brain development and function. In neuronal stems cells, the loss of RNA methyltransferase NSUN2 results in decreased m\(^5\)C-tRNA levels, leading to aggregation of tRNA fragments, activation of a cellular stress response, and impairment of neuronal stem cell differentiation [13, 14]. In mouse models, NSUN2 loss inhibits the neurogenesis of upper-layer cortical neurons inducing microcephaly and motor defects [13, 15]. Later, targeted knockout of NSUN2 in the mouse prefrontal cortex was shown to have reduced levels of m\(^5\)C-tRNA and resulted in the disruption of neuronal synaptic signaling patterns and behavior [16]. Although the influence of NSUN2 loss on mRNA methylation remains unclear, methylated mRNAs in brain tissue were found to be enriched for genes involved in ion transport and synapse function [17]. In neuronal oxidative stress models, an elevation in mRNA methylation level was associated with the stress response and regulation of apoptosis [18]. In glioblastoma cells, methylated transcripts mediated by the m\(^5\)C writer NSUN6 were linked to the regulation of transcriptional and
translational processes after alkylation treatment [19]. These findings indicate that m$^5$C-modified RNAs may influence brain cell differentiation and function. However, no previous study has attempted to characterize m$^5$C dynamics during mammalian brain development.

Brain development is driven by drastic transcriptome changes accompanied by an increase in cellular diversity and neuronal connectivity [20]. The expansion and migration of neuronal lineages occur at the pre-natal stage while post-natal brain development is characterized by active gliogenesis and rapid neuronal diversification [21, 22]. Around postnatal day 17 (P17), the mouse brain enters a critical period of cortical plasticity that shapes neuronal circuits in response to early life experiences [23]. In this study, we performed RNA BS-seq to survey the landscape of mRNAs with methylated cytosines in the mouse brain at critical developmental time points, as well as neural stem cells and E16.5 cortical neuron cultures (Figure 2-1A). In combination with bulk RNA-seq and single-cell RNA-seq data, we aimed to explore the links between the patterns of mRNA cytosine methylation and gene expression during brain development.

2.3 Materials and Methods

Mice
C57BL/6 mice were maintained and bred in a 12-hour light/dark cycle under standard pathogen-free conditions. Mouse brains were harvested at postnatal day 0, 17, and 6wk for total RNA extraction. Adult female mice were used for setting up timed pregnancy. Embryos were timed by checking virginal plugs daily in the morning. Positive plugs were designated as E0.5. The experiments were approved by the Institutional Animal Care and Use Committee (IACUC) of Virginia Tech.

Mouse neural stem cell (NSC) culture
Mouse NSCs were isolated from the subventricular zone (SVZ) of the lateral ventricles as described previously [24]. NSCs were seeded on poly-ornithine and laminin-coated plates. The culture media was prepared by mixing Dulbecco's Modified Eagle Medium (DMEM) and Ham’s F12 media at 1:1 volume ratio, supplemented with 2% B27, 2 mmol/L L-glutamine, 1x penicillin-streptomycin, 20 ng/ml epidermal growth factor (EGF, PeproTech), and 20 ng/ml basic fibroblast growth factor (bFGF, PeproTech).
Primary mouse cortical neuronal culture
Primary mouse cortical neurons were prepared as previously described [25] with some modifications. Briefly, C57BL/6 E16.5 mouse embryos were micro-dissected for cortex tissues and the cortex tissues were dissociated into single-cell suspensions using the neural tissue dissociation kit (Cat# 130-092-628) according to manufacturer’s instructions. After dissociation, neuronal cells were filtered through a 70-μm strainer (Falcon) and spun at 300g for 10 min. The cell pellet was resuspended in neuronal culture medium (Neurobasal medium containing 2% B27 supplement (Invitrogen), 1% Glutamax (ThermoFisher), and 1% penicillin-streptomycin (ThermoFisher) and seeded on laminin and poly-ornithine coated 10-cm dishes. Neurons were grown in vitro for 7 days with fresh medium changed on Day 3 and Day 6.

Immunostaining
Immunostaining was performed as previously described [26]. Briefly, NSCs or E16.5 mouse cortical neurons were seeded onto an 8-well chamber. The neurons were fixed with 4% paraformaldehyde in phosphate buffered saline (PBS) for 15 min and permeabilized with 0.2% TritonX-100 in PBS for 10 min. After being blocked with 5% Normal Goat Serum (Thermo Fisher) at room temperature (RT) for 1 h, the cells were incubated with mouse anti-Nestin antibody (Millipore, MAB353) and rabbit anti-Sox2 antibody (Abcam, ab97959) for NSCs, or with mouse anti-Tuj1 antibody (Biolegend, 801201) and rabbit anti-GFAP antibody (Sigma, HPA056030) for E16.5 cortical neurons at 4 °C overnight. Then, the cells were incubated with Cy3 conjugated anti-rabbit IgG (A10520, Invitrogen) and Alexa Fluor 488 conjugated anti-mouse IgG (A10680, Invitrogen) secondary antibodies at RT in darkness for 1 h. After washing 3 × 5 min with 1×PBS, cells were mounted with DAPI-Fluoromount-G™ Clear Mounting Media (Southern Biotech, 010020). Fluorescent images were acquired using a confocal microscope.

RNA BS-seq library construction
RNA bisulfite conversion was performed as previously described [27] with minor modifications. Poly(A) RNA was first mixed with spiked-in Xef1 unmethylated RNA at a ratio of 0.5%. The spiked-in unmethylated mRNA was transcribed from the pTRI-Xef plasmid supplied by the MEGAscript™ T7 Transcription Kit (Invitrogen) according to manufacturer’s instructions. RNA bisulfite conversion was performed with an initial denaturation at 95°C for 1min, followed by
three cycles of 70°C for 10min and 64°C for 45min using the EZ RNA methylation Kit (Zymo Research). The bisulfite converted RNA was subjected to the stranded RNA-seq library construction procedure using the TruSeq Stranded mRNA Library Preparation Kit (Illumina). We modified the procedure to skip the RNA fragmentation step and supply both random and ACT random hexamers during the first strand cDNA synthesis.

**RNA BS-seq data analysis**

RNA BS-seq data analysis was performed as previously described [27]. Raw sequencing reads were trimmed at the 5’ and 3’ ends by 6 bp to account for methylation bias and then filtered for low quality bases and adaptor sequences. After the removal of reads with short lengths, clean reads were mapped to the mm10 genome (Ensembl v.79) using meRanGh [28]. To exclude partially unconverted reads, mapped reads were further filtered by removing reads with more than 3 “C”s (“G”s on the cDNA strand). Methylation calling was performed using meRanCall and subjected to a series of filters to reduce false positive signals [27]. In addition to the thresholds applied on the methylation level (≥ 0.1) and read coverage (≥ 20) of m\(^5\)C sites, additional filters were adopted including signal-to-noise ratio and the maximum number of m\(^5\)C sites detected in a read to reduce false positive methylation calling [27]. The C-coverage threshold ranging from 6 to 10 was determined using Gini Index [8]. High-confidence sites were defined as the m\(^5\)C sites passed all filters in both biological replicates.

**Differential methylation analysis**

Differential methylation analysis was performed using Fisher’s exact test on high-confidence sites present in at least one sample. Benjamini-Hochberg p-value correction was used to correct for multiple comparisons, and sites with p-adjusted < 0.05 were reported.

**RNA-seq library construction**

Stranded RNA-seq libraries were constructed using the TruSeq Stranded mRNA Library Preparation Kit (Illumina) following manufacturer’s instructions. Briefly, after two rounds of poly(A) selection, the mRNA samples were fragmented and primed to synthesize first strand cDNA, followed by synthesis of the second strand cDNA. After Ampure XP bead purification, dA tailing was performed, and indexed adapters were ligated to both ends of the double-stranded
cDNA. Adapter-ligated DNA fragments were enriched by PCR amplification for 12 cycles. After Ampure XP bead purification, the PCR products were size-selected with a range from 350bp to 550bp on 2% dye-free agarose gel using the pippin recovery system (Sage Science). The recovered libraries were sequenced on a HiSeq 4000 platform in the 150bp paired end mode (Illumina).

**RNA-seq data analysis**

Trim Galore (version 0.6.5) was used to filter short reads, low quality reads, and trim adapter sequences from raw reads (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Clean reads were mapped to the mm10 genome and expression quantified using STAR (version 2.7.3a) [29]. Differentially expressed genes were identified using DESeq2 [30] using a fold change greater than 1.5 and an adjusted p-value of 0.01 as cutoffs.

**Gene Ontology (GO) analysis**

GO analysis was performed using DAVID (Database for Annotation, Visualization, and Integrated Discovery) [31]. The GO Direct terms, Biological Process (BP), cellular component (CC), and molecular function (MF) were identified. Significantly enriched terms were identified using Benjamini-Hochberg p-adjusted < 0.05.

**Clustering Analysis**

The z-score of bulk RNA-seq expression of transcript per million (TPM) and differentially methylated site methylation levels were used as input for clustering analysis. Clustering analysis was performed with the Python package sklearn’s Agglomerative Clustering. The ‘complete’ method was used with the number of clusters set to 5 for RNA-seq and RNA BS-seq datasets. Error bars represent the standard deviation of the z-score.

**Availability of data and software**

Data generated in this study were submitted to the NCBI Gene Expression Omnibus under accession number GSE207092. Analyses in this study was performed using the R v4.1.1, and Python 3.9.4 packages Biopython v1.78, matplotlib v3.3.4, Seaborn v0.11, and Pysam v0.16.
The software package developed in this study is available in the GitHub repository (https://github.com/zaustinj33/BrainDev).

2.4 Results

2.4.1 Identification of high-confidence m\textsuperscript{5}C sites in mouse NSCs, neurons, and brain tissues

To obtain RNA expression and m\textsuperscript{5}C epitranscriptome profiles during brain development, we performed RNA-seq and RNA BS-seq for NSCs (Figure 2-1B), neurons in culture (Figure 2-1C), and brain tissues isolated from postnatal day 0 (P0), postnatal day 17 (P17), and 6-week-old (6W) mice. Two biological replicates were generated for each condition with a total of 10 libraries obtained for RNA-seq and RNA BS-seq. Approximately 20 million paired-end reads were generated for each RNA-seq library and aligned to the mouse reference genome (mm10). For these libraries, the average percentage of sequences uniquely mapped to the reference was around 92%. Since methylation calling requires a much higher read depth compared with those of RNA-seq libraries, five times more sequences were generated for RNA BS-seq libraries (~100 million pair-end reads for each library). Sequencing reads derived from RNA BS-seq libraries were processed following the procedure described in our recent study [27]. In brief, bisulfite sequencing reads were first trimmed to remove adaptors and bases with low quality scores. After quality control, reads were mapped to the C2T converted mm10 genome. Around 85% of bisulfite sequencing reads were uniquely aligned (Table 2-1). Methylation calling was performed using meRanCall [28] and putative m\textsuperscript{5}C sites were subjected to a series of filters to reduce false positive signals [27].
Figure 2-1. Experimental design. (A) Overall experimental design of this study. (B) Mouse neural stem cells (NSCs) were double stained with the neural progenitor markers Nestin (cytoplasmic, green) and Sox2 (nuclear, red); nuclei were counterstained with DAPI (blue). Scale bar: 50μm. (C) E16.5 mouse cortical neuronal culture was double stained with the neuronal marker Tuj1 (green) and glial marker GFAP (red); nuclei were counterstained with DAPI (blue). Scale bar: 50μm.

A total of 2,259 m^5C sites were identified from the ten RNA BS-seq libraries. Among these m^5C sites, 1,758 sites passed all filters in one biological replicate but failed in the other; thus, these were denoted as “low-confidence” m^5C sites. The remaining 501 m^5C sites shared by both biological replicates were considered to be “high-confidence” in this study (Figure 2-2A). The methylation levels of the high-confidence m^5C sites were highly correlated between the biological replicates and had Spearman R values between 0.82 and 0.92 (Figure 2-2B). Notably, 76.4% of
these sites had a methylation level lower than 0.3 (Figure S1A). This result was consistent with previous reports that only a small percentage of RNA copies are methylated for a given gene [8, 27, 32].

To explore the factors contributing to the level of confidence in methylation calling, we first compared methylation levels and read coverages of the high-confidence and low-confidence sites. Across five conditions, the high-confidence sites showed average methylation levels in the range of 0.18-0.21, while the low-confidence sites had average methylation levels in the range of 0.06-0.12 (Figure 2-2C). The average read coverages for the high-confidence sites were between 109-227 reads per site, while the average for the low-confidence sites was only 24-48 reads per site (Figure 2-S2A & 2-S2B). We next examined the relationship between read coverage and methylation level (Figure 2-S2C). A very weak positive correlation was observed between the methylation level and the read coverage for both high-confidence (Spearman R < 0.19) and low-confidence sites (Spearman R < 0.08). Altogether, the high-confidence m\textsuperscript{5}C sites tended to be found with a higher methylation level and more read coverage to enable them to survive all the stringent filters in both biological replicates.

Previous studies suggested that the process of library construction, the bisulfite conversion step in particular, could have a significant impact on methylation calling [8, 27, 32]. In this study, we generated both RNA-seq and RNA BS-seq data for the same pool of RNA samples. This enabled us to perform correlation analyses for the read coverage of each methylated transcript in the paired libraries for RNA-seq and RNA BS-seq. Interestingly, compared with the low-confidence sites, the high-confidence sites showed higher correlations in read coverage with or without bisulfite conversion, except for P17 brain samples in which the two correlations are similar (Figure 2-S3A). We further examined the average length of methylated transcripts and the distribution of m\textsuperscript{5}C sites in transcripts. No significant difference in transcript length was detected for transcripts containing either high or low-confidence sites (Figure 2-S3B). The distribution of high-confidence sites was generally biased towards the 3’UTR of transcripts, in particular for the P17 brain samples (Figure 2-S3C). This result indicates that low-confidence sites tend to reside in a transcript sensitive to bisulfite treatment, but such sensitivity may not have a link with transcript length.
Figure 2.2. Reproducibility of replicates among RNA BS-seq samples. (A) Overlap of m^5C sites identified between biological replicates. The red circle indicates replicate 1 and the blue circle indicates replicate 2. (B) Methylation correlation (Spearman’s R) of the overlapping m5C sites between biological replicates. Color code indicates the density of sites, ranging from red (high density) to blue (low density). (C) Distribution of methylation levels for the overlapping and non-overlapping m5C sites identified between biological replicates.

2.4.2 High-confidence m^5C sites differentially methylated in mouse NSCs, neurons, and brain tissues

We next focused on the 501 high-confidence sites to examine their sequence features and distribution across five conditions. Two common features of m^5C sites on mRNAs have been reported: 1) they often localize upstream of the translation initiation sites (TISs), and 2) a “GGG” motif is frequently observed downstream from the m^5C sites[8, 27, 32, 33]. To determine the distribution of high-confidence m^5C sites along the transcripts, we binned the lengths of methylated transcripts into the 5’-untranslated region (5’UTR: bins 1-5), coding sequence (CDS: bins 6-28), and 3’-untranslated region (3’UTR: bins 19-46). Two density peaks surrounding TISs and an additional strong peak right before the termination of the coding region were observed for
m^5C sites in all five conditions (Figure 2-S3A). A downstream “GGG” motif was determined for the high-confidence m^5C sites in all five conditions (Figure 2-S3B), which may serve as potential targets of the NSUN2 enzyme [8].

Out of 501 high-confidence m^5C sites, a total of 31 sites were identified as methylated in all conditions and 253 sites (50.8%) were shared by at least two conditions (Figure 2-3A). Such a distribution indicates that at least some mRNAs are consistently methylated across multiple time points throughout brain development. Only 82 m^5C sites were identified in NSCs while the libraries derived from the three-stage brain tissues reported a similar number of high-confidence sites ranging from 222 to 269. In addition, m^5C sites in undifferentiated NSCs were largely overlapped with those in neurons and brain tissues. To determine the effect of sequencing depth on methylation calling, we calculated the yield of m^5C sites after normalizing to the number of total mapped reads for each library (Table 2-S1). Despite more reads generated for NSC libraries, both NSC replicates yielded the fewest m^5C sites regardless of high-confidence or low-confidence. The methylation level of high-confidence sites reported in the P17 brain were significantly lower than sites reported in the P0 brain (Figure 2-3B). These results suggested that the diversity of mRNA methylation may increase during brain cell specification.

Since brain development is accompanied with substantial changes in gene expression, we selected a set of 50 high-confidence m^5C sites that possessed at least 20x read coverage in all conditions. According to their average methylation levels, five clusters were identified with unique methylation patterns (Figure 2-3C). The m^5C sites in three clusters had relatively consistent methylation levels throughout all conditions (at high, low, and medium methylation levels, respectively). Such stable methylation patterns suggest that cells from neural lineage may share a regulatory mechanism to control the methylation of a small set of transcripts. Cluster C0 and C4 showed high methylation levels in neurons and P0 brain samples, respectively (Figure 2-3D). We further determined their associated transcripts for GO enrichment analysis. Interestingly, transcripts highly methylated in neuron culture (cluster C0) were associated with transcription, chromatin organization, and stem cell maintenance (Figure 2-3E). For instance, genes including Foxo3, RelA, and Rptor are important mediators of neuronal cell functions including reprogramming and differentiation [34] and synaptic formation [35].
Figure 2-3. Characteristics of high-confidence m^5C sites identified in NSCs, neuron, and brain samples. (A) Methylation distribution of the high-confidence m^5C sites. Wilcoxon rank-sum test was performed to assess significance (*p<0.05). (B) High-confidence m^5C sites shared among conditions. (C) Clustering of m^5C sites according to methylation changes. (D) Average methylation levels of clusters identified in Figure 3C. Bars indicate standard deviation. (E) Gene ontology analysis of methylated transcripts.

<table>
<thead>
<tr>
<th>Term [Biological Process]</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1 [C1, C2, C3]</td>
<td>PTFRN [C2], SRSF6 [C3], ZBED5 [C2]</td>
</tr>
<tr>
<td>response to insulin</td>
<td>MYOS [C3], CACNA1A [C2]</td>
</tr>
<tr>
<td>synapse assembly</td>
<td></td>
</tr>
<tr>
<td>Group 2 [C0]</td>
<td>SMARCDS, MED24, WAC, FOXO5, RELA</td>
</tr>
<tr>
<td>positive regulation of transcription, DNA-templated</td>
<td>SMARCDS, WAC, RELA</td>
</tr>
<tr>
<td>chromatin organization</td>
<td>FOXO5, RELA</td>
</tr>
<tr>
<td>tumor necrosis factor-mediated signaling pathway</td>
<td></td>
</tr>
<tr>
<td>positive regulation of pri-mRNA transcription from RNA polymerase II promoter</td>
<td>FOXO5, RELA</td>
</tr>
<tr>
<td>regulation of autophagy</td>
<td>RPTOR, WAC</td>
</tr>
<tr>
<td>stem cell population maintenance</td>
<td>MED24, MIAT</td>
</tr>
</tbody>
</table>
2.4.3 Gene expression dynamics of methylated mRNAs, m⁵C readers, writers, and erasers during brain development

We extended the analysis to identify differential methylation sites (DMSs) across five conditions. DMSs were defined as m⁵C sites with a methylation difference ≥ 0.05 between two conditions and a p-adjusted value ≤ 0.05. Pairwise comparisons determined 83, 176, 62, and 172 DMS sites for NSCs vs neurons, P0 vs P17, P17 vs 6W, and P0 vs 6W, respectively (Figure 2-4A). Interestingly, of the 83 DMSs identified between NSCs vs neurons, 96% showed increased methylation in neurons. Gene Ontology analyses indicated that differentially methylated sites in neurons compared to NSCs are significantly enriched for positive regulation of DNA-templated transcription, including RNA polymerase II transcription (Figure 2-4B). The differentially methylated sites identified in postnatal brain tissues were significantly enriched for the regulation of GTPase activity and synaptic plasticity associated with brain development. For instance, the transcripts Atxn1, Fgfr1, and Itsn1, which are known to be critical for brain development, were found to be hyper-methylated in the P0 brain.

To examine the relationship between RNA methylation and gene expression, we performed differential expression analysis of sample-matched RNA-seq libraries. These differentially expressed genes (DEGs) were defined as transcripts with an expression difference ≥ 1.5-fold between two conditions and a p-adjusted value ≤ 0.01. The majority of transcripts carrying DMSs were not differentially expressed between P17 and 6W (Figure 2-4C), while a number of genes showed significant changes in both gene expression and RNA methylation during early brain development. A moderate positive correlation (R = 0.337 in NSCs vs neurons) was observed between the changes in gene expression and RNA methylation (Figure 2-4D). As aforementioned, transcripts carrying m⁵C sites tend to have a higher methylation level in neurons compared to those in NSCs. Of those differentially methylated transcripts, 52.3% exhibited an increased expression level in neurons. Notably in neuronal samples, the transcript Psd, which codes for a Plekstrin homology and SEC7 domain-containing protein, was over-expressed and hyper-methylated. Sec7 domains were conserved throughout many proteins and served to catalyze the guanine nucleotide exchange factor initiating the formation of vesicle coating [36]. However, both positive and negative correlations between gene expression and methylation were observed for pairwise comparisons of NSCs vs neurons, P0 vs P17, and P0 vs 6W. For instance, the differentially methylated transcripts Csf1r and Endod1, which were hyper-methylated and under-expressed in
the P0 brain, are markers for microglia and endothelial cell types, suggesting cell-type specific methylation of transcripts. Conversely, two transcripts, Map6 and Sdc3, were consistently hyper-methylated and over-expressed in P0 samples compared to P17 and 6W samples. Map6 infers structural stability in developing neuronal microtubules, promoting interconnectivity and signaling [37, 38], while Sdc3 is a well-known mediator of neuronal lineage, promoting neuronal migration and synapse formation [39, 40]. This result indicated that the regulation of RNA methylation and gene expression may be cell-type specific and is correlated for some transcripts but may be two independent events for others.
Figure 2-4. Methylation dynamics throughout brain development. (A) Differential methylation analysis of neuron and brain samples. Blue and red points represent significantly hyper and hypo methylated transcripts, respectively. Significance was determined using Fisher exact test followed by Benjamini-Hochberg correction for multiple comparison ($p < 0.05$), requiring a minimum difference of 0.05 in methylation. (B) Gene Ontology analysis of differentially methylated transcripts identified in (A). (C) Correlation analysis of changes in RNA expression and methylation for transcripts carrying differentially methylated sites. Blue and red bars represent positive and negative correlation between methylation level difference and expression fold-change, respectively. Green bars represent transcripts without expression fold-changes above 2 and $p$-adjusted less than 0.01. (D) Dotplot representation of transcripts to indicate the changes in methylation and expression. Overall correlation was determined using Spearman’s $R$. Transcripts with methylation difference over 0.3 are notated.

Recently, significantly increased understanding has been gained about the writers, readers, and erasers of m$^5$C in mRNAs [41-44]. To examine the expression patterns of these factors, we
first identified the top 1,000 variably expressed genes in our RNA-seq dataset and performed clustering analysis (Figure 2-5A). Not surprisingly, NSCs showed distinct gene expression profiles from other samples. However, the expression profiles of the factors regulating RNA methylation were found to be similar between NSCs and neurons (Figure 2-5B). Interestingly, all factors related to m⁵C regulation except for RAD52 were minimally expressed in 6W and P17 samples. On the other hand, the P0 brain displayed the highest expression of all m⁵C writers except for NSUN4. Worthy of mentioning, P0 brain has the highest number of m⁵C sites identified exclusive from other samples (Figure 2-3B). We further performed pairwise comparisons to determine the transitions in gene expression (Figure 2-5C). Thousands of DEGs were identified in between NSCs and neurons, while three times fewer DEGs were found in P17 vs 6W comparison. Despite most m⁵C writers being reported to be differentially expressed in at least one of the comparisons, the NSUN2 gene is stably expressed during brain development (Figure 2-5D). Remarkably, NSUN2 is known as an m⁵C RNA methyltransferase that plays a key role in tRNA m⁵C methylation in neurons. Defects in the NSUN2 gene have been associated with neurological disorders [45, 46]. The P17 vs 6W comparison yielded the fewest differentially expressed factors related to RNA m⁵C regulation. This suggests that m⁵C regulation is highly dynamic during early brain development and may reach homeostasis throughout the maturation process.
2.4.4 Differentially methylated transcripts have cell type-specific expression patterns and are temporally regulated in developing brains

During brain cell specification, more m$^5$C sites (Figure 2-3B) and increased methylation (Figure 2-4D) were observed in differentiated neurons and maturing brains. This inspired us to further examine the temporal expression patterns of differentially methylated transcripts using bulk RNA-seq data publicly available for mouse embryonic/postnatal brain tissues including different
neuronal subsets. We focused on 42 differentially methylated transcripts identified with our RNA BS-seq dataset, which also showed expression fold-change greater than 2 across our five RNA-seq conditions (Figure 2-6A). According to their expression profiles in bulk RNA-seq, four distinct clusters were identified (Figure 2-6B). Cluster 1 and cluster 3 were characterized with high expression levels in the adult brain and fetal brain, respectively. Of the Cluster 1 transcripts, Endod1, Wbp2, and Rapgef4 were highly expressed in the mature 22-month-old brain samples and were found to be both over-expressed and hyper-methylated in our 6W brain samples compared to the P0 brain. ENDOD1, an endothelial marker, was found in a previous study to have high expression in the cerebral cortex and hippocampus, while WW domain binding protein 2 (WBP2) and Rap guanine nucleotide exchange factor 4 (RAPGEF4) were mainly localized to the cerebral cortex [47, 48]. WBP2 is a transcriptional co-activator of estrogen receptor α (ESR1), of which mutations induce abnormal glutamatergic synapse development [49].

Methylated transcripts in cluster 2 showed neuronal specific gene expression patterns. To further examine the neuronal specific expression patterns of these methylated transcripts, we integrated the epitranscriptome profiles generated in this study with an scRNA-seq dataset recently published for the P0 cortex [22] (Figure 2-6C). Previous studies reported that the methylation of tRNAs is essential for proper development of cortical layers [13, 15]. Intriguingly, scRNA-seq data also suggested that the neuronal specific gene expression profile was associated with a high level of mRNA methylation for some transcripts (Figure 2-6D). In neurons, the Psd, Runc3a, and Smard3 transcripts were hyper-methylated, and their expressions were also broadly enriched for many neuronal subpopulations in cortical layers I-VI, interneurons 1-3, and striatal inhibitor neurons 1 and 2. For these genes, such a cell-type-specific expression was also reported in previous studies. For example, localized expression of Runc3a was previously found to be concentrated within noradrenergic populations of multiple brain regions [47, 48]. Taken together, these results indicate that highly methylated transcripts in mature neuron populations have temporal and cell-type specific expression patterns.
Figure 2-6. Temporal and spatial expression of differentially methylated transcripts. (A) Clustering analysis of expression profiles for differentially methylated transcripts. Z-score of TPM was used to perform clustering. (B) Linear representation of clusters identified in (A). Error bars represent standard deviation. (C) tSNE clustering of scRNA-seq P0 brain tissue from Loo et al., 2019 [22] (D) TPM z-score of differentially methylated transcripts using scRNA-seq data generated for P0 brain. Color of bubble plot indicates methylation level differences between NSC and neuron RNA BS-seq samples.
2.5 Discussion

Cytosine methylation of mRNA has emerged as a critical regulator of mRNA transportation and stability. Despite distinct mRNA methylation profiles reported in brain and NSCs [17], no attempt has been made to characterize mRNA m^5C methylation patterns in the developing mammalian brain. In this study, we provide both gene expression and mRNA methylation profiles for NSCs, mature neurons, and postnatal brain tissues.

RNA methylation profiling can be achieved with m^5C-specific antibodies to enrich methylated transcripts followed by deep-sequencing [50, 51]. Such an approach has resolution limitations and is highly dependent on the quality of antibody used. Despite the fact that bisulfite sequencing remains the gold standard for RNA methylation detection, our recent experiments [27], together with reports from other labs [8, 52], indicated that both experimental procedure and methylation calling have a significant impact on the detection of RNA methylation. Thus, inconsistent methylation sites may be observed between biological replicates. In this study, we found that high-confidence m^5C sites shared by two biological replicates tend to have a higher methylation level and more read coverage than those of low-confidence m^5C sites identified in one biological replicate only. In addition, low-confidence sites frequently reside in a transcript sensitive to bisulfite treatment.

Notwithstanding current technical challenges in the determination of RNA methylation, our analyses on high-confidence m^5C sites led to a few interesting findings. Consistent with previous studies [7, 8, 17], m^5C sites are densely located near the translation initiation sites of mRNAs and usually have a methylation level between 20-30% with a downstream “GGG” motif. Approximately 6% of high-confidence m^5C sites identified in this study were methylated across all samples. However, the diversity of mRNA methylation increases in differentiated neurons. In addition, increased methylation in neurons were observed in 96% of the DMSs identified between NSCs vs neurons. Previous studies linked the methylation of both tRNA and mRNA to synapse formation [13, 15]. Intriguingly, the DMSs identified in P17 vs P0 and 6W brain tissues were significantly enriched for synaptic plasticity. This suggests that mRNA methylation may also play important roles during brain development.
In this study, we examined the expression profiles of RNA methylation readers, writers, and erasers. As a result of tRNA degradation, neuronal synapse function was impaired in the NSUN2 knockout mouse model [16]. Interestingly, the NSUN2 enzyme is the only one that showed no significant change in expression throughout development. This result supported the critical function of the NSUN2 enzyme and also suggested that other enzymes may contribute to RNA methylation dynamics in the developing brain. The integration of RNA-seq data with RNA BS-seq data allowed us to explore the correlation between the levels of gene expression and RNA methylation. A weak to moderate positive correlation was observed in general, but such a trend may not be true for all transcripts. This suggests that the regulation of RNA expression and methylation might be two independent events, at least for some transcripts. However, the methylated transcripts could be cell-subset specific, spatially, and/or temporally regulated. Future studies are needed to explore how cell type specific mRNA methylation may contribute to neuronal differentiation. Collectively, our study provided insight into the m\(^5\)C methylation dynamics of the developing brain and is a rich resource to facilitate further investigation on the role of brain mRNA methylation.

**Author Contributions**

H.X. conceived and designed the study; X.X. isolated and cultured E16.5 mouse cortical neurons and neural stem cells, constructed libraries for RNA-seq and RNA bisulfite sequencing; Z.J. and Y.L. performed the bioinformatic analyses; X.X., Z.J., Y.L., and H.X. interpreted results and wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

**Acknowledgments**

This study was supported by NIH grant ES031521, NS094574, MH120498, NSF1922428, the Center for One Health Research at the Virginia-Maryland College of Veterinary Medicine and the Edward Via College of Osteopathic Medicine, the Center for Engineered Health, the Virginia-Maryland College of Veterinary Medicine at Virginia Tech, and the Fralin Life Sciences Institute faculty development fund for H.X. We thank Dr. Janet Webster for English language editing.

**Competing financial interests**

The authors declare no competing financial interests.
2.6 References


36. Wright J, Kahn RA, Sztul E: Regulating the large Sec7 ARF guanine nucleotide exchange factors: the when, where and how of activation. *Cellular and Molecular Life Sciences* 2014, 71:3419-3438.


Chapter 3 - Transcript specific targets of m\(^5\)C methylation in \textit{Fmr1}-KO and \textit{Tet1}-KO models

Zachary Johnson\(^{1,2\#}\), Xiguang Xu\(^{1,3,4\#}\), Hehuang Xie\(^{1,2,3,4,5,6\ast}\)

1. Epigenomics and Computational Biology Lab, Fralin Life Sciences Institute of Virginia Tech, VA 24061, USA
2. Genetics, Bioinformatics and Computational Biology program, Virginia Tech, Blacksburg, VA 24061, USA
3. Department of Biological Sciences, College of Science, Virginia Tech, Blacksburg, VA 24061, USA
4. Department of Biomedical Sciences and Pathobiology, Virginia-Maryland College of Veterinary Medicine, Virginia Tech, VA 24061, USA
5. Translational Biology, Medicine and Health Program, Virginia Tech, Blacksburg, VA 24061, USA
6. School of Neuroscience, Virginia Tech, Blacksburg, VA 24061, USA

\(^\#\)These authors have contributed equally to this work.

\(^\ast\)Corresponding author: Email: davidxie@vt.edu

\textbf{Status:} In progress
3.1 Abstract

RNA modifications are gradually becoming well known as modulators of RNA stability and export, thus influencing the differentiation and development of neurons. While the mechanisms of m$^\text{5}$C-mRNA methylation are largely unknown, recent studies have demonstrated that the fragile X mental retardation protein, FMRP, and the m$^\text{5}$C eraser, TET1, cooperate to demethylate regions of transcriptional stalling, promoting the proliferation of embryonic stem cells. However, no study has investigated the m$^\text{5}$C methylation profiles of FMRP and TET1 knockout models in vivo. In this study, we profile the m$^\text{5}$C landscape of complete Fmr1-KO and Tet1-KO brain tissue using RNA BS-seq. By integrating methylation data with expression and protein pulldown assays, we found that select few transcripts’ methylation and expression statuses are significantly affected by the ablation of these proteins. One such transcript, Ralbp1, is a critical regulator of long term depression in proliferating neurons. This transcript was identified as a binding partner of FMRP and exhibited hypermethylation and overexpression in Fmr1-KO mouse brain samples. We provide a first insight into the effects of FMRP and TET1 ablation on m$^\text{5}$C-mRNA methylation in the brain, further supporting links between the m$^\text{5}$C modification and brain development.

**Keywords:** Brain development, Fragile X mental retardation protein, ten-eleven translocation methylcytosine dioxygenases, RNA cytosine-5 methylation, RNA bisulfite sequencing, RNA-seq, RIP-seq
3.2 Introduction

The *Fmr1* gene codes for the Fragile X mental retardation protein (FMRP), an RNA-binding protein which increases mRNA stability and reduces protein translation efficiency through binding in a transcript-specific manner. Absence and mutations of *Fmr1* cause fragile X syndrome (FXS), autism, and other intellectual disabilities [1]. Such defects result from the dysregulation of protein translation critical for cognitive function including synapse formation and long term depression (LTD) in neurons [2, 3]. Recently, studies have shown RNA modifications may facilitate the binding of FMRP to mRNAs, promoting neuronal proliferation and differentiation through the export of certain mRNAs [4, 5]. RNA modifications, including N⁶-methyladenosine (m⁶A) and 5-methylcytosine (m⁵C), are emerging as critical regulators of RNA export and stability, and play essential roles in neuronal differentiation [6-8].

Recently, m⁵C-RNAs have also been found to regulate the DNA damage repair process via recruitment of FMRP and the m⁵C eraser ten-eleven translocation protein 1 (TET1). During the transcription of template DNA, nascent RNA can rehybridize with single-strand DNA to form RNA:DNA hybrids, also known as R-loops. R-loops are important regulators of gene expression near the transcription start site (TSS) and termination site. R-loops can also trigger the DNA damage response by promoting DNA double strand breaks (DSBs) [9]. Elevated levels of DNA damage and defects in the DNA damage repair process are common to many neurodegenerative diseases [10, 11]. The m⁵C writer tRNA aspartic acid methyltransferase 1 (TRDMT1) locally induces RNA m⁵C modifications to stabilize R-loops in response to DSBs and promotes homologous recombination (HR) repair events via the m⁵C reader RAD52 [12]. FMRP is also recruited to sites of DNA damage by TRDMT1 further facilitating the recruitment of TET1 to oxidize and demethylate m⁵C sites. This interaction resolves the DSB repair of the damaged site, resulting in genomic stability of the area [13]. One previous study has identified the mRNA targets of TET1 and TET2 using methylated RNA immunoprecipitation [14]; however, transcript-specific m⁵C-mRNA targets of the TET enzymes have yet to be determined.

To date, no studies have reported transcript-specific m⁵C binding targets of FMRP and TET1 using *in vivo* knockout (KO) models. In this study, we generated *in vivo Fmr1*-KO and *Tet1*-KO mouse models to identify significant m⁵C methylation changes in the brain caused by FMRP and TET1 ablation (Figure 3-1A). By integrating RNA immunoprecipitation sequencing (RIP-seq) data, we identified *Ralph1*, a transcript essential for neuronal LTD, as a potential m⁵C binding
target of FMRP. In \textit{Fmr1}-KO brains, \textit{Ralbp1} was both hypermethylated and overexpressed compared to wildtype (WT) tissue, indicating FMRP may be recruited to this specific site. This study provides a useful resource for understanding the transcript-specific changes in methylation levels related to \textit{Fmr1}-KO and \textit{Tet1}-KO models.

\section*{3.3 Materials and Methods}

\textbf{Mice}
These experiments were approved by the Institutional Animal Care and Use Committee (IACUC) of Virginia Tech. Mice were maintained and bred in a 12-hour light/dark cycle under standard pathogen-free conditions. The mouse lines \textit{Tet1}-knockout mice (stock number: 017358) and \textit{Fmr1}-knockout mice (stock number: 003025) were obtained from Jackson Laboratory. Heterozygous mice were used to breed mutant mice and wild type controls. Young adult (6 weeks) male mouse brain samples were used for experiments.

\textbf{Genotyping}
Crude DNA was extracted from tail biopsies using Direct PCR tail lysis buffer (Viagen Biotech Inc, cat# 101T) supplemented with Proteinase K solution (Thermo Fisher, cat# EO0491) and genotyped by PCR according to Jackson Laboratory’s protocols. The following three primers were used for \textit{Tet1} genotyping PCR: wild type forward primer (TCA GGG AGC TCA TGG AGA CTA), mutant forward primer (AAC TGA TTC CCT TCG TGC AG) and common reverse primer (TTA AAG CAT GGG TGG GAG TC). Mutant (\textit{Tet1}^{-/-}) mice showed bands at 650bp, wild type (\textit{Tet1}^{+/+}) mice showed bands at 300bp, and heterozygote (\textit{Tet1}^{+/-}) mice showed bands at both 650bp and 300bp.

The following three primers were used for \textit{Fmr1} genotyping PCR: wild type forward primer (TGT GAT AGA ATA TGC AGC ATG TGA), mutant forward primer (CAC GAG ACT AGT GAG ACG TG), common reverse primer (CTT CTG GCA CCT CCA GCT T). Mutant (\textit{Fmr1}^{-/-}) mice showed bands at 400bp, wild type (\textit{Fmr1}^{+/+}) mice showed bands at 131bp, and heterozygote (\textit{Fmr1}^{+/-}) mice showed bands at both 400bp and 131bp.
**In vivo RIP-seq assay**

RIP was performed as previously described [15] with minor modifications. Briefly, NSC cultures from two 15cm plates were washed with PBS, detached from the plates by plastic scraper, and pooled as one sample. NSC pellets were resuspended with 1 mL lysis buffer (150mM KCl, 10mM HEPES pH 7.6, 2mM EDTA, 0.5% NP-40, 0.5mM DTT, 1:100 protease inhibitor cocktail, 0.4U/μl RNase inhibitor) by rotating at 4°C for 30 min and centrifuged at 15,000g for 15 min. The supernatant was collected and pre-cleared with 60μl Dynabeads protein A (ThermoFisher) by rotation at 4°C for 1h. 50μL pre-cleared cell lysate was saved as Input and the remaining were subject to IP. 10μg rabbit polyclonal FMRP antibody (Abcam, ab171722) was incubated with 60μL Dynabeads Protein A at 4°C for 4h and washed twice with binding/blocking buffer to remove free antibodies before IP. The pre-cleared cell lysate was incubated with antibody-coupled beads at 4°C for 4h. After washing with 600μL ice-cold NT2 buffer (200 mM NaCl, 50mM HEPES pH 7.6, 2mM EDTA, 0.05% NP-40, 0.5mM DTT, 0.4U/μl RNase inhibitor) for eight times and once with 1mL ice-cold 1 × PK buffer (100mM Tris-HCl pH 7.4, 50mM NaCl, 10mM EDTA, 0.2% SDS), the beads were treated in 240μl PK buffer containing 10μl proteinase K solution (ThermoFisher, 20mg/mL) for 1 h at 55°C. The Input and IP RNA were extracted by using TRizol™ LS Reagent and digested with TURBO DNase (Invitrogen) to eliminate DNA contamination. RNA samples recovered from Input and IP were further subjected to mRNA purification by rRNA removal (RiboMinus Eukaryote Kit v2, Ambion) and used to generate RNA-BisSeq and RNA-Seq library.

**RNA BS-seq library construction**

RNA bisulfite conversion was performed as previously described [16] with minor modifications. Briefly, poly(A) RNA was first mixed with spiked-in Xef1 unmethylated RNA at a ratio of 0.5%. The spiked-in unmethylated mRNA was transcribed from the pTRI-Xef plasmid supplied by the MEGAscript™ T7 Transcription Kit (Invitrogen). The linearized pTRI-Xef plasmid was transcribed in-vitro in a reaction with MEGAscript T7 RNA polymerase (Ambion) at 37°C for 4h. Further, DNase treatment was applied at 37°C for 15 min to remove the DNA template and then purified using the RNeasy Mini Kit (QIAGEN). RNA bisulfite conversion was performed with an initial denaturation at 95°C for 1min, followed by three cycles of 70°C for 10 min and 64°C for 45 min using the EZ RNA Methylation Kit (Zymo Research). The bisulfite converted RNA was
subjected to stranded RNA-seq library construction using TruSeq Stranded mRNA Library Preparation Kit (Illumina). We modified the procedure to skip the RNA fragmentation step and to supply both random and ACT random hexamers during the first strand cDNA synthesis.

**RNA BS-seq data analysis**
RNA bisulfite analysis was performed as previously reported [17]. Briefly, raw sequencing reads were filtered for short length, low quality, adaptor sequences, and were trimmed by 6bp at the 5’ and 3’ ends to account for methylation bias. Clean reads were mapped to the mm10 genome (Ensembl v.79) using meRanGh [18]. Mapped reads were further filtered by removing reads with more than 3 “C”s (“G”s on the cDNA strand) to remove partially unconverted reads. Methylation calling was performed using meRanCall, and sites were filtered using a series of filters. Sites were further required to have m\(^5\)C depth of 6 or more. Sites which passed all filters in both biological replicates were reported as high-confidence sites and used for downstream analysis.

**Differential methylation analysis**
Differential methylation analysis was performed using Fisher’s exact test on high-confidence sites present in at least one sample. Benjamini-Hochberg p-value correction was used to correct for multiple comparisons, and sites with p-adjusted < 0.05 were reported.

**RNA-seq library construction**
Stranded RNA-seq libraries were constructed as previously described [19]. Libraries were sequenced on a Hiseq 4000 platform in the 150bp paired end mode (Illumina).

**RNA-seq data analysis**
Trim Galore (version 0.6.5) was used to filter short reads, low quality reads, and trim adapter sequences from raw reads ([https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/](https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)). Clean reads were mapped to the mm10 genome and expression quantified using STAR (version 2.7.3a) [20]. Differential expressed genes were identified using DESeq2 [21] using fold change greater than 1.5 and an adjusted p-value of 0.01 as cutoffs.

**Gene Ontology (GO) analysis**
GO analysis was performed using DAVID (Database for Annotation, Visualization, and Integrated Discovery) [22]. The GO Direct terms, Biological Process (BP), cellular component (CC), and molecular function (MF), were identified. Significantly enriched terms were identified using Benjamini-Hochberg $p$-value correction, $p$-adjusted < 0.05.

**Clustering Analysis**

The $z$-score of bulk RNA-seq expression TPM and differentially methylated sites’ methylation level were used as input for clustering analysis. Clustering analysis was performed with the Python package sklearn’s Agglomerative Clustering. The ‘complete’ method was used with the number of clusters set to 5 for RNA-seq and RNA BS-seq datasets. Error bars represent the standard deviation of $z$-scores.

**3.4 Results**

As discussed in Chapter 2, the post-natal brain undergoes dramatic methylation changes from post-natal day 0 to post-natal day 17. To uncover whole brain m$^5$C methylation profiles of FMRP and TET1 complete knockout models during this critical development window, we performed RNA BS-seq on whole brain tissue of complete knockout P17 mice (Figure 3-1B, Figure 3-1C). Approximately 100 million paired end reads were generated per library. Sequencing reads were processed as described in our previous study using the meRanTk software suite [17]. Briefly, low quality sequencing reads were removed, and clean reads were trimmed by 6bp at read ends to account for methylation bias caused by residual primers. Clean reads were then mapped to the C2T converted mm10 genome using bowtie2 within meRanGh. Reads were mapped with ~90% accuracy. Methylation sites were called using meRanCall and potential false positive sites were subsequently removed with a battery of filters. These filters include a maximum 3C-count per read, minimum 6C-depth per site, signal-to-noise ratio, secondary structure prediction, and conversion rate efficiency (Table 3-S1).
Figure 3-1: Experimental design. (A) Schematic describing the overall experimental design. (B-C) Genotyping results for Tet1 wild type (Tet1^+/_) and mutant mice (Tet1^-/-) (B), and Fmr1 wild type (Fmr1^+/_) and mutant mice (Fmr1^-/-) (C).

Of the six RNA BS-seq libraries, between 473 and 1,106 sites were called, resulting in a total of 3,245 m5C unique sites called. To further refine our m5C site calling, we defined sites found in both replicates as “high-confidence” sites. We reported 269, 264, and 278 high-confidence sites in the WT, Fmr1-KO, and Tet1-KO BS-seq samples respectively (Figure 3-2A). These sites were highly correlated by methylation level in between biological replicates, with a Spearman R values between 0.83 and 0.92 (Figure 3-2B). High confidence sites reported higher methylation levels on average (0.23) compared to low confidence sites (0.12) (Figure 3-2C). However, 72.1% of high confidence sites reported methylation levels below 0.3.
3.4.1 High-confidence m^5C sites are shared among WT, *Fmr1*-KO, and *Tet1*-KO models

We identified a total of 382 high-confidence sites among WT, *Fmr1*-KO, and *Tet1*-KO models. Features common to these sites are a dominant “GGG” motif downstream of the m^5C site (Figure 3-S1A). These characteristics have been identified by previous reports [17, 23-25], and serve as an indicator of potential NSUN2 activity [25]. 97% of sites identified were located on
mRNAs, while 3 lncRNA sites and 1 miRNA site were identified in each condition (Figure 3-S1B).

It is currently unknown if FMRP or TET1 ablation will affect specific m^5^C sites or brain-wide methylation levels. We found 180 high-confidence sites (47.5%) were shared by WT, Fmr1-KO, and Tet1-KO conditions, and 249 sites (65.2%) were shared by at least two samples (Figure 3-3A). We identified sites shared by conditions. To identify potential changes in methylation level among conditions, we performed pair-wise analysis of 180 shared high-confidence sites’ methylation levels, however we found no significant difference among the samples (Figure 3-3B). To identify distinct patterns in methylation levels, we performed clustering analysis of sites with at least 20x read coverage in all samples (Figure 3-3C). We identified six distinct clusters (C0-C5) of methylation level. These clusters were defined as consistently methylated: high (C0), medium (C2), and low (C1) methylation level, or as methylation sites distinct to WT (C4), Fmr1-KO (C5), and Tet1-KO (C3) conditions (Figure 3-3D). The larger abundance of consistently methylated sites suggests roles of FMRP and TET1 related to m^5^C methylation may be compensated by other enzymes. The distinct presence or lack of high-confidence sites in Fmr1-KO and Tet1-KO tissues indicates possible methylation-specific roles played by FMRP and TET1. To investigate this, we performed Gene Ontology analysis of sites enriched in samples by clustering analysis (Figure 3-3E). In transcripts consistently methylated in all conditions, protein binding and calcium signaling pathways were significantly enriched. While not significantly enriched, pathways related to calcium ion binding and signaling were identified in WT and TET1-KO clusters. Calcium signaling is an essential function of neuronal differentiation and migration. Interestingly, transcripts methylated in the Fmr1-KO cluster were associated with neuronal projection and positive regulation of transcription. These findings suggest transcripts specifically methylated in Fmr1-KO and Tet1-KO models may serve functional roles in neuronal development.
Figure 3-3: Methylation level changes in Fmr1-KO and Tet1-KO models. (A) High-confidence m\(^5\)C sites shared among conditions. (B) Methylation levels of m\(^5\)C sites. Wilcoxon rank-sum test was performed to assess significance (*\(p < 0.05\)). (C) Clustering of m\(^5\)C sites according to methylation level. Clustering was performed using the “ward” method of SciPy’s “Agglomerative Clustering” package with number of clusters set to \(n = 5\). (D) Average methylation levels of clusters identified in A. Bars indicate standard deviation. (E) Gene Ontology analysis of clustered m\(^5\)C-bound transcripts identified in D. Biological Process (BP), Molecular Function (MF), and KEGG pathways are listed. Significantly enriched pathways are denoted by *\(p\)-adjusted \(\leq 0.05\).

3.4.2 Transcript-specific changes in methylation and expression levels related to Fmr1-KO and Tet1-KO models

To identify transcript expression related to Fmr1 and Tet1 KO models, we performed RNA-seq analysis of WT, Fmr1-KO, and Tet1-KO whole brain samples. On average ~25 million read pairs were generated per sequencing library (Table 3-S1). Sequencing reads were trimmed using then mapped to the mm10 reference genome using STAR with a unique mapping rate of...
Differential expression analysis revealed 1,776 and 1,829 differentially expressed genes (DEGs) under-expressed in Fmr1-KO and Tet1-KO conditions respectively, while less than 60 transcripts were over-expressed in both conditions (Figure 3-4A). To identify enriched gene and protein expression pathways, Gene Ontology of DEGs was performed for Biological Pathway, Molecular Function, and KEGG protein pathways. Both Fmr1-KO and Tet1-KO were most significantly enriched for protein binding gene expression pathways. Interestingly, many neurodegenerative diseases were also enriched in both conditions by KEGG pathway analysis including Alzheimer’s disease, Prion disease, Huntington’s disease, and Parkinson’s disease (Figure 3-S2).

To uncover the effect of Fmr1-KO and Tet1-KO models on methylation changes in the brain, we performed differential methylation analysis (Figure 3-4B). To our surprise, we identified very few differentially methylated transcripts: 17 in Fmr1-KO and 9 in Tet-KO. All but 2 transcripts were hypermethylated in the Fmr1-KO condition, and 4 were hypermethylated in the Tet1-KO condition. These results suggest that FMRP and TET1 may not affect the methylation status of transcripts generally but may influence the methylation of select few transcripts instead. We found no correlation between RNA-seq expression data and the methylation status of differentially methylated sites in Fmr1-KO and Tet1-KO models (Figure 3-4C).

FMRP is an RNA-binding protein [27] as well as transcription factor [28]. In addition, FMRP is a reader of the m^6A epitranscriptomic modification, influencing the neuronal differentiation of neuronal precursor cells [5]. To identify a potential link between FMRP-dependent m^5C sites and neuronal differentiation, we performed FMRP-bound RNA immunoprecipitation sequencing (FMRP RIP-seq) on neuronal stem cells and integrated peak calling data with differentially methylated sites. On average, ~20 million reads were generated for FMRP-IP and input-IP libraries. Clean reads were uniquely mapped to the mm10 genome with ~76% efficiency, and ~3,000 peaks were called per sample using input normalization (Table 3-S1). 2,686 peaks were common to both biological replicates resulting in an 82.8% peak overlap rate, demonstrating high reproducibility (Figure 3-S3A). To confirm the gene annotation of FMRP peaks, we cross-referenced Darnell et al., 2011 [29] FMRP pulldown data with 54% overlap rate (Figure 3-S3B). The majority of peaks were identified downstream of the promoter region, followed by peaks annotated in the 3’-untranslated region (Figure 3-S3C). To support evidence for FMRP modulation of specific m^5C sites, we integrated FMRP RIP-seq peak calling data with
RNA BS-seq methylation calling and RNA-seq expression data. Three differentially methylated sites overlapped with FMRP RIP-seq datasets (Figure 3-4D). The m\(^5\)C site chr17:65,848,931 located in the 3’UTR of the transcript Ralbp1 was found to be enriched in FMRP-IP peak calling data (Figure 3-S3E). This site was hypermethylated in the Fmr1-KO BS-seq dataset with a methylation difference of +0.233 and had increased expression by a fold-change of +0.18. Ralbp1 has been found to directly regulate long term depression (LTD) in neurons via promotion of synaptic AMPA glutamate receptor endocytosis [30]. LTD promotes neuronal migration and differentiation through elevated expression of proliferation markers. These findings suggest few transcripts are exclusively reliant on FMRP or TET1 for methylation. However, through the integration of RNA-seq and RIP-seq datasets, we identified Ralbp1 as a hypermethylated transcript in the Fmr1-KO brain model and may have implications on neuronal development.
Figure 3-4: Differentially methylated and expressed transcripts in Fmr1-KO and Tet1-KO models. (A) Differential methylation analysis of neuron and brain samples. Blue and red points represent significantly hyper and hypo methylated transcripts, respectively. Significance was determined using Fisher exact test after Benjamini-Hochberg correction for multiple comparison ($p \leq 0.05$), requiring a minimum methylation level difference of 0.05. (B) Gene Ontology analysis of differentially methylated transcripts identified in (A). (C) Correlation of RNA-seq expression analysis with differentially methylated sites. Blue and red bars represent positive and negative correlation between methylation level difference and expression fold-change, respectively. Green bars represent DMSs without any noticeable correlation. (D) Table denoting differentially methylated transcripts identified within peak calling regions of the FMRP pulldown assay. Information on transcript position, RNA type, methylation level difference, and expression difference is shown.
3.5 Discussion

RNA modifications are emerging as critical regulators of neuronal differentiation and proliferation. Recently, FMRP has been shown to regulate synaptic development via modified RNA export [5]. Further, the m$^5$C eraser TET1 plays functional roles in embryonic stem cell development via m$^5$C hydroxymethylation (hm$^5$C) [14]. In this study, we analyzed the transcriptome-wide effects of FMRP and TET1 ablation on brain tissue methylation status.

Detection of RNA-m$^5$C has been a significant challenge due to the low signal-to-noise ratio of the m$^5$C modification. Recent studies have demonstrated improvements to RNA BS-seq computational analysis methods, improving the reproducibility of high-confidence sites; however, improvements to the chemical conversion are to be desired [17, 25]. With these improvements, we identified more than 150 high-confidence m$^5$C sites common to Fmr1-KO and Tet1-KO brain tissues. Sites which were commonly methylated in WT and KO models were significantly enriched for calcium ion binding pathways, a process critical for LTD functions, while sites exclusively methylated in Fmr1-KO brain tissue were associated with positive transcriptional regulation. Previous studies have shown FMRP to recruit TET1 to locations of transcriptional stalling marked by m$^5$C sites [13]. The transcripts identified here may provide potential targets to explore this effect in the brain.

While FMRP and TET1 are known to be recruited to m$^5$C sites [13], their transcript-specific binding targets are still unknown. After we performed differential methylation of analysis of KO models, we found fewer than 10% of identified sites were differentially methylated in these respective models. Additionally, differentially methylated transcripts were not correlated with expression non-bisulfite converted transcripts despite more than 15% of transcripts being under expressed in both KO models. One previous study has shown hm$^5$C binding targets of TET1 are largely overlapped with TET2, suggesting the loss of these enzymes may be compensated for by other TET enzymes with regards to m$^5$C methylation [14]. To identify sites which may influence the differentiation of neuronal stem cells, we performed RIP-seq of FMRP and TET1 in NSC culture. Of RIP-seq peaks overlapped with RNA BS-seq sites, the transcript Ralbp1 was identified to be hypermethylated and overexpressed in Fmr1-KO RNA BS-seq data. RalBP1 interacts with RalA, a GTPase required for synaptic branching, to complete NMDAR-dependent endocytosis of AMPA receptors and engaging in LTD. Our findings indicate that FMRP may influence the methylation status of Ralbp1 transcripts, furthering evidence for m$^5$C’s potential role in neuronal
differentiation and migration. Further biological validation of this m$^5$C site and its protein interactions is needed.

Altogether, our findings show a transcript-specific influence of m$^5$C methylation due to the absence of FMRP and TET1 in the brain. We identified a common set of methylated RNAs in these knockout models; however, few transcripts’ methylation levels were significantly affected by either knockout. Despite this, we identified a hypermethylated transcript, Ralbp1, whose methylation status may be an important factor in neuronal differentiation and LTD.

Acknowledgements
This work was supported by NIH grant ES031521, NS094574, MH120498, NSF1922428, the Center for One Health Research at the Virginia-Maryland College of Veterinary Medicine and The Edward Via College of Osteopathic Medicine, and the Fralin Life Sciences Institute faculty development fund for H.X. We recognize The Center for Engineered Health, the Virginia-Maryland College of Veterinary Medicine at Virginia Tech, and VT’s Open Access Subvention Fund. We thank Dr. Janet Webster for English language editing.

Author Contributions
H.X. conceived and designed the study; X.X. isolated and cultured neural stem cells, isolated brain tissue, and constructed libraries for RNA-seq and RNA bisulfite sequencing; Z.J. performed the bioinformatic analyses; Z.J., X.X., and H.X. interpreted results and wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Competing financial interests
The authors declare no competing financial interests.

3.6 References


Chapter 4 - Conclusions and Future Directions

The \( m^5C \) modification of RNAs has been demonstrated to be a novel factor in controlling gene expression and protein translation by regulating the stability and export of RNAs. The field of epitranscriptomics has revealed \( m^5C \)’s relationship to cancer progression, stem cell development, and neuronal functions; however, we have yet to elucidate the mechanisms behind these relationships. Recent advances in sequencing technologies have allowed researchers to perform transcriptome-wide searches for \( m^5C \) at single nucleotide resolution using RNA bisulfite sequencing strategies. However, the lack of standardized methods for performing, analyzing, and interpreting such experiments has led to extreme variations in the total number and feature set describing \( m^5C \) sites [1-4].

In **Chapter 1**, we sought to improve upon existing RNA bisulfite preparation methodologies to identify mitochondrial mRNAs that may be methylated. We combined these approaches with a proposed “gold-standard” computation pipeline which employed a combination of processing and filtering strategies to eliminate sequencing artifacts. By utilizing a novel combination of UMI adapters and ERCCs control transcripts, we were able to determine increased rates of PCR-generated false positive sites were correlated with PCR cycle number. In addition, we found that certain ERCCs were more susceptible to non-conversion events than others; however, we were unable to identify the commonality between these transcripts. We also found that the addition of a high temperature denaturation step during bisulfite conversion resulted in an increased overall conversion rate for transcripts derived from the mitochondrial genome. Due to this and other site filters, all mitochondrial mRNA \( m^5C \) sites were eliminated from consideration as “high-confidence” sites. Our findings demonstrate a need for improved chemical approaches for RNA bisulfite conversion to appropriately measure mitochondrially derived mRNA \( m^5C \) sites, if such sites exist.

In **Chapter 2**, we evaluated the \( m^5C \) landscape in the developing mammalian brain focusing on neuronal stem cells, E16.5 cortical neuron culture, P0, P17, and 6W old mouse brain samples. 501 “high-confidence” sites were identified among all samples, with whole brain samples containing more total sites than cell culture samples. We found methylation levels of P0 and P17
m$^5$C to be significantly different from the 6W condition, indicating the most dynamic changes in m$^5$C methylation occur early in life. Transcripts carrying highly differential methylation were significantly enriched for gene pathways regulating synaptic plasticity, brain development, and neuronal differentiation. The results support evidence that m$^5$C may play a role in cognitive development.

In Chapter 3, we investigated the methylation changes related to the ablation of m$^5$C reader FMRP and the m$^5$C eraser TET1 in the P17 brain. After identifying a common set of 157 high-confidence m$^5$C sites, we found that fewer than 10 sites were differentially methylated in both datasets. This result suggests that m$^5$C may rely on redundant reader and erasure systems. Using a FMRP pulldown assay, we integrated FMRP-binding peak calling data with differentially methylated site locations. With these results, we identified RalBP1, a regulator of neuronal LTD to by both bound by FMRP and hypermethylated in Fmr1-KO brain samples. Previous studies have found m$^5$C-tRNAs to regulate neuronal stem cell differentiation and motility [5] as well as control protein expression critical for synaptic firing [6]. Our findings suggest that m$^5$C-bound mRNAs in addition to tRNAs may be essential for neuronal differentiation.

**4.1 Future Directions**

The characterization of low abundance m$^5$C-modified mRNAs remains as an elusive problem for the field of epitranscriptomics, only exacerbated by inconsistent analytical approaches and artifacts generated by bisulfite sequencing approaches. Our efforts here, along with other groups, have made progress towards identifying and eliminating sources of artifacts. Zhang et al., has shown in publications that did not use extensive computational filters or high-stringency preparation conditions, over 90% of m$^5$C sites identified were likely to be false positive sites [4]. Huang et al., found that by using the Gini coefficient to determine an appropriate C-cutoff, variations in library-specific preparation conditions could be controlled for [2]. Together with these studies, we argue for the need to find improved RNA-specific chemical approaches to better understand the true context of m$^5$C.

Other methods used for the detection of m$^5$C include physiochemical-based methods such as liquid chromatography mass spec (LC-MS) and pulldown-based methods including CLIP-seq, meRIP-seq, and 5-Aza-seq. While these approaches have found success in identifying m$^5$C-tRNAs
and synthetic oligo-m^5C transcripts, the identification of m^5C-mRNAs using these methods is limited by their inability to achieve single nucleotide resolution [7]. In addition, pulldown-based methods are largely dependent on antibody quality, and a reputable m^5C-RNA antibody has yet to hit the market. Despite this claim, researchers continue to progress in the development of these assays to be essential validation strategies for RNA bisulfite sequencing experiments [8-10].

Regardless of limitations in their detection, RNA modifications continue to surprise us with the diversity of functions and biological processes they help to regulate. The rapidly expanding field of epitranscriptomics continues to understand the mechanisms behind reading, writing, and erasing of m^5C with new targets of m^5C being discovered at a rapid rate. Modifications including m^6A and m^5C are critical for neuronal differentiation, synaptic function, and brain development with far-reaching implications for cognitive disabilities and neurodegenerative diseases. Further development of detection and analytical methods are needed to discover the true relevance of m^5C in biology.[11]

4.2 References


