THE UNIVERSITY OF CHICAGO

OME-WIDE ILLUMINATION OF UNPRODUCTIVE SPLICING IN BRAINS AND DNA N^6-METHYLDEOXYADENOSINE IN MAMMALS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF HUMAN GENETICS

BY

XINRAN FENG

CHICAGO, ILLINOIS

AUGUST 2023

DEDICATION

To my father, Wei Feng,

and to all those who have held my wild dreams in their tender embrace.

Table of Contents

List of Figures

List of Tables

Acknowledgement

First and foremost, I would like to express my sincere gratitude to my thesis supervisor, Dr. Chuan He, for his invaluable guidance and support throughout my Ph.D. journey. Dr. He's expertise, encouragement, and unwavering commitment to my research have been instrumental in shaping my academic and professional development. Before joining his group in 2020, I had no knowledge of chemical biology techniques such as mass spectrometry. I wouldn't have been able to accomplish what I have today without him.

Likewise, I would like to express my sincere gratitude to my other two mentors, Dr. Xiaochang Zhang and Dr. Xiaoxi Zhuang, for their unwavering mentorship and support throughout my Ph.D. journey. The period during which I studied and conducted research in their labs was the time when my research skills improved the most rapidly. It was during this time that I gained proficiency in bioinformatics, mouse experiments, and scientific figure creation that were previously less familiar to me. Even to this day, I still often think of the convivial atmosphere and the tasty food provided in their group meetings.

I would also like to thank my committee members, Dr. Marcelo Nobrega, Dr. Xiaoxi Zhuang, Dr. Tao Pan, and Dr. Xin He for their insightful feedback and constructive criticism, which have significantly contributed to the quality of my research and thesis.

I am grateful to the faculty and staff at the Department of Human Genetics at UChicago for providing a stimulating academic environment and excellent resources that have facilitated my research. Especially Dr. Anna Di Rienzo, Dr. Marcelo Nobrega, Dr. Francois Spitz, and Ms. Susan Levison. They have always been so helpful when I need it.

I would like to express my gratitude to the Biological Sciences Division and the Dean's Council at the University of Chicago. There were so many informative and enjoyable courses,

workshops, and social activities, such as BSD Quantitative Biology Bootcamp at MBL and Graduate Student Seminars. These activities added vibrancy and joy to the challenging and monotonous Ph.D. life.

My sincere appreciation goes to my wonderful colleagues in Zhang Lab, Zhuang Lab, and He Lab, particularly Dr. Qi Cai and Dr. Li-Sheng Zhang, who have been great sources of inspiration, motivation, and support. Their friendship, encouragement, and assistance have been invaluable. Here, I extend my sincere wishes to all my colleagues for success in their academic and professional endeavors and a promising future ahead.

I also owe a tremendous thank you to Dr. Boxun Lu, my supervisor in college, as well as my teachers in high school, Mr. Xuexian Mei, and Mr. Lei Cheng. Your teachings, guidance, and the belief in my potiential have been instrumental in my transformation from an average student to an excellent one, ultimately paving the way for me to embark on a career in scientific research.

And finally, I would like to express my heartfelt gratitude to my family and my beloved one for their unconditional love, encouragement, and support. In my days of despondency, their unwavering belief in me has been a constant source of motivation and strength. Three years ago, my father's health took a serious turn for the worse, and regrettably, I was unable to be by his side. I am profoundly thankful to those relatives and friends who graciously stepped in, assuming the responsibility of caring for my father and liaising with doctors on my behalf. Ultimately, my utmost wish is for my father's prompt and complete recovery.

As Micheal Faraday noted, “It is the great beauty of our science, chemistry, that advancement in it, whether in a degree great or small, instead of exhausting the subjects of research, opens the doors to further and more abundant knowledge, overflowing with beauty and utility”.[1]

I'm grateful for the invaluable contributions of you all, which have enabled me to make the trivial yet meaningful “advancement” to the field of science.

Abstract

Gene expression is controlled through both transcriptional regulation and post-transcriptional events. Pre-mRNA alternative splicing (AS) represents a crucial post-transcriptional mechanism that regulates the spatial and temporal expression of genes. Despite its significance, the roles of AS in brain development remain largely unknown. The first chapter of this thesis is dedicated to examining the unproductive splicing events in neurodevelopment, through the use of RNA-Seq, which generated a long list of unproductive splicing events dynamically regulated during neurogenesis. Thereinto, an alternative 3' splice site (A3SS) located in the *SYNGAP1* gene was identified as an important regulator of its expression, playing a vital role in neurodevelopment. The use of A3SS in non-neuronal cells results in nonsense-mediated decay (NMD) and guarantees neuron-specific expression of SYNGAP1 protein. This mammal-specific AS event is controlled by PTBP1 and PTBP2, which repress SYNGAP1 expression by promoting the usage of this A3SS. Manipulating this AS event through the use of splice-switching antisense oligonucleotides (SSOs) can restore SYNGAP1 expression, presenting a potential therapeutic strategy for non-syndromic intellectual disability (NSID) caused by *SYNGAP1* haploinsufficiency. Overall, this part of the study sheds light on the previously unknown role and mechanism of cell-type-specific unproductive AS in brain development.

On another level, DNA methylations, one type of main epigenetic modifications, usually serve as a repressive or activation mark for gene expression through transcriptional regulation. As the predominant DNA modification found in prokaryotic genomes, DNA N^6-methyldeoxyadenosine (6mA) plays crucial roles in the restriction modification system, DNA repair, and also gene expression regulation. However, the frequency, dynamics, distribution patterns, effector proteins, and biological functions of 6mA in eukaryotic cells remain subject to

controversy due to the extremely low abundance and the lack of effective and sensitive base-resolution detection approaches. The second chapter of this thesis presents an antibody-independent single-nucleotide-resolution 6mA sequencing method, namely Direct-Read 6mA Sequencing (DR-6mA-seq). This novel approach capitalizes on the misincorporation tendency of deoxythymidine triphosphate (dTTP) analog at 6mA sites, enabling quantitative mapping of 6mA in various genomic DNA species in a fast and cost-effective manner. By using this method, we were able to characterize the presence and elevated level of 6mA in the nuclear DNA of glioblastoma model cells. DR-6mA-seq can serve as a gold standard for quantitative Next Generation Sequencing (NGS) methods of 6mA, facilitating further biological studies and inspiring the development of new mapping methods for other DNA/RNA modifications.

List of Publications Based on Work Presented in this Thesis[†]

1. Yang, R.*, **Feng, X.** *, Arias-Cavieres, A., Mitchell, R.M., Polo, A., Hu, K., Zhong, R., Qi, C., Zhang, R.S., Westneat, N., et al. (2023). Upregulation of SYNGAP1 expression in mice and human neurons by redirecting alternative splicing. Neuron. 10.1016/j.neuron.2023.02.021.
2. **Feng, X.** *, Cui, X.*, Zhang, L.-S.*, Ye, C., Wang, P., Zhong, Y., Zheng, Z., and He, C. (2023). Sequencing of N^6-methyl-deoxyadenosine at single-base resolution across the mammalian genome. bioRxiv. 10.1101/2023.01.16.524325.
3. **Feng, X.**, and He, C. (2023). Mammalian DNA N^6-methyladenosine: Challenges and new insights. Mol Cell 83, 343–351. 10.1016/j.molcel.2023.01.005.

* Co-authors contributed equally.

[†] The following chapters of the dissertation contain sections and figures adopted from the listed pubilications with modifications. Chapter 1: publication 3; Chapter 2: publication 1; Chapter 3: publication 2; Chapter 4: publication 3.

Chapter 1

Introduction: Pre-mRNA Splicing and DNA Epigenetic Modifications

1.1 Pre-mRNA splicing and its regulation

In contrast to prokaryotic cells, eukaryotic cells harbor a substantial abundance of non-coding sequences within their genomes.[2] Thus, pre-mRNA splicing becomes a crucial process in eukaryotic gene expression, where the initial transcript of a gene, called pre-mRNA, undergoes precise removal of non-coding sequences, i.e., introns, and joining of coding sequences, i.e., exons, to produce mature mRNA.[3] This process occurs in the nucleus of eukaryotic cells before the mRNA is transported to the cytoplasm for translation into proteins.[3] The splicing of pre-mRNA is facilitated by a complex macromolecular machinery called the spliceosome, which consists of small nuclear ribonucleoproteins (snRNPs) and other associated proteins. The spliceosome, which is assembled by five small nuclear ribonucleoprotein (snRNP) (U1, U2, U4/U6, and U5), recognizes specific sequence elements at the exon-intron boundaries, including the 5' splice site (by U1 snRNP) as well as polypyrimidine tract (by U2 auxiliary factor), and the branch point sequence (by splicing factor 1, or mammalian branch point binding protein) at the 3' splice site, to accurately remove the introns and ligate the exons together (Figure 1.1A).[4]

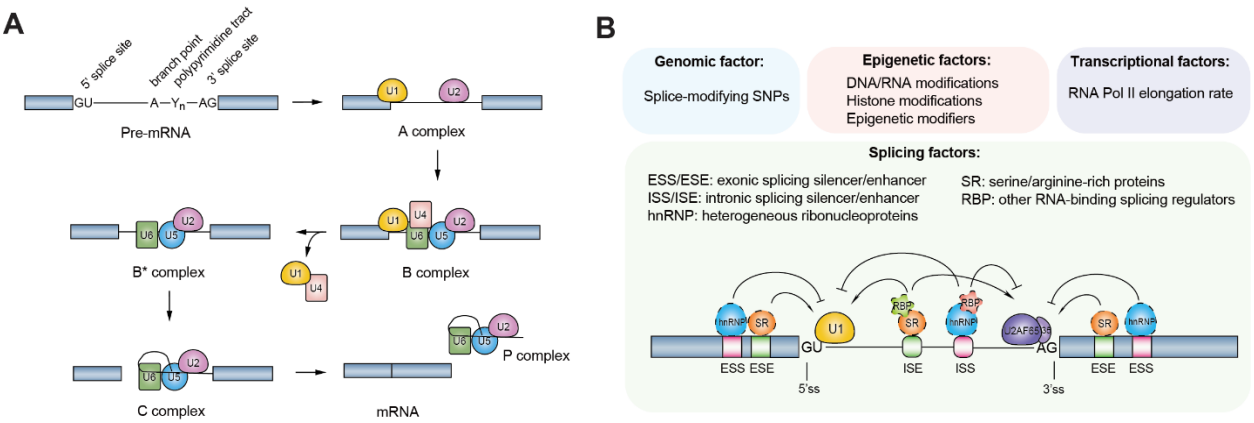


Figure 1.1 The mechanism and regulation of pre-mRNA splicing
A) The diagram illustrating the sequential assembly of the spliceosome, a piece of complex molecular machinery, through the integration of its constituent small nuclear ribonucleoprotein (snRNP) components. The initial step involves the binding of U1 to the 5' splice site and U2 to the branch point, leading to the formation of the A complex. Subsequently, the tri-snRNP U4/U6.U5 associates with the prespliceosome (A complex), resulting in the formation of the precatalytic spliceosome (B complex). The removal of U1 and U4 triggers the activation of the spliceosome, leading to its functional state. Following the completion of the first splicing step, the catalytic step 1 spliceosome (C complex) is formed. Finally, after the second splicing step, the spliced mRNA product is released from the post-splicing (P) complex. **B)** Schematic representation of splicing regulation mechanisms. Alternative splicing exhibits intricate associations with splicing factors, transcriptional machinery (RNA polymerase II elongation rates), and epigenetic modifications (histone marks, epigenetic modifiers, and DNA methylation) at both post- and co-transcriptional levels. During the splicing process, splicing factors selectively target and interact with spliceosome components to regulate the recognition of 5' and 3' splice sites flanking the alternative exon. Serine/arginine-rich (SR) proteins and heterogeneous ribonucleoproteins (hnRNPs) are examples of such factors. SR proteins function as general splicing activators by binding to exonic/intronic splicing enhancers (E/ISEs), thereby facilitating exon formation. Conversely, hnRNPs act as general splicing inhibitors by binding to exonic/intronic splicing silencers (E/ISSs), thereby interfering with splice site recognition.

The regulation of pre-mRNA splicing is essential for generating protein diversity and controlling gene expression. It allows alternative splicing, a process where different combinations of exons can be included or excluded from the final mRNA, leading to the production of multiple protein isoforms from a single gene.[4] The regulation of alternative splicing can occur concomitantly with mRNA transcription and is governed by a combination of epigenetic modifications and RNA Pol II elongation (Figure 1.1B). During the splicing process, the regulation of alternative splicing involves a complex interplay between splice-modifying SNPs, the splicing

machinery, the regulatory elements within the pre-mRNA sequence, and various RNA-binding proteins (Figure 1.1B).[5,6] First of all, regulatory RNA sequences within the transcripts, known as exonic or intronic splicing enhancers (ESEs or ISEs) and exonic or intronic splicing silencers (ESSs or ISSs), can modulate the splicing process. ESEs and ISEs promote splice site recognition, while ESSs and ISSs inhibit or suppress splice site usage. These regulatory elements can be bound by specific RNA-binding proteins (RBPs) that either enhance or repress splicing.[7] Therefore, RBPs play a critical role in alternative splicing regulation and can act as either splicing activators or repressors, depending on their binding locations and the specific context of the pre-mRNA. Examples of such RBPs include the serine/arginine-rich (SR) proteins, heterogeneous nuclear ribonucleoproteins (hnRNPs), and many other splicing factors.[8] Since the expression levels and cellular localization of RBPs are usually dynamically regulated, changes in their levels or subcellular distribution can impact alternative splicing patterns. Signaling pathways and developmental cues can influence the expression, phosphorylation, or subcellular localization of these splicing regulators, thereby modulating their activity and splicing outcomes.[9]

Beyond that, chromatin structure and epigenetic modifications can also affect alternative splicing. DNA methylation, histone modifications, and chromatin remodeling complexes can regulate the accessibility of splicing regulatory elements and influence the recruitment of splicing factors.[10] Similarly, the secondary structure of pre-mRNA can also impact splice site recognition and alternative splicing outcomes because stem-loop structures or RNA-binding proteins that stabilize or disrupt RNA structures may influence the accessibility of splice sites and regulatory elements.[11] Other secondary factors, such as RNAPII elongation rate, splicing factor recruitment efficiency, and binding of snRNAs at the splice site may also potentially modulate the alternative splicing.[5] Overall, the regulation of alternative splicing is a complex process involving the

interaction and coordination of multiple factors that determine the specific splice site selection and alternative splicing patterns observed in different cellular contexts and conditions.

1.2 The types and functions of alternative splicing

Pre-mRNA alternative splicing plays crucial roles in expanding the coding potential of the genome and regulating gene expression. It greatly expands the coding potential of the genome and plays a critical role in various biological processes, including tissue-specific gene expression, development, and disease.[12] Based on the location of alternative splice sites, pre-mRNA alternative splicing can be categorized into seven types: exon skipping (SE), intron retention (RI), alternative 5' splice site (A5SS), alternative 3' splice site (A3SS), mutually exclusive exons (MXE), alternative promoters (AP), and alternative polyadenylation terminators (AT) (Figure 1.2A).[13] Otherwise, based on the functional outcome, alternative splicing can be simply divided into two categories: those that alter protein sequence and those that modify protein expression levels.

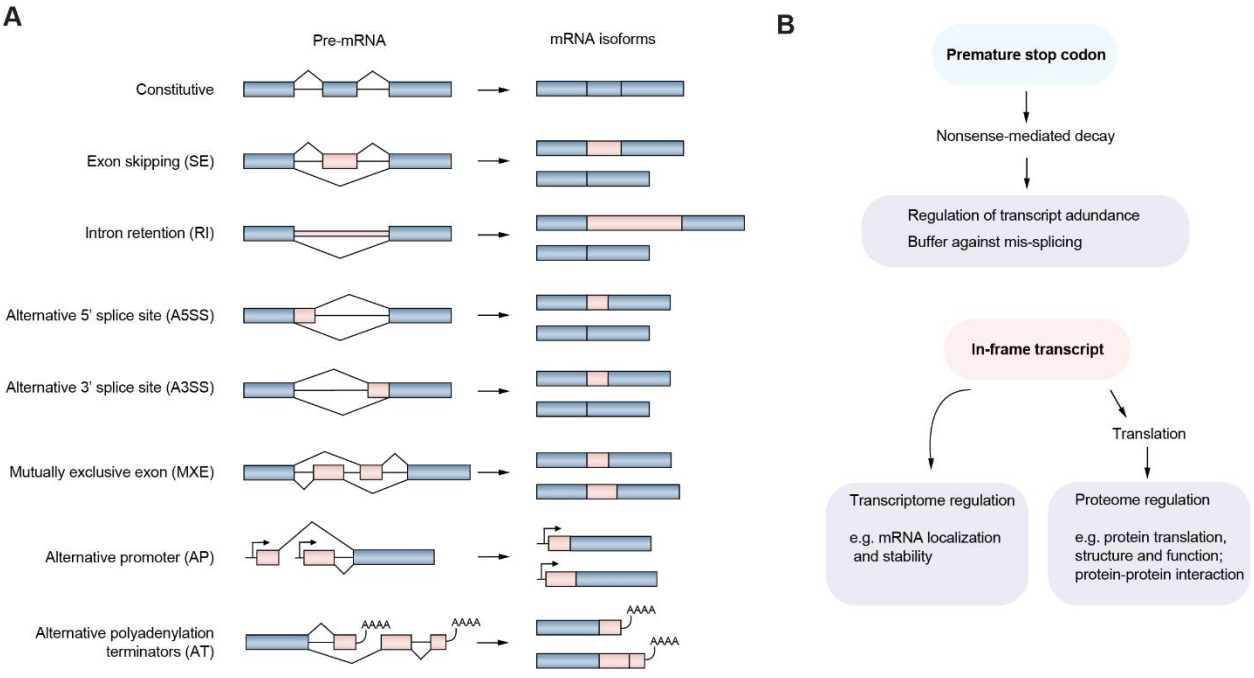


Figure 1.2 The types and functional outcomes of alternative splicing

(**Figure 1.2, continued**) **A)** The schematic diagram illustrating seven distinct modes of alternative splicing in pre-mRNA: exon skipping (cassette exon), intron retention, alternative 5' splice site, alternative 3' splice site, mutually exclusive exons, alternative promoters, and alternative polyadenylation. The exons are represented by boxes and the introns by lines. **B)** The two molecular outcomes of alternative spliced transcripts: transcripts harboring premature stop codons are subject to degradation via nonsense-mediated decay, while in-frame alternative splicing modifies the functionality of the transcripts at the post-transcriptional level (mRNA localization and stability) and/or the post-translational level (protein sequence and modifications).

The altered protein sequence generated by alternative splicing significantly increases the diversity of protein isoforms that can be produced from a limited number of genes. This enables cells and organisms to have a greater repertoire of protein variants to perform specialized functions (Figure 1.2B). Occasionally, the newly generated protein isoforms exhibit distinct functions compared to the wild-type variant (gain of function, GOF). For instance, an alternatively spliced form of tissue factor (TF), wherein a portion of the TF extracellular domain, transmembrane domain, and cytoplasmic domain is substituted with a unique C-terminus, largely enhances angiogenesis in comparison to the full-length TF.[14] The synaptic Ras GTPase-activating protein 1 gene (*SYNGAP1*) has four splice variants at the C-terminal, namely α1, α2, β, and γ. These variants possess distinctive biochemical properties and subcellular localization, each performing unique functions in neuronal development and synaptic plasticity.[15] However, in many other cases, protein variants generated through alternative splicing lead to the loss of function (LOF), thereby serving as a regulatory mechanism for protein activity. The alternatively spliced isoform of thyrotropin-releasing hormone degrading enzyme (Trhde) lacking part of the C-terminal domain, for example, is enzymatically inactive and has a dominant-negative effect.[16] At times, alternative protein variants can even exert inhibitory effects on the activity of the wild-type isoform. For instance, Pro-Casp6b directly interacts with pro-Casp6a and functions as an inhibitor of pro-Casp6a activation.[17] Another example is observed in the alternative splicing of DAPK1 at the kinase domain, which generates a novel isoform capable of destabilizing the wild-type isoform.[18]

In a distinct subset of scenarios, alternative splicing does not give rise to protein variants but instead induces the degradation of the respective mRNA transcripts, consequently lowering the expression levels of the relevant protein. (Figure 1.2B) In such cases, alternative splicing events introduce in-frame premature termination codons (PTC), leading to the degradation of the mRNA through a process called nonsense-mediated decay (NMD).[19] As a quality control mechanism to prevent the accumulation of potentially harmful truncated proteins that could disrupt cellular processes, NMD begins with the recognition of PTCs by a group of specialized proteins. Among these proteins, Up-frameshift 1 (UPF1) plays a key role as it associates with the exon junction complex (EJC) formed during the splicing process. When PTCs are present, they disrupt the distribution of the EJC, thereby triggering the recruitment of UPF1. Upon PTC recognition, UPF1 recruits additional NMD factors, forming the core NMD complex. This complex includes UPF2 and UPF3, which interact with UPF1, enhancing its stability and activity. The NMD complex then interacts with the mRNA, leading to the recruitment of endonucleases and exonucleases that initiate the degradation of the targeted mRNA. This degradation process proceeds in a 5'-to-3' direction, resulting in the rapid turnover of the aberrant mRNA molecule.[20] NMD predominantly occurs in a translation-dependent manner, where the ribosomes engaged in the pioneer round of translation become stalled near the PTC and trigger NMD.[21] In mammals, the PTCs usually trigger NMD only when they are positioned more than 50-55 nucleotides (nt) upstream of the most downstream EJC.[22] The occurrence of alternative splicing events that result in the inclusion of PTCs and subsequently trigger NMD of the transcripts, commonly known as unproductive splicing, adds an additional level of post-transcriptional regulation and plays a critical role in diverse biological processes. One notable example involves the alternative skipping of exon 11 in polypyrimidine tract binding proteins 1 (PTBP1, aka hnRNP-1), resulting in the generation of a

transcript targeted for degradation by NMD. This pathway accounts for the degradation of at least 20% of *PTBP1* transcripts in HeLa cells. Interestingly, the skipping of exon 11 is facilitated by PTBP1 itself, establishing a negative feedback loop.[23]

Alternative splicing is a prevalent and intricate process that significantly enhances protein diversity, fine-tunes gene expression, and regulates cellular functions. In fact, it takes place in around 95% of multi-exon genes in humans and is linked to approximately 15% of human inherited diseases and cancers.[24] Gaining a comprehensive understanding of the functions and regulatory mechanisms governing alternative splicing is of utmost importance in unraveling the intricacies of gene expression and comprehending its profound impact on the biology of cells and organisms.

1.3 Differential gene expression regulation is vital in brain development

Similar to most biological processes, brain development requires spatially and temporally regulated differential gene expression to guarantee the maturation of neurons, accurate formation of neural circuits, and the establishment of functional connectivity.[25,26] During early brain development, neural stem cells (NSCs) differentiate into various neurons and glial cells. The expression of specific genes guides the fate of these stem cells, directing them toward distinct cell lineages. Transcription factors and signaling molecules play crucial roles in regulating gene expression patterns that determine the types of neural cells.[27] For instance, during the initiation of neuronal differentiation in NSCs, the expression of Hes1 transitions from oscillatory to a repressed state, while the expression of Ascl1 shifts from oscillatory to sustained.[28,29] Later on, proper neuronal migration is vital for establishing the brain's layered structure. Differential expression of the genes involved in cytoskeletal dynamics, adhesion molecules, and guidance cues guide the movement of neurons along specific pathways to reach their appropriate positions in the developing brain.[27] To be specific, migrating interneurons heading to the cortex relies on attractive

cues, particularly neuregulin1/ErbB4, to navigate the corticostriatal notch and enter the cortical mantle. Therefore, genes involved in interneuron migration, like ErbB4, are expressed at higher levels in the cortical interneuron population compared to the ganglionic eminences (GE).[30,31] Once neurons reach their target destinations, they either differentiate into mature neurons with axons and dendrites or undergo apoptosis. Axon guidance molecules, such as netrins, semaphorins, and ephrins, provide cues for axons to navigate through the developing brain and establish proper synaptic connections.[27] At this stage, differential gene expression, again, governs the guidance of dendrites, axons, and the formation of synapses. For instance, synaptic proteins like postsynaptic density protein-95 (PSD-95) encoded by *DLG4* and SYNGAP1 must be translated and effectively transported to synaptic terminals in order to fulfill their synaptic functions during synaptogenesis. In some cases, they may even undergo liquid-liquid phase separation to increase the local protein concentrations.[32] In addition to synaptogenesis, some specific genes play a crucial role in enabling synaptic plasticity, which involves the strengthening and pruning of synapses based on neural activity and experience. The overexpression of PSD-95, for example, enhances synaptic strength and obstructs long-term potentiation (LTP), while the knockdown of PSD-95 reduces the surface expression of AMPARs and weakens synaptic strength.[33] This precise adjustment of neural circuits is vital for ensuring proper sensory processing, facilitating learning, and promoting the formation of memories.[27] The final stage of brain development is myelination, which involves the formation of a myelin sheath around axons. The regulated expression of genes involved in myelin synthesis, axon-glial interactions, and signal transduction pathways regulate myelination, ensuring efficient electrical signal transmission and insulation of neural circuits.[27]

Therefore, as a pivotal mechanism that governs brain development, differential gene expression plays a crucial role in determining neural stem cell fate, orchestrating neuronal

migration, guiding axon growth, facilitating synapse formation, refining neural circuits, promoting myelination, and governing other critical processes. Genetic mutations or alterations that impact the expression of genes involved in brain development can disrupt normal processes, resulting in abnormal brain structure and function. This, in turn, can contribute to the development of neurodevelopmental disorders such as autism spectrum disorders, intellectual disabilities, and epilepsy.[32] Gaining a comprehensive understanding of the complex dynamics of gene expression during brain development is crucial for unraveling the mechanisms that govern normal brain function, as well as the underlying causes and potential therapeutic approaches for neurodevelopmental disorders.

1.4 DNA modifications and their functions

Covalent DNA modifications play a critical role in gene regulation and cellular processes.[34] For more than half a century, the most abundant DNA modification in the mammalian genome, 5-methylcytosine (5mC), has undergone extensive research and emerged as one of the extensively studied epigenetic modifications in mammals.[35] In recent years, significant progress in scientific investigations has shed light on a wide array of other base modifications found within DNA. These newly identified modifications encompass but are not limited to, 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), 5-carboxycytosine (5caC), N^4-methylcytosine (4mC), N^6-methyldeoxyadenine (6mA), N^6-hydroxymethyladenine (6hmA), 8-oxo-7,8-dihydroguanine (8-oxo-G), and 5-glyceryl-methylcytosine (5gmC) (Figure 1.3).[34,36] Many of the covalent DNA modifications, including 8-oxoguanine, 1-methyladenine, 3-methylcytosine, and 6-O-methylguanine, are actually a result of DNA damage caused by factors such as oxidation, alkylation, free radicals, ultraviolet, and ionizing radiation (Figure 1.3).[37,38] These modifications are usually cytotoxic, alter the coding properties or normal functions of DNA during transcription

or replication, and may contribute to various aberrant processes such as aging, neurodegeneration, and cancer.[37,39] Efficient DNA repair mechanisms are essential for removing these harmful DNA damages and maintaining the integrity and normal functioning of the genome.[39,40]

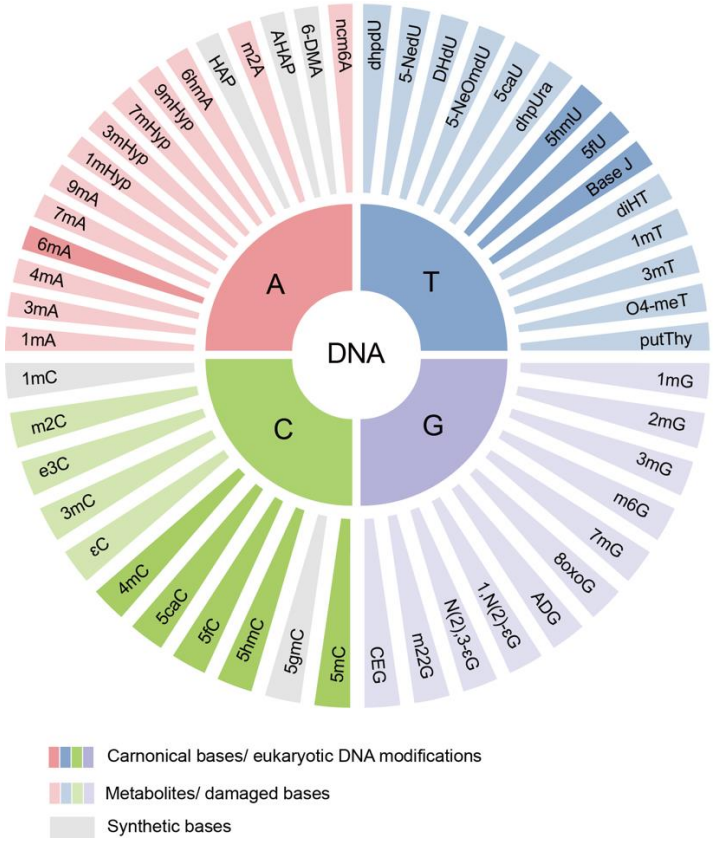


Figure 1.3 The known covalent DNA modifications

Pie diagram showing the known modified DNA bases (modified from dnamod.hoffmanlab.org). The inner ring comprises the four canonical DNA bases. The outer ring encompasses modified bases present in DNA, whereby the natural bases are labeled in color and the synthetic bases are labeled in grey.

On the other hand, there are several directed DNA modifications that serve as epigenetic marks, capable of modifying gene expression by effectively switching gene transcription on or off or functioning in normal cellular processes. Furthermore, these modifications are usually maintained and erased through enzymatic processes and can be inherited by subsequent generations of cells.[41] Such examples are 5mC, 5hmC, 5fC, 5caC, 5-hydroxymethyluracil (5hmU),

5-formyluracil (5fU), β-d-glucosyl-5-hydroxymethyluracil (Base J), 4mC, and 6mA (Figure 1.3).[42,43] 5mC is a well-known DNA modification and is prevalent in mammalian genomes. It plays a crucial role in gene expression regulation, genomic stability, and epigenetic inheritance. 5mC is primarily associated with gene silencing and is recognized by specific binding proteins that recruit chromatin modifiers to establish repressive chromatin states. The Ten-Eleven Translocation (TET) family enzymes, including TET1, TET2, and TET3, are responsible for catalyzing the oxidation of 5mC, leading to the generation of other modified cytosine derivatives. This conversion process is implicated in active DNA demethylation and is vital for proper development and cellular differentiation.[44] 5hmC is a DNA modification that has gained significant attention in recent years. It is formed by the oxidation of 5mC and is prevalent in various tissues, particularly in the brain. 5hmC is involved in gene regulation and is thought to be associated with gene activation. It acts as an epigenetic mark that influences chromatin structure and can recruit specific proteins to modulate gene expression. The TET enzymes, particularly TET1 and TET2, are responsible for the conversion of 5mC to 5hmC. 5hmC has been implicated in neurodevelopment, neurodegenerative diseases, and cancer, highlighting its importance in cellular processes and disease mechanisms.[45] 5fC and 5caC are oxidation products derived from 5hmC. These modifications are present at lower abundance compared to 5mC and 5hmC but have significant regulatory roles. They are believed to be involved in active DNA demethylation processes, influencing gene expression and genome stability. The TET enzymes play a crucial role in the conversion of 5mC to 5fC and 5caC. The base excision repair (BER) pathway, involving specific glycosylases and other enzymes, is responsible for the removal and replacement of 5fC and 5caC, contributing to the maintenance of proper DNA methylation patterns.[46,47] 5hmU and 5fU are modified bases that can be found in DNA. They are formed through the oxidative modification of

thymine and are associated with certain biological processes and stress conditions. 5hmU and 5fU have been observed in DNA from different organisms, including bacteria and mammals. While their functions and regulatory enzymes are not yet fully understood, they are believed to contribute to DNA repair, DNA damage response, and adaptation to environmental stressors. Further research is needed to elucidate the precise roles and regulatory mechanisms of 5hmU and 5fU in DNA.[48,49] Base J is a DNA modification found predominantly in the genomes of kinetoplastid parasites, including trypanosomes. It replaces thymine in DNA and is involved in gene expression regulation and genome organization in these organisms. Base J is introduced into DNA by a series of enzymatic steps, including hydroxylation, glycosylation, and deamination. Its presence influences chromatin structure, transcriptional regulation, and DNA replication. Base J-binding proteins and other associated factors recognize and interact with this modified base to exert functional effects on gene expression, and other biological processes.[50] 4mC is a DNA modification that occurs in addition to 5mC. It is mainly found in thermophilic bacteria and archaea, but also plants and mammals, although its prevalence and regulatory roles differ among species. 4mC can be involved in gene regulation, genome stability, and other biological processes. The enzymes responsible for establishing and removing 4mC can vary, depending on the organism. In bacteria, N4C-MTases catalyze the addition of methyl groups to cytosine, while specific DNA glycosylases are involved in the removal of 4mC during DNA repair processes.[51,52]

In addition to the aforementioned DNA modifications, 6mA is another prominent directed DNA epigenetic modification that will be thoroughly discussed in the subsequent sections.

1.5 N^6-methyldeoxyadenosine and its detection method

N^6-methyldeoxyadenine, or 6mA for short, was initially uncovered in the 1950s in *Bacterium coli*.[53] Subsequently, over the course of several decades, 6mA was acknowledged as a

crucial DNA modification present in the genomes of both prokaryotes and protists.[54] It has also been reported to be present in higher eukaryotes, such as plants and animals, although in much lower abundance.[55,56] 6mA is known to play various roles in different cellular pathways across different species.

As bacteria do not possess histones and nucleosomes, DNA methylation serves as their primary mode of epigenetic gene regulation.[57] Among the various types of DNA modifications found in bacteria, 6mA modification is the most abundant, sometimes can reach 10% 6mA/dA. It has been found to play important roles in various cellular processes, including DNA replication, DNA repair, gene expression, restriction–modification (RM) system, and regulates cell division, virulence, and phase variation.[57–62] In protists such as green algae (*C. reinhardtii*) and tetrahymena, the abundance of 6mA is comparatively lower than that in bacteria, with levels ranging from 0.4% to 0.6% 6mA/dA. However, it has been reported that this modification has a significant impact on gene expression and biological outcomes.[63,64] In plants, 6mA modification is generally less abundant compared to other DNA modifications such as 5mC and 5hmC, ranging from 0.1% to 0.7% 6mA/dA in rice seedlings, wheat, maize, sorghum, *Setaria italica*, and Medicago.[65] Nevertheless, recent studies have shown that 6mA modification may play important roles in regulating gene expression and stress responses in plants.[65] This suggests that 6mA modification may be involved in the plant's defense mechanisms against biotic and abiotic stresses such as pathogen infection, salt stress, and drought stress. In vertebrates, especially mammals, the levels of 6mA on genomic DNA and mitochondrial DNA are extremely low, generally below 1000 ppm (parts per million). In these species, 5mC has replaced 6mA as the most abundant DNA modification and plays a primary role in cellular processes. Therefore, in order to demonstrate the

presence and significance of 6mA in vertebrate cells, particularly mammalian cells, more sensitive and accurate techniques are required, as well as more rigorous and in-depth research.[55]

Ultra-high performance liquid chromatography triple quadrupole mass spectrometry (UHPLC-QqQ-MS/MS) and antibody-based dot blotting are the most commonly used methods to quantify the modified nucleosides in DNA/RNA samples. Although both of these methods are quick and efficient, they do not offer precise information about the location of modified sites on DNA/RNA, and their outcomes can be influenced by external DNA/RNA species. Using 6mA as an example, since it is highly prevalent in prokaryotic genomes, when utilizing UHPLC-QqQ-MS/MS and dot blot techniques, bacterial contamination can lead to an overestimation of 6mA levels in multicellular eukaryotes.[66,67] In recent years, more and more NGS- and TGS (third-generation sequencing, aka long-read sequencing)-based methods have been developed to characterize 6mA modification profiles at the genome-wide level. These methods can be categorized into two groups: those that rely on antibodies (antibody-dependent methods) and those that do not require antibodies (antibody-independent methods). The antibody-dependent methods, such as DIP-seq, ChIP-exo (aka 6mACE-seq), SMRT-ChIP, and MM-seq, utilize antibodies specific to N^6-methyladenosine/N^6-methyldeoxyadenosine to identify the modified sites.[68–72] Despite their efficiency, the lack of specificity and sensitivity of these antibodies is often regarded as a potential cause of confounding factors, which may lead to systematic false positives and false positives. Using an IgG-immunoprecipitated control instead of input DNA or RNA can be helpful in reducing the antibody non-specificity.[67,73] The antibody-independent methods, i.e., 6mA-RE-seq, NT-seq, and SMRT, also show different limitations, although they are not affected by the non-specific binding of antibodies and usually detect 6mA at single base resolution.[63,74–76] The 6mA-RE-seq method, which utilizes restriction enzymes like DpnI, is only capable of detecting 6mA at

certain motifs such as GATC. NT-seq is a newly developed technique that has only been applied to detect bacterial 6mA so far. Its efficiency and accuracy in detecting 6mA in other species still need validation. SMRT-based methods have been widely used to detect 6mA in bacterial genomes. However, when applied to mammalian DNA, the SMRT signals can be easily affected by other modifications, such as 5mC, and bacterial DNA contamination. This can result in a high false positive rate when detecting 6mA sites with low modification fractions. Additionally, as a TGS method, SMRT has a higher technical threshold than NGS methods for most labs and requires a high DNA input because it can only detect 6mA from unamplified DNA. In summary, despite the development and improvement of various 6mA detection methods over the years, there is still a lack of a robust, cost-efficient, single-base resolution, highly accurate, and sensitive 6mA detection method at present.

1.6 Mammalian N^6-methyldeoxyadenosine: challenges and controversies

Unlike in bacteria, where the level of 6mA can reach up to 100, 000 ppm, the abundance of 6mA in mammals is significantly lower and has been a topic of debate for a long time.[58] The application of the highly sensitive detection and quantification method, UHPLC-QqQ-MS/MS, has resulted in widely varying estimations of 6mA abundance, ranging from 1 ppm to thousands of ppm in different mammalian genomes.[66,77–81] These discrepancies suggest that 6mA may be dynamically regulated across different types of mammalian cells and, on the other hand, raise concerns that prokaryotic DNA contamination may have led to the overestimation of 6mA in certain LC-MS/MS experiments. Due to its low abundance and the absence of a dominant consensus motif, many 6mA mapping approaches that work effectively in bacterial DNA, such as DIP-seq, 6mA-RE-seq, and SMRT, are less efficient when applied to mammalian DNA.[73] Studies conducted so far generally agree that 6mA is present in specific types of mammalian cells and

tissues, such as embryonic cells (e.g., stem cells) and tumors (e.g., glioblastoma), and enriched in non-coding regions, such as LINE, SINE elements, and intergenic regions.[70,80–82] The low abundance and the unique genomic distribution of mammalian 6mA, in turn, makes it difficult to study its cellular functions. Most studies about mammalian 6mA suggest that it fine-tunes gene expression, while some others believe that it is simply a byproduct of DNA/RNA metabolism.[55,83]

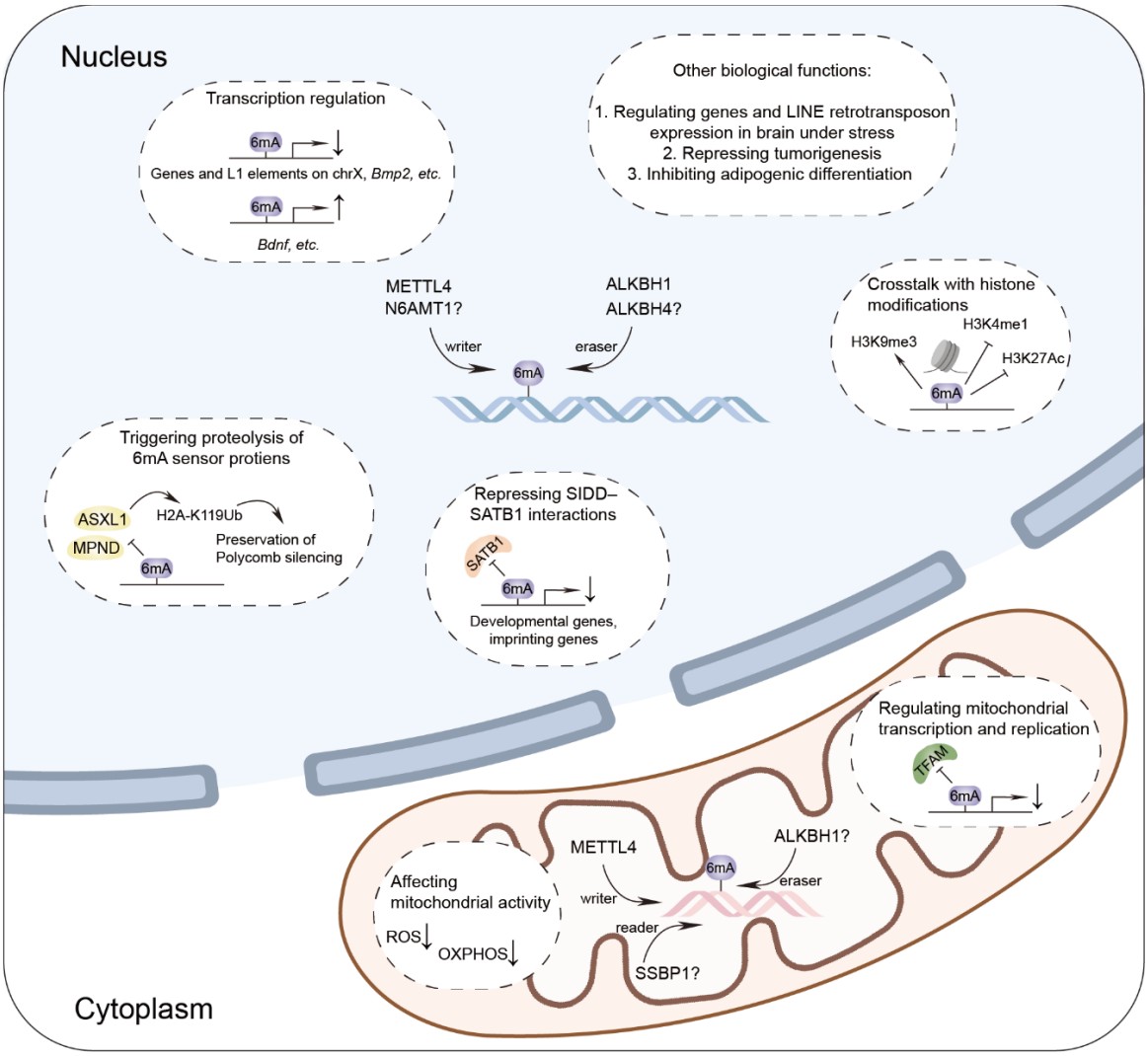


Figure 1.4 Proposed functions and effector proteins for 6mA on mammalian genomic DNA and mtDNA.

The downward arrows refer to the repression of transcription or the decrease of level. The upward arrows refer to the upregulation of transcription. The question marks refer to the putative writer, eraser, and reader protein but with conflicting data or a lack of either *in vivo*, *in vitro*, or genetic evidence.

While the biological functions of 6mA in mammalian cells are still a topic of debate, identifying the regulatory and effector proteins associated with this modification is a crucial step toward establishing 6mA as a directed and functional epigenetic mark in mammals.[54] So far, only two methyltransferases, N6AMT1 and METTL4, along with two demethylases, ALKBH1 and ALKBH4, have been reported as putative 6mA writer and eraser proteins in mammalian cells (Figure 1.4).[54] N6AMT1, also known as N^6-adenine-specific DNA methyltransferase, has been reported to have a regulatory role in 6mA levels in multiple mammalian cells, such as human liver cancer cells and mouse primary cortical neurons.[84,85] However, its methyltransferase activity has been questioned due to its lack of activity even *in vitro*.[81] In contrast, the methyltransferase activity of METTL4 on 6mA has been confirmed both *in vitro* and *in vivo*.[69] Studies have shown that METTL4 regulates 6mA modification on mtDNA and thus affects mitochondrial replication, transcription, mitochondrial activities, and adipogenesis in certain types of human/mouse cells.[69,86] As for demethylases, ALKBH1 has been found to erase 6mA *in vitro* and in various cell types, such as mESCs, patient-derived glioblastoma models, and human cancer cell lines.[70,71,81,85] It has also been suggested that ALKBH1 acts as a 6mA demethylase in human mtDNA, affecting oxidative phosphorylation.[70] However, ALKBH1 has multiple catalytic substrates; it also mediates tRNA oxidation and demethylation in mammalian cytosol and mitochondria.[54,87] In addition, it only exhibits demethylation activity on ssDNA but not on dsDNA, which raises concerns about whether its biological effects mostly arise from tRNA oxidation or DNA 6mA demethylation.[54,86] Regarding ALKBH4, it has shown a low level of demethylation activity on dsDNA *in vitro* and needs more *in vivo* and cell-based evidence to support its function.[86] It is not difficult to see that, based on current research progress, there is no methyltransferase or demethylase that has been fully validated to regulate 6mA modification on mammalian genomic DNA. Therefore, whether

6mA modification exists on mammalian gDNA, whether it is a directed epigenetic mark, and what biological functions it plays in cells are still difficult questions to answer. Developing more accurate, sensitive mapping methods that are suitable for detecting 6mA in mammalian genomes holds the promise of providing definitive and compelling answers to these questions.

1.7 Scope of this thesis

My thesis primarily delves into two projects that revolve around either the application or the development of NGS-based methods to unveil the cellular functions of two crucial gene expression regulation processes: mRNA alternative splicing and DNA methylation. By examining *SYNGAP1* exon 11 A3SS and genomic 6mA modification in glioblastoma cells, respectively, this thesis exemplifies the pivotal roles played by these two processes in neural development and oncogenesis.

Chapter 2 presents the transcriptome-wide survey of unproductive splicing events regulated by polypyrimidine tract-binding (PTB) proteins and dynamically regulated in neural development. Among the various splicing events, *SYNGAP1* exon 11 A3SS stands out as the most significant. Sequentially, this chapter describes the investigation of its regulation mechanism, evolutionary aspects, functional implications, and genetic manipulation by antisense oligonucleotides (ASOs).

In Chapter 3, the focus shifts toward the development, validation, and practical application of DR-6mA-seq. This cutting-edge method utilizes Illumina sequencing at a base-resolution level and is antibody-independent and cost-effective, offering a genome-wide analysis of 6mA modifications in multiple species.

Chapter 4 provides a comprehensive summary and discussion of the pivotal discoveries presented in this thesis. It also offers insights into potential inspirations for future studies, the

application prospects of the developed techniques, and an in-depth exploration of the limitations encountered during the research. Furthermore, this chapter raises thought-provoking questions that require further investigation to expand our understanding of the subject matter.

Chapter 2

Unraveling the Role of Unproductive Pre-mRNA Splicing in Brain Development: A Focus on *SYNGAP1* A3SS

2.1 Introduction: pre-mRNA alternative splicing in brain development

Pre-mRNA alternative splicing is a critical process that increases the diversity of transcriptomic and proteomic profiles while also post-transcriptionally regulating mRNA levels. It is highly prevalent in developing brains and significantly impacts various stages of nervous system development, including cell fate determination, neuronal migration, axon guidance, and synaptogenesis.[88] In fact, as cells advance along the neuronal lineage, alternative splicing patterns undergo significant changes. These alternative splicing events, especially those unproductive ones, precisely control the expression of neural-enriched alternative splicing factors, neural development regulators, and neuronal components.

The brain-development-related alternative splicing events are mostly modulated by the alterations in the expression of specific RBPs, including PTBP1, PTBP2, RBFOX1, RBFOX2, SRSF2, SBM4, NOVA1, NOVA2, nSR100, and so on (Figure 2.1A).[88,89] Taking PTB proteins as an example, a significant shift occurs in the predominant expression pattern during the differentiation of progenitor cells into mature neurons, transitioning from PTBP1 to its neuronal homolog, PTBP2 (nPTB). This transition plays a crucial role in the transformation of neural stem cells (NSCs) into neurons (Figure 2.1B).[89,90] Notably, PTBP1 functions by suppressing the inclusion of exon 10 in *PTBP2*, resulting in the generation of an mRNA variant that skips exon 10 and is susceptible to NMD.[91] Consequently, PTBP1 acts as a suppressor for the expression level of PTBP2 in non-neuronal cells and neural progenitor cells (NPCs). PTBP1 and PTBP2 share

similar domain organization and exert control over overlapping yet distinct sets of splicing events.[92] While PTBP1 inhibits the splicing of specific neural targets, thereby impeding neuronal differentiation, PTBP2 expression increases during the differentiation of neuronal cells and activates certain neural targets that facilitate the differentiation process (Figure 2.1B).[93] However, as cells mature and undergo synaptogenesis, PTBP2 is subsequently downregulated. This sequential downregulation of PTBP1 and PTBP2 plays a vital role in two transitions of splicing regulation throughout neuronal differentiation and maturation, as well as the functional expression of PSD-95 (*DLG4*), a vital postsynaptic scaffolding protein in excitatory neurons, through the splicing control at its exon 18 (Figure 2.1B).[94,95]

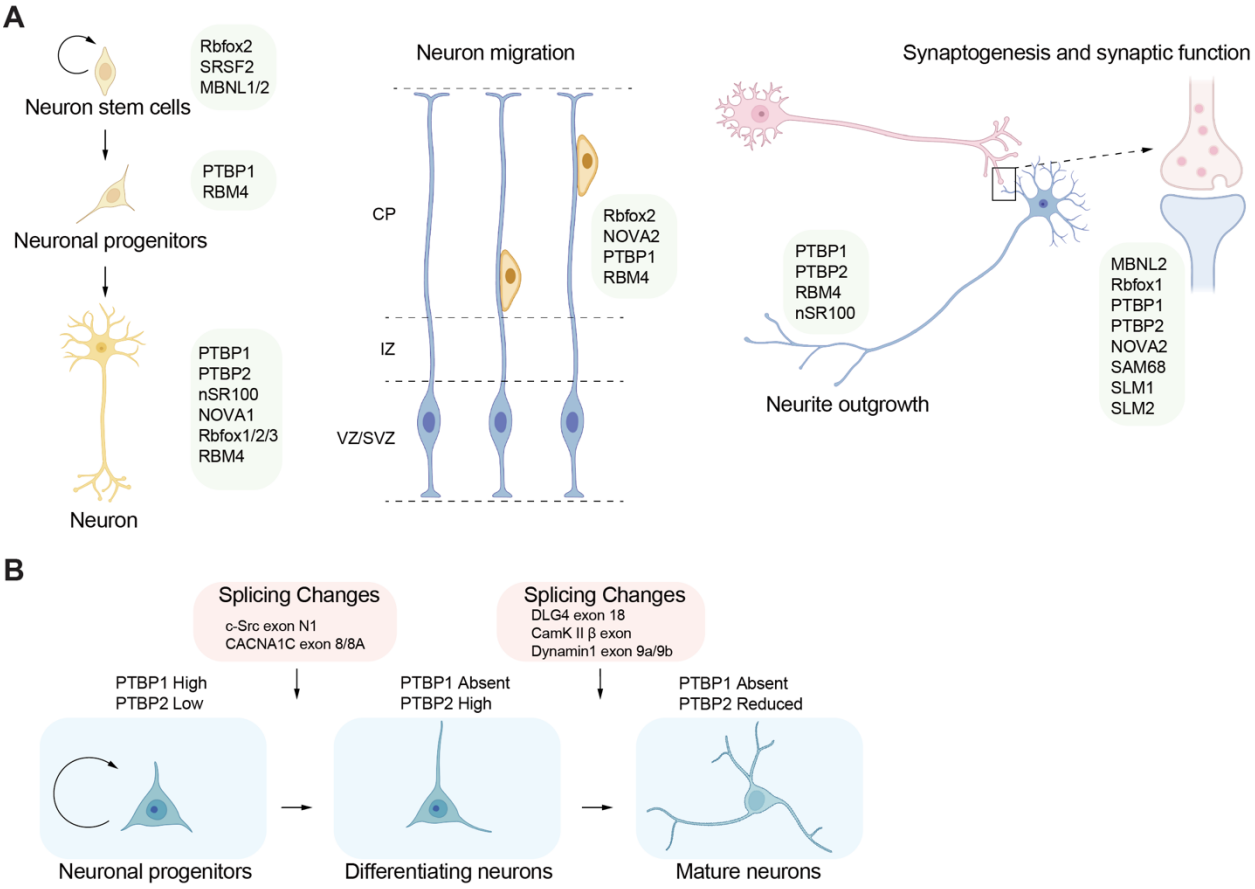


Figure 2.1 The critical splicing regulators in brain development

(**Figure 2.1, continued**) **A)** The schematic diagram showing the representative splicing factors (depicted in green boxes) that play crucial roles in various processes during brain development, encompassing self-renewal and fate determination of neural stem cells, neuronal cell differentiation, neuron migration during corticogenesis, synaptogenesis, and neurite outgrowth. **B)** Changes in the expression of PTBP1 and PTBP2 proteins define three splicing regulatory states during brain development and facilitate neuronal differentiation.

The dysregulation of these splicing events is associated with various brain developmental disorders. The most striking evidence is that the conditional knockout or haploinsufficiency of many neural-enriched alternative splicing factors, including PTBP1 and nSR100, all leads to developmental defects such as hydrocephaly and autistic-like phenotypes in mice.[96,97] There is compelling evidence demonstrating the impact of intronic mutations in neural genes on aberrant splicing, as evidenced by their discovery in patients with various brain developmental disorders. One such example is the c.1429+182 G>A mutation identified in *FLNA*, which results in a cell-type- and tissue-specific partial LOF of this actin-binding protein and gives rise to an atypical periventricular nodular heterotopia (PVNH) syndrome in affected patients.[98] The presence of heterozygous mutations, such as c.277dupA and c.1672+2T-C, in *DLG4* has been reported in patients with autosomal dominant intellectual developmental disorder-62 (MRD62). This mutation has been predicted to induce aberrant splicing, leading to a frameshift and premature termination of the protein.[99,100]

In recent years, significant progress has been made in leveraging advanced technologies such as RNA-Seq and single-cell RNA-Seq to explore alternative splicing events within the transcriptome. These breakthroughs have provided us with the means to investigate splicing patterns under varying conditions or in diverse tissues. To identify and quantify differentially splicing events from (single-cell) RNA-Seq data, diverse computational approaches have been developed.[101] By uncovering cryptic splicing events, we gain valuable insights into the intricate molecular mechanisms underlying brain development.

2.2 Results

2.2.1 Transcriptome-wide survey of differential and unproductive splicing events during brain development

Alternative splicing can result in the exclusion of a canonical exon or the inclusion of an extra exon, leading to a PTC in the transcript. When a PTC is located over 55 bp away from a downstream exon-exon junction, it triggers mRNA degradation during protein translation.[102] To gain a comprehensive understanding of the unproductive splicing events that introduce PTCs and their involvement in brain development, we treated E14 DIV1 (embryonic day 14, 1 day *in vitro*) mouse cortex primary neurons isolated on with the translation inhibitor cycloheximide (CHX). This treatment was expected to increase the abundance of unproductive splicing isoforms in the CHX-treated RNA library compared to the control group treated with DMSO. We performed bulk RNA-Seq on these samples and conducted the differential splicing analysis.

Simultaneously, we analyzed long-read RNA-Seq data obtained from the E11.5 (embryonic day 11.5) and E18 (embryonic day 18) mouse central neural system (CNS) to capture differentially spliced (DS) isoforms during neuronal development. Using DS analysis, we identified more than 500 isoforms that were both classified as unproductive transcripts that are subject to NMD and differentially spliced during neuronal maturation (p-value <0.001, |IncLevelDifference|>0.1) (Figure 2.2A). To functionally categorize these unproductive splicing events, we performed Gene Ontology analysis for the DS isoforms identified in the CHX-treated and control RNA-Seq datasets (Figure 2.2B). Z-scores were computed by evaluating the fold change in exon inclusion rates between E11.5 and E18 CNS, thereby offering insights into the enrichment of isoforms within certain clusters during early or late brain development stages. Negative values indicated more genes in a cluster undergoing unproductive splicing in the early

brain developmental stage (E11.5), while positive values suggested that genes in a cluster tended to undergo unproductive splicing in the late brain developmental stage (E18) (Figure 2.2B).

To understand the regulatory factors facilitating these unproductive splicing events during development, we analyzed the binding sites of over 100 known RNA-binding proteins (RBPs) around the differentially spliced regions (Figure 2.2C). Splicing events were categorized into five major types, and motif enrichment scores were determined in different regions adjacent to the affected splice sites. We identified motifs of 37 RBPs that were enriched in the vicinity of these DS exons (p-value <0.001), promoting either exon inclusion (pink) or exclusion (green) (Figure 2.2C).

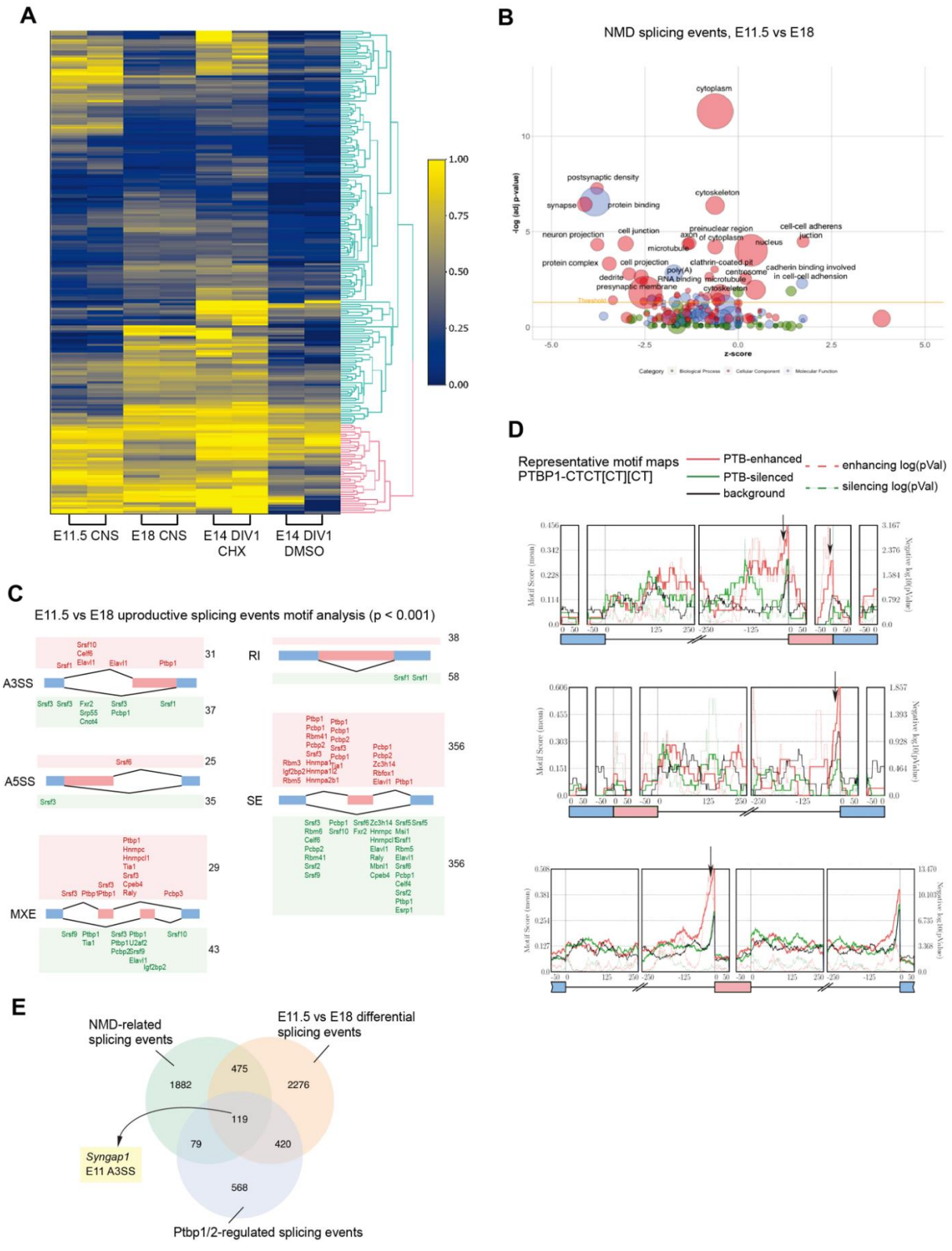


Figure 2.2 RNA-Seq reveals dynamic unproductive alternative splicing regulated by Ptbp1 in mouse brain development

(**Figure 2.2, continued**) **A)** Heatmap showing the inclusion ratios of differentially spliced isoforms in brain developmental stages (E11.5 vs. E18 mouse CNS) and the mRNA isoforms that are subject to NMD (CHX-treated vs. DMSO-treated E14 neuron culture). The inclusion ratios of alternative exon/intron are indicated by color. **B)** Bubble plot showing the Gene Ontology enrichment of unproductive splicing events, which differentially occur in E11.5 vs. E18 mouse CNS. The top clusters are annotated. The size of the bubble indicates the number of genes in certain cluster. The Z-scores indicate whether the biological process/molecular function/cellular components is more likely to have less (negative value) or more (positive value) unproductive gene isoforms in E18 stage compared to E11.5 stage. **C)** The motif analysis revealing the potential splicing regulators of E18-enriched unproductive splicing events. The numbers on right count the differential splicing events. **D)** The motif maps of the top candidate regulator, Ptbp1. The arrows point to the regions highly enriched with Ptbp1-binding motifs near either the enhanced or silenced splicing sites. **E)** Venn diagram showing the overlapped splicing events among differentially spliced isoforms in early vs. late brain developmental stages, CHX-treated vs. control, and Ptbp1/2 double knockdown vs. control RNA-Seq datasets.

2.2.2 Ptbp1 is a top splicing regulator of differential unproductive splicing events in brain development

Based on our data, we observed that Ptbp1 is the key splicing regulator, potentially controlling alternative splicing choices of numerous unproductive isoforms during brain development. For example, the Ptbp1 binding motif CTCT[CT][CT] was enriched in the 3' splice site of these DS exons (Figure 2.2D). Notably, the expression of PTBP1 itself dynamically changed during development, likely due to self-regulation involving an unproductive splicing event in one of its exons (Figure 2.1B). To gain further insight into the splicing events controlled by Ptbp1, we used lentivirus-mediated delivery of short hairpin RNA (shRNA) to knock down Ptbp1 or/and its homolog Ptbp2 in Neuro2a, a mouse neural crest-derived cell line that has been extensively used to study neuronal differentiation.[103] PTBP1 and PTBP2 have redundant functions in regulating pre-mRNA splicing and repressing each other's expression through unproductive splicing.[92] We observed that double knockdown of Ptbp1 and Ptbp2 resulted in a significant depletion of most PTBP proteins compared to single knockdown of PTBP1 or PTBP2 (Figure 2.5B). By performing DS analysis using scramble and double knockdown data, we identified over 1100 isoforms that

exhibited splicing patterns predicted to be regulated by Ptbp1/Ptbp2 with high confidence (P-vaule <0.001, |IncLevelDifference|>0.1). Notably, more than 100 isoforms were found to be overlapping DS events in all three RNA-Seq datasets, suggesting that these unproductive splicing events are differentially regulated during early and late brain developmental stages by PTB proteins (Figure 2.2E).

With a comprehensive overview of unproductive splicing events in the mouse brain, the question of the specific role these events play in neuronal development becomes critical. Deciphering the significance of producing numerous "useless" transcripts in terms of energy and material costs for cells requires understanding whether unproductive splicing serves as a post-transcriptional regulatory pathway to facilitate protein differential expression (DE) during brain development. To explore this function, we focused on a top unproductive DS exon, alternative 3' splice site (A3SS) of *Syngap1* exon 11, as an example (Figure 2.2E).

2.2.3 Alternative 3' splicing event regulates SYNGAP1 expression

Syngap1 is a critical synaptic Ras GTPase-activating protein (GAP) that plays a significant role in neurons.[104] Our analysis of mouse CNS RNA-Seq data revealed the utilization of a cryptic alternative 3' splice site (A3SS) located 207 bp upstream of the canonical splice site in mouse *Syngap1* intron 10 during early developmental stages (E11.5). In contrast, the canonical 3' splice site is predominantly used by the spliceosome during late stages (E18), resulting in the removal of the additional 207 bp cryptic exon. To gain insights into the spatiotemporal dynamics of this A3SS and its conservation in humans, we obtained RNA-Seq data from the ventricular zone (VZ) and cortical plate (CP) of both E14.5 mouse cortex and human fetal cortex (Figure 2.3A-B). Consistent with our previous findings, we observed significant inclusion of *Syngap1* exon 11 A3SS in mRNA from neural progenitor cells in the VZ, while it was excluded from neuronal mRNA in the CP, in

both species. This suggests that the splicing of *Syngap1* exon 11 A3SS is not only temporally regulated but also spatially controlled during brain development. Additionally, we performed semi-quantitative RT-PCR using human fetal brain cDNA and mouse dorsal cortex cDNA to confirm these observations (Figure 2.3C-D). The exclusion and inclusion isoforms of *SYNGAP1/Syngap1* A3SS were simultaneously amplified, represented by the top and bottom bands in the results, respectively. A striking splicing switch from a high percent splicing-in (PSI) to a low percent splicing-in was observed when comparing early developmental stages (GW15 for human; E12 for mouse) with late developmental stages (GW18 for human; P40 for mouse). Similar splicing switches were also observed when comparing the A3SS inclusion ratio between human VZ tissue and CP tissue from the same brain.

We further examined the splicing pattern of *Syngap1* exon 11 in non-central nervous system (non-CNS) tissues. Analysis of ENCODE/CSH Long RNA-Seq panels using the UCSC Genome Browser revealed predominant insertion of the A3SS of *Syngap1* exon 11 in non-CNS such as the heart, kidney, lung, limb, ovary, and testis (Figure 2.3E). In contrast, this A3SS was almost entirely spliced out in adult mouse cortex and frontal lobe mRNAs (Figure 2.3E). Collectively, these findings indicate that *Syngap1* exon 11 A3SS is predominantly included, resulting in a longer exon 11, in non-neural tissues and early developing fetal brain or brain regions enriched with neural progenitor cells (NPCs). As more mature neurons emerge, the A3SS is predominantly excised during late embryonic stages and in the adult brain, leading to the production of the canonical coding exon 11.

To explore the molecular function of this alternative splicing event in *Syngap1/SYNGAP1*, we examined the mRNA sequences and identified an in-frame stop codon (TGA) within exon 11 A3SS in both human and mouse (Figure 2.4C). This premature stop codon resides in the coding

region of the RasGAP domain and is expected to either trigger nonsense-mediated mRNA decay (NMD) or produce a truncated Syngap1/SYNGAP1 protein with a disrupted core domain. (Figure 2.3H) To distinguish between these possibilities, we employed two methods to block the NMD process. The brief treatment of Neuro2a cells with CHX, the protein translation inhibitor, resulted in the accumulation of the A3SS-harboring isoform, indicating NMD exemption for transcripts with premature termination codons. The results of semi-quantitative RT-PCR showed a marked increase in the abundance of the A3SS-harboring isoform after CHX inhibition, supporting the role of NMD in regulating this isoform (Figure 2.3F). Consistently, knockdown of Upf1, an essential RNA helicase involved in NMD, also led to an increase in the A3SS-harboring isoform compared to the A3SS-exclusion isoform (Figure 2.3F).[105]

Strikingly, despite the remarkable levels of the A3SS-harboring mRNA, we did not detect the hypothetical truncated Syngap1 protein in Neuro2a cells or early-stage fetal cortex (Figure 2.3G). These results suggest that the *Syngap1* A3SS-harboring isoform is subject to NMD, leading to the degradation of the unproductive mRNA isoform. Consequently, this post-transcriptional regulation serves as a mechanism for controlling the expression of the Syngap1 protein across different tissues and during brain development.

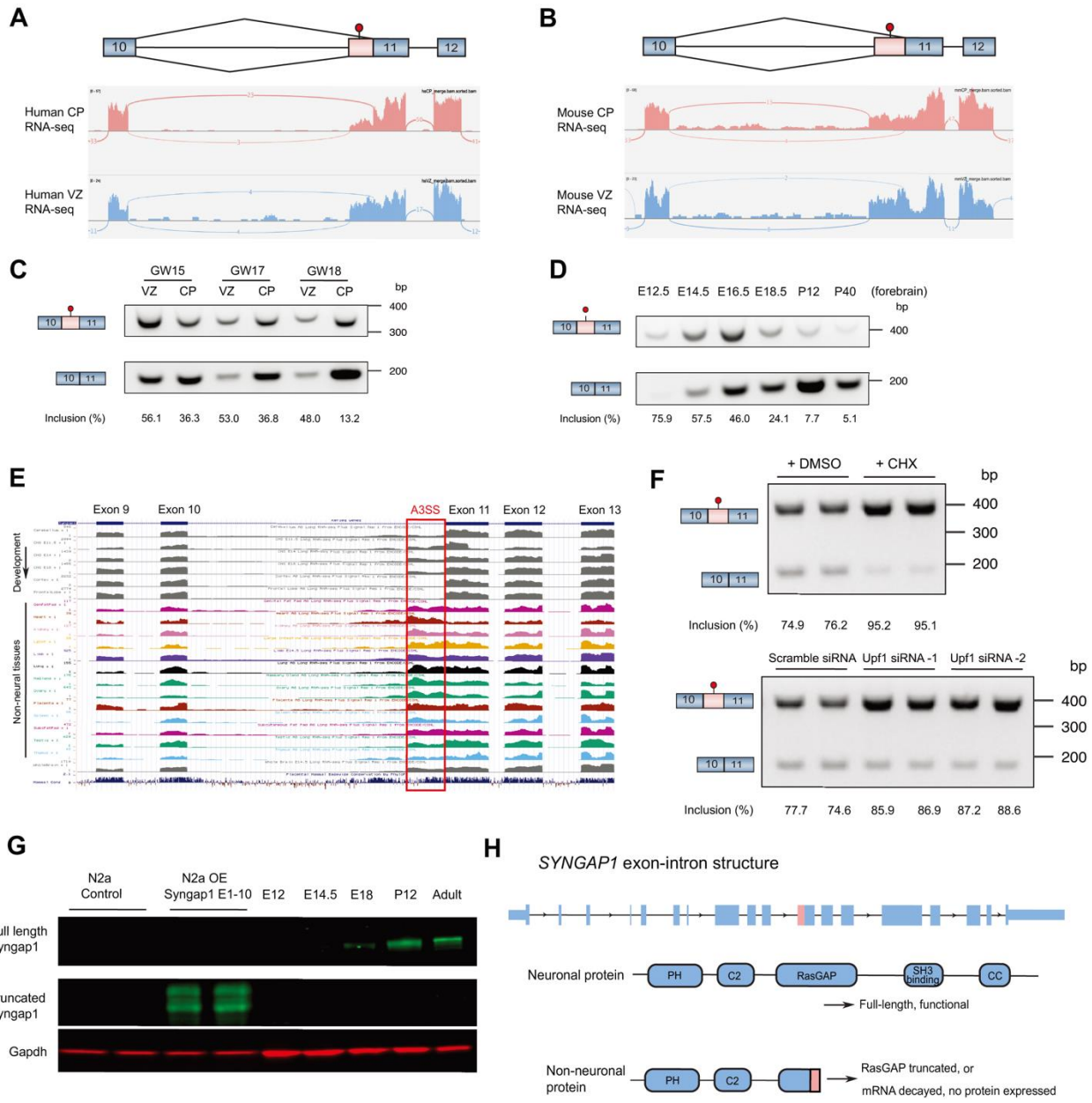


Figure 2.3 Human and mouse *SYNGAP1* exon 11 PTC-harboring A3SS are spliced-out specifically in neurons and subject to NMD

A)-B) The RNA-Seq data showing the splicing patterns of human *SYNGAP1* and mouse *Syngap1* E11 A3SS in ventricular zone (VZ) and cortical plate (CP). Exon–intron structures of the splicing isoforms are given on top. **C)-D)** The gel images of RT-PCR showing that human *SYNGAP1* and mouse *Syngap1* E11 A3SS are dynamically regulated during brian development. **E)** RNA-Seq data (ENCODE/CSHL) showing splicing patterns of *Syngap1* E11 A3SS in different brain developmental stages and across multiple tissues. **F)** The gel images of RT-PCR showing the splicing patterns of *Syngap1* A3SS in Neuro2a cells under CHX treatment and Upf1 knockdown. **G)** Western-blot showing Syngap1 protein expression increases during brain development with the putative truncated protein isoform undetectable. **H)** The schematic diagram illustrating the

(**Figure 2.3, continued**) molecular fate of *Syngap1* isoforms in neuron and non-neuronal cells. The PTC-harboring exon 11 A3SS is marked in pink.

2.2.4 *Syngap1* A3SS is conserved in mammals

Syngap1/SYNGAP1 is believed to have originated from gene and whole genome duplication events in the vertebrate lineage.[106] This raises the question of whether the unproductive splicing event involving exon 11 of *Syngap1/SYNGAP1* is conserved across vertebrate species. To address this question, we conducted a detailed examination of the sequential features and splicing patterns within *Syngap1* intron 10 in multiple vertebrate species. Multiple sequence alignment data of *Syngap1* genes were obtained from the UCSC Genome Browser, while RNA-Seq data revealing splicing patterns were obtained from the NCBI Genome Data Viewer. Our analysis revealed that while exon 10 and exon 11 are highly conserved among vertebrates, the length and structure of intron 10 vary significantly (Figure 2.4A). In mammalian *Syngap1* intron 10, a poly-pyrimidine region is located near the canonical 3' end, with shorter lengths observed in lower mammals (e.g., opossum) and longer lengths observed in higher mammals (e.g., mouse) (Figure 2.4A). Additionally, an ultra-conserved long intronic region specific to mammals is situated upstream of the alternative exon (A3SS). The regulatory function of these regions in facilitating the occurrence of the A3SS alternative splicing event will be explored in greater depth in subsequent sections (Figure 2.5). Furthermore, the premature stop codon (TGA) introduced by the A3SS in humans and mice is also conserved in other mammalian species. In contrast, none of these features are observed in intron 10 of non-mammalian vertebrates, such as fishes, frogs, and birds (Figure 2.4A). Consistently, the A3SS of *Syngap1* intron 10 was only detected in mammalian RNA-Seq data. Summarizing these observations, we can conclude that the A3SS harboring a PTC in *Syngap1/SYNGAP1* is mammal-specific.

After identifying the conservation of *Syngap1* exon 11 A3SS in mammals, we conducted multiple sequence alignment for ten representative mammalian species using the evolutionarily constrained region of *Syngap1* intron 10. This region spans from the mammalian ultra-conserved long intronic region to the canonical 3' splice site (Figure 2.4B). Interestingly, the genomic proximity of species on the gene or whole genome phylogenetic tree does not necessarily correlate with their proximity in terms of the evolutionarily constrained region of *Syngap1* intron 10. For example, the dog (*Canis lupus familiaris*) is genetically closer to the pig (*Sus scrofa*) than to the human (*Homo sapiens*); however, the intron 10 structure of the dog and human shows striking similarity, distinguishing them from the pig. This observation is consistent with the previous report that the constitutive and alternative regions in mammalian genes evolve in different ways.[107]

Another noteworthy observation is that in mammals, *Syngap1* intron 10 utilizes various "AG" sites for the A3SS, depending on the variable intron sequences, while the introduced PTC (TGA) remains in-frame and highly conserved across species (Figure 2.4C). For instance, in Figure 2.4C, all ten mammalian species examined possess the "NNNCAG" alternative splice site used by the opossum (*Monodelphis domestica*) and guinea pig (*Cavia porcellus*), while the other eight species have evolutionarily acquired and adopted other upstream cryptic splice sites. These subsequently emerged alternative splice sites are likely to be stronger than the primitive ones utilized by opossum and guinea pig. Interestingly, the ultra-conserved PTC (TGA) always maintains the correct reading frame, regardless of the specific A3SS utilized across species. This sequence variability and functional conservation indicate that although noncoding regions exhibit loose constraints, the mutations facilitating the *Syngap1* unproductive splicing event have been positively selected during evolution. This emphasizes the likely involvement of this splicing event in mammalian-specific adaptation, e.g. the formation of the unique brain features and functions.

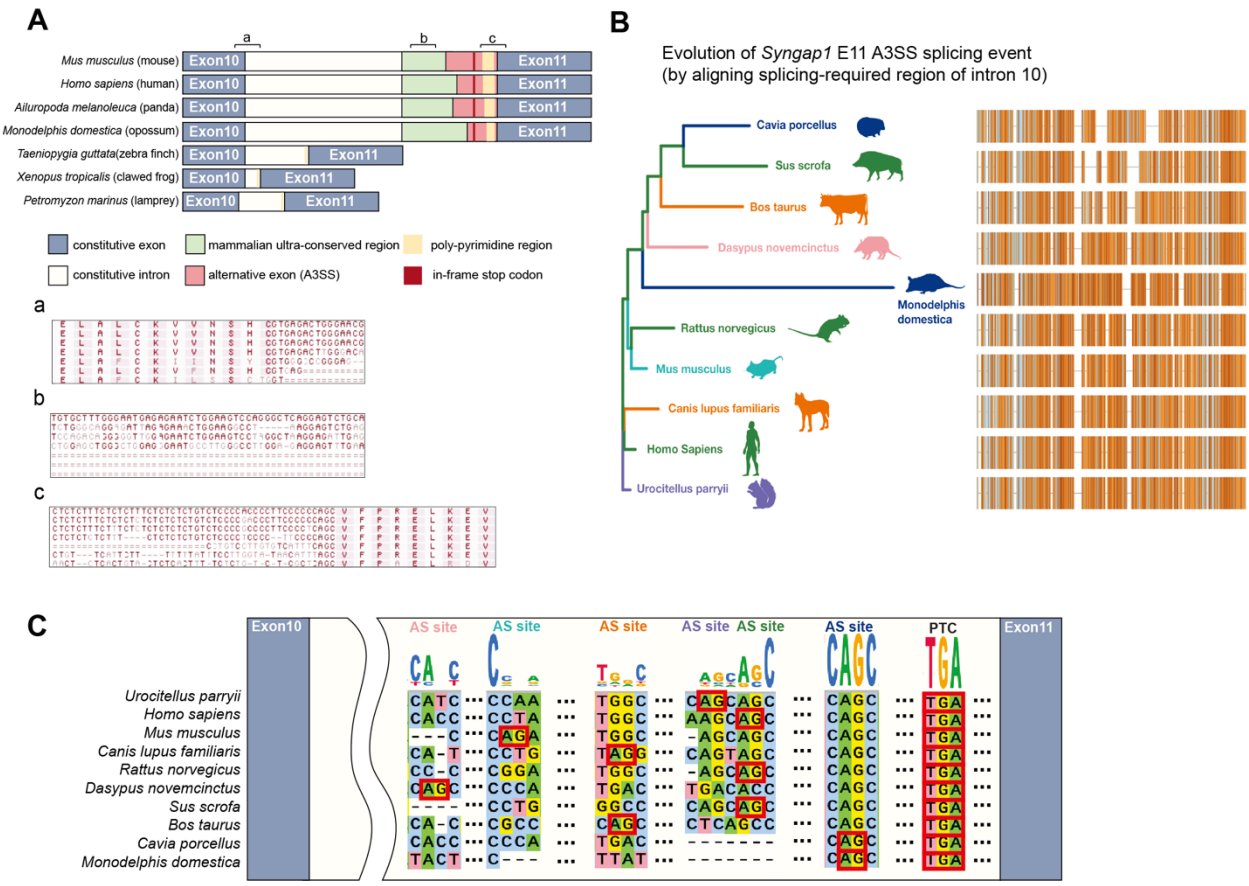


Figure 2.4 *SYNGAP1* unproductive splicing event is mammal-specific, variable in sequence but functionally conserved

A) Schematic diagram showing the multiple sequence alignments of *SYNGAP1* exon 10, intron 10 and exon 11, for seven representative vertebrates. a, b, and c zoom in and show the sequence alignment at the 5' splice site, mammal-conserved AS-required region in intron 10, and 3' splice site, respectively. **B)** Phylogenic tree built by aligning mammal-conserved AS-required region of *SYNGAP1* intron 10, with the alignment shown on the right. Mammals in the same colors use the same alternative splicing sites. **C)** Sequence alignment showing that the premature stop codon (TGA) in *SYNGAP1* A3SS is conserved in mammals while the splice acceptor sites (AG) are variable. The AG sites were annotated according to RNA-Seq results of corresponding species in NIH Genome Data Viewer.

2.2.5 PTBP1/2 regulates *SYNGAP1* splicing by interfering the function of spliceosome

How does this mammalian-specific alternative 3' splicing event occur, and how is it controlled spatiotemporally? Alternative splicing is often regulated by the binding of RNA binding proteins (RBPs) to cis-regulatory RNA sequences. Based on the analysis of Ptbp1/2 double-knockdown DS analysis results, we hypothesized that Ptbp1/2 may regulate the splicing choice at

Syngap1/SYNGAP1 intron 10. Consistent with the findings from RNA-Seq, semi-quantitative RT-PCR experiments confirmed that Ptbp1 and Ptbp2 double knockdown significantly reduces the insertion of *Syngap1* A3SS (Figure 2.5B). Considering that the RNA-binding protein hnRNP C has been reported to repress alternative exons by interfering with the recognition of core splicing factors[108], we investigated whether the competition between Ptbp1/2 and the 3' splice site recognition factor is involved in the regulation of *Syngap1* A3SS. To examine the binding pattern of Ptbp1/2 on *Syngap1* transcripts and their competition with the 3' splice site recognition factor, we analyzed published CLIP-Seq datasets for Ptbp1/2 and U2af65, a core splicing factor essential for 3' splice site recognition. Co-visualization of these datasets with mouse fetal brain RNA-Seq data revealed the binding of all three proteins at and near the *Syngap1* exon 11 A3SS (Figure 2.5A). Remarkably, the binding of U2af65 at the canonical 3' splice site of exon 11 overlapped with a strong peak of Ptbp1 binding, suggesting a potential competition between these two proteins at this site, leading to weakened splicing.

The competition between Ptbp1 (and Ptbp2) and U2af65 is possibly due to their similar binding motifs, as both proteins recognize polypyrimidine tracts. To investigate the detailed regulatory function of Ptbp1 on *Syngap1* A3SS, we first determined the precise positions of cis-regulatory elements that could potentially bind to Ptbp1 and U2af65. We generated multiple minigene constructs containing a genomic fragment of human *SYNGAP1* exon 9-12, with deletions of potential regulatory regions identified by CLIP-Seq and conserved regions across vertebrates (Figure 2.5C). These minigene constructs were transfected into Neuro2a cells, which are functionally similar to neuronal progenitors, and mRNA was subsequently harvested. Interestingly, the removal of either the mammalian ultra-conserved long intronic region (Site #1) or the potential U2af65 binding site (Site #2) upstream of *SYNGAP1* A3SS, as revealed by semi-quantitative RT-

PCR, resulted in the enhanced exclusion of this alternative exon from the transcripts. Conversely, the depletion of the potential U2af65/Ptbp1 competitive binding site (Site #3) upstream of the canonical 3' splice site increased the usage and inclusion of the A3SS (Figure 2.5C). These two potential U2af65 binding sites (Site #2 and #3) upstream of the 3' splice sites were both enriched with polypyrimidine tracts. To examine the affinity of PTBP1 for the U2af65 binding sites, we performed electrophoretic mobility shift assays (EMSAs) using PTBP1 protein and RNA probes corresponding to the upstream and downstream binding sites (Figure 2.5D). A noticeable band shift was observed with increasing amounts of PTBP1, and the shift decreased when cold competitors were added, suggesting that PTBP1 can bind to both sites. Interestingly, when adding cold probes of the other site as competitors, we observed that PTBP1 binding to the upstream site was completely abolished by the addition of the cold probe for the downstream site. In contrast, the band shift remained when using the cold probe of the upstream site to interfere with the binding of the upstream site. These results demonstrate that PTBP1 can tightly bind to the U2af65 binding site near the canonical intron 10 splice site, and it also exhibits a weaker binding to the alternative U2af65 binding site near the cryptic intron 10 splice site.

Based on the aforementioned findings, we propose a comprehensive model that elucidates the mechanism by which PTBP1 promotes the inclusion of *SYNGAP1* A3SS by competing with U2AF65 (Figure 2.5F). In cells characterized by high expression levels of PTBP1 and/or PTBP2, such as non-neural tissues and early embryonic neural tissues, PTBP1/2 competes with U2AF65 at the canonical splice site and exhibits a higher binding affinity. This competition effectively hinders U2AF65 from binding to the downstream polypyrimidine tract near the canonical splice site. Consequently, U2AF65 is compelled to recognize and bind to an alternative upstream cryptic splice site, which exhibits weaker affinity with PTBP1. This dynamic interaction ultimately leads

to the inclusion of the additional intronic fragment (A3SS) within the mature mRNA transcripts of *SYNGAP1*. Conversely, in cells lacking PTBP1/2, i.e., the mature neurons, the entire intron is successfully excised from *SYNGAP1* transcripts without interference from PTBP1/2 hindering U2AF65 recognition. Notably, similar splicing regulation mechanisms involving competitive binding between U2AF65 and other RBPs have been previously observed in other genes and elements, such as cardiac troponin T (cTNT) and *Alu* elements.[108,109]

Based on our proposed model, we conclude that PTBP1/2 exerts a post-transcriptional repressive effect on SYNGAP1 expression through its promotion of the inclusion of a PTC-harboring A3SS. This promotion is facilitated by PTBP1/2's competitive binding at the U2AF65-bound polypyrimidine tract. To further substantiate the impact of Ptbp1 in regulating Syngap1 expression, we conducted experiments involving the transfection of E14 DIV3 primary neurons derived from the mouse cortex. One week post-transfection, we observed a significant reduction in Syngap1 protein levels in neurons that re-expressed Ptbp1 when compared to untransfected neurons (Figure 2.5E). Collectively, these findings strongly support the conclusion that PTBP1/2 negatively regulates SYNGAP1 expression by promoting a mammalian-specific unproductive splicing event.

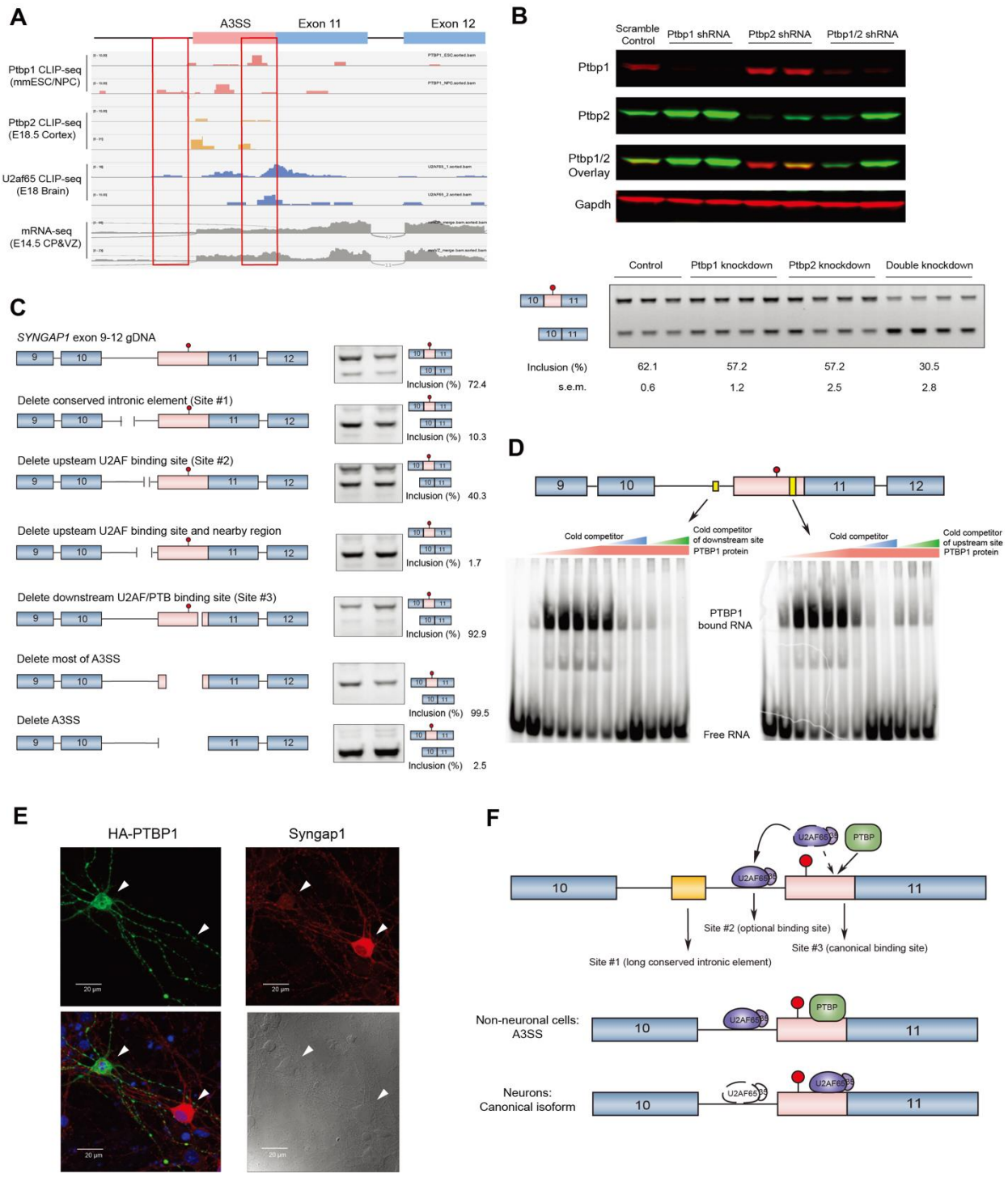


Figure 2.5 *Syngap1* unproductive splicing event is regulated by PTB proteins

A) CLIP-Seq analyses showing Ptbp1, Ptbp2, and U2af65 binding peaks in the A3SS region. **B)** Western blot results showing shRNA knockdown of Ptbp1 and Ptbp2 in Neuro2a cells. Two different shRNAs were used for each of Ptbp1 and Ptbp2. RT-PCR results showing that loss of Ptbp1/2 promoted splicing of the canonical/productive Syngap1 isoform. **C)** Minigene constructs

(**Figure 2.5, continued**) of the human *SYNGAP1* showing deletion of predicted splicing elements and RT-PCR results showing their effects on A3SS insertion (two replicates). Noticeably, the intronic element (site #1) and predicted upstream U2AF65-PTBP binding site (site #2) were required for A3SS inclusion; the predicted U2AF65-PTBP binding site (site #3) was required for canonical splicing and A3SS skipping. **D)** Electrophoretic mobility shift assays showing that PTBP1 binds strongly to the polypyrimidine region near the constitutive splicing site and also the upstream U2AF65 binding site. **E)** Immunofluorescence staining of E14 DIV11 primary cortical neurons showing that ectopic expression of PTBP1 (green) decreased Syngap1 protein level (red). **F)** Current working model: in neural progenitors and differentiating neurons, PTBP proteins bind to site #3 in A3SS (red) and suppress the canonical/neuronal 3' splice site; in neurons, PTBP proteins are turned down/off, and site #3 is exposed for splicing machinery (U2AF65) recognition, which promotes neuronal isoform expression.

2.2.6 Disrupting *SYNGAP1* splicing causes deficient neuronal development

After comprehensively elucidating the dynamic changes and regulatory mechanism of *Syngap1/SYNGAP1* exon 11 A3SS, we became intrigued by the significance of precise timing and controlling this unproductive splicing event in mammals. To address this question, I investigated the potential outcomes of constitutive inclusion or exclusion of the A3SS in *Syngap1/SYNGAP1* transcripts.

Permanent inclusion of the intron 10 A3SS results in nonsense-mediated mRNA decay (NMD) of *Syngap1/SYNGAP1* mRNA during the first round of translation, essentially abolishing protein expression. Mouse models with constitutive inclusion should exhibit phenotypes similar to those observed in *Syngap1* knockout mice. Homozygotes of these mice die a few days after birth, while heterozygotes display accelerated synapse maturation and cognitive impairments. Notably, haploinsufficiency of *SYNGAP1* due to loss-of-function (LOF) mutations is a dominant cause of non-syndromic intellectual disability (NSID) in the population. To explore whether mutations leading to aberrant exonization of the intron 10 A3SS or mutations within intron 10 itself could give rise to SYNGAP1-related NSID, we examined reported cases of causal mutations near *SYNGAP1* intron 10 splice sites. Interestingly, we identified two cases in which female patients exhibited similar phenotypes of mental retardation and epilepsy, possibly stemming from

disrupted splicing at *SYNGAP1* intron 10.[110,111] One NSID patient, aged 34, harbored a G to A mutation at the +5 position of the intron 10 splice donor site, while another 8-year-old patient carried a 9-bp deletion at the 3' splice site of intron 10 (Figure 2.6A-B). To investigate the impact of these mutations, we constructed minigenes for both mutations and transfected them into Neuro2a cells, subsequently harvesting RNA for analysis. Our RT-PCR results validated our hypothesis: the two mutations, respectively, led to the retention of the entire intron 10 and elevated A3SS inclusion. In both scenarios, the exonized intron 10 introduced a premature termination codon (PTC), triggering NMD of additional *SYNGAP1* mRNA molecules (Figure 2.6C). Therefore, the aberrant inclusion of *Syngap1/SYNGAP1* A3SS can give rise to parallel defects in brain development, akin to the insufficiency resulting from *Syngap1/SYNGAP1* haploinsufficiency.

After this, I sought to understand whether the constitutive exclusion of *Syngap1* A3SS affects mouse development by genetically deleting this NMD event without affecting the protein-coding isoform. Based on the results above (Figure 2.5C), I deleted the intronic elements and the splice acceptor site required for A3SS-NMD insertion with CRISPR-Cas9 and created a mouse strain named *Syngap1*-NISO (standing for Neuronal ISOform, N allele, 268 bp deletion). With the help from the collaborators, we observed that homozygous deletion of mouse *Syngap1* A3SS (N/N) led to a modest reduction in the magnitude of long-term potentiation (LTP), but it did not overtly impact behavioral performance in the Barnes maze or Rotarod assays (data not shown here). Also, we were unable to detect Syngap1 protein in E18.5 heart or lung tissues from *Syngap1* N/+ animals, and there was no significant cortical neurogenesis defect in *Syngap1* N/N mice (data not shown here). Nevertheless, we found that LTP was impaired in Syngap1 cKO/+ mice and rescued in Syngap1 cKO/N animals, indicating that deletion of the *Syngap1* A3SS- NMD alleviated the LTP deficits caused by the Syngap1 knockout allele (data not shown here). These findings indicate that

the *Syngap1* A3SS possesses the ability to suppress Syngap1 expression; however, it is not the sole mechanism responsible for the downregulation of Syngap1 in non-neuronal cells. Plus, the modestly elevated Syngap1 level in non-neuronal cells does not result in apparent brain developmental abnormalities in normal animals but will rescue *Syngap1* haploinsufficiency.

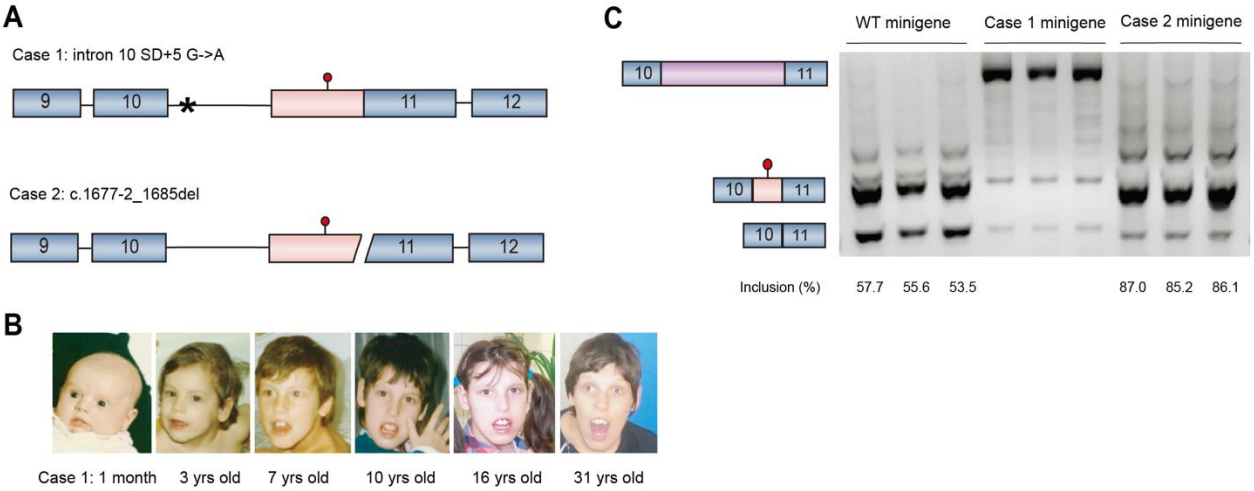


Figure 2.6 *SYNGAP1* intronic mutations cause intellectual disability in patients through unproductive splicing

A) Schematic diagram showing *SYNGAP1* exon 10-exon 11 gene structures of two human *SYNGAP1* intronic mutations identified in autism and ID patients. Case 1 is a 33-year-old female with severe intellectual disability, epilepsy, and autistic features. Case 2 is a 7-year-old female with severe intellectual disability, ASD and hypotonia. **B)** Facial photographs of Case 1 patient at different ages, modified from D. Prchalova *et al.* BMC Medical Genetics (2017). **C)** RT-PCR results of *SYNGAP1* minigene constructs transfected in Neuro2a cells showing that human *SYNGAP1* intronic mutations identified in autism and ID patients led to intron 10 retention (c.1676 +5 G>A, NM_006772.2) or abnormal A3SS inclusion (c.1677 -2_1685del) (three replicates).

2.2.7 SSOs suppress *SYNGAP1* A3SS-NMD in human iPSC- derived neurons

Single-stranded oligonucleotides (SSOs) have demonstrated successful advancements in the treatment of neurological disorders, such as spinal muscular atrophy, with attempts made toward personalized medicine.[112–114] Furthermore, they have shown potential for addressing haploinsufficiency in human diseases. Notably, the increased levels of Syngap1 protein resulting from the exclusion of *Syngap1* A3SS in NISO mice and the conserved *SYNGAP1* A3SS in human brains have prompted an investigation into the possibility of upregulating the productive isoform

of *SYNGAP1* in human cells through SSO-mediated suppression of *SYNGAP1* A3SS-NMD. SSOs incorporating 2'-O-methoxyethyl (2-MOE) chemistry have been successfully developed for splicing redirection, gene expression manipulation, and disease treatment.[115] In order to identify SSOs capable of suppressing human *SYNGAP1* A3SS-NMD, we performed ASO walk on the sequences within intron 10 that encompassed critical splice elements identified through serial deletion of the human *SYNGAP1* minigene (Figure 2.5C), predicted splicing regulatory sequences such as branch points, and predicted stem-loop structures and conserved sequences that overlapped with experimentally identified splice elements (Figure 2.7A).

Eleven SSOs (Supplemental Table 1) were synthesized using a phosphorothioate backbone with 2-MOE modified residues and subsequently evaluated in human induced pluripotent stem cells (iPSCs) (Figure 2.7A-B). Among these SSOs, CH933 and CH937 displayed the highest efficiency in suppressing A3SS-NMD inclusion in a human iPSC line (PGP1-iNGN) (Figures 2.7B). Subsequently, CH933 and CH937 SSOs, along with a scrambled control and a previously reported ASO71, were delivered to human iPSC-derived neurons.[116] The results indicated that CH937 exhibited the most significant decrease in A3SS-NMD inclusion (Figure 2.7C-D). Furthermore, the productive/functional *SYNGAP1* transcript was quantified using qPCR primers specific to the non-NMD isoform, revealing that CH937 induced a 2.5-fold increase in functional *SYNGAP1* mRNA (Tukey's multiple comparisons tests, adj.p-value <0.001, Figure 2.7E). In contrast, the ASO71 developed in HEK293 cells exhibited lower efficiency in upregulating *SYNGAP1* transcript in human iPSC-derived neurons (1.5-fold, adj.p-value =0.029, Figure 2.7E). These findings demonstrate that the leading SSO, CH937, effectively suppresses human A3SS-NMD and elevates the level of canonical *SYNGAP1* transcripts in both human iPSCs and iPSC-derived neurons.

Furthermore, we conducted additional investigations to assess the impact of SSO CH937 on two additional control human iPSC lines (NA19101 and 28126). The results confirmed that CH937 exhibited greater effectiveness than ASO71 in reducing *SYNGAP1* A3SS inclusion (Figure 2.7F-K). Importantly, CH937 significantly increased the levels of functional *SYNGAP1* transcript to 6.3-fold and 3.6-fold of the non-treated controls in NA19101 and 28126, respectively (Figure 2.7H-K). Subsequently, we administered CH937 to an iPSC line derived from a SYNGAP1 patient carrying a heterozygous frameshift mutation (Lys114SerfsX20). The findings demonstrated that SSO CH937 decreased *SYNGAP1* A3SS inclusion (Figure 2.7L-M) and significantly increased the levels of functional *SYNGAP1* transcript to 2.6-fold of the non-treated controls (Figure 2.7N). These results provide support for the effectiveness of the lead SSO CH937 in suppressing *SYNGAP1* A3SS and increasing the levels of functional SYNGAP1 isoform in SYNGAP1 patient-derived iPSCs.

To investigate whether suppression of *SYNGAP1* A3SS can lead to an upregulation of SYNGAP1 protein, we evaluated the effects of SSO CH937 in iPSC-derived brain organoids. Cerebral organoids were generated from human iPSCs (28126) following a well-established protocol.[117] Subsequently, they were transfected with CH937 on post-induction days 133, 135, and 137, and harvested on day 139 for protein analysis (Figure 2.7O-P). The results showed that CH937 significantly increased SYNGAP1 protein levels compared to control organoids (83% ± 28%, p-value < 0.05 by unpaired t-test). Taking together, we identified an SSO, CH937, which effectively upregulates SYNGAP1 protein expression in human iPSC-derived cerebral organoids by suppressing the insertion of *SYNGAP1* exon 11 A3SS.

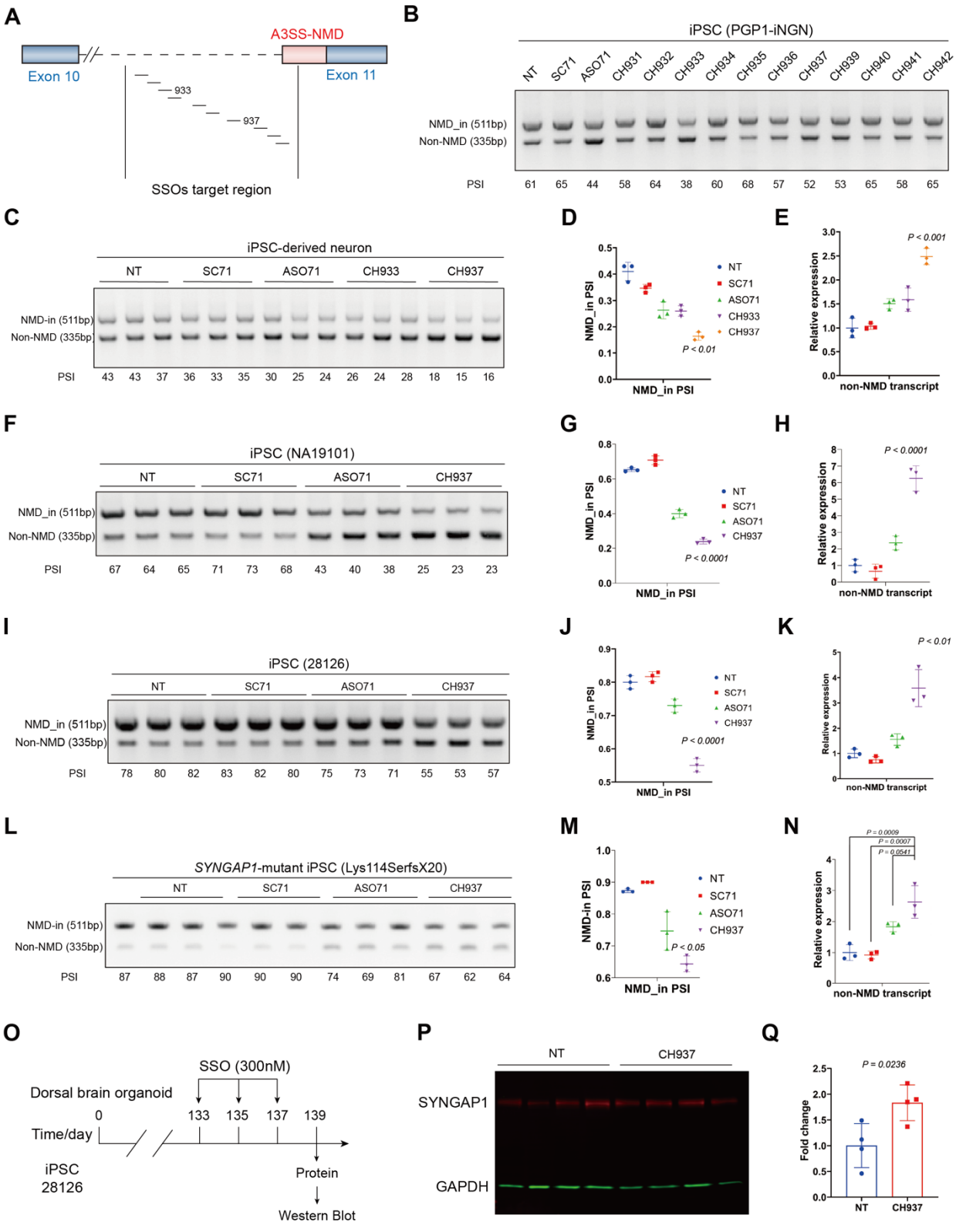


Figure 2.7 The lead SSO upregulates SYNGAP1 expression in human iPSCs and iPSC-derived neurons

(**Figure 2.7, continued**) **A)** Schematic illustration of the SSO design targeting the *SYNGAP1* A3SS. **B)** RT-PCR results showing the screening of SSOs in iPSCs (PGP1-iNGN). SC71 (scrambled control) and ASO71 were reported before.28 One biological sample per lane. **C)**–**E)** Identification of the lead SSO in iPSC-derived neurons. RT-PCR results (C) and quantification (D) showing that CH937 suppresses *SYNGAP1* A3SS in iPSC- derived neurons. Q-PCR results (E) showing that the productive *SYNGAP1* transcript was upregulated in CH937-treated human iPSC-derived neurons. **F)**–**K)** The lead SSO suppresses *SYNGAP1* A3SS in two additional human iPSC lines. RT-PCR results (F and I) and quantification (G and J) showing that CH937 suppressed *SYNGAP1* A3SS in human iPSCs (NA19101 and 28126). Q-PCR results (H and K) showing that the CH937 significantly increased the productive *SYNGAP1* transcript levels in human iPSC lines. **L)**–**N)** The lead SSO suppresses *SYNGAP1* A3SS-NMD in SYNGAP1 patient-derived iPSCs. RT-PCR results (L) and quantification (M) showing that CH937 suppressed *SYNGAP1* A3SS in SYNGAP1-patient-derived iPSCs (333del, Lys114SerfsX20). Q-PCR results (N) showing that the CH937 significantly increased the productive *SYNGAP1* mRNA level. **O)**–**Q)** Application of CH937 to human iPSC-derived cerebral organoids (O) led to increased SYNGAP1 protein expression (83% ± 28%, P and Q). P-value < 0.05 by unpaired t-test.

2.3 Discussion and Conclusion

Unproductive splicing events play a role in the selective degradation of mRNA transcripts through spatiotemporally controlled alternative splicing. This mechanism has been suggested as a means of regulating the expression of specific proteins, such as PSD-95, during neural development.[94] In this chapter, I conducted a comprehensive survey and cataloged numerous unproductive splicing events that trigger nonsense-mediated decay in the brain. Interestingly, in addition to previously reported synaptic proteins, I found that proteins involved in cell-cell adhesion undergo unproductive splicing in later stages of neural development. This novel observation warrants further investigation.

In the subsequent sections of this chapter, I focused on SYNGAP1, a synaptic protein associated with mental retardation. Different SYNGAP1 protein isoforms exist and are generated through alternative splicing and alternative promoter usage.[118] One of the promoters of *SYNGAP1* is likely shared with the neighboring gene, *CUTA*, which is widely expressed in multiple tissues (The Eukaryotic Promoter Database). However, SYNGAP1 protein is hardly detectable or weakly

expressed in non-neuronal tissues, indicating a missing link between the ubiquitous transcription and neuron-specific protein expression. We demonstrated that mammalian *Syngap1* genes adopt an unproductive splicing event to reduce transcript abundance and protein expression in non-neuronal tissues. Furthermore, we showed that this AS-NMD-dependent regulation is positively selected during mammalian evolution and likely plays a beneficial and critical role in brain development. Interestingly, perpetuation or abolishment of the alternative splicing event leads to defective neural development, albeit with varying severity.

A striking coincidence is the remarkably similar unproductive splicing of PSD-95, which directly interacts with SYNGAP1 in the postsynaptic density (PSD). This splicing event is regulated by PTBP1/2 and occurs in early brain developmental stages as well as in the non-neural tissues.[94] The occurrence of these two splicing events in the major components of the PSD also shows evolutionary concurrence. Previous studies have shown that SYNGAP1 and PSD-95 exhibit a near stoichiometric ratio in the PSD, and this protein ratio is critical for the formation of liquid-like droplets containing SYNGAP1 and PSD-95.[119] It can be hypothesized that abolishing either of these two NMD-splicing events may disrupt the protein ratio and local concentration, thereby affecting the formation of the SYNGAP1/PSD-95 liquid phase. This may explain the coexistence of these two unproductive splicing events in mammals.

In addition to elucidating the functional significance of developmentally regulated and neuron-specific alternative splicing programs, our research also lays the foundation for utilizing splicing-switch oligonucleotides (SSOs) to treat a wide range of haploinsufficiency-related diseases. As an example, loss-of-function mutations in one allele of *SYNGAP1* lead to mental retardation in humans, often accompanied by epilepsy and autism. We propose that targeting *SYNGAP1* intron 10 with SSOs could increase the expression of SYNGAP1 from the unaffected

allele by skipping the NMD-associated alternative 3' splice site (A3SS-NMD), thereby rescuing SYNGAP1 haploinsufficiency during brain development. Our lead SSO, CH937, effectively suppressed *SYNGAP1* A3SS in human induced pluripotent stem cells (iPSCs) and iPSC-derived neurons, resulting in a significant increase in functional SYNGAP1 isoforms. CH937 also enhanced SYNGAP1 protein expression in cerebral organoids derived from two different human iPSC lines.

Despite these promising findings, several key questions remain to be addressed. Firstly, it is crucial to determine the specific time points and brain regions where upregulation of Syngap1 is sufficient to rescue haploinsufficiency. Understanding whether deleting or suppressing the A3SS-NMD exon at postnatal stages can rescue the heterozygous knockout allele is of utmost importance. Additionally, it is essential to systematically characterize neural circuits and mouse behaviors, as SYNGAP1 patients exhibit a wide range of symptoms.[120] Secondly, it is necessary to investigate whether the SSO can effectively increase Syngap1 protein levels *in vivo* and rescue the heterozygous knockout mouse. Although the lead SSO CH937 show promise as a therapeutic target, rigorous *in vivo* studies are required to fully comprehend their therapeutic potential.

Hundreds of genes have been reported to undergo alternative splicing followed by nonsense-mediated decay (AS-NMD) during cortical development, with significant enrichment of chromatin regulators.[121] These AS-NMD exons are regulated by splicing factors, and mutations in the host genes are frequently associated with neurodevelopmental disorders, including autism and epilepsy.[89,98,122,123] While AS-NMD exons have shown promise as therapeutic targets for diseases such as Dravet Syndrome, the biological functions of these AS-NMD events remain unclear.[114] Furthermore, comprehensive studies using animal models and human neurons are necessary to fully explore the dosage effects of their host genes and their therapeutic potential. Genetic

investigations of AS-NMD exons will shed light on their organismal functions and contribute to the development of therapeutic targets and strategies.

In summary, this study conducted a thorough genomic analysis of unproductive splicing transcripts that trigger nonsense-mediated decay, significantly expanding our understanding of the widespread and crucial role of alternative splicing in modulating protein expression dynamics. By focusing on the specific case of *SYNGAP1*, we have provided detailed insights into the precise regulation of unproductive splicing events by RNA-binding proteins, which exhibit evolutionary significance and exert a significant impact on neuronal differentiation and functioning. We anticipate that the dissection and manipulation of these splicing "switches" for protein expression will have profound implications for unraveling the development of complex structures within the mammalian brain and hold potential for the treatment of various neurological and other pathological conditions.

2.4 Methods

2.4.1 Materials and resource

Table 1 Reagents and resources for 2.4

Reagent/Resource	Source	Identifier
Anti-Syngap1 (C terminal)	Thermo-Fisher	Cat# PA1-046
Anti-Syngap1 (N terminal)	Thermo-Fisher	Cat# PA5-37909
Anti-HA	Roche	12158167001
Anti-Ptbp1	Abcam	Cat# ab5642
Anti-Ptbp2	Millipore	Cat# ABE431
Anti-Sox2	Santa Cruz	Cat# sc-17320
Anti-Rbfox1	Millipore	Cat# MABE159
Anti-Gapdh	Millipore	Cat# AB2302
Cycloheximide	Sigma-Aldrich	Cat# C4859-1ML
TnT SP6 High-Yield Wheat Germ Protein	Promega	Cat# L3260

(**Table 1**, continued) **Expression System**		
Quick-RNA MiniPrep	Zymo	Cat# R1054
GeneJET Genomic DNA Purification Kit	Thermo-Fisher	Cat# K0721
TruSeq RNA Library Prep Kit	Illumina	Cat# RS-122-2001
Alt-R Genome Editing Detection Kit	IDT	Cat# 1075931
SuperScript IV First-Strand Synthesis System	Thermo-Fisher	Cat# 18091050
Novex 4-12% Tris-Glycine Mini Gels	Thermo-Fisher	Cat# XP04122BOX
Novex TBE Gels, 8%	Thermo-Fisher	Cat# EC6215BOX
Gibson Assembly Master Mix	NEB	Cat# E2611S
Quick Ligation Kit	NEB	Cat# M2200S
NotI-HF	NEB	Cat# R3189L
AscI	NEB	Cat# R0558L
Phusion Hot Start II DNA Polymerase	Thermo-Fisher	Cat# F537S
FuGENE HD Transfection Reagent	Promega	Cat# E2311
Lipofectamin 2000 CD Transfection Reagent	Thermo-Fisher	Cat# 12566014
Neuro2a	ATCC	Cat# CCL-131
STAR 2.7	PMID: 23104886	https://github.com/alexdobin/STAR
rMATS 3.2.5	PMID: 25480548	http://rnaseq-mats.sourceforge.net/rmats3.2.5/
rMAPS2	PMID: 27174931	http://rmaps.cecsresearch.org/
CLIPSeqTools	PMID: 26577377	http://mourelatos.med.upenn.edu/clipseqtools/
MAFFT v7	PMID: 28968734	https://mafft.cbrc.jp/alignment/server/
GOPlot	PMID: 25964631	https://wencke.github.io/
NMD RNA-Seq	This paper	NCBI(PRJNA930469:SRR23308049,SRR23308050)

(**Table 1**, continued) **PTB RNA-Seq**	This paper	NCBI(PRJNA930469:SRR23308049,SRR23308050)
Mouse CNS RNA-Seq rare data	Encode CSHL Long RNA-Seq	http://genome.ucsc.edu/cgi-bin/hgFileUi?db=mm9&g=wgEncodeCshlLongRnaSeq
PTBP1/2, U2AF65 CLIP-Seq rare data	ArrayExpress https://www.ebi.ac.uk/arrayexpress/	SRR2121761, SRR2121762; SRR871026, SRR871030; ERR208893, ERR208897
Mouse tissue RNA-Seq	UCSC Genome Browser UCSC Genome Browser on Mouse July 2007 (NCBI37/mm9)	CSHL Long RNA-Seq
Vertebrate RNA-Seq	NIH Genome Data Viewer	RNA-Seq tracks for each species
Human mutation	PMID: 30541864 PMID: 28576131	
CRISPOR	PMID: 27380939	http://crispor.tefor.net/crispor.py
Nucleotide synthesis	IDT	https://www.idtdna.com/

2.4.2 Cell culture

Mouse Neuro-2a (N2a) neuroblastoma cells (ATCC® CCL-131™) were purchased from the American Type Culture Collection and cultured in Dulbecco's modified Eagle medium (DMEM) with high glucose (Gibco 11965092) supplemented with 10% (v/v) fetal bovine serum (Gibco 26140079). Neuro2a cells were incubated at 37 °C in a 5% CO2 humidified atmosphere.

The iPSC lines utilized in this study included PGP1-iNGN,51 28126, 21792 (Yoav Gilad lab at The University of Chicago), and NA19101 (Marcelo A. Nobrega lab at The University of Chicago). These iPSC lines were cultured in Essential 8 medium (Thermo Fisher, A1517001) supplemented with penicillin-streptomycin (100 U/mL, Thermo Fisher, 15140122) in 10-cm dishes coated with GelTrex (Thermo Fisher, A1413301). The cultures were maintained in a 37°C incubator with 5% CO2. In the case of SYNGAP1-mutant (Lys114SerfsX20, Simons Foundation) iPSCs, StemFlex medium (Thermo Fisher, A3349401) was used instead of Essential 8. Following passage, the culture media were supplemented with 10 mM Rock inhibitor Y-27632 dihydrochloride (Tocris, 1254) for 24 hours. For neuron induction, PGP1-iNGN iPSCs were

cultured in Essential 8 medium supplemented with doxycycline (1 mg/mL, Sigma-Aldrich, D9891) and penicillin-streptomycin (100 U/mL) for 4 days. To assess the NMD-transcript of *SYNGAP1*, iPSCs and iPSC-derived neurons were treated with cycloheximide (200 μg/mL dissolved in DMSO) for 5 hours in 12-well plates with 1 mL of culture media in each well. Subsequently, RNA extraction and analyses were performed.

2.4.3 Cerebral organoids

The induction of brain organoids was conducted following the established protocol.[117] In brief, 3.3 x 10^6 human iPSCs (28126 and 21792) were seeded per AggreWell 800 well (STEMCELL Technologies, 34815) in Essential 8 medium supplemented with Y-27632 dihydrochloride (10 mM) (Tocris, 1254) and penicillin-streptomycin (100 U/mL). After 24 hours, iPSC spheroids were transferred to ultra-low attachment 6-well plates and incubated in Essential 6 medium supplemented with dorsomorphin (2.5 mM) (Sigma-Aldrich, P5499) and SB-431542 (10 mM) (Tocris, 1614) for 6 days to induce neural spheroids. Subsequently, the neural spheroids were cultured in Neurobasal A medium (Thermo Fisher Scientific, 10888022) supplemented with B-27 without vitamin A (1:50) (Thermo Fisher Scientific, 12587010), GlutaMax (1:100) (Thermo Fisher Scientific, 35050-061), epidermal growth factor (EGF) (20 ng/mL) (R&D Systems, 236-EG), basic fibroblast growth factor (bFGF) (20 ng/mL) (R&D Systems, 233-FB), and penicillin-streptomycin (100 U/mL) for 19 days. Afterward, EGF and bFGF were replaced with brain-derived neurotrophic factor (BDNF) (20 ng/mL) and NT-3 (20 ng/mL) for an additional 18 days. Starting from day 43, the brain organoids were cultured long-term in Neurobasal A medium supplemented with B-27 without vitamin A (1:50), GlutaMax (1:100), and penicillin-streptomycin (100 U/mL).

2.4.4 Splice-switching oligonucleotides

The SSOs used in this study were synthesized by Integrated DNA Technologies (IDT). For SSO transfection, 43105 cells were seeded in GelTrex-coated (Thermo Fisher, Gibco, A1413301) 12-well plates. Transfection of 200nM SSOs into the cells was carried out using TransIT-LT1 Transfection Reagent (Mirus, MIR2300) following the manufacturer's instructions. Brain organoids derived from iPSC 28126 were treated with three doses of SSOs (300nM) from day 133 to 137. On day 139, RNA and protein were extracted from the brain organoids for PCR and western blot analyses. Similarly, brain organoids derived from iPSC 21792 were treated with five doses of SSOs (200nM) from day 169 to 173, and RNA and protein were extracted on day 174. Total RNA extraction was performed using TRIzol reagent (Thermo Fisher, 15596018) and the Direct-zol RNA Purification Kit (Zymo Research, R2060) 24 hours after transfection. Subsequently, cDNA synthesis was carried out using the SuperScript IV Reverse Transcriptase kit (Thermo Fisher, 18090050). Quantitative PCR (Q-PCR) was performed using the SYBR Green PCR Master Mix (Thermo Fisher, 4344463) on the QuanStudio Real-Time PCR Systems (Thermo Fisher, ZG11CQS3STD) following the manufacturers' instructions..

2.4.5 RT-PCR and Western blot

To perform RNA extraction, brain tissues or cultured cells were thoroughly dissolved in TRIzol by pipetting vigorously. Subsequently, the dissolved samples were processed either through precipitation or by employing the Direct-zol RNA Purification Kit. Primers for RT-PCR are listed in Supplementary Table 1. In the case of Western blotting, protein lysates were obtained by extracting proteins using RIPA buffer (Thermo Fisher, PI89901), supplemented with proteinase inhibitors (Sigma-Aldrich, 11836170001). The resulting protein samples were loaded onto SDS-PAGE gels, transferred onto PVDF membranes, and consecutively incubated with primary and

secondary antibodies (as specified in the Key Resources Table). Finally, the samples were imaged using the LI-COR Odyssey system (LI-COR, 9142).

2.4.6 EMSA

Cy5 probes and cold competitors of potential Ptbp1 binding sites were synthesized by IDT. PTBP1 protein was produced by TnT SP6 High-Yield Wheat Germ Protein Expression System. Added different volumes of protein mixture into the RNA-protein binding solution, together with either Cy5 probes or extra cold competitors. Load the RNA-protein binding mixture to 8% TBE gel and perform electrophoresis. The gel was visualized by Typhoon imaging system.

2.4.7 Primary neuron culturing, transfection, and immunostaining

Mouse cortex was dissected from E16-E18 embryos and disassociated by Papain. Cells were suspended and cultured in Neurobasal medium with GlutaMax, N2 and B27 supplements, as well as 1 μM AraC. Primary neurons were transfected by Lipofectamin 2000 at the highest concentration recommended by the manual.

For immunostaining, primary neurons were fixed at 4°C in 4% paraformaldehyde (PFA) for 10 minutes. They were then rinsed with 1x PBS and incubated with blocking buffer (1x PBS containing 0.03% Triton X-100 and 5% normal donkey serum) at room temperature for 30 minutes. Subsequently, the neurons were further incubated overnight at 4°C with primary antibodies diluted in PBST buffer (1x PBS containing 0.03% Triton X-100). After three washes with 1x PBS, the slides were incubated in the dark at room temperature for one hour with secondary antibodies conjugated with fluorophores. Finally, the slides were scanned using a Leica SP8 confocal microscope.

2.4.8 Sequencing data analysis

RNA-Seq: Neuro2a cells were infected with Ptbp1/2 KO shRNA lenti-virus. Harvest the cells 5 days after. Primary DIV0 neurons from E16 mouse cortex were treated by 50 μg/mL CHX and DMSO. Harvest the cell after 8 hours. Total RNA was prepared by Quick-RNA MiniPrep Kit. Build RNA-Seq library with TruSeq RNA Library Prep Kit. Paired-end stranded sequencing was performed, with the read length 150bp.

Alternative Splicing Analysis: RNA-Seq data was QC and trimmed by cutadapt and aligned to mm10 genome by STAR. Differential alternative splicing analysis was performed with rMATS with default parameters. The differentially spliced isoforms with p-value smaller than 0.005 and inclusion rate change larger than 10% were considered to be significant.

GO Analysis and Motif Analysis: GO analysis was carried out by GOPlot with the differentially spliced isoforms which were significant in both CHX versus control treatment and E11 versus E18 mouse CNS datasets. Motif maps for different RBPs were generated by rMAPS near the splice sites (50 bp in exon and 250bp in intron).

CLIP-seq Analysis: CLIP-seq data were aligned and mapped to hg19 genome by CLIPSeqTools, using the default parameters. The .bam files were visualized by IGV.

Multiple Sequence Alignment: Genome sequences of multiple vertebrates were obtained from NCBI HomoloGene database. Multiple sequence alignment and phylogeny analysis was performed by MAFFT over the conserved region of *Syngap1* intron 10. The splice site usage was validated using the deposited RNA-Seq data of non-neuronal tissues from the corresponding vertebrates.

2.4.9 Data availability

All sequencing data are available at NCBI Gene Expression Omnibus with the accession number PRJNA930469: SRR23308049, SRR23308050.

2.5 Acknowledgements of work performed

I would like to thank all of our collaborators for their contributions to this work, especially: Dr. Xiaochang Zhang for the preliminary experiments, e.g., the production of Ptbp1/2 shRNA lentivirus, as well as supervising the whole project. Dr. Runwei Yang for his contribution to the SSO-related experiments (data in Figure 2.7). Ms. Alejandra Arias-Cavieres for the electrophysiology experiments (data not shown here).

Chapter 3

Quantitative base-resolution mapping of N^6-methyldeoxyadenosine using DR-6mA-seq

3.1 Introduction:

Covalent modifications of DNA play a crucial role as epigenetic marks in various biological systems.[124] One such modification, N^6-methyldeoxyadenosine (6mA), is commonly found in the genomes of bacteria and protists, where it is involved in important processes such as the restriction-modification system, DNA repair, replication, transcription, nucleoid segregation, and gene expression regulation.[55] However, in higher eukaryotes, particularly mammals, 6mA is considered a rare DNA modification, and its presence and functional significance in the mammalian genome have remained unclear.[51,54] This is largely due to the lack of sensitive and accurate methods to detect 6mA at a genome-wide scale.

Detecting 6mA in mammalian genomes is challenging due to its low abundance. Therefore, highly sensitive and accurate sequencing methods are required.[51,73] One commonly used method is ultra-high performance liquid chromatography coupled with mass spectrometry (UHPLC-MS/MS), which allows for the quantification of global 6mA levels. However, this method is susceptible to contamination from bacterial DNA, which can arise from cell culture infection, plasmids, and reagents.[73,76] Moreover, the low concentrations of 6mA in mammalian genomic DNA (gDNA) samples make it difficult to detect using UHPLC-MS/MS, especially in the presence of bacterial DNA contamination.[66,125] Therefore, there is a critical need to develop high-sensitivity, base-resolution, and whole-genome mapping methods for 6mA sequencing.

DNA immunoprecipitation sequencing (DIP-seq), often combined with exonuclease digestion (6mA crosslinking exonuclease sequencing; 6mACE-seq or ChIP-exo), has been extensively used to generate genome-wide profiles of 6mA.[63,69,70] However, concerns have been raised about the potential for off-target binding of antibodies, sequence misalignment, and RNA contamination, which may lead to false positive detection of 6mA using these antibody-based methods.[126] Another approach, restriction enzyme digestion-based methods such as 6mA-RE-seq and DpnI-seq, can provide single-nucleotide resolution profiles of 6mA modifications. However, these methods are limited to detecting specific sequence motifs and cannot provide a comprehensive view of 6mA distribution in the genome.[74,127]

A recent study introduced a base-resolution 6mA sequencing method that relies on the nitrite-mediated deamination of unmodified A sites.[128] Although this strategy is innovative, further improvements are needed to enhance the deamination rate and minimize the occurrence of false negative or false positive sites. In addition to second-generation sequencing-based techniques, third-generation SMRT sequencing has enabled direct detection of DNA methylation, including 6mA, in bacterial genomes with high confidence.[57] However, when applied to mammalian DNA, SMRT sequencing has yielded high false discovery rates, likely due to various factors such as the low abundance of 6mA compared to 5mC, the absence of consistent sequence context for 6mA, and the presence of bacterial contaminants containing 6mA.[73,129,130]

The limitations associated with each of these methods have contributed to the controversial findings regarding the presence and functions of 6mA in the mammalian genome, leaving this topic subject to ongoing debate. For example, the quantification of the 6mA to A ratio in mitochondrial DNA from human cell lines has yielded varying results, ranging from approximately 15 ppm to 130-400 ppm, which can be attributed to differences in the purity of the isolated mtDNA

samples analyzed.[55,69,70] Interestingly, despite the detection of 6mA in mitochondrial DNA using UHPLC-QqQ-MS/MS, the same cell line showed no enrichment of 6mA in reads mapped to mtDNA when analyzed using 6mASCOPE, a newly developed 6mA deconvolution pipeline based on SMRT sequencing.[76] Given the limitations of current detection methods, we aimed to develop novel base-resolution sequencing approaches to overcome existing challenges and gain a better understanding of the genomic features of 6mA in mammals.

Our approach capitalizes on the destabilizing effect of the methyl group present in 6mA on Watson-Crick base-pairing. Despite its ability to pair with deoxythymidine (dT) similar to unmodified deoxyadenosine (dA), we hypothesized that using a thymidine dTTP analog that forms unstable base-pairing with dA could further weaken the base-pairing with 6mA.[131,132] This would result in the generation of misincorporation signatures opposite 6mA sites during DNA replication using selected DNA polymerases, enabling the detection of 6mA at the base-resolution level. In this study, we present the development and application of DR-6mA-seq, a method that allows for high-confidence, mutation-based, base-resolution detection of 6mA throughout the entire genome. Our findings reveal distinct genomic patterns of 6mA in various types of mammalian genomic DNAs. Consistent with previous research, we observed minimal levels of 6mA in most mammalian gDNAs. However, we did identify significantly elevated levels of 6mA methylation in specific cell types.

3.2 Results

3.2.1 The principle and development of DR-6mA-seq

To establish a new antibody-independent, single-base-resolution, NGS-based mapping method for 6mA, we utilized a synthetic biotinylated 100-mer DNA oligo containing an NN-6mA-NN modification site as our model system. Our objective was to investigate whether 6mA could

be detected as misincorporation signatures under specific primer extension conditions. To accomplish this, we tested various commercially available non-high-fidelity DNA polymerases with different buffer systems. By sequencing the newly synthesized DNA strand after strand separation, we were able to determine the ratios of 6mA-to-T/G/C mutations at the 6mA sites. Notably, *Bst* 2.0 DNA and EpiMark Hot Start *Taq* DNA Polymerase exhibited detectable mutation ratios in certain motif contexts, consistent with recent findings suggesting that N^6-methylated adenosine could be identified as mutations.[133] Subsequently, we employed the buffer system of EpiMark Hot Start *Taq* DNA Polymerase in conjunction with *Bst* 2.0 DNA, resulting in an improved mutation proficiency. However, the mutation rates remained too low to be practically useful.

We postulated that amplifying the difference in base pair stability between 6mA and dTTP, as well as between dA and dTTP, could enhance the mutation frequency opposite 6mA. To test this hypothesis, we screened all commercially available dTTP analogs and evaluated their performance in the primer-extension system, replacing unmodified dTTP. Among the analogs tested, we identified 2-thiothymidine triphosphate (2-thio-dTTP), which incorporates a sulfur substitution at the 2-position of thymine, significantly increasing the misincorporation rate at the 6mA site while leaving unmethylated dA sites unaffected. This observation aligns with the reduced base-pairing strength between 6mA and 2-thio-dTTP, as well as between dA and 2-thio-dTTP. Specifically, thione does not effectively serve as a hydrogen bond acceptor in Watson-Crick base pairing.[134] Furthermore, the presence of the methyl group on 6mA weakens its base pairing with canonical 2-thio-dTTP during extension, allowing dATP/dCTP/dGTP to compete with 2-thio-dTTP for base pairing with 6mA (Figure 3.1A), thereby resulting in an increased mutation rate at

the 6mA sites. Conversely, dA sites can still form base pairs with 2-thio-dTTP when using *Bst* 2.0 polymerase, leading to significantly fewer mutations (Figure 3.1A).[135]

After identifying a thymidine analog that significantly increases misincorporation opposite 6mA sites, we proceeded to optimize the primer-extension conditions to maximize the mutation ratio. It has long been acknowledged that the addition of excessive magnesium ions and molecular crowders, such as gelatins, can enhance PCR mismatches.[136] Biased dNTP ratios and unconventional incubation temperatures have also been shown to promote mutations.[137] We conducted experiments using various combinations of magnesium ion concentrations, molecular crowder addition, dNTP ratios, polymerase concentrations, and incubation temperatures, ultimately establishing an optimal condition for generating mutation signatures at 6mA sites. Under this condition, the average mutation ratio across 256 NN-6mA-NN motifs is 16%, with the highest mutation ratio observed in a specific motif reaching 50%. This finding suggests that the mutation patterns at 6mA sites are influenced by the sequence context (Figure 3.1B, Supplementary Table 2). The majority of mutations involved 6mA-to-T conversions, although 6mA-to-C and 6mA-to-G conversions were also detected (Figure 3.1C). These moderate to high mutation rates enable the sensitive detection of 6mA in DNA.

Due to the absence of any conversion or addition of bulky adduct on 6mA sites and any other bases, this approach induces mismatches directly at 6mA sites by solely adjusting the buffer system and supplements of the primer extension reaction. This strategy is highly concise and rapid, and thus, we have named it Direct-Read 6mA sequencing (DR-6mA-seq).

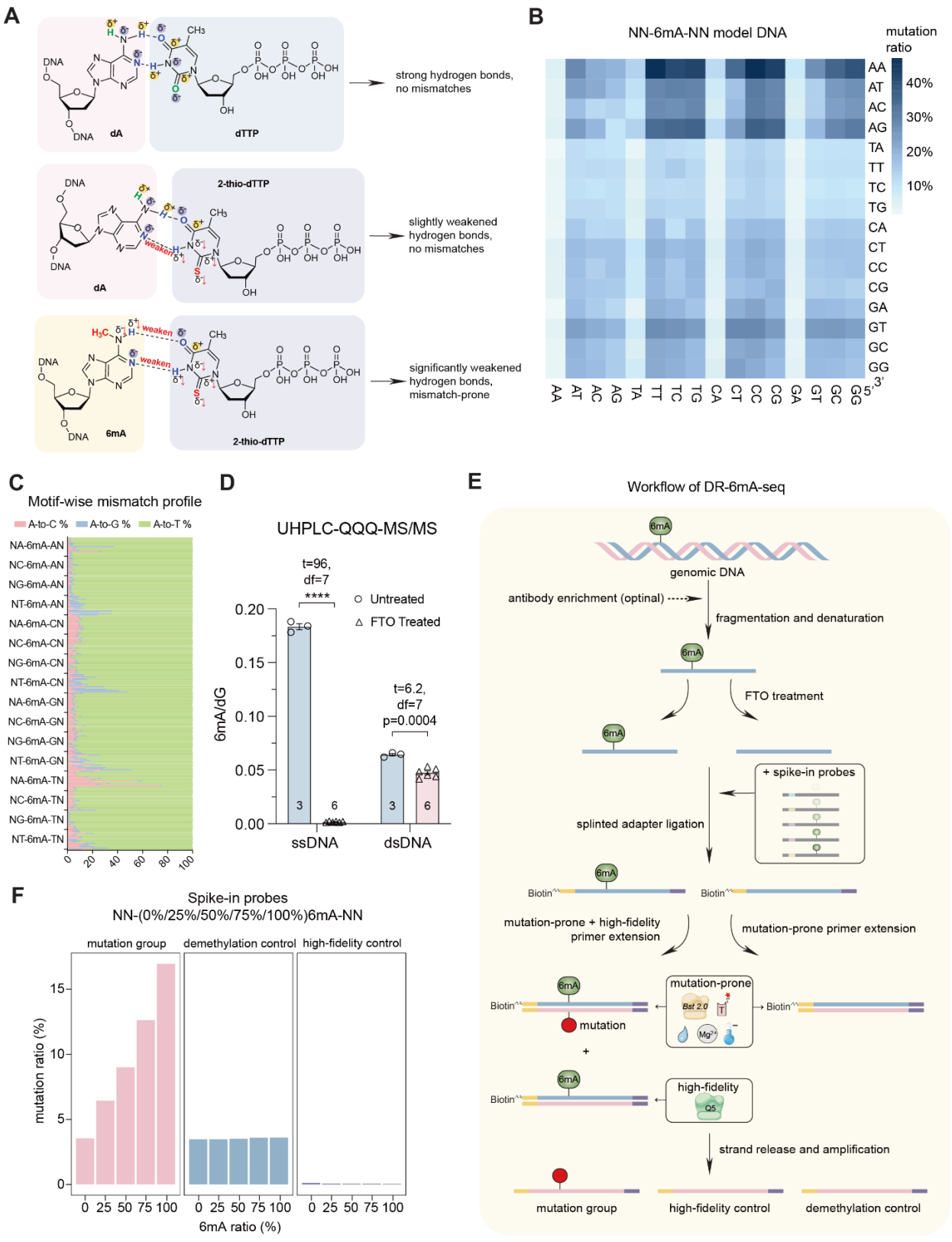


Figure 3.1 Development and validation of DR-6mA-seq

(**Figure 3.1, continued**) **A)** The 6mA-2-thio-dTTP base pair forms weaker hydrogen bonds compared to dA-dTTP and dA-2-thio-dTTP base pairs. **B)** Heatmap plot for the 6mA-to-T/C/G mutation ratios at 256 motifs (NN-6mA-NN) on the single-stranded model DNA after mutation-prone primer extension. **C)** The mutation patterns (6mA to T, C, and G, respectively) at different motif contexts (NN-6mA-NN) on the single-stranded model DNA. **D)** FTO demethylates over 99% methylated base of NN-6mA-NN model ssDNA *in vitro* and shows weaker activity on dsDNA. Data are mean ± s.e.m.; analyzed by two-tailed unpaired t-tests. The numbers of independently repeated reactions are shown in each plot. **** p-value <0.0001. **E)** Schematic diagram of DR-6mA-seq. Genomic DNA is sequentially denatured, mixed with spike-in probes, treated with FTO (or not), ligated to biotin-tagged adaptors, and extended by *Bst* 2.0 DNA polymerase or Q5 high-fidelity DNA polymerase. The synthesized DNA strand is enriched and amplified for next-generation sequencing. 6mA-to-T/C/G mutations were counted for defining 6mA sites and calculating 6mA methylation stoichiometry using spike-in calibration curves. **F)** Mutation profiles (including 6mA to T, C, and G) of mutation-prone primer extension on the 6mA spike-in probes harboring 0%, 25%, 50%, 75%, and 100% 6mA. Primer extension on FTO-treated probes and primer extension by Q5 high-fidelity DNA polymerase serve as two control groups and do not respond to 6mA fractions.

3.2.2 FTO is an efficient 6mA demethylase on single-stranded DNA

In order to effectively analyze the genomic features of 6mA, it is crucial to eliminate any false positives or mutations that may be caused by other potential modifications. The fat mass and obesity-associated protein (FTO) is recognized for its ability to efficiently remove RNA N^6-methyladenosine (m6A) and catalyze mRNA m6A demethylation in an iron(II)- and α-ketoglutarate-dependent manner. Previous studies have also demonstrated its biochemical demethylation activity towards 6mA on single-stranded DNA (ssDNA), which serves as a crucial control for identifying 6mA in DR-6mA-seq. To investigate this further, we purified recombinant human FTO protein and treated both the aforementioned NN-6mA-NN model ssDNA probes and their double-stranded form (formed by annealing with the unmodified complementary strand) with the FTO protein in the presence of iron(II) and α-ketoglutarate (α-KG). After performing nucleotide digestion and UHPLC-QqQ-MS/MS analysis, we observed that more than 99% of the 6mA on the ssDNA probes were removed by the FTO treatment, while 26.4% of the modifications on the double-stranded DNA (dsDNA) probes were eliminated (Figure 3.1D). It is possible that

some regions of the dsDNA containing 6mA may have partially unwound to adopt a single-stranded conformation during the FTO treatment, resulting in the observed demethylation activity on dsDNA. Nevertheless, we have confirmed that FTO is a potent demethylase of DNA 6mA on single-stranded DNA *in vitro* without significant motif bias. In comparison, the reported *in vitro* efficiency of ALKBH1 as an ssDNA demethylase is approximately 50%.[81,138] Therefore, the FTO treatment can serve as a crucial control for specifically identifying and mapping 6mA in DNA.

In order to comprehensively investigate the presence of 6mA in mammalian genomes, we have developed a robust workflow termed DR-6mA-seq, which incorporates essential controls to ensure the accuracy and reliability of our findings (Figure 3.1E). Our approach involves the fragmentation of RNA-free genomic DNA (gDNA) into short single-stranded fragments. Subsequently, half of the DNA samples are subjected to separation, denaturation, and demethylation using FTO, serving as a control for demethylation. The remaining untreated DNA samples, as well as the demethylated DNA, are then ligated and treated with *Bst* 2.0 for primer extension. To further validate our results, the other half of the untreated DNA is subjected to the primer extension using Q5 high-fidelity DNA polymerase, serving as an additional control. The newly synthesized DNA strands are then enriched and amplified to generate sequencing libraries. By comparing the mutation profiles of *Bst* 2.0-extended DNA from methylated DNA and FTO-demethylated DNA, and by eliminating background noise using the high-fidelity control, we can confidently identify candidate sites exhibiting 6mA modification.

To accurately estimate the level of 6mA modification based on the ratio of 6mA-to-T/G/C mutations, we synthesized five spike-in calibration probes containing known fractions of 6mA (0%, 25%, 50%, 75%, and 100%) at the NN-6mA-NN modification site (Supplementary Table 3). By applying DR-6mA-seq to these probes, we observed significant linear correlations ($R^2 > 0.98$,

p-value <0.001) between average mutation frequencies and 6mA fractions. This provides a reliable method to quantify 6mA modification levels in different sequence contexts. It is important to note that the average mutation rate on unmodified spike-in probes was approximately 3.5%, which closely matches that of the fully modified spike-in probes after FTO treatment (Figure 3.1F). This confirms that FTO treatment effectively removes 6mA. To minimize any batch effects, these calibration probes, each with unique barcodes, were added to each real biological DNA sample during the DA-6mA-seq procedures for constructing sample-specific calibration curves. The modification fractions for each high-confidence 6mA site were then estimated using the corresponding linear regression model based on the motif context surrounding the 6mA sites.

3.2.3 Quantitative 6mA maps of *E. coli* gDNA using DR-6mA-seq

To validate the performance of DR-6mA-seq, we initially examined the genomic DNA (gDNA) of *Escherichia coli* K-12, a well-characterized wild-type strain with a known 6mA methylome. In order to ensure accurate computational analysis for 6mA detection, we developed a False Discovery Rate (FDR)-based analysis pipeline. This pipeline replaces the conventional cutoff-based analysis and focuses on the statistical properties of 6mA methylomes. Our FDR-based approach takes into consideration various factors, including sequencing depth, mutation count, mutation rate, and background noise at each 6mA candidate site. By implementing this novel pipeline, we successfully identified 6mA sites with high confidence while effectively filtering out most false positives.

To ensure the reliability of our 6mA site assignments, we obtained DR-6mA-seq results from three biological replicates. By comparing the results of the *Bst* 2.0 polymerase group with those of the FTO-treated and Q5 high-fidelity groups, we observed significant mutations at 6mA sites (Figure 3.2A), which were accompanied by a well-distributed FDR value (Figure 3.2B).

Using the FDR-based analysis pipeline, we identified approximately 21,500 6mA sites in each of the three biological replicates of *E. coli* genomic DNA. In contrast, fewer than 10 putative 6mA sites were detected in each replicate of the control synthetic *E. coli* genomic DNA lacking 6mA (Figure 3.2C). Furthermore, we identified a total of 16,852 overlapping 6mA sites among the three replicates of *E. coli* genomic DNA (Figure 3.2D, Supplementary Table 4). The methylation stoichiometry profiles exhibited high correlation and reproducibility across different biological replicates (Figure 3.2E). All 6mA sites in *E. coli* genomic DNA exhibited methylation fractions above 30%, with the majority of detected sites displaying a ~100% 6mA modification level, consistent with a previous study (Figure 3.2D).[139]

The distribution of 6mA sites in the genomic DNA of *E. coli* is predominantly observed in coding regions, with a smaller fraction found in intergenic regions, non-coding RNA regions, pseudogenes, and tRNA regions (Figure 3.2F-G). The level of 6mA methylation varies across different annotated regions in the *E. coli* genome, ranging from 30% to 100% (Figure 3.2H). Notably, approximately 34% of the coding regions modified by 6mA in the *E. coli* genome exhibit more than five 6mA sites per gene, and around 7% of these genes have more than ten 6mA sites per gene (Figure 3.2I). The analysis further revealed that 97.6% of the identified 6mA sites correspond to the G(6mA)TC motif, which is recognized by the Dam methylase (Figure 3.2J-K).[139–141] The second most prevalent motif, accounting for 2.2% of 6mA sites, is GC(6mA)CNNNNNNGTT, which is the consensus sequence recognized by hsdM, the methylase subunit of EcoKI (Figure 3.2J, L).[142,143] Notably, Dam and hsdM are the only two 6mA methyltransferases recognized in *E. coli* K-12.[144] Thus, the accuracy of the method and analysis pipeline employed is quite high, as also demonstrated by the fact that 99.9% of the 6mA sites

identified through DR-6mA-seq overlap with those detected by SMRT sequencing when using the FDR-based analysis pipeline (Figure 3.2M).[145]

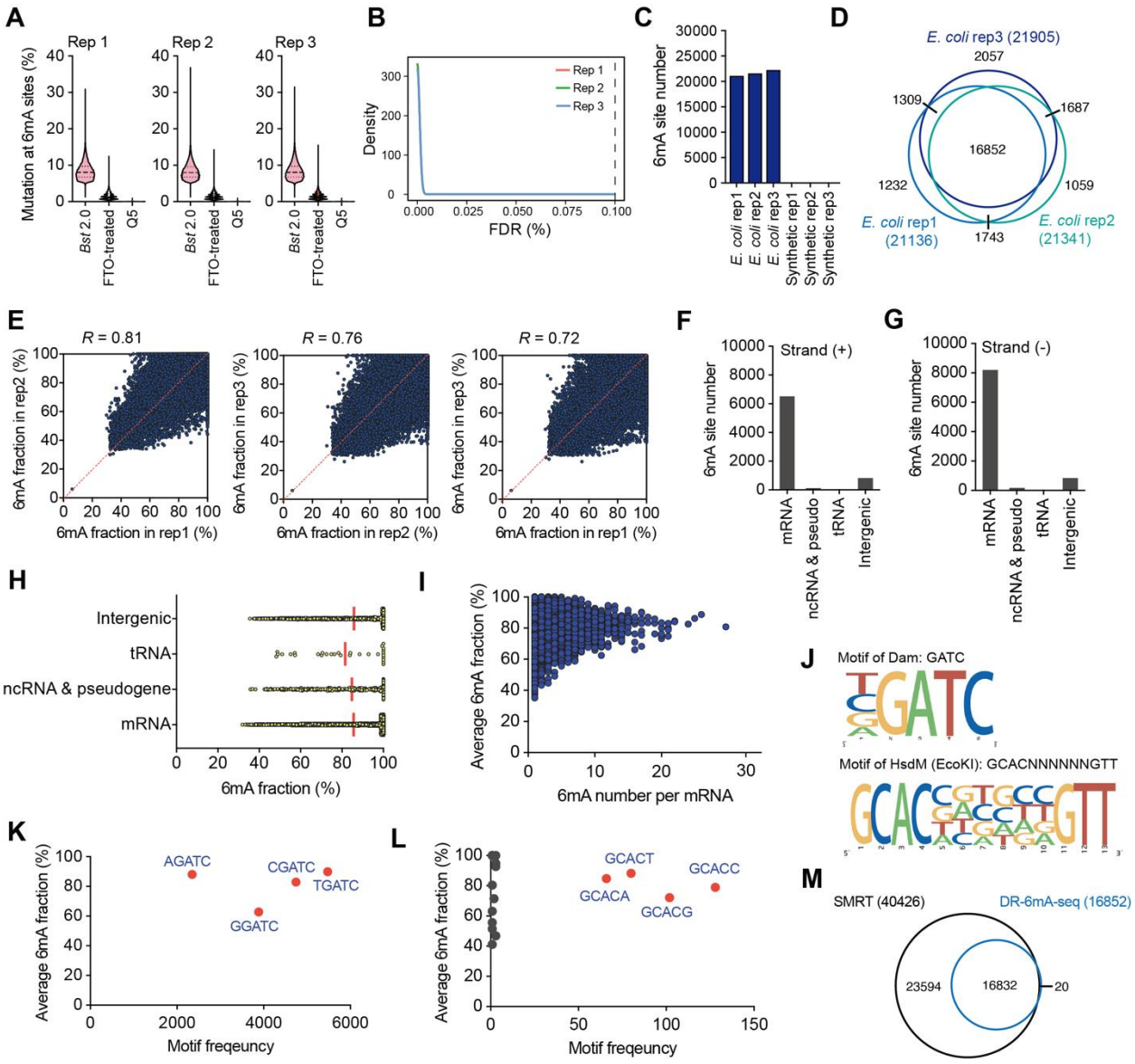


Figure 3.2 DR-6mA-seq uncovers the quantitative base-resolution 6mA map in *E. coli* genome

A) The violin plot of mutation ratio distribution at 6mA sites detected in *E. coli* K-12 gDNA, revealed by DR-6mA-seq. The mutation ratios of three groups are shown, i.e., *Bst* 2.0 extended DNA from untreated DNA (mutation group), *Bst* 2.0 extended DNA from FTO-treated DNA (demethylation control), and Q5-extended DNA from untreated DNA (high-fidelity control). **B)** The distribution of FDR of all the identified 6mA sites in *E. coli* gDNA. The dashed line refers to the 0.1% FDR cutoff. **C)** The bar plot of numbers of identified 6mA sites in *E. coli* gDNA and the synthetic *E. coli* gDNA prepared by highly uniform whole genome amplification. **D)** Venn

(**Figure 3.2, continued**) diagram showing the overlapped 6mA sites among three biologically independent replicates, detected by DR-6mA-seq. **E)** Correlation analysis of methylation fractions in *E. coli* K-12 genome, indicating a high correlation between each pair of biological replicates of DR-6mA-seq. **F)** The bar plot showing the numbers of overlapped 6mA sites (among three replicates) in each annotated region on the plus strand of *E. coli* gDNA. **G)** The bar plot showing the numbers of overlapped 6mA sites (among three replicates) in each annotated region on the plus strand of *E. coli* gDNA. **H)** The scatter plot of the modification fractions of overlapped 6mA sites (among three replicates) in each annotated region on *E. coli* gDNA. **I)** The 2-D plot of 6mA site number and their average methylation fraction within one individual mRNA. Each dot represents a single annotated mRNA region in *E. coli* gDNA. **J)** The motif sequence logo of the overlapped 6mA sites identified in *E. coli* gDNA, uncovered by DR-6mA-seq. **K)** The plot of the average 6mA fractions at each consensus 6mA motif (motif frequency >150) of the overlapped 6mA sites identified in *E. coli* gDNA. **L)** The plot of the average 6mA fractions at each consensus 6mA motif (motif frequency <150) of the overlapped 6mA sites identified in *E. coli* gDNA. **M)** Venn diagram showing the excellent overlap between DR-6mA-seq-detected 6mA sites and SMRT-detected 6mA sites in *E. coli* K-12 genome.

3.2.4 6mA is upregulated in *Erbb2/neu**-transformed NIH/3T3 cells

While previous studies have provided compelling evidence for the presence and functional role of 6mA in bacterial genomes, its existence in the mammalian genome has been a topic of controversy.[71,73,76,77,80–82,84–86,126] To conduct a more comprehensive investigation of 6mA methylomes in mammalian genomic DNA using DR-6mA-seq, we sought suitable cell line systems for 6mA sequencing. A recent study utilizing mass spectrometry and antibody-dependent enrichment and sequencing demonstrated relatively high levels of 6mA in glioblastoma cancer stem cells (GCSs) and primary glioblastoma patient tumors.[81] As a result, we selected this system for further experimentation with DR-6mA-seq. Among the available glioblastoma models, we specifically chose the B104-1-1 cell line, which is commonly used and was established by transforming NIH/3T3 embryonic fibroblast cells with the activated *Erbb2/neu* oncogene (Erbb2 with p.Val661Glu).[146–151] Given their nearly identical genetic background, we decided to investigate potential changes in 6mA levels during gliomagenesis using these two cell lines. Initially, we performed UHPLC-QqQ-MS/MS analysis on genomic DNA isolated from bacteria-

and mycoplasma-free NIH/3T3 and B104-1-1 cell cultures to assess their overall levels of 6mA (Figure 3.3A, Supplementary Table 5). The glioblastoma cell line, B104-1-1, exhibited approximately four-fold higher levels of 6mA (~80 ppm) compared to its parental cell line, NIH/3T3. Subsequently, we enriched 6mA-containing DNA fragments from both B104-1-1 and NIH/3T3 using the anti-6mA antibody and performed DR-6mA-seq.

To begin with, we determined the sex of the cell lines as female before conducting data analysis (Figure 3.4A).[152] Using an analysis pipeline based on false discovery rate (FDR), we obtained confident 6mA sites (FDR <0.1%) from biological replicates using DR-6mA-seq with FTO-treated and Q5 high-fidelity groups as controls (Figure 3.3B-E, Supplementary Table 6, 7). For the two biological replicates, ~68% and ~40% of the 6mA sites overlapped very well within the ±500 bp flanking windows (Figure 3.3F). In NIH/3T3 cells, DR-6mA-seq detected around 639~843 gDNA 6mA sites, whereas B104-1-1 gDNA exhibited around 4,709~5,355 6mA sites (mtDNA sites not included) (Figure 3.3F). Notably, around 80% of the 6mA sites in NIH/3T3 gDNA were also detected in B104-1-1 gDNA within the ±500 bp flanking windows (Figure 3.3G). The identification of approximately 5-fold more 6mA sites in B104-1-1 cells is consistent with our mass spectrometry results (Figure 3.3A).

Another observation is that the nuclear DNA 6mA sites tend to cluster together (Figure 3.4O). To be specific, 80% of these overlapped 6mA sites between biological replicates were found within clusters that are smaller than 200 bp (Figure 3.3H-I). We further noticed that the majority of 6mA modification clusters in gDNA contain less than 10 6mA sites, although there are a few heavily methylated clusters that have more than 10 methylated sites (Figure 3.3J).

B104-1-1 cells exhibited a significantly higher number of genomic DNA (gDNA) 6mA sites on specific chromosomes, such as 1, 4, 6, 10, and 19, compared to NIH/3T3 cells (Figure

3.4B-C). The genomic 6mA sites identified in glioblastoma cells were found to be enriched in regions associated with repeat-associated RNA, intergenic regions, and introns (Figure 3.3K-L). Even with antibody enrichment, approximately 94% and 95% of the 6mA sites showed methylation fractions above 25% (Figure 3.4D-E). In contrast, all 6mA sites identified in bacterial gDNA exhibited methylation fractions greater than 30% (Figure 3.2H). These findings suggest that the 6mA sites in mammalian cells tend to show much lower stoichiometry than that in prokaryotic cells.

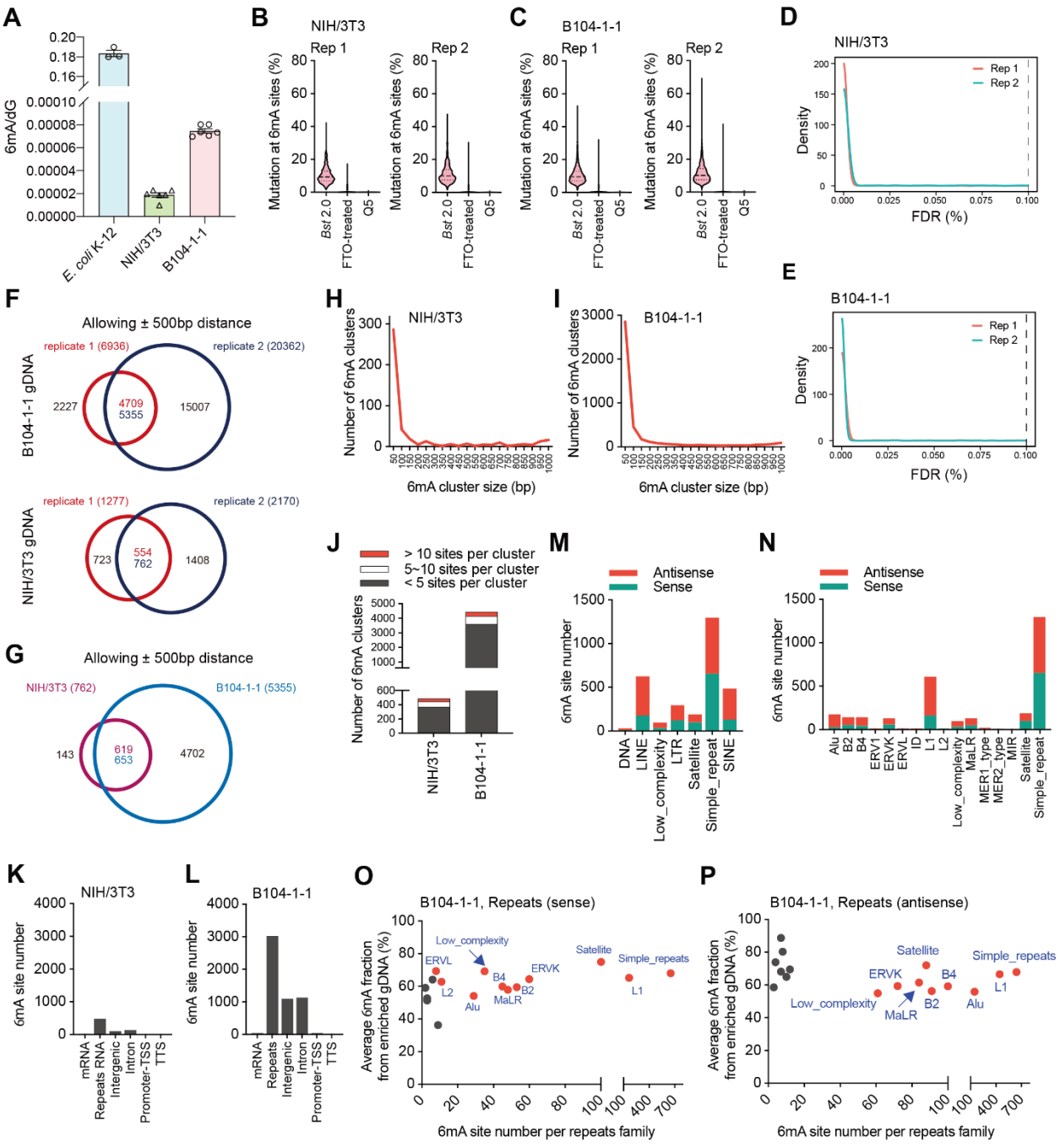


Figure 3.3 The prevalence of 6mA modification is significantly elevated during the transition from mouse embryonic cells to glioblastoma cells

A) Quantification of 6mA level in bacterial-DNA-free gDNA samples from cultured NIH/3T3 and B104-1-1 cells, by LC-MS/MS, with *E. coli* gDNA as a positive control. Data are mean ± s.e.m. **B)** The violin plot of mutation ratio distribution at 6mA sites detected in NIH/3T3 gDNA, revealed by DR-6mA-seq. The mutation ratios of three groups are shown, i.e., *Bst* 2.0 extended DNA from untreated DNA (mutation group), *Bst* 2.0 extended DNA from FTO-treated DNA (demethylation control), and Q5-extended DNA from untreated DNA (high-fidelity control). **C)** The violin plot of

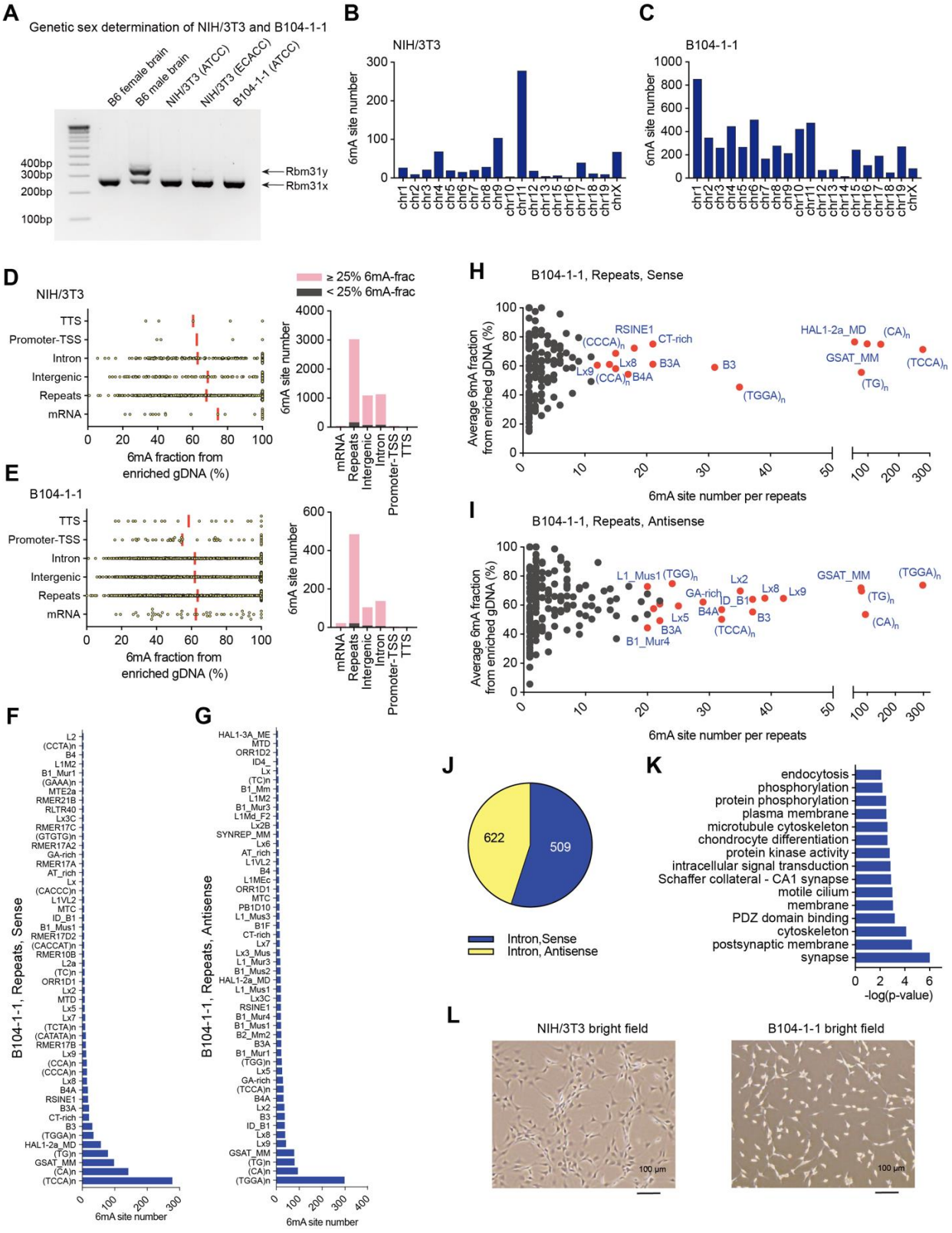
(**Figure 3.3, continued**) mutation ratio distribution at 6mA sites detected in B104-1-1 gDNA, revealed by DR-6mA-seq. The mutation ratios of three groups are shown, i.e., *Bst* 2.0 extended DNA from untreated DNA (mutation group), *Bst* 2.0 extended DNA from FTO-treated DNA (demethylation control), and Q5-extended DNA from untreated DNA (high-fidelity control). **D)** The distribution of FDR of all the identified 6mA sites in NIH/3T3 gDNA. The dashed line refers to the 0.1% FDR cutoff. **E)** The distribution of FDR of all the identified 6mA sites in B104-1-1 gDNA. The dashed line refers to the 0.1% FDR cutoff. **F)** Venn diagram showing the overlapped gDNA 6mA sites detected in NIH/3T3 and B104-1-1, respectively, allowing ± 500 bp window, between two biological replicates. **G)** Venn diagram showing the overlapped gDNA 6mA sites between NIH/3T3 and B104-1-1 cells, allowing ± 500 bp window. **H)** The histogram showing the frequency distribution of the size of 6mA clusters in NIH/3T3 gDNA. **I)** The histogram showing the frequency distribution of the size of 6mA clusters in B104-1-1 gDNA. **J)** The bar plots showing the number of clusters that are categorized according to the numbers of 6mA sites within one cluster, in both cell lines. **K)** The bar plot showing the numbers of overlapped 6mA sites (allowing ± 500 bp window) in each annotated genomic element in NIH/3T3 genome. **L)** The bar plot showing the numbers of overlapped 6mA sites (allowing ± 500 bp window) in each annotated genomic element in B104-1-1 genome. **M)** The bar plot showing the numbers of overlapped 6mA sites (allowing ± 500 bp window) on the antisense DNA (the template strand) vs. sense DNA (the coding strand) of each annotated general repeats class (based on RepeatMasker) in B104-1-1 genome. **N)** The bar plot showing the numbers of overlapped 6mA sites (allowing ± 500 bp window) on the antisense DNA (the template strand) vs. sense DNA (the coding strand) of each annotated repeats family (based on RepeatMasker) in B104-1-1 genome. **O)** The 2-D plot of the numbers of overlapped B104-1-1 6mA sites (allowing ± 500 bp window) on the sense DNA (the coding strand) of each annotated repeats family and their average 6mA modification fractions (after antibody enrichment). **P)** The 2-D plot of the numbers of overlapped B104-1-1 6mA sites (allowing ± 500 bp window) on antisense DNA (the template strand) of each annotated repeats family and their average 6mA modification fractions (after antibody enrichment).

We then conducted a comprehensive analysis of 6mA sites in B104-1-1 gDNA and made several interesting observations. Firstly, we found that these sites tend to accumulate in regions that encode repeat-associated RNAs, including LINE, LTR, satellite, simple repeats, and SINE (Figure 3.3M). Interestingly, more 6mA was displayed on the antisense strand (the template strand) of LINE and SINE regions (Figure 3.3M). To gain a more detailed understanding of the specific repeat families modified by 6mA in glioblastoma cells, we conducted further analysis. We observed dense 6mA modifications on *Alu* elements, B2, B4, ERVK, LINE1, low complexity, MaLR, satellite, and simple repeats (Figure 3.3N, 3.4F-G). Among these, more 6mA was displayed on the antisense strand (the template strand) of *Alu* elements, B2, B4, ERVK, and LINE1 regions

(Figure 3.3N). Moreover, the 6mA methylomes on repeat-associated RNA regions exhibited high methylation fractions above 50% after antibody enrichment (Figure 3.3O-P, 3.4H-I). In contrast, the distribution pattern of 6mA sites found in intron regions of glioblastoma cells was distinct from those in repeat-associated regions. Specifically, we observed a nearly equal distribution of 6mA modifications on both the coding and template strands (Figure 3.4J). Furthermore, functional enrichment analysis of specific intron 6mA sites in the B104-1-1 cell line revealed a significant enrichment of synapse-, neuron-, and membrane-associated genes (Figure 3.4K). This finding is generally consistent with the previous report that the human glioblastoma stem cell 6mA peaks in the gene body exhibit an enrichment in neurogenesis and neuronal development pathways[81], and may reflect morphological differences between the two cell lines (Figure 3.4L).

The motif contexts surrounding 6mA sites in non-cancerous NIH/3T3 cells were found to be similar to those in B104-1-1 glioblastoma cells (Figure 3.4M). In mammalian genomic DNA, a wider range of motif contexts was observed compared to the limited number of predominant motifs found in bacterial genomic DNA (Figure 3.2K-L). To ensure the accuracy of the 6mA calling statistics obtained from the FDR-based analysis pipeline, we conducted a statistical analysis to compare them with random overlaps across the genome. We found that the 6mA sites identified by the FDR-based analysis pipeline showed a high level of statistical significance compared to random overlaps (Figure 3.4N).

In summary, our method not only confirmed the previous findings of the elevated 6mA prevalence in the nuclear DNA of glioblastoma cells but also identified highly confident mammalian 6mA sites that can be further investigated to understand the precise roles of 6mA in mammals.



(Figure 3.4 to be continued)

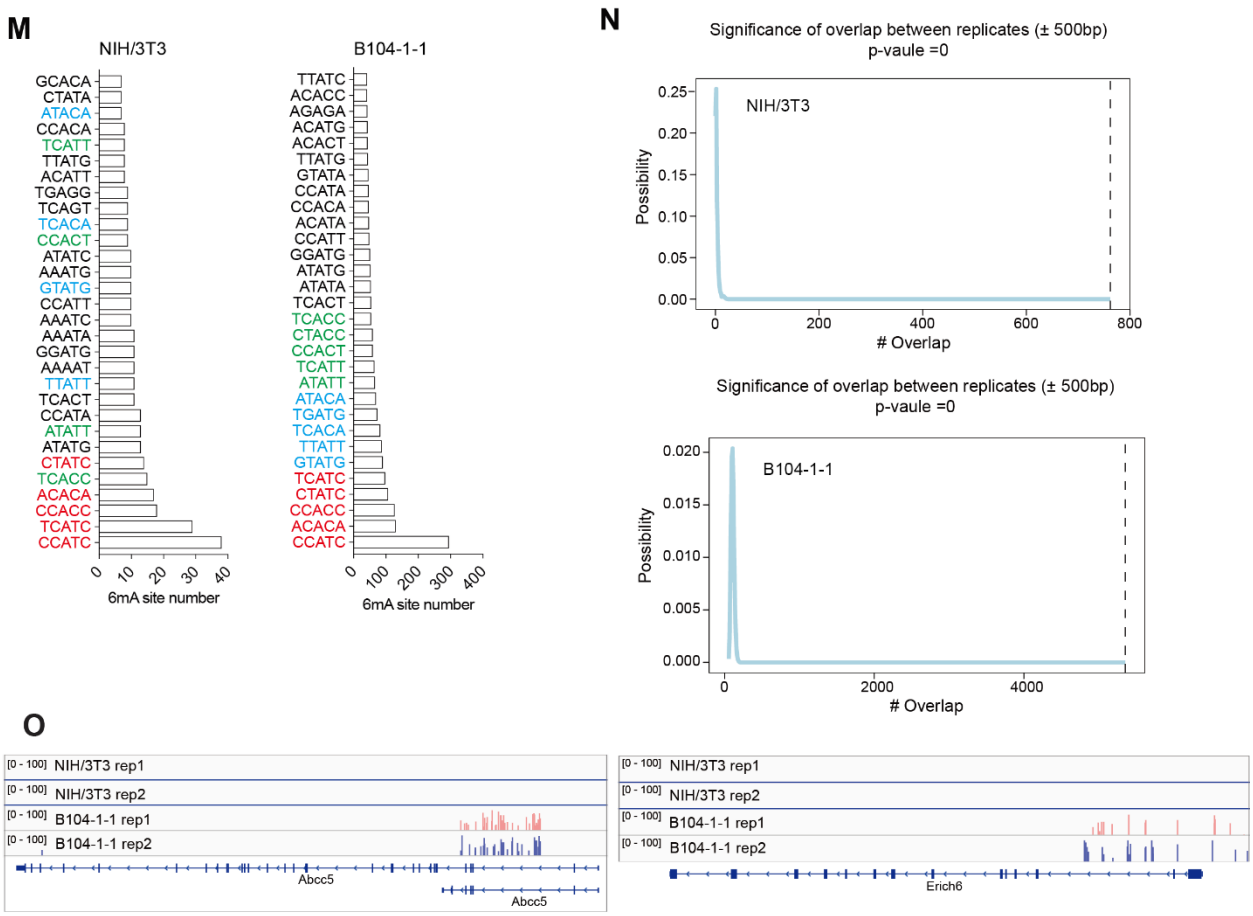


Figure 3.4 Features and statistics of gDNA 6mA sites identified by DR-6mA-seq in NIH/3T3 and B104-1-1 cells

A) Genetic sex determination of NIH/3T3 and B104-1-1 cell lines by amplification of *Rbm31x* and *Rbm31y* by simplex PCR, with female and male mouse brains as controls. **B)** Chromosome-wide distributions of the overlapped 6mA sites (allowing ± 500 bp window) in gDNA from NIH/3T3. **C)** Chromosome-wide distributions of the overlapped 6mA sites (allowing ± 500 bp window) in gDNA from B104-1-1. **D)** The scatter plot of the modification fractions (after antibody enrichment) of overlapped 6mA sites (allowing ± 500 bp window) in each annotated region on NIH/3T3 gDNA. The bar plot showing the distribution of the 6mA sites with fractions higher than 25% or below 25% at different genomic elements. **E)** The scatter plot of the modification fractions (after antibody enrichment) of overlapped 6mA sites (allowing ± 500 bp window) in each annotated region on B104-1-1 gDNA. The bar plot showing the distribution of the 6mA sites with fractions higher than 25% or below 25% at different genomic elements. **F)** The bar plot showing the numbers of overlapped 6mA sites (allowing ± 500 bp window) on the sense DNA (the coding strand) of each exact repeat annotation (based on RepeatMasker) in B104-1-1 genome. **G)** The bar plot showing the numbers of overlapped 6mA sites (allowing ± 500 bp window) on the antisense DNA (the template strand) of each exact repeat annotation in B104-1-1 genome. **H)** The 2-D plot of the numbers of overlapped B104-1-1 6mA sites (allowing ± 500 bp window) on the sense DNA (the coding strand) of each exact repeat annotation and their average 6mA modification fractions (after

(**Figure 3.4, continued**) antibody enrichment). **I)** The 2-D plot of the numbers of overlapped B104-1-1 6mA sites (allowing ± 500 bp window) on the antisense DNA (the template strand) of each exact repeat annotation and their average 6mA modification fractions (after antibody enrichment). **J)** Pie chart showing the numbers of B014-1-1 6mA sites (overlapped in replicates, allowing ± 500 bp window) across the sense and antisense strands of intronic regions. **K)** The enriched GO clusters of the genes with 6mA modification at the intronic regions in B104-1-1 cells. **L)** Bright-field photographs of NIH/3T3 and B104-1-1 cells. **M)** Plots showing the top 30 consensus motifs containing 6mA sites in gDNA from NIH/3T3 (ATCC) and B104-1-1 cells, uncovered by DR-6mA-seq. Some of the representative consensus motifs identified in both cells were labeled with the same colors. **N)** Plots showing the distribution of possibilities for different numbers of overlaps for NIH/3T3 and B104-1-1 gDNA 6mA sites, respectively. Possibilities were generated by randomly shuffling the ± 500 bp region sets across the genome 1,000 times and counting the number of overlaps. The dashed line refers to the actual numbers of overlaps in Figure 3.3. **O)** Two representative cell-line-specific gDNA 6mA clusters located at *Abcc5* and *Erich6* in B104-1-1 cells.

3.2.5 6mA methylations in mouse glioblastoma model cells overlap with heterochromatic histone modifications

It is known that the crosstalk between histone modifications and DNA modifications plays a crucial role in regulating gene expression and chromatin structure. Histone modifications, including acetylation, methylation, phosphorylation, ubiquitination, and others, can influence DNA modifications such as DNA methylation and hydroxymethylation. Conversely, DNA modifications can also impact the deposition and stability of histone modifications. This intricate interplay between histone and DNA modifications adds an additional layer of complexity to the regulation of gene expression. For instance, there is an agreement on the universal coexistence of both histone deacetylation and 5mC at silenced loci.[153]

Xie *et al.* (2018) reported that 6mA methylations in mouse glioblastoma cells can co-localize with heterochromatic histone modifications, predominantly H3K9me3 and H3K27me3.[81] In order to further investigate this phenomenon, we performed chromatin immunoprecipitation sequencing (ChIP-seq) experiments for H3K9me3, H3K27me3, and H3K4me3 marks in two biological replicates of NIH/3T3 and B104-1-1 cell lines (Supplementary Table 8).

Although the 6mA methylated sites in NIH/3T3 genomic DNA sites did not show a significant overlap with the ChIP-seq peaks of three tested histone modifications (Figure 3.5B, 3.6A-C), we observed that approximately 25% of the 6mA sites in B104-1-1 (glioblastoma) genomic DNA exhibited a striking overlap with H3K27me3, followed by H3K9me3 (Figure 3.5A, 3.6D-F), which is consistent with the previous report.[81] We did not observe a strong correlation between 6mA sites that overlap with H3K9me3 peaks and those overlap with H3K27me3 peaks (Figure 3.5C). The further annotation analysis revealed that the B104-1-1 6mA sites overlapped with H3K27me3 peaks were primarily located within regions encoding repeat-associated RNAs (Figure 3.5D), with a stronger accumulation in the antisense strand (the templated strand) of LINE1, *Alu* elements, B2, B4, ERVK, and MaLR regions (Figure 3.5E-F). Notably, these 6mA sites mostly fall on the centers of H3K27me3 ChIP-seq peaks (Figure 3.5G).

In conclusion, our findings have underscored the potential interplay between genomic 6mA modifications and heterochromatic histone modifications, specifically within glioblastoma-associated cells, but not in the parental non-cancerous cells. Further investigations are warranted to elucidate the mechanisms underlying the epigenetic crosstalk between histone modifications and mammalian 6mA, as well as to determine the functional implications of this interplay.

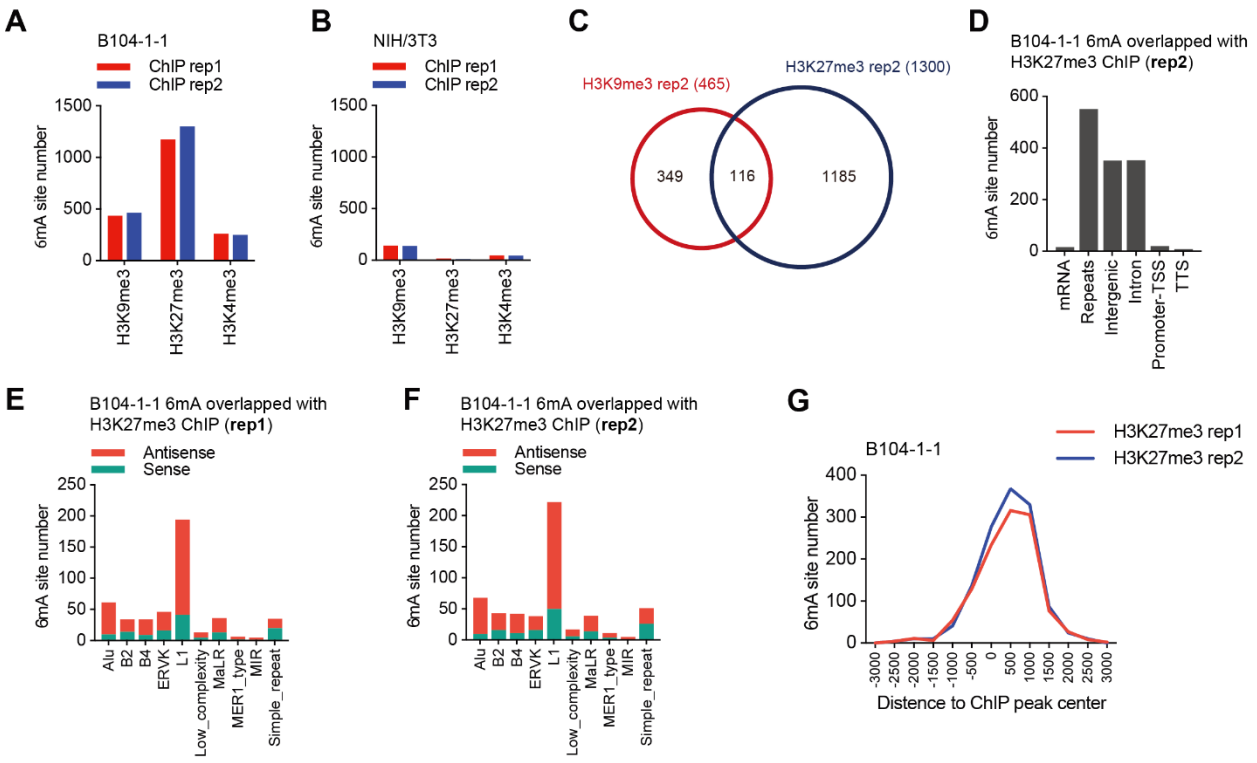


Figure 3.5 6mA in B104-1-1 cells significantly overlap with certain histone modifications

A) The bar plot showing the numbers of 6mA sites (the sites that overlapped ± 500 bp in two DR-6mA-seq replicates, with flanking 1 kb regions) that fall into H3K3me3, H3K27me3, and H3K4me3 ChIP-seq peaks in B104-1-1 cells. **B)** The bar plot showing the numbers of 6mA sites (the sites that overlapped ± 500 bp in two DR-6mA-seq replicates, with flanking 1 kb regions) that fall into H3K9me3, H3K27me3, and H3K4me3 ChIP-seq peaks in NIH/3T3 cells. **C)** Venn diagram showing the overlap of B104-1-1 6mA sites (the sites that overlapped ± 500 bp in two DR-6mA-seq replicates, with flanking 1 kb regions) that respectively fall into H3K9me3 and H3K27me3 ChIP-seq (replicate 2) peaks. **D)** The bar plot showing the numbers of B104-1-1 6mA sites (the sites that overlapped ± 500 bp in two DR-6mA-seq replicates, with flanking 1 kb regions) that fall into H3K27me3 ChIP-seq (replicate 2) peaks, in each annotated genomic element. **E)** The bar plot showing the numbers of B104-1-1 6mA sites (the sites that overlapped ± 500 bp in two DR-6mA-seq replicates, with flanking 1 kb regions) that fall into H3K27me3 ChIP-seq (replicate 1) peaks, in the antisense (the template strand) vs. sense (the coding strand) DNA of each annotated repeats family. **F)** The bar plot showing the numbers of B104-1-1 6mA sites (the sites that overlapped ± 500 bp in two DR-6mA-seq replicates, with flanking 1 kb regions) that fall into H3K27me3 ChIP-seq (replicate 2) peaks, in the antisense (the template strand) vs. sense (the coding strand) DNA of each annotated repeats family. **G)** The histogram indicating the distribution of B104-1-1 6mA sites (the sites that overlapped ± 500 bp in two DR-6mA-seq replicates, with flanking 1 kb regions) regarding their distance to the center of H3K27me3 ChIP-seq peaks.

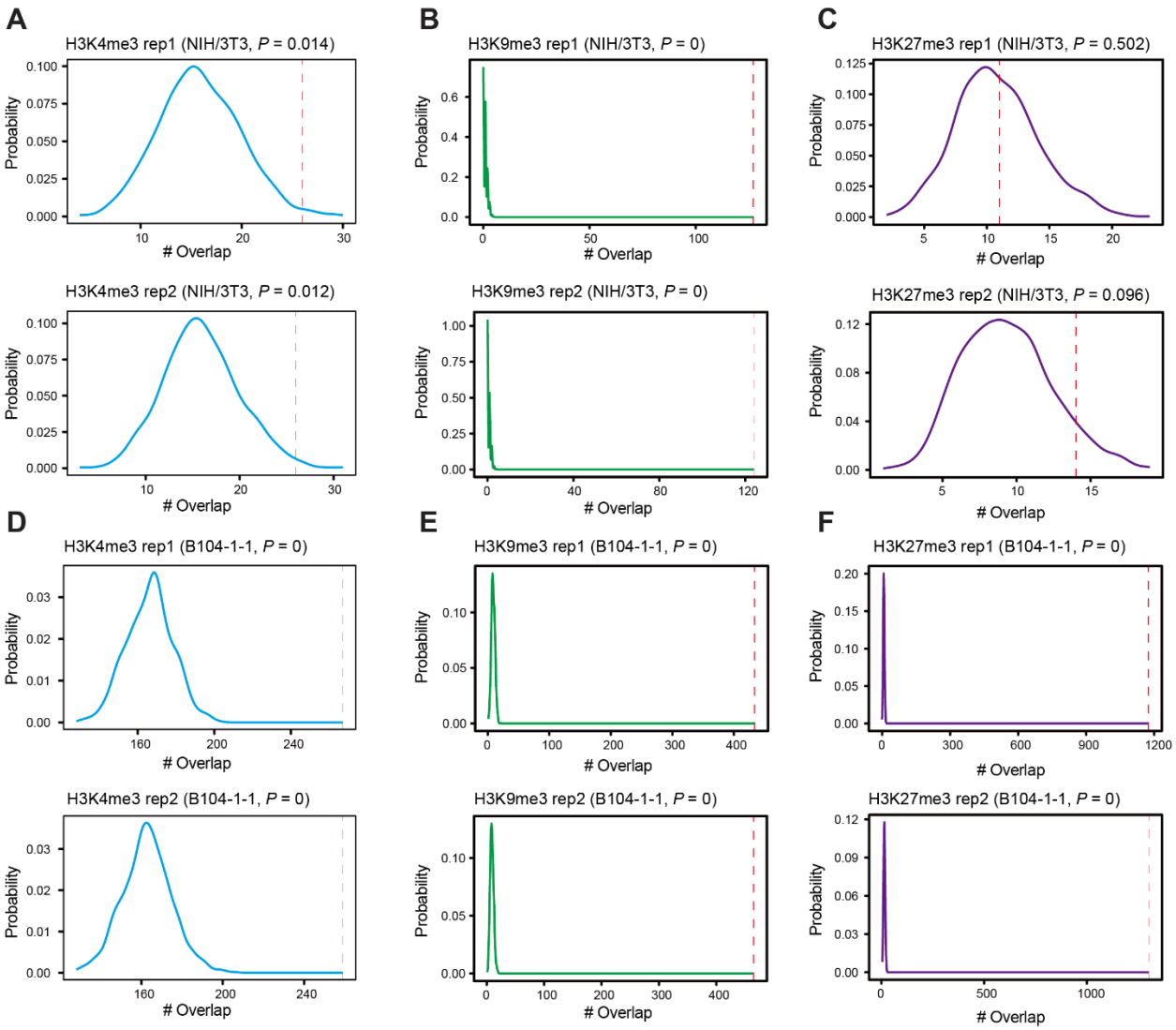


Figure 3.6 Statistics of the overlap between 6mA sites identified in B104-1-1 cells and multiple histone modification ChIP-seq peaks

A) Plot showing the distribution of possibilities for different numbers of overlapped sites with H3K4me3 ChIP-seq peaks (two replicates plotted separately) for NIH/3T3 dataset. Possibilities were generated by randomly shuffling region-level 6mA sites (flanking 1-kb regions) across the genome for 1,000 times and counting the number of sites overlapped with the ChIP-seq peaks. The dashed line refers to the actual numbers of overlapped sites in Figure 3.5. **B)** Plot showing the distribution of possibilities for different numbers of overlapped sites with H3K9me3 ChIP-seq peaks (two replicates plotted separately) for NIH/3T3 dataset. Possibilities were generated by randomly shuffling region-level 6mA sites (flanking 1-kb regions) across the genome for 1,000 times and counting the number of sites overlapped with the ChIP-seq peaks. The dashed line refers to the actual numbers of overlapped sites in Figure 3.5. **C)** Plot showing the distribution of possibilities for different numbers of overlapped sites with H3K27me3 ChIP-seq peaks (two replicates plotted separately) for NIH/3T3 dataset. Possibilities were generated by randomly shuffling region-level 6mA sites (flanking 1-kb regions) across the genome for 1,000 times and

(**Figure 3.6, continued**) counting the number of sites overlapped with the ChIP-seq peaks. The dashed line refers to the actual numbers of overlapped sites in Figure 3.5. **D)** Plot showing the distribution of possibilities for different numbers of overlapped sites with H3K4me3 ChIP-seq peaks (two replicates plotted separately) for B104-1-1 dataset. Possibilities were generated by randomly shuffling region-level 6mA sites (flanking 1-kb regions) across the genome for 1,000 times and counting the number of sites overlapped with the ChIP-seq peaks. The dashed line refers to the actual numbers of overlapped sites in Figure 3.5. **E)** Plot showing the distribution of possibilities for different numbers of overlapped sites with H3K9me3 ChIP-seq peaks (two replicates plotted separately) for B104-1-1 dataset. Possibilities were generated by randomly shuffling region-level 6mA sites (flanking 1-kb regions) across the genome for 1,000 times and counting the number of sites overlapped with the ChIP-seq peaks. The dashed line refers to the actual numbers of overlapped sites in Figure 3.5. **F)** Plot showing the distribution of possibilities for different numbers of overlapped sites with H3K27me3 ChIP-seq peaks (two replicates plotted separately) for B104-1-1 dataset. Possibilities were generated by randomly shuffling region-level 6mA sites (flanking 1-kb regions) across the genome for 1,000 times and counting the number of sites overlapped with the ChIP-seq peaks. The dashed line refers to the actual numbers of overlapped sites in Figure 3.5.

3.2.6 Mammalian 6mA validation

In order to detect and characterize the 6mA modification in mammalian nuclear DNA (nDNA), we implemented a two-step approach. Firstly, due to the low levels of 6mA in mammalian nDNA, we employed an antibody-based enrichment. This was followed by DR-6mA-seq to identify and locate the presence of 6mA modification. Our results obtained from the IP-enriched gDNA, thus, only provided information of the locations of highly-confident 6mA sites, without indicating the exact modification fraction at each individual site. To accurately quantify the actual modification fractions at the 6mA sites identified by DR-6mA-seq, we subsequently conducted an amplicon sequencing assay targeting representative 6mA sites. Unlike the standard DR-6mA-seq library construction process, the amplicon assay required consideration of the extension stop that occurred at the 6mA sites in order to calculate the precise mutation ratio at each 6mA site (Figure 3.7A).

We selected 16 6mA sites in NIH/3T3 cells and B104-1-1 cells from the detected gDNA 6mA sites by DR-6mA-seq. These sites had suitable surrounding sequences for designing primers

for the amplicon assay, ensuring there were no repeat sequences, and showed up in both two biological replicates, allowing us to obtain dual 6mA fraction values for each modified site (Supplementary Table 9). To determine the actual 6mA modification fractions, we performed amplicon sequencing using gDNA without antibody enrichment. The results showed that most 6mA sites had mutation ratios below 5%, indicating low 6mA fractions (Figure 3.7B). By converting the mutation ratios to 6mA modification fractions using calibration curves (Figure 3.1B, F), we identified two sites with approximately 40-60% 6mA fractions, four sites with approximately 5-20% 6mA fractions, and the remaining sites with less than 2% 6mA fractions (Figure 3.7C). Interestingly, we found the highest modified 6mA site (~87%, chr3:141676536) in the B104-1-1 nuclear genome, specifically in the *Bmpr1b* gene, which was not present in the NIH/3T3 genome (Figure 3.7G). It is worth noting that *Bmpr1b* is closely associated with the tumorigenicity of glioblastoma and is exclusively expressed in B104-1-1 but not in NIH/3T3 (Figure 3.7F). This observation suggests a potential regulatory role of 6mA in the tumorigenicity of B104-1-1.[154,155] Additionally, the orthogonal amplicon assay provided further evidence supporting the reliability of DR-6mA-seq and confirmed the presence of 6mA in mammalian gDNA. However, it is important to note that only a small number of sites showed high fractions of 6mA, indicating potential functionality.

To establish an independent and reliable method for validating DNA 6mA modifications, we made modifications to a previously reported technique that utilizes a silver-ion-mediated base-pairing affinity assay.[156] This assay relies on the selective stabilization of the dA-dCTP mismatch by a suitable concentration of Ag^+ ions, which promotes primer extension. In contrast, the complex formed between 6mA, Ag^+, and dCTP is unstable and leads to the termination of primer extension. To validate the presence of 6mA modifications, we selected the 6mA site with the highest

modification fraction, as determined by the amplicon assay, in glioblastoma gDNA (located at chr3:141676536, *Bmpr1b*). We then performed the silver-ion-mediated base-pairing affinity assay using biotinylated dCTP on both untreated gDNA and FTO-treated gDNA samples (Figure 3.7D). Through RT-qPCR quantification, we observed a slightly higher level of biotinylated DNA in the FTO-treated samples, indicating a higher dCTP incorporation ratio and confirming the presence of 6mA methylation at mammalian genomic 6mA sites (Figure 3.7E).

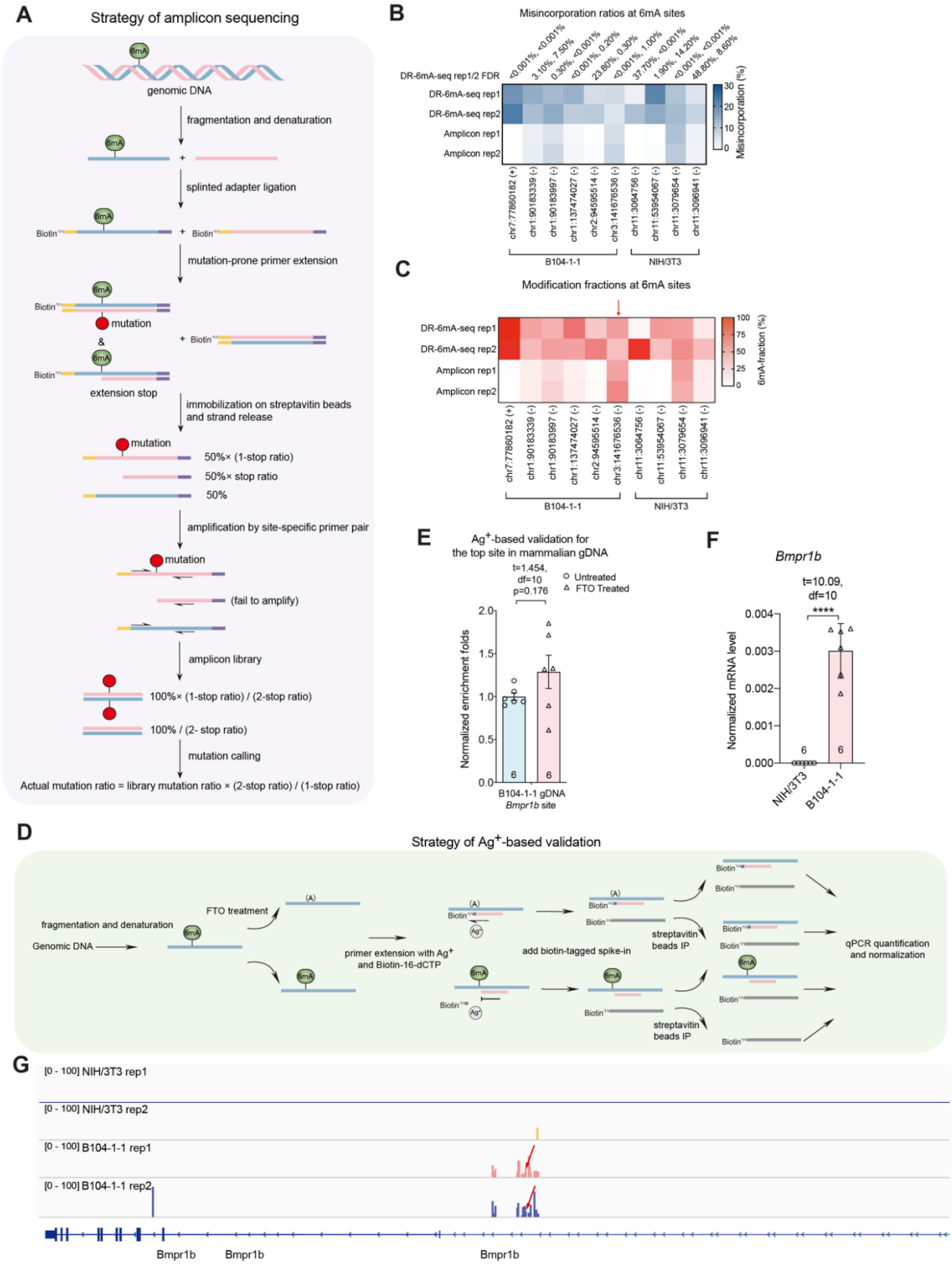


Figure 3.7 Validation of mammalian 6mA sites by amplicon sequencing and silver-ion-mediated base-paring affinity assay

(**Figure 3.7, continued**) **A)** A flowchart of amplicon assay for gDNA 6mA validation, revealing 6mA fraction by adjusted misincorporation ratios. **B)** The heatmap showing the actual misincorporation ratios obtained in amplicon assay versus the misincorporation calculated from DR-6mA-seq data. The FDR for each site calculated by DR-6mA-seq was marked on the top. **C)** The heatmap showing the 6mA methylation fractions by amplicon assay versus 6mA fractions calculated from DR-6mA-seq data. The gDNA 6mA site showing the highest estimated modification in amplicon assay was marked by red arrow. For B) and C), 10 gDNA 6mA sites were investigated, with 4 from NIH/3T3 cells and 6 from B104-1-1 cells. **D)** Schematic diagram of silver-ion-mediated base-paring affinity assay for 6mA site validation. **E)** The normalized enrichment folds at the 6mA site of the highest modification fraction (chr3:141676536) identified in C), measured by RT-qPCR-assisted silver-ion-mediated base-paring affinity assay, in the presence and absence of FTO treatment. Data are mean ± s.e.m.; analyzed by two-tailed unpaired t-test. The number of independently repeated reactions is shown in the plot. **F)** RT-qPCR quantification of *Bmpr1b* expression level in NIH/3T3 and B104-1-1 cells, normalized to *Actb*. Data are mean ± s.e.m.; analyzed by two-tailed unpaired t-tests. The number of independently repeated reactions is shown in the plot. **** p-value <0.0001. **G)** A representative 6mA cluster located at *Bmpr1b*, with the arrow marking the 6mA site of the highest modification fraction (chr3:141676536) identified in C).

3.3 Discussion and Conclusion

6mA was initially discovered in *Bacterium coli* during the mid-twentieth century.[53] Since then, it has been increasingly recognized as a crucial DNA modification in prokaryotes and protists, although its presence and functions in mammals remain less understood. While 6mA is found in the genomes of lower eukaryotes and invertebrates, exhibiting distinct distribution patterns and playing regulatory roles in transcription, its detection and characterization in mammalian genomic DNA has been a subject of intense debate primarily due to the lack of highly sensitive detection methods.[63,157–161] We have previously suggested that 6mA might have limited yet regulatory functions in specific cells or during specific biological processes, given its extremely low abundance in mammalian genomes.

In this chapter, we present DR-6mA-seq, a highly sensitive sequencing method capable of probing 6mA modifications in genomic DNA at base resolution, without the need for antibodies. This method also demonstrates high specificity to 6mA, as validated by the FTO-mediated 6mA

demethylation control. To validate the efficacy of DR-6mA-seq, we utilized *E. coli* K-12 gDNA, a well-established system for analyzing the 6mA methylome. Our LC-MS/MS analysis revealed that the levels of gDNA 6mA in most human tissues were below the detectable limit. However, we successfully detected visible levels of 6mA in mouse testis gDNA and mouse glioblastoma cell gDNA through mass spectrometry. By employing DR-6mA-seq, we generated the initial maps of 6mA distribution in these gDNA samples. Furthermore, we employed amplicon sequencing to precisely determine the fraction of 6mA modification at specific sites. As anticipated, most sites exhibited low levels of modification, but we did observe certain sites with modification fractions ranging from 70% to 80%, indicating potential functional significance. Additionally, we optimized an orthogonal silver-ion-mediated base-pairing affinity assay and confirmed the findings obtained from our DR-6mA-seq method.

Our results from DR-6mA-seq have demonstrated the enrichment of 6mA modification in various regions of the mammalian genome, including intergenic regions, intronic regions, LINE elements, SINE elements, and LTR (Figure 3.3M), which is consistent with previous studies using DIP-seq.[71] The abundance of 6mA in mammalian genomes varies significantly across different tissues, and its levels can be increased in glioblastoma cells upon the expression of a single oncogene, leading to high modification fractions at specific sites. These findings provide evidence for the presence of 6mA modification, although its occurrence is limited to specific sites and with varying modification fractions, suggesting potential functional relevance in specific cell types (such as glioblastoma) or biological processes (such as development). Moreover, our results support the hypothesis of crosstalk between 6mA and heterochromatic histone modifications, particularly H3K27me3. We anticipate that future applications of DR-6mA-seq will not only enable the characterization of 6mA distribution in different genomes of interest but also facilitate

the identification of dynamic changes at modification sites, thereby unraveling their functional implications. Lastly, the utilization of modified dNTPs to induce elevated misincorporation opposite a modified base through weakened Watson-Crick base pairing represents a promising strategy for the detection of other nucleic acid base modifications with enhanced sensitivity.

3.4 Methods

3.4.1 Materials and resource

Table 2 Reagents and resources for 3.4

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-N^6-Methyladenosine (m6A) antibody	Sigma	ABE572
Anti-N^6-Methyladenosine (m6A) antibody	Abcam	ab151230
Anti-H3K9me3 antibody	Cell Signaling Technology (CST)	13969
Anti-H3K27me3 antibody	Cell Signaling Technology (CST)	9733
Bacterial and Virus Strains		
***E. coli* BL21 (DE3)**	Thermo Scientific	EC0114
Biological Samples		
***E. coli* K-12 gDNA**	ATCC	10798D-5
Human brain gDNA	Zyagen	HG-201
Human liver gDNA	Zyagen	HG-314
Human heart gDNA	Zyagen	HG-801
Human testis gDNA	Zyagen	HG-401
Mouse Genomic DNA (control)	Sigma	69239
Human Genomic DNA (control)	Sigma	69237
Chemicals, Peptides, and Recombinant Proteins		
2-Thio-dTTP	Trilink	N-2035
Biotin-dCTP	Lumiprobe	2715
Recombinant human FTO protein	This study	This study
Ammonium iron(II) sulfate hexahydrate	Sigma	7783-85-9
L-Ascorbic acid	Sigma	50-81-7
α-Ketoglutaric acid	Sigma	328-50-7
Silver nitrate solution	Sigma	7761-88-8
Deposited Data		
***E. coli* K-12 SMRT sequencing data**	Gene Expression Omnibus (GEO)	GSE69872
Experimental Models: Cell Lines		
HepG2	ATCC	HB-8065
NIH/3T3	ATCC	CRL-1658
B104-1-1	ATCC	CRL-1887

(**Table 2**, continued) **NIH/3T3 (ECACC)**	Sigma	93061524-1VL
Experimental Models: Organisms/Strains		
C57BL/6J mice	The Jackson Laboratory	IMSR_JAX:000664
Oligonucleotides		
Model DNAs, adaptors, and primers used in this study	See Supplementary Table 3	
Software and Algorithms		
Trim Galore!		https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
Bowtie2	https://www.nature.com/articles/nmeth.1923	https://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Samtools	https://academic.oup.com/bioinformatics/article/25/16/2078/204688?login=true	http://www.htslib.org/
VarScan	https://genome.cshlp.org/content/22/3/568	http://dkoboldt.github.io/varscan/
Bedtools	https://academic.oup.com/bioinformatics/article/26/6/841/244688	https://bedtools.readthedocs.io/en/latest/
Homer	https://www.sciencedirect.com/science/article/pii/S1097276510003667?via%3Dihub	http://homer.ucsd.edu/homer/
STAR 2.7.10a	https://academic.oup.com/bioinformatics/article/29/1/15/272537	https://github.com/alexdobin/STAR
MACS2 2.2.7.1	ttps://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-9-r137	https://pypi.org/project/MACS2/
DeepTools v3.5.0	https://academic.oup.com/nar/article/44/W1/W160/2499308	https://deeptools.readthedocs.io/en/develop/
IGV (version 2.6.3)	Broad Institute	https://software.broadinstitute.org/software/igv/download
Prism (version 8.4.0)		www.graphpad.com

3.4.2 Cell culture

The HepG2, NIH/3T3, and B104-1-1 cell lines were obtained from ATCC and cultured in DMEM (Gibco, 11965092) supplemented with 10% FBS (Gibco, 26140079). The cells were incubated at 37°C with 5% CO2. The NIH/3T3 cells obtained from ECACC were maintained under the same conditions but were only used for sex determination purposes. To ensure purity of the isolated genomic DNA used in this study, mycoplasma contamination tests was performed using the LookOut Mycoplasma PCR Kit (Sigma, MP0035).

3.4.3 Development of DR-6mA-seq protocol using synthetic DNA.

The 100-mer biotin-tagged NN-6mA-NN model DNA oligo and primer-extension primer (Supplementary Table 9) were synthesized by Integrated DNA Technologies (IDT). Initially, the primer-extension primer was denatured at 70°C for 2 minutes and immediately cooled on ice. In a total volume of 50 μL, 50 ng of the model DNA and 2.5 μL of 2 μM oligo-primer-extension primer were mixed with various DNA polymerases, including *OneTaq* DNA Polymerase, Vent DNA Polymerase, phi29 DNA Polymerase, Pwo SuperYield DNA Polymerase, EpiMark Hot Start *Taq* DNA Polymerase, *Bst* DNA Polymerase-full length, *Bst* 2.0 DNA Polymerase, *Bst* 2.0 WarmStart DNA Polymerase, *Bst* 3.0 DNA Polymerase, and Recombinant HIV reverse transcriptase. The reactions were performed in the supplied buffers, with the addition of different concentrations of magnesium sulfate, manganese chloride, polyethylene glycol (PEG8000, PEG4000, PEG1000, PEG600, PEG600, PEG200), dATP/dCTP/dGTP, and modified dTTP/UTP (dUTP, UTP, 2-Thio-dTTP, 4-Thio-dTTP, 5-Aminoallyl-dUTP, Fluorescein-12-dUTP). Primer extension tests were conducted at different temperatures. The newly synthesized strands resulting from the extension reaction were subsequently released from the biotinylated template strands immobilized on Streptavidin C1 beads (Thermo Fisher) by incubation in 150 mM NaOH. The released strands were neutralized, cleaned up using the Oligo Clean & Concentrator Kit (Zymo Research, D4061), and incubated with Streptavidin C1 beads again to completely remove the template strands. The supernatant was purified using the Oligo Clean & Concentrator Kit and amplified for 10 cycles using the NEBNext Ultra II Q5 Master Mix (NEB, M0544S) and NEBNext Multiplex Oligos for Illumina (NEB, E7335S). The resulting libraries were purified by gel electrophoresis and recovery (NEB, T1020S). All libraries were sequenced at a depth of 2 million reads in single-ended 100 bp mode using the NovaSeq 6000 sequencer (Illumina). The final primer extension conditions were as follows: 1500 units of *Bst* 2.0 DNA polymerase (NEB, M0537M), 1× EpiMark Hot Start *Taq*

Reaction Buffer (NEB, M0490S), 8 mM dATP (NEB, N0446S), 2 mM dCTP (NEB, N0446S), 2 mM dGTP (NEB, N0446S), 0.02 mM 2-Thio-dTTP (Trilink, N-2035), 6 mM magnesium sulfate (NEB), and 8% PEG4000 (Rigaku, #25322-68-3). All reactions were incubated at 60°C for 1 hour.

3.4.4 Expression and purification of recombinant human FTO protein.

Previously, the human FTO gene (GenBank Accession No. NP_001073901.1) was subcloned into the pET28a vector (Novagen) and transformed into the BL21 (DE3) *Escherichia coli* strain.[10.1016/bs.mie.2015.03.013] To induce protein expression, when the optical density at 600 nm reached 1.0, the cells were cooled to 16 °C. For a 1L bacterial culture, 1 mL of 100 mM IPTG and 200 µL of a 4 µg/µL $(NH_4)_2Fe(SO_4)_2$ solution were added, and the cells were cultured at 16 °C for an additional 16 hours. The cells from the 1L culture were pelleted and resuspended in 40 mL of 1X Lysis Buffer (20 mM Tris-HCl pH 7.5, 300 mM NaCl, 5 mM imidazole), which contained 1 tablet of Roche cOmplete Protease Inhibitor Cocktail (EDTA-free). The cells were then sonicated and centrifuged. The supernatant containing the soluble recombinant protein was filtered through 0.22 µm syringe filters and purified using Ni Sepharose 6 Fast Flow (GE Healthcare). After washing the agarose once with 50 mL of 1X Lysis Buffer, the agarose was washed with two different washing buffers. The first wash buffer (Wash Buffer A) contained 20 mM Tris-HCl pH 7.5, 500 mM NaCl, 5 mM imidazole, and 1 tablet of Roche cOmplete Protease Inhibitor Cocktail (EDTA-free). The second wash buffer (Wash Buffer B) contained 20 mM Tris-HCl pH 7.5, 300 mM NaCl, 25 mM imidazole, and 1 tablet of Roche cOmplete Protease Inhibitor Cocktail (EDTA-free). The protein was eluted using 14 mL of Elution Buffer (20 mM Tris-HCl pH 7.5, 300 mM NaCl, 250 mM imidazole). The collected flowthrough was initially concentrated by centrifugation and then diluted with 1X Resuspension Buffer (20 mM Tris-HCl pH 7.5, 300 mM NaCl). The diluted protein was re-concentrated five times to remove imidazole. Finally, the

protein was concentrated to approximately 22 mg/mL using Ultra 0.5 Centrifugal Filters (Millipore, UFC503096). Glycerol was added to the enzyme to achieve a final concentration of 25%, and the protein was stored at -80 °C for future use..

3.4.5 Preparation of DNA samples.

The *Escherichia coli* strain K-12 genomic DNA was acquired from ATCC (10798D-5). Genomic DNA samples from the human brain, liver, heart, and testis were obtained from Zyagen (HG-201; HG-314; HG-801; HG-401). Genomic DNA from NIH/3T3 cells, B104-1-1 cells, 4-month-old male C57BL/6J mouse whole testis, 4-month-old male C57BL/6J mouse whole brain, and 2.5-month-old female C57BL/6J mouse whole brain were all prepared using the Monarch Genomic DNA Purification Kit (NEB, T3010S). Prior to library construction, all DNA samples underwent a rigorous treatment with RNase A (NEB, T3018L) to eliminate any RNA contaminants..

3.4.6 Preparation of synthetic unmodified DNA samples.

The unmodified *E. coli* K-12 DNA was synthesized using the REPLI-g Mini Kit (Qiagen, 150023). A total of 1 ng of *E. coli* K-12 DNA from each of the three biological replicates was used as the template, resulting in a yield of over 1 μg of synthetic DNA.

3.4.7 Detection of bacterial DNA contamination in genomic DNA samples.

For the survey of the 16S rRNA gene against diverse bacterial species, a qPCR protocol was employed, as described in a previous publication.[162] The primer pair 799f/1193r was utilized to ensure high phylogenetic richness and low non-specificity. To serve as positive controls, previously reported primer pairs targeting the mouse and human 18S rDNA were used (Supplementary Table 3).[163,164] As negative controls, standard mouse and human genomic DNA were purchased from Sigma (cat# 69239, 69237). A quantity of 1 ng of genomic DNA was used

per qPCR reaction. The qPCR reactions were set up using SYBR™ Select Master Mix (Applied Biosystems, 4472908), and 40 cycles of PCR were performed.

3.4.8 LC-MS/MS analysis.

Validation of FTO demethylation activity on ssDNA:

The 100-mer biotin-tagged NN-6mA-NN model DNA oligo was hybridized with its complementary oligo (Supplementary Table 3) in Nuclease Free Duplex Buffer (IDT, 11-05-01-12) at a mole ratio of 1:3. A total of 300 ng denatured NN-6mA-NN model single-stranded DNA (ssDNA) oligo, as well as 300 ng annealed NN-6mA-NN model double-stranded DNA (dsDNA), were incubated at 37℃ for 4 hours in a 500 μL reaction solution containing 50 mM HEPES pH 7.0, 60 mM KCl, 75 μM ammonium iron(II) sulfate hexahydrate, 2 mM L-ascorbic acid, 300 μM α-ketoglutarate, and 2.4 μM recombinant FTO protein. The demethylation reactions were subsequently purified using the Oligo Clean & Concentrator Kit (Zymo research, D4061) for ssDNA and the DNA Clean & Concentrator Kits (Zymo research, D4003) for dsDNA, followed by digestion with the Nucleoside Digestion Mix (NEB, M0649S) at 37 ℃ for 2 hours. The digested DNA was then filtered through a 0.22 μm filter (Millipore, SLGVR04NL).

For UHPLC-QqQ-MS/MS analysis, a 10 μL sample was injected, and the nucleosides were separated using reverse-phase UHPLC on a C18 column (Agilent, 927,700-092) followed by MS detection using an Agilent 6460 QQQ-MS/MS set to multiple reaction monitoring (MRM) in positive electrospray ionization mode. Nucleosides were quantified using the nucleoside precursor ion to base ion mass transitions of 266.1-150.0 for 6mA and 268.1-152.0 for dG. The concentration of nucleosides was determined using calibration curves generated from nucleoside standards run under the same conditions.

Quantification of gDNA 6mA in biological samples:

Cell cultures were treated with plasmocin (InvivoGen, ant-mpt) to eliminate any potential mycoplasma contamination. Genomic DNA extracted from various cell lines (such as NIH/3T3 and B104-1-1 cells) and tissues (including *E. coli* K-12, mouse testis, mouse E15.5 embryo, human liver, human brain, and human testis) was subjected to digestion using Nucleoside Digestion Mix (NEB, M0649S) with 1X Reaction Buffer (50 mM potassium acetate pH 5.4, 1 mM ZnCl2) at 37 ℃ for 2 hours. For UHPLC-QqQ-MS/MS analysis, a 10 μL sample was injected and analyzed using the standard procedures with a modified UHPLC protocol. To enhance the detection of 6mA signal and achieve better separation from other peaks, we adjusted the UHPLC protocol as follows: 98% A + 2% B at 0 minutes; 97% A + 3% B at 8 minutes; 93% A + 7% B at 9 minutes; 89% A + 11% B at 10 minutes; 50% A + 50% B from 10.25 to 12.50 minutes; 98% A + 2% B at 12.75 minutes; where A = H2O + 0.1% HCOOH and B = MeOH + 0.1% HCOOH.

3.4.9 Optimized DR-6mA-seq protocol for biological DNA samples.

For *E. coli* DNA and mtDNA, we diluted 400 ng~1 μg of RNA-free DNA in TE buffer (Invitrogen, AM9849) with 0.1 M NaOH and sonicated it to a length of 100-300 nt using the Bioruptor Pico sonication device. In the case of mammalian gDNA, antibody enrichment was performed before library construction. We diluted 50 to 100 μg of RNA-free DNA in TE buffer (Invitrogen, AM9849) with 0.1 M NaOH and sonicated it to a length of 100-300 nt. The fragmented DNA was then immunoprecipitated using 25 μg of anti-6mA antibody (Sigma, ABE572 or Abcam, ab151230) by gently rotating it at 4°C overnight. We washed Pierce Protein A Magnetic Beads (Thermo Scientific, 88845) and used them to pull down the antigen-antibody complex. Subsequently, we treated the beads with proteinase K (NEB, P8111S), and the 6mA-enriched DNA was purified using the DNA Clean & Concentrator Kit-25 (Zymo Research, D4033).

The fragmented DNA was mixed with 15% (mass percentage) spike-in probes (Supplementary Table 3) and then divided into three parts in a ratio of 3:4:1. Each part was used for the mutation group, demethylation control, and high-fidelity control. For the demethylation control, the DNA fragments were denatured and incubated at 37°C for 4 hours in a 300 μL reaction containing 50 mM HEPES (pH 7.0), 60 mM KCl, 75 μM ammonium iron (II) sulfate hexahydrate, 2 mM L-ascorbic acid, 300 μM α-ketoglutarate, and 2.4 μM FTO purified protein. The demethylation reaction was then treated with proteinase K (NEB, P8111S) and cleaned up using the Oligo Clean & Concentrator Kits (Zymo Research, D4061).

The three groups of DNA were denatured and subjected to splint ligation following a previous protocol (Supplementary Table 3). The ligation products of the mutation group and demethylation control underwent primer extension at 60°C for 3 hours in a 50 μL reaction containing 6 μM biosample-primer-extension primer (Supplementary Table 3), 1500 units of *Bst* 2.0 DNA polymerase (NEB, M0537M), 1× EpiMark Hot Start *Taq* Reaction Buffer (NEB, M0490S), 8 mM dATP (NEB, N0446S), 2 mM dCTP (NEB, N0446S), 2 mM dGTP (NEB, N0446S), 0.02 mM 2-thio-dTTP (Trilink, N-2035), 6 mM magnesium sulfate (NEB), and 8% PEG4000 (Rigaku, # 25322-68-3). Meanwhile, the ligation product of the high-fidelity control underwent primer extension in a 50 μL reaction containing 6 μM biosample-primer-extension primer and 1× Q5 High-Fidelity 2X Master Mix (NEB, M0492S) with 6 thermal cycles.

The synthesized strands for each group were subsequently released from the biotin-tagged template strands immobilized on Dynabeads MyOne Streptavidin C1 (Invitrogen, 65001) by incubation in 0.15 M sodium hydroxide. The released strands were neutralized, cleaned up with Oligo Clean & Concentrator Kits, and incubated again with Dynabeads MyOne Streptavidin C1 to completely remove the template strands. The supernatant was cleaned up with Oligo Clean &

Concentrator Kits and amplified for 10 to 12 cycles using the KAPA HiFi Uracil+ Master Mix (Roche, KK2801) and NEBNext Multiplex Oligos for Illumina (NEB, T1020S). The resulting libraries were purified by gel electrophoresis and recovery and finally sequenced at a depth of 60 to 100M with single-ended 100 bp mode.

3.4.10 Genetic sex determination of NIH/3T3 and B104-1-1 cell culture.

The methodology used for determining the sex of mouse cell lines was adapted from a previous study.[165] Genomic DNA was extracted from female and male brain samples of C57BL/6 mice, as well as from NIH/3T3 cell cultures (ATCC, CRL-1658, and Sigma 93061524-1VL) and B104-1-1 cell culture (CRL-1887). The DNA was then amplified using Q5 High-Fidelity 2X Master Mix (NEB, M0492S) and *Rbm31x/y*-F/R primers (Supplementary Table 3).

3.4.11 Quantification of 6mA fraction on specific sites by amplicon sequencing.

Genomic DNA was isolated from NIH/3T3 cells and B104-1-1 cells and subjected to a modified version of the DR-6mA-seq protocol to induce mutations at 6mA sites during extension. Adaptor ligation was not performed on the treated genomic DNA, which was then utilized for the amplicon assay. For 6mA candidate sites with high mutation ratios or high modification fractions in NIH/3T3 and B104-1-1 cell lines, a 500-nt window was defined for designing amplicon primers. These primers covered a region of 150-300 nt, including the target 6mA site.

Illumina sequencing barcodes were incorporated into the amplicon primers. Approximately 1-2 ng of the treated genomic DNA was used for amplicon PCR amplification, consisting of 40 cycles performed at the optimized annealing temperature for each primer pair. The PCR reaction mixture was subsequently purified using agarose gel, and size selection was employed to isolate the target band representing the PCR product from the specific 6mA site region. All libraries were sequenced on the NextSeq 500 platform using paired-end 300 bp mode.

The sequencing reads were aligned to the 500-nt region, allowing for a maximum of 20 mismatches per read. The stop ratio at the 6mA site was estimated by comparing the read coverage of the two strands, which were differentiated by Illumina sequencing. The actual mutation ratio at the 6mA site was determined by comprehensively calculating the mutation ratio and stop ratio observed in the amplicon assay. This was done using the formula depicted in Figure 6A. In summary, the actual mutation ratio was calculated as follows: (mutation ratio in amplicon assay) × (2.0 – stop ratio) / (1.0 – stop ratio).

3.4.12 Library construction for ChIP-seq

NIH/3T3 and B104-1-1 cells were cultured in 10-cm dishes with an initial confluency of 30% in antibiotic-free medium. After 48 hours for NIH/3T3 cells and 24 hours for B104-1-1 cells, the cell pellet containing approximately 7 million cells per dish was collected. Since this study involved testing three histone modifiers, six dishes of NIH/3T3 cells and six dishes of B104-1-1 cells were prepared, with two biological replicates for each cell line. Each ChIP-seq reaction was initiated with approximately 7 million cells, following the standard protocol of the iDeal ChIP-seq Kit for Histones (Diagenode, C01010051). The ChIP-seq grade antibody H3K4me3 (1 μg/μl) was provided by the iDeal ChIP-seq Kit for Histones, while H3K9me3 (D4W1U, Rabbit mAb, #13969, diluted 1:50) and H3K27me3 (C36B11, Rabbit mAb, #9733, diluted 1:50) antibodies were separately purchased from Cell Signaling. The immunoprecipitated DNA was further processed using the NEBNext® Ultra™ II DNA Library Prep Kit (NEB, E7645S) for library preparation, followed by sequencing on the NextSeq 2000 platform in paired-end 75 bp mode.

3.4.13 Validation of 6mA sites in the mammalian genome by Ag^+ based method.

SsDNA was extracted from HepG2 mtDNA, B104-1-1 total DNA, and NIH/3T3 total DNA. The ssDNA samples were divided into two groups: one group underwent FTO demethylation

treatment, while the other served as the control. For each sample, 10 μL of ssDNA (~ 150 ng/μL) was combined with 1 μL of a 50 μM extension primer, 1 μL of 2 mM Biotin-dCTP (Lumiprobe, 2715), 5 μL of 10X Standard *Taq* Reaction Buffer (NEB, M0273S), 5 μL of 10 μL freshly prepared AgNO3, 0.25 μL of *Taq* DNA Polymerase (NEB, M0273S), and 27.75 μL of nuclease-free water. The mixture was thoroughly mixed and incubated at 95℃ for 30 seconds. This step was repeated for 20 cycles, with incubation at 95℃ for 15 seconds, 62℃ for 1 minute, and 68℃ for 10 minutes. The reaction mixture was then purified using the Oligo Clean & Concentrator kit (Zymo Research, D4061), and the DNA was eluted with 40 μL of nuclease-free water. 5 μL of a 10 nM biotinylated ssDNA spike-in oligo was then added to the eluted DNA. The mixture was thoroughly mixed, and 2 μL was saved as the 'Input' sample. The remaining 43 μL of DNA was combined with 30 μL of Streptavidin C1 beads and incubated at 4℃ for 30 minutes. The supernatant was then removed, and the beads were washed seven times with C1-Wash Buffer (50 mM HEPES pH 7.3, 300 mM NaCl, 0.05% NP40). Proteinase K treatment was performed to release the ssDNA from the beads. The eluted ssDNA was further purified using the Oligo Clean & Concentrator kit (Zymo Research, D4061), resulting in the 'Pulldown' sample. Both the 'Input' and 'Pulldown' samples were subjected to RT-qPCR assays to quantify the enrichment folds of the target 6mA region in the FTO demethylation group compared to the control.

3.4.14 Data processing for whole-genome DR-6mA-seq.

The raw sequencing reads were initially processed using Trim_Galore to eliminate adaptors and low-quality nucleotides. To evaluate the library quality and mutation levels in various sequence contexts, the high-quality reads were aligned to spike-in sequences using bowtie2. Subsequently, the reads from high-quality libraries were aligned to reference genomes using bowtie2, and only uniquely mapped reads were retained for subsequent analyses.

In the case of mutation calling for mitochondrial DNA (mtDNA), only the reads that mapped to mtDNA were kept. Conversely, for mutation calling in genomic DNA, the reads aligned to mtDNA were excluded from further analyses.

3.4.15 Statistical calling of 6mA and assessing FDR of whole-genome DR-6mA-seq.

Mutated A or T sites were identified using VarScan software with default parameters. To minimize potential artifacts resulting from low sequencing depth, only mutations with a depth of at least 20x in all replicate groups were considered. Candidate 6mA sites were defined based on the following criteria: 1) absence of mutations in the control group, 2) baseline mutations in the demethylated group, and 3) mutations in the error-prone mutation group in all replicates. To remove potential noise, an FDR-based filter was applied to our 6mA lists. Specifically, the FDR was determined by analyzing the mutation distribution in the synthetic *E. coli* negative control group. By examining the mutation ratios in the negative control group, null mutation ratio distributions were generated for each motif, representing the expected background noise in the absence of specific mutation effects. Mutation cutoffs were then determined for each motif based on a desired FDR threshold of less than 0.1% (or 0.001), considering 99.9% of mutations in the null distributions as background noise. These mutation cutoffs were subsequently applied to filter out potential noise in the biological samples. The resulting 6mA site lists were further calibrated to obtain the actual modification levels.

For 6mA calibration, a total of 256 linear regression models were generated, each corresponding to a specific nucleotide context, based on the mutation ratios of spike-in sequences. Methylation levels were predicted using these linear models according to the corresponding sequence context.

Comparison of different replicates was performed using Bedtools. To compare regions, 6mA sites were first extended by 5 bp for mtDNA and 1000 bp for gDNA using Bedtools, and then compared using the intersectBed function.

To assess the significance of overlaps between different sets of regions, we randomly shuffled the regions across the genome 1,000 times and recorded the number of overlaps in these null distributions. P-values for specific numbers of overlaps were calculated as the probability of obtaining an equal or greater number of overlaps in the null distributions. Visualization of 6mA clusters was done using IGV (version 2.6.3).

3.4.16 Data processing for ChIP-seq.

The initial step in our analysis involved trimming the raw sequencing reads to eliminate adaptors and low-quality nucleotides, using Trim_Galore. Subsequently, we aligned the high-quality reads to reference genomes using bowtie2, retaining only uniquely mapped reads for subsequent analyses. To remove PCR duplicates, we employed Samtools. The identification of peaks was performed using the MACS software.

To assess the significance of overlaps between sets of histone modification peaks and 6mA sites, we employed a random shuffling approach. Specifically, we randomly shuffled the 1-kb flanking regions of the region-level 6mA sites across the entire genome, repeating this process 1,000 times. We then determined the number of overlaps between the shuffled region sets and the histone modification peak sets, creating null distributions. The p-values for specific numbers of overlaps were defined as the probabilities of obtaining equal or greater numbers of overlaps in the null distributions.

3.4.17 Statistics

Statistical comparisons between two groups were performed using unpaired two-tailed t-tests. For comparisons among multiple groups, one-way ANOVA tests were employed. All statistical analyses and data visualization were conducted using Prism software (version 8.4.0).

3.4.18 Data availability

All sequencing data are available at NCBI Gene Expression Omnibus with the accession number: GSE213876.

3.5 Acknowledgements of work performed

I would like to thank all of our collaborators for their contributions to this work: Dr. He for supervising the whole project. Dr. Xiaolong Cui and Dr. Li-Sheng Zhang for their contribution to data analysis and 6mA site validation (data in Figure 3.7).

Chapter 4

Summary and perspectives

4.1 Coordinated transcriptional regulation and alternative splicing modulate the brain development

In Chapter 2, I conducted a comprehensive RNA-Seq analysis of alternative splicing events that exhibit differential regulation during brain development, leading to NMD of the transcripts, and are controlled by PTB proteins in mouse brains. The study identified a total of 2555 NMD-related splicing events, 3290 differential splicing events, and 1186 splicing events that are regulated by Ptbp1/2. Remarkably, 119 splicing events were found to be present in all three categories. These splicing events were found to be significantly enriched in gene clusters associated with synapse, postsynaptic density, and neuronal functions. Notably, the A3SS of *Syngap1* exon 11 was identified as an example that is mammal-specific, regulated by Ptbp1, and subject to NMD, leading to reduced expression of Syngap1 in non-neuronal cells.

As early as 2009, researchers conducted whole-genome exon-level analysis and observed that during the development of the human brain, a significant proportion of the total annotated genes, up to 76% (13223), were expressed. Among these expressed genes, 33% (4369) exhibited differential expression, while 28% (3755) underwent differential alternative splicing. It is worth noting that 44% of the expressed genes in the developing human brain demonstrated evidence of some form of differential regulation, while 17% (2260) displayed both differential expression and splicing.[166] Such a complicated and highly coordinated gene expression regulation network facilitates the unique feature of mammalian corticogenesis and advanced brain functions. Notably, in many cases, including *SYNGAP1*, alternative splicing events collaborate with transcriptional

regulation or other splicing events within the same gene to synergistically regulate gene expression. For example, in the non-neuronal cell types where the unproductive A3SS of *Syngap1* is spliced in, the transcription activity of *Syngap1* is also reduced (data not shown). This also suggests that during evolution, mammalian cells gradually acquired regulatory mechanisms that are more efficient in conserving material and energy consumption. One of the future research directions is to investigate, at the genomic level, how these transcriptional regulatory events and alternative splicing events, which play important roles in brain development, have emerged one by one during evolution and finally coordinated with each other to ensure normal neurogenesis. Studying this issue will provide important molecular biology evidence for our understanding of the evolution of advanced brain functions in mammals, especially primates.

4.2 Splice-switching ASOs targeting unproductive splicing holds promise for disease therapy

In Chapter 2, we developed and assessed the efficacy of antisense oligonucleotides (ASOs) that target the upstream intronic region of *SYNGAP1* exon 11 A3SS. Our results demonstrate that at least one splice-switching ASO (CH937) can effectively suppress the inclusion of exon 11 A3SS and significantly increase the production of productive *SYNGAP1* transcripts in human induced pluripotent stem cell (iPSC) lines, iPSC-derived neurons, and cerebral organoids. Notably, reduced expression of SYNGAP1 resulting from genomic mutations is associated with developmental abnormalities in the neural system, including nonsyndromic intellectual disability, autism spectrum disorder, and epilepsy.[167,168] Therefore, targeting *SYNGAP1* exon 11 A3SS with splice-switching ASOs represents a promising therapeutic strategy to restore SYNGAP1 expression levels by reducing the nonsense-mediated decay of transcripts produced by the normal allele.

Nevertheless, further *in vivo* studies are necessary to fully comprehend its therapeutic potential since the inclusion rate of *SYNGAP1* exon 11 A3SS in the brain during the postnatal stage is low.

There exist more than 10,000 human disorders that are attributed to a single gene mutation, also referred to as monogenic diseases.[169] Among these, a substantial proportion of monogenic disorders arise from mutations that lead to decreased expression of particular proteins. Many of these disorders arise when one copy of a gene is mutated, inactivated, or deleted, and the remaining functional copy of the gene is insufficient to produce the necessary protein to maintain normal cellular function. This condition, referred to as haploinsufficiency, arises from dominant mutations that lead to the functional loss of one allele. Haploinsufficient genes are linked to various diseases, including but not limited to neurological disorders and cognitive impairment.[170] For instance, cystic fibrosis is a genetic disease that arises from mutations in the *CFTR* gene, which encodes a protein that regulates salt and water movement in and out of cells. These mutations can lead to reduced expression or function of the CFTR protein, resulting in the hallmark symptoms of the disease. Likewise, sickle cell anemia is caused by mutations in the *HBB* gene, which encodes the beta-globin protein. These mutations can cause haploinsufficiency and reduced expression of beta-globin protein, leading to the characteristic abnormal shape of red blood cells in the disease.[171]

There are several treatment strategies available for diseases caused by haploinsufficiency. One approach involves introducing a functional copy of the affected gene into the patient's cells to restore normal protein expression by gene therapy. Another strategy is to use drugs to increase the expression of the remaining functional copy of the gene, or to compensate for the loss of protein function, such as enzyme replacement therapy for lysosomal storage disorders. Stem cell therapy involves replacing the patient's defective cells with healthy cells that can produce the missing protein. In some cases, modifying the patient's diet can help to compensate for the loss of protein

function. For example, in phenylketonuria (PKU), a genetic disorder caused by pathogenic variants in the phenylalanine hydroxylase (*PAH*) gene, patients need to avoid foods that contain phenylalanine to prevent the buildup of toxic substances in their bodies.[172] However, due to the fact that genetic diseases caused by haploinsufficiency are typically rare and involve distinct pathogenic genes, the development of relevant therapies is fraught with significant challenges, and the efficacy of such therapies varies widely.

Our research in Chapter 2 provides new insights into the treatment of haploinsufficiency-related diseases. Not only those caused by mutations resulting in unproductive splicing but also any normal transcripts carrying unproductive splicing events can be targeted for splice out using splice-switching ASOs, thereby increasing protein expression from the normal allele and restoring relevant biological functions, ultimately leading to alleviation and treatment of the disease. It has been observed that a significant number of genes identified in our unproductive splicing list are associated with haploinsufficiency, such as *PDK1* in the context of polycystic kidney disease. Considering this, it may be worthwhile to explore the possibility of restoring protein expression through splicing-switching of NMD exons as a promising novel therapeutic approach that warrants further validation in future research.

4.3 Analogs of deoxynucleoside triphosphates offer a widely applicable strategy for mapping DNA and RNA modifications

In Chapter 3, I developed a novel approach to map 6mA at a single-base resolution using mutation calling. This method involves inducing mutations at 6mA sites during DNA primer extension, which is achieved through the use of 2-thio-dTTP instead of dTTP. It is widely recognized that thiones, such as thioT, have limited capacity as hydrogen bond acceptors in Watson-Crick pairing. 2-thio-dTTP has been shown to pair normally with dA during regular DNA

amplification, albeit with a slower incorporation rate and reduced stability.[135] Additionally, it can be utilized to decrease the mispair between thymidine (T) and the minor tautomer of isoG, consequently increasing replication fidelity.[134] Meanwhile, extensive research has been carried out to investigate the properties of the 6mA base. This base is typically present in two conformations, which differ by a 180° rotation of the methyl group. In its free form, the 6mA base exhibits a preference ratio of 20:1 towards the Watson-Crick edge over the Hoogsteen edge of the base. However, only the second conformation is compatible with the Watson-Crick pairing of bases in double-stranded DNA. The presence of 6-methyl groups in adenine necessitates the less favored conformation of 6mA for base pairing, which thus acts as a destabilizing factor for double-stranded DNA.[173] The presence of both the 2-thio group on T and the N^6-methyl group on A can have a destabilizing effect on base pairing stability, which is further exacerbated when they coexist. This effect can be compounded under conditions that promote mutation, such as the use of a DNA polymerase with relatively low fidelity and the addition of magnesium ions. Consequently, mispairing at the 6mA sites can be introduced while the normal A sites remain relatively unaffected.

Analogs of deoxynucleoside triphosphates are molecules that are structurally similar to natural nucleotides but contain modifications in their chemical structure. These modifications can alter their function and can make them useful for various applications, including DNA sequencing, labeling, and modification studies, as well as the development of antiviral and anticancer drugs. They are also used in research to study DNA replication, repair, and recombination, as well as to investigate the mechanisms of genetic diseases. For instance, 5-bromo-2'-deoxyuridine (BrdU) is used in DNA synthesis studies to label newly synthesized DNA.[174] 2',3'-dideoxynucleoside triphosphates (ddNTPs) lack the 3'-OH group required for DNA synthesis and are used in DNA

sequencing techniques such as the Sanger method.[175] 2-aminopurine (2-AP) can replace adenine in DNA and RNA and is useful for studying base-pairing interactions.[176]

In an increasing number of DNA/RNA modification sequencing methods, the mutation-based approaches that introduce mutations/deletions/stops at the modified sites during DNA amplification and RNA reverse transcription processes is gradually demonstrating many advantages in sensitivity, resolution, and quantifiability.[177–179] Prior to this, this type of mutation-based sequencing methods mainly relied on adding bulky chemical groups to the modified bases and evolving the DNA polymerases or RNA reverse transcriptases through molecular evolution, with supplementary approaches including the use of biased dNTP ratios. In our study in Chapter 3, we have made a groundbreaking discovery that analogs of deoxynucleoside triphosphates can also introduce mutations at modified bases while maintaining fidelity at unmodified A/T/C/G bases. This is enabled by the differential stabilities of Watson–Crick base pairs between the analog of deoxynucleoside triphosphate and mofidied vs. unmodified bases. Notably, this strategy can theoretically be applied to any DNA/RNA modification. By synthesizing and screening analogs of deoxynucleoside triphosphates that meet the criteria, combined with specific or evolved DNA polymerases and RNA reverse transcriptases, convenient, rapid, single-base resolution, and quantitative whole-genome/transcriptome DNA/RNA modification sequencing can be achieved.

4.4 Mammalian 6mA: where are we, and what comes next?

In Chapter 3, I employed UHPLC-QqQ-MS/MS and the newly developed sequencing method, DR-6mA-seq, to discover and characterize a mammalian system with enriched 6mA levels. Specifically, I focused on the B104-1-1 mouse glioblastoma model cell line, which was generated through transfection of the embryonic mouse fibroblast cell line, NIH/3T3, with the digested DNA from the rat neuroblastoma cell line B-104. This cell line harbors the neu

transforming gene, which is homologous to the *Erbb2* oncogene and encodes for a 185000-dalton antigen known as p185.[81,147] Based on LC-MS/MS analysis, the level of 6mA in B104-1-1 gDNA was found to be approximately four times higher than that in its parental cell line, NIH/3T3, with a concentration of approximately 80 ppm. Interestingly, the DR-6mA-seq results identified newly emerged 6mA sites in B104-1-1 that were not present in NIH/3T3, and these sites exhibited unique distribution patterns. These findings, which agree with the previous report, suggest a strong association between increased 6mA levels and tumorigenesis in glioblastoma and highlight the potential of B104-1-1 as an excellent model for studying the processing and function of 6mA in mammalian cells.[81]

Since the detection of 6mA in human tissues using LC-MS/MS, there has been a sharp increase in publications on mammalian 6mA, utilizing various mapping methods, with a peak in 2018-2020 (Figure 4.1).[54,180] As an epigenetic mark, it is important to identify the writers, erasers, and readers of mammalian 6mA. However, determining the regulators of 6mA in mammals is a challenging task due to the low modification level and the absence of a homolog of the bacterial 6mA methyltransferase (Dam) in mammals.[180] Some tissues or cells, including B104-1-1 cell line I characterized in Chapter 3, display significantly higher abundance of 6mA, and it shows dynamic changes in certain biological processes and conditions, which may provide clues to future screening assays of 6mA regulators. Additionally, the identification of the consensus 6mA motif using advanced mapping tools in the future will provide some enlightenment regarding the 6mA processors and binding proteins.

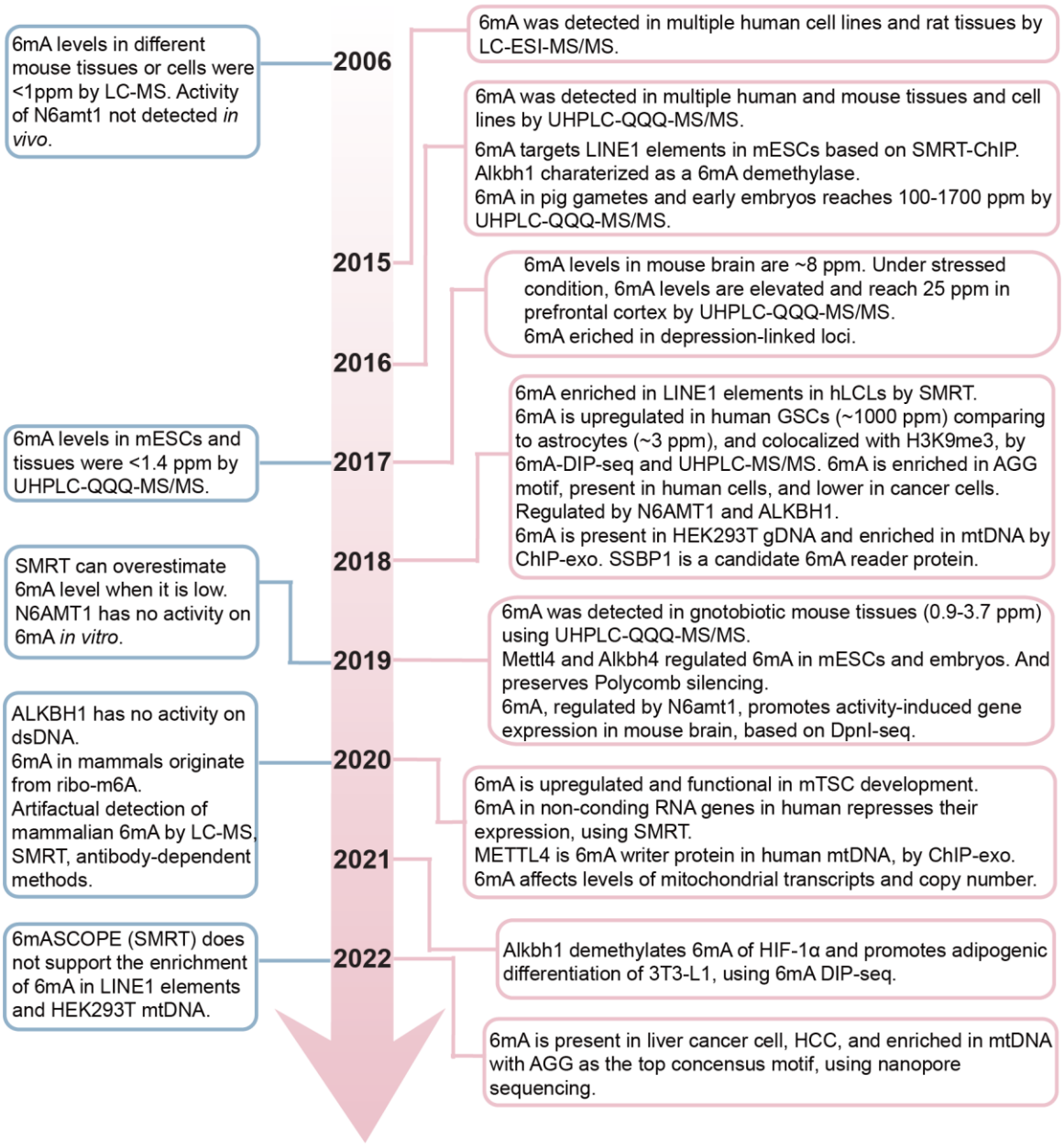


Figure 4.1 Timeline of key discoveries on mammalian 6mA.

mESC, mouse embryonic stem cell. LINE-1, long interspersed nuclear element-1. hLCL, human lymphoblastoid cell. GSC, glioblastoma stem cell. mTSC, mouse trophoblast stem cell. HCC, hepatocellular carcinoma.

However, functional studies of 6mA cannot solely rely on manipulating the putative effector proteins of 6mA. It is crucial to acknowledge that current studies investigating the function of 6mA often involve the manipulation of specific 6mA methyltransferase or demethylase

enzymes. This approach raises concerns as many of these enzymes have additional substrates apart from 6mA. For instance, METTL4 serves as an snRNA m6Am methyltransferase that regulates RNA splicing, ALKBH1 has been shown to preferentially demethylate m1A and m5C on tRNAs, ALKBH4 was initially identified as a lysine demethylase, and N6AMT1 was recently discovered to be a protein methyltransferase.[87,181–184] Therefore, it is imperative to eliminate any effects mediated by other substrates in future 6mA functional studies. Ideally, a targeted DNA demethylation system that employs dCas9 fused with the catalytic domain of 6mA methyltransferase and demethylase should be developed and utilized to manipulate specific 6mA loci.[54] This would enable the direct functional effects of specific methylation events to be demonstrated. Currently, it remains unclear whether the processing of 6mA is functional or simply incidental to the processing of other substrates, emphasizing the need for further research in this area.

There is concern among some researchers that the scarcity of mammalian 6mA may limit its ability to control transcription.[173] However, it is important to note that many covalent modifications with similarly low abundance, such as 5fC on DNA and m1A and m7G on RNA, have been shown to play critical roles in mammalian cells.[185–187] Additionally, the level of 6mA is much higher and more likely to have functions in early embryonic and tumor cells. It is plausible that 6mA on the mammalian nuclear genome plays a specific role in certain biological processes, such as tumorigenesis and early development. Therefore, 6mA has the potential to become a novel pathological marker and even a druggable target for cancer treatment.

List of references

1. Faraday, M. (1834). XXXVII. Experimental researches in electricity.—Seventh series. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science *5*, 252–264. 10.1080/14786443408648457.

2. Shabalina, S.A., and Spiridonov, N.A. (2004). The mammalian transcriptome and the function of non-coding DNA sequences. Genome Biol *5*, 105. 10.1186/gb-2004-5-4-105.

3. Han, J., Xiong, J., Wang, D., and Fu, X.-D. (2011). Pre-mRNA splicing: where and when in the nucleus. Trends Cell Biol *21*, 336–343. 10.1016/j.tcb.2011.03.003.

4. Gehring, N.H., and Roignant, J.-Y. (2021). Anything but Ordinary – Emerging Splicing Mechanisms in Eukaryotic Gene Regulation. Trends in Genetics *37*, 355–372. 10.1016/j.tig.2020.10.008.

5. Dvinge, H. (2018). Regulation of alternative mRNA splicing: old players and new perspectives. FEBS Lett *592*, 2987–3006. 10.1002/1873-3468.13119.

6. Chao, Y., Jiang, Y., Zhong, M., Wei, K., Hu, C., Qin, Y., Zuo, Y., Yang, L., Shen, Z., and Zou, C. (2021). Regulatory roles and mechanisms of alternative RNA splicing in adipogenesis and human metabolic health. Cell Biosci *11*, 66. 10.1186/s13578-021-00581-w.

7. Barash, Y., Calarco, J.A., Gao, W., Pan, Q., Wang, X., Shai, O., Blencowe, B.J., and Frey, B.J. (2010). Deciphering the splicing code. Nature *465*, 53–59. 10.1038/nature09000.

8. Fu, X.-D., and Ares, M. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. Nat Rev Genet *15*, 689–701. 10.1038/nrg3778.

9. Baralle, F.E., and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol *18*, 437–451. 10.1038/nrm.2017.27.

10. Brown, S.J., Stoilov, P., and Xing, Y. (2012). Chromatin and epigenetic regulation of pre-mRNA processing. Hum Mol Genet *21*, R90–R96. 10.1093/hmg/dds353.

11. Buratti, E., and Baralle, F.E. (2004). Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process. Mol Cell Biol *24*, 10505–10514. 10.1128/MCB.24.24.10505-10514.2004.

12. Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., and Stamm, S. (2013). Function of alternative splicing. Gene *514*, 1–30. 10.1016/j.gene.2012.07.083.

13. Xing, Y., and Lee, C. (2006). Alternative splicing and RNA selection pressure — evolutionary consequences for eukaryotic genomes. Nat Rev Genet *7*, 499–509. 10.1038/nrg1896.

14. van den Berg, Y.W., van den Hengel, L.G., Myers, H.R., Ayachi, O., Jordanova, E., Ruf, W., Spek, C.A., Reitsma, P.H., Bogdanov, V.Y., and Versteeg, H.H. (2009). Alternatively spliced tissue factor induces angiogenesis through integrin ligation. Proceedings of the National Academy of Sciences *106*, 19497–19502. 10.1073/pnas.0905325106.

15. Araki, Y., Hong, I., Gamache, T.R., Ju, S., Collado-Torres, L., Shin, J.H., and Huganir, R.L. (2020). SynGAP isoforms differentially regulate synaptic plasticity and dendritic development. Elife *9*. 10.7554/eLife.56273.

16. Chavez-Gutierrez, L., Bourdais, J., Aranda, G., Vargas, M.A., Matta-Camacho, E., Ducancel, F., Segovia, L., Joseph-Bravo, P., and Charli, J.-L. (2005). A truncated isoform of pyroglutamyl aminopeptidase II produced by exon extension has dominant-negative activity. J Neurochem *92*, 807–817. 10.1111/j.1471-4159.2004.02916.x.

17. Lee, A.W., Champagne, N., Wang, X., Su, X.-D., Goodyer, C., and LeBlanc, A.C. (2010). Alternatively Spliced Caspase-6B Isoform Inhibits the Activation of Caspase-6A. Journal of Biological Chemistry *285*, 31974–31984. 10.1074/jbc.M110.152744.

18. Lin, Y., Stevens, C., Harrison, B., Pathuri, S., Amin, E., and Hupp, T.R. (2009). The alternative splice variant of DAPK-1, s-DAPK-1, induces proteasome-independent DAPK-1 destabilization. Mol Cell Biochem *328*, 101–107. 10.1007/s11010-009-0079-4.

19. Mironov, A., Petrova, M., Margasyuk, S., Vlasenok, M., Mironov, A.A., Skvortsov, D., and Pervouchine, D.D. (2023). Tissue-specific regulation of gene expression via unproductive splicing. Nucleic Acids Res *51*, 3055–3066. 10.1093/nar/gkad161.

20. Brogna, S., and Wen, J. (2009). Nonsense-mediated mRNA decay (NMD) mechanisms. Nat Struct Mol Biol *16*, 107–113. 10.1038/nsmb.1550.

21. Karousis, E.D., and Mühlemann, O. (2019). Nonsense-Mediated mRNA Decay Begins Where Translation Ends. Cold Spring Harb Perspect Biol *11*, a032862. 10.1101/cshperspect.a032862.

22. Maquat, L.E. (2004). Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. Nat Rev Mol Cell Biol *5*, 89–99. 10.1038/nrm1310.

23. Wollerton, M.C., Gooding, C., Wagner, E.J., Garcia-Blanco, M.A., and Smith, C.W.J. (2004). Autoregulation of Polypyrimidine Tract Binding Protein by Alternative Splicing Leading to Nonsense-Mediated Decay. Mol Cell *13*, 91–100. 10.1016/S1097-2765(03)00502-1.

24. Jiang, W., and Chen, L. (2021). Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing. Comput Struct Biotechnol J *19*, 183–195. 10.1016/j.csbj.2020.12.009.

25. Dillman, A.A., and Cookson, M.R. (2014). Transcriptomic Changes in Brain Development. In, pp. 233–250. 10.1016/B978-0-12-801105-8.00009-6.

26. Li, M., Santpere, G., Imamura Kawasawa, Y., Evgrafov, O. V., Gulden, F.O., Pochareddy, S., Sunkin, S.M., Li, Z., Shin, Y., Zhu, Y., et al. (2018). Integrative functional genomic analysis of human brain development and neuropsychiatric risks. Science (1979) *362*. 10.1126/science.aat7615.

27. Jiang, X., and Nardelli, J. (2016). Cellular and molecular introduction to brain development. Neurobiol Dis *92*, 3–17. 10.1016/j.nbd.2015.07.007.

28. Shimojo, H., Ohtsuka, T., and Kageyama, R. (2008). Oscillations in Notch Signaling Regulate Maintenance of Neural Progenitors. Neuron *58*, 52–64. 10.1016/j.neuron.2008.02.014.

29. Imayoshi, I., Isomura, A., Harima, Y., Kawaguchi, K., Kori, H., Miyachi, H., Fujiwara, T., Ishidate, F., and Kageyama, R. (2013). Oscillatory Control of Factors Determining Multipotency and Fate in Mouse Neural Progenitors. Science (1979) *342*, 1203–1208. 10.1126/science.1242366.

30. Flames, N., Long, J.E., Garratt, A.N., Fischer, T.M., Gassmann, M., Birchmeier, C., Lai, C., Rubenstein, J.L.R., and Marín, O. (2004). Short- and Long-Range Attraction of Cortical GABAergic Interneurons by Neuregulin-1. Neuron *44*, 251–261. 10.1016/j.neuron.2004.09.028.

31. Faux, C., Rakic, S., Andrews, W., Yanagawa, Y., Obata, K., and Parnavelas, J.G. (2009). Differential gene expression in migrating cortical interneurons during mouse forebrain development. J Comp Neurol, NA-NA. 10.1002/cne.22271.

32. Qi, C., Luo, L.-D., Feng, I., and Ma, S. (2022). Molecular mechanisms of synaptogenesis. Front Synaptic Neurosci *14*. 10.3389/fnsyn.2022.939793.

33. Citri, A., and Malenka, R.C. (2008). Synaptic Plasticity: Multiple Forms, Functions, and Mechanisms. Neuropsychopharmacology *33*, 18–41. 10.1038/sj.npp.1301559.

34. Sood, A.J., Viner, C., and Hoffman, M.M. (2019). DNAmod: the DNA modification database. J Cheminform *11*, 30. 10.1186/s13321-019-0349-4.

35. Luo, C., Hajkova, P., and Ecker, J.R. (2018). Dynamic DNA methylation: In the right place at the right time. Science (1979) *361*, 1336–1340. 10.1126/science.aat6806.

36. Yuan, B.-F. (2020). Assessment of DNA Epigenetic Modifications. Chem Res Toxicol *33*, 695–708. 10.1021/acs.chemrestox.9b00372.

37. Clark, T.A., Spittle, K.E., Turner, S.W., and Korlach, J. (2011). Direct Detection and Sequencing of Damaged DNA Bases. Genome Integr *2*, 10. 10.1186/2041-9414-2-10.

38. Trewick, S.C., Henshaw, T.F., Hausinger, R.P., Lindahl, T., and Sedgwick, B. (2002). Oxidative demethylation by Escherichia coli AlkB directly reverts DNA base damage. Nature *419*, 174–178. 10.1038/nature00908.

39. Martin, L.J. (2008). DNA Damage and Repair. J Neuropathol Exp Neurol *67*, 377–387. 10.1097/NEN.0b013e31816ff780.

40. Carusillo, A., and Mussolino, C. (2020). DNA Damage: From Threat to Treatment. Cells *9*, 1665. 10.3390/cells9071665.

41. Kim, M., and Costello, J. (2017). DNA methylation: an epigenetic mark of cellular memory. Exp Mol Med *49*, e322–e322. 10.1038/emm.2017.10.

42. Raiber, E.-A., Hardisty, R., van Delft, P., and Balasubramanian, S. (2017). Mapping and elucidating the function of modified bases in DNA. Nat Rev Chem *1*, 0069. 10.1038/s41570-017-0069.

43. Yu, M., Ji, L., Neumann, D.A., Chung, D., Groom, J., Westpheling, J., He, C., and Schmitz, R.J. (2015). Base-resolution detection of N4-methylcytosine in genomic DNA using 4mC-Tet-assisted-bisulfite- sequencing. Nucleic Acids Res, gkv738. 10.1093/nar/gkv738.

44. Wu, X., and Zhang, Y. (2017). TET-mediated active DNA demethylation: mechanism, function and beyond. Nat Rev Genet *18*, 517–534. 10.1038/nrg.2017.33.

45. Shi, D.-Q., Ali, I., Tang, J., and Yang, W.-C. (2017). New Insights into 5hmC DNA Modification: Generation, Distribution and Function. Front Genet *8*. 10.3389/fgene.2017.00100.

46. Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., and Zhang, Y. (2011). Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. Science (1979) *333*, 1300–1303. 10.1126/science.1210597.

47. Branco, M.R., Ficz, G., and Reik, W. (2012). Uncovering the role of 5-hydroxymethylcytosine in the epigenome. Nat Rev Genet *13*, 7–13. 10.1038/nrg3080.

48. Pfaffeneder, T., Spada, F., Wagner, M., Brandmayr, C., Laube, S.K., Eisen, D., Truss, M., Steinbacher, J., Hackner, B., Kotljarova, O., et al. (2014). Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. Nat Chem Biol *10*, 574–581. 10.1038/nchembio.1532.

49. Hofer, A., Liu, Z.J., and Balasubramanian, S. (2019). Detection, Structure and Function of Modified DNA Bases. J Am Chem Soc *141*, 6420–6429. 10.1021/jacs.9b01915.

50. van Luenen, H.G.A.M., Farris, C., Jan, S., Genest, P.-A., Tripathi, P., Velds, A., Kerkhoven, R.M., Nieuwland, M., Haydock, A., Ramasamy, G., et al. (2012). Glucosylated Hydroxymethyluracil, DNA Base J, Prevents Transcriptional Readthrough in Leishmania. Cell *150*, 909–921. 10.1016/j.cell.2012.07.030.

51. Boulias, K., and Greer, E.L. (2022). Means, mechanisms and consequences of adenine methylation in DNA. Nat Rev Genet *23*, 411–428. 10.1038/s41576-022-00456-x.

52. Rodriguez, F., Yushenova, I.A., DiCorpo, D., and Arkhipova, I.R. (2022). Bacterial N4-methylcytosine as an epigenetic mark in eukaryotic DNA. Nat Commun *13*, 1072. 10.1038/s41467-022-28471-w.

53. DUNN, D.B., and SMITH, J.D. (1955). Occurrence of a New Base in the Deoxyribonucleic Acid of a Strain of Bacterium Coli. Nature *175*, 336–337. 10.1038/175336a0.

54. Feng, X., and He, C. (2023). Mammalian DNA N6-methyladenosine: Challenges and new insights. Mol Cell *83*, 343–351. 10.1016/j.molcel.2023.01.005.

55. Shen, C., Wang, K., Deng, X., and Chen, J. (2022). DNA N6-methyldeoxyadenosine in mammals and human disease. Trends in Genetics *38*, 454–467. 10.1016/j.tig.2021.12.003.

56. Liang, Z., Shen, L., Cui, X., Bao, S., Geng, Y., Yu, G., Liang, F., Xie, S., Lu, T., Gu, X., et al. (2018). DNA N-Adenine Methylation in Arabidopsis thaliana. Dev Cell *45*, 406-416.e3. 10.1016/j.devcel.2018.03.012.

57. Beaulaurier, J., Schadt, E.E., and Fang, G. (2019). Deciphering bacterial epigenomes using modern sequencing technologies. Nat Rev Genet *20*, 157–172. 10.1038/s41576-018-0081-3.

58. O'Brown, Z.K., and Greer, E.L. (2016). N6-Methyladenine: A Conserved and Dynamic DNA Mark. In, pp. 213–246. 10.1007/978-3-319-43624-1_10.

59. Tronche, F., Rollier, A., Bach, I., Weiss, M.C., and Yaniv, M. (1989). The Rat Albumin Promoter: Cooperation with Upstream Elements Is Required when Binding of APF/HNF1 to the Proximal Element Is Partially Impaired by Mutation or Bacterial Methylation. Mol Cell Biol *9*, 4759–4766. 10.1128/mcb.9.11.4759-4766.1989.

60. Razin, A. (1984). DNA Methylation Patterns: Formation and Biological Functions. In, pp. 127–146. 10.1007/978-1-4613-8519-6_7.

61. Messer, W., and Noyer-Weidner, M. (1988). Timing and targeting: The biological functions of Dam methylation in E. coli. Cell *54*, 735–737. 10.1016/S0092-8674(88)90911-7.

62. LU, M. (1994). SeqA: A negative modulator of replication initiation in E. coli. Cell *77*, 413–426. 10.1016/0092-8674(94)90156-2.

63. Fu, Y., Luo, G.-Z., Chen, K., Deng, X., Yu, M., Han, D., Hao, Z., Liu, J., Lu, X., Doré, L.C., et al. (2015). N6-Methyldeoxyadenosine Marks Active Transcription Start Sites in Chlamydomonas. Cell *161*, 879–892. 10.1016/j.cell.2015.04.010.

64. Luo, G.-Z., Hao, Z., Luo, L., Shen, M., Sparvoli, D., Zheng, Y., Zhang, Z., Weng, X., Chen, K., Cui, Q., et al. (2018). N6-methyldeoxyadenosine directs nucleosome positioning in Tetrahymena DNA. Genome Biol *19*, 200. 10.1186/s13059-018-1573-3.

65. Zhang, Q., Liang, Z., Cui, X., Ji, C., Li, Y., Zhang, P., Liu, J., Riaz, A., Yao, P., Liu, M., et al. (2018). N6-Methyladenine DNA Methylation in Japonica and Indica Rice Genomes and Its Association with Gene Expression, Plant Development, and Stress Responses. Mol Plant *11*, 1492–1508. 10.1016/j.molp.2018.11.005.

66. O'Brown, Z.K., Boulias, K., Wang, J., Wang, S.Y., O'Brown, N.M., Hao, Z., Shibuya, H., Fady, P.-E., Shi, Y., He, C., et al. (2019). Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. BMC Genomics *20*, 445. 10.1186/s12864-019-5754-6.

67. Kong, Y., Mead, E.A., and Fang, G. (2023). Navigating the pitfalls of mapping DNA and RNA modifications. Nat Rev Genet *24*, 363–381. 10.1038/s41576-022-00559-5.

68. Koziol, M.J., Bradshaw, C.R., Allen, G.E., Costa, A.S.H., Frezza, C., and Gurdon, J.B. (2016). Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. Nat Struct Mol Biol *23*, 24–30. 10.1038/nsmb.3145.

69. Hao, Z., Wu, T., Cui, X., Zhu, P., Tan, C., Dou, X., Hsu, K.-W., Lin, Y.-T., Peng, P.-H., Zhang, L.-S., et al. (2020). N6-Deoxyadenosine Methylation in Mammalian Mitochondrial DNA. Mol Cell *78*, 382-395.e8. 10.1016/j.molcel.2020.02.018.

70. Koh, C.W.Q., Goh, Y.T., Toh, J.D.W., Neo, S.P., Ng, S.B., Gunaratne, J., Gao, Y.-G., Quake, S.R., Burkholder, W.F., and Goh, W.S.S. (2018). Single-nucleotide-resolution sequencing of human N6-methyldeoxyadenosine reveals strand-asymmetric clusters associated with SSBP1 on the mitochondrial genome. Nucleic Acids Res *46*, 11659–11670. 10.1093/nar/gky1104.

71. Wu, T.P., Wang, T., Seetin, M.G., Lai, Y., Zhu, S., Lin, K., Liu, Y., Byrum, S.D., Mackintosh, S.G., Zhong, M., et al. (2016). DNA methylation on N6-adenine in mammalian embryonic stem cells. Nature *532*, 329–333. 10.1038/nature17640.

72. Chen, L.-Q., Zhang, Z., Chen, H.-X., Xi, J.-F., Liu, X.-H., Ma, D.-Z., Zhong, Y.-H., Ng, W.H., Chen, T., Mak, D.W., et al. (2022). High-precision mapping reveals rare N6-deoxyadenosine methylation in the mammalian genome. Cell Discov *8*, 138. 10.1038/s41421-022-00484-1.

73. Douvlataniotis, K., Bensberg, M., Lentini, A., Gylemo, B., and Nestor, C.E. (2020). No evidence for DNA N6-methyladenine in mammals. Sci Adv *6*. 10.1126/sciadv.aay3335.

74. Luo, G.-Z., Wang, F., Weng, X., Chen, K., Hao, Z., Yu, M., Deng, X., Liu, J., and He, C. (2016). Characterization of eukaryotic DNA N6-methyladenine by a highly sensitive restriction enzyme-assisted sequencing. Nat Commun *7*, 11301. 10.1038/ncomms11301.

75. Li, X., Guo, S., Cui, Y., Zhang, Z., Luo, X., Angelova, M.T., Landweber, L.F., Wang, Y., and Wu, T.P. (2022). NT-seq: a chemical-based sequencing method for genomic methylome profiling. Genome Biol *23*, 122. 10.1186/s13059-022-02689-9.

76. Kong, Y., Cao, L., Deikus, G., Fan, Y., Mead, E.A., Lai, W., Zhang, Y., Yong, R., Sebra, R., Wang, H., et al. (2022). Critical assessment of DNA adenine methylation in eukaryotes using quantitative deconvolution. Science (1979) *375*, 515–522. 10.1126/science.abe7489.

77. Liu, J., Zhu, Y., Luo, G.-Z., Wang, X., Yue, Y., Wang, X., Zong, X., Chen, K., Yin, H., Fu, Y., et al. (2016). Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. Nat Commun *7*, 13052. 10.1038/ncomms13052.

78. Liang, D., Wang, H., Song, W., Xiong, X., Zhang, X., Hu, Z., Guo, H., Yang, Z., Zhai, S., Zhang, L.-H., et al. (2016). The decreased N6-methyladenine DNA modification in cancer cells. Biochem Biophys Res Commun *480*, 120–125. 10.1016/j.bbrc.2016.09.136.

79. Ratel, D., Ravanat, J.-L., Charles, M.-P., Platet, N., Breuillaud, L., Lunardi, J., Berger, F., and Wion, D. (2006). Undetectable levels of N6-methyl adenine in mouse DNA: Cloning and analysis of PRED28, a gene coding for a putative mammalian DNA adenine methyltransferase. FEBS Lett *580*, 3179–3184. 10.1016/j.febslet.2006.04.074.

80. Yao, B., Cheng, Y., Wang, Z., Li, Y., Chen, L., Huang, L., Zhang, W., Chen, D., Wu, H., Tang, B., et al. (2017). DNA N6-methyladenine is dynamically regulated in the mouse brain following environmental stress. Nat Commun *8*, 1122. 10.1038/s41467-017-01195-y.

81. Xie, Q., Wu, T.P., Gimple, R.C., Li, Z., Prager, B.C., Wu, Q., Yu, Y., Wang, P., Wang, Y., Gorkin, D.U., et al. (2018). N6-methyladenine DNA Modification in Glioblastoma. Cell *175*, 1228-1243.e20. 10.1016/j.cell.2018.10.006.

82. Li, Z., Zhao, S., Nelakanti, R. V., Lin, K., Wu, T.P., Alderman, M.H., Guo, C., Wang, P., Zhang, M., Min, W., et al. (2020). N6-methyladenine in DNA antagonizes SATB1 in early development. Nature *583*, 625–630. 10.1038/s41586-020-2500-9.

83. Musheev, M.U., Baumgärtner, A., Krebs, L., and Niehrs, C. (2020). The origin of genomic N6-methyl-deoxyadenosine in mammalian cells. Nat Chem Biol *16*, 630–634. 10.1038/s41589-020-0504-2.

84. Li, X., Zhao, Q., Wei, W., Lin, Q., Magnan, C., Emami, M.R., Wearick-Silva, L.E., Viola, T.W., Marshall, P.R., Yin, J., et al. (2019). The DNA modification N6-methyl-2'-deoxyadenosine (m6dA) drives activity-induced gene expression and is required for fear extinction. Nat Neurosci *22*, 534–544. 10.1038/s41593-019-0339-x.

85. Xiao, C.-L., Zhu, S., He, M., Chen, D., Zhang, Q., Chen, Y., Yu, G., Liu, J., Xie, S.-Q., Luo, F., et al. (2018). N6-Methyladenine DNA Modification in the Human Genome. Mol Cell *71*, 306-318.e7. 10.1016/j.molcel.2018.06.015.

86. Kweon, S.-M., Chen, Y., Moon, E., Kvederaviciutė, K., Klimasauskas, S., and Feldman, D.E. (2019). An Adversarial DNA N6-Methyladenine-Sensor Network Preserves Polycomb Silencing. Mol Cell *74*, 1138-1147.e6. 10.1016/j.molcel.2019.03.018.

87. Liu, F., Clark, W., Luo, G., Wang, X., Fu, Y., Wei, J., Wang, X., Hao, Z., Dai, Q., Zheng, G., et al. (2016). ALKBH1-Mediated tRNA Demethylation Regulates Translation. Cell *167*, 816-828.e16. 10.1016/j.cell.2016.09.038.

88. Su, C.-H., D, D., and Tarn, W.-Y. (2018). Alternative Splicing in Neurogenesis and Brain Development. Front Mol Biosci *5*. 10.3389/fmolb.2018.00012.

89. Vuong, C.K., Black, D.L., and Zheng, S. (2016). The neurogenetics of alternative splicing. Nat Rev Neurosci *17*, 265–281. 10.1038/nrn.2016.27.

90. Boutz, P.L., Stoilov, P., Li, Q., Lin, C.-H., Chawla, G., Ostrow, K., Shiue, L., Ares, M., and Black, D.L. (2007). A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. Genes Dev *21*, 1636–1652. 10.1101/gad.1558107.

91. Spellman, R., Llorian, M., and Smith, C.W.J. (2007). Crossregulation and Functional Redundancy between the Splicing Regulator PTB and Its Paralogs nPTB and ROD1. Mol Cell *27*, 420–434. 10.1016/j.molcel.2007.06.016.

92. Vuong, J.K., Lin, C.-H., Zhang, M., Chen, L., Black, D.L., and Zheng, S. (2016). PTBP1 and PTBP2 Serve Both Specific and Redundant Functions in Neuronal Pre-mRNA Splicing. Cell Rep *17*, 2766–2775. 10.1016/j.celrep.2016.11.034.

93. Boutz, P.L., Stoilov, P., Li, Q., Lin, C.-H., Chawla, G., Ostrow, K., Shiue, L., Ares, M., and Black, D.L. (2007). A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. Genes Dev *21*, 1636–1652. 10.1101/gad.1558107.

94. Zheng, S., Gray, E.E., Chawla, G., Porse, B.T., O'Dell, T.J., and Black, D.L. (2012). PSD-95 is post-transcriptionally repressed during early neural development by PTBP1 and PTBP2. Nat Neurosci *15*, 381–388. 10.1038/nn.3026.

95. Yoo, K.-S., Lee, K., Oh, J.-Y., Lee, H., Park, H., Park, Y.S., and Kim, H.K. (2019). Postsynaptic density protein 95 (PSD-95) is transported by KIF5 to dendritic regions. Mol Brain *12*, 97. 10.1186/s13041-019-0520-x.

96. Shibasaki, T., Tokunaga, A., Sakamoto, R., Sagara, H., Noguchi, S., Sasaoka, T., and Yoshida, N. (2013). PTB Deficiency Causes the Loss of Adherens Junctions in the Dorsal Telencephalon and Leads to Lethal Hydrocephalus. Cerebral Cortex *23*, 1824–1835. 10.1093/cercor/bhs161.

97. Quesnel-Vallières, M., Dargaei, Z., Irimia, M., Gonatopoulos-Pournatzis, T., Ip, J.Y., Wu, M., Sterne-Weiler, T., Nakagawa, S., Woodin, M.A., Blencowe, B.J., et al. (2016). Misregulation of an Activity-Dependent Splicing Network as a Common Mechanism Underlying Autism Spectrum Disorders. Mol Cell *64*, 1023–1034. 10.1016/j.molcel.2016.11.033.

98. Zhang, X., Chen, M.H., Wu, X., Kodani, A., Fan, J., Doan, R., Ozawa, M., Ma, J., Yoshida, N., Reiter, J.F., et al. (2016). Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex. Cell *166*, 1147-1162.e15. 10.1016/j.cell.2016.07.025.

99. Moutton, S., Bruel, A.-L., Assoum, M., Chevarin, M., Sarrazin, E., Goizet, C., Guerrot, A.-M., Charollais, A., Charles, P., Heron, D., et al. (2018). Truncating variants of the DLG4 gene are responsible for intellectual disability with marfanoid features. Clin Genet *93*, 1172–1178. 10.1111/cge.13243.

100. Lelieveld, S.H., Reijnders, M.R.F., Pfundt, R., Yntema, H.G., Kamsteeg, E.-J., de Vries, P., de Vries, B.B.A., Willemsen, M.H., Kleefstra, T., Löhner, K., et al. (2016). Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. Nat Neurosci *19*, 1194–1196. 10.1038/nn.4352.

101. Mehmood, A., Laiho, A., Venäläinen, M.S., McGlinchey, A.J., Wang, N., and Elo, L.L. (2020). Systematic evaluation of differential splicing tools for RNA-seq studies. Brief Bioinform *21*, 2052–2065. 10.1093/bib/bbz126.

102. Popp, M.W., and Maquat, L.E. (2016). Leveraging Rules of Nonsense-Mediated mRNA Decay for Genome Engineering and Personalized Medicine. Cell *165*, 1319–1322. 10.1016/j.cell.2016.05.053.

103. Tremblay, R.G., Sikorska, M., Sandhu, J.K., Lanthier, P., Ribecco-Lutkiewicz, M., and Bani-Yaghoub, M. (2010). Differentiation of mouse Neuro 2A cells into dopamine neurons. J Neurosci Methods *186*, 60–67. 10.1016/j.jneumeth.2009.11.004.

104. Jeyabalan, N., and Clement, J.P. (2016). SYNGAP1: Mind the Gap. Front Cell Neurosci *10*. 10.3389/fncel.2016.00032.

105. Kim, Y.K., and Maquat, L.E. (2019). UPFront and center in RNA decay: UPF1 in nonsense-mediated mRNA decay and beyond. RNA *25*, 407–422. 10.1261/rna.070136.118.

106. Kozol, R.A., Cukier, H.N., Zou, B., Mayo, V., De Rubeis, S., Cai, G., Griswold, A.J., Whitehead, P.L., Haines, J.L., Gilbert, J.R., et al. (2015). Two knockdown models of the autism genes SYNGAP1 and SHANK3 in zebrafish produce similar behavioral phenotypes associated with embryonic disruptions of brain morphogenesis. Hum Mol Genet *24*, 4006–4023. 10.1093/hmg/ddv138.

107. Ermakova, E.O., Nurtdinov, R.N., and Gelfand, M.S. (2006). Fast rate of evolution in alternatively spliced coding regions of mammalian genes. BMC Genomics *7*, 84. 10.1186/1471-2164-7-84.

108. Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N.M., and Ule, J. (2013). Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements. Cell *152*, 453–466. 10.1016/j.cell.2012.12.023.

109. Warf, M.B., Diegel, J. V., von Hippel, P.H., and Berglund, J.A. (2009). The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. Proceedings of the National Academy of Sciences *106*, 9203–9208. 10.1073/pnas.0900342106.

110. Prchalova, D., Havlovicova, M., Sterbova, K., Stranecky, V., Hancarova, M., and Sedlacek, Z. (2017). Analysis of 31-year-old patient with SYNGAP1 gene defect points to importance of variants in broader splice regions and reveals developmental trajectory of SYNGAP1-associated phenotype: case report. BMC Med Genet *18*, 62. 10.1186/s12881-017-0425-4.

111. Vlaskamp, D.R.M., Shaw, B.J., Burgess, R., Mei, D., Montomoli, M., Xie, H., Myers, C.T., Bennett, M.F., XiangWei, W., Williams, D., et al. (2019). SYNGAP1 encephalopathy: A distinctive generalized developmental and epileptic encephalopathy. Neurology *92*, e96–e107. 10.1212/WNL.0000000000006729.

112. Finkel, R.S., Mercuri, E., Darras, B.T., Connolly, A.M., Kuntz, N.L., Kirschner, J., Chiriboga, C.A., Saito, K., Servais, L., Tizzano, E., et al. (2017). Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy. New England Journal of Medicine *377*, 1723–1732. 10.1056/NEJMoa1702752.

113. Kim, J., Hu, C., Moufawad El Achkar, C., Black, L.E., Douville, J., Larson, A., Pendergast, M.K., Goldkind, S.F., Lee, E.A., Kuniholm, A., et al. (2019). Patient-Customized Oligonucleotide Therapy for a Rare Genetic Disease. New England Journal of Medicine *381*, 1644–1652. 10.1056/NEJMoa1813279.

114. Han, Z., Chen, C., Christiansen, A., Ji, S., Lin, Q., Anumonwo, C., Liu, C., Leiser, S.C., Meena, Aznarez, I., et al. (2020). Antisense oligonucleotides increase Scn1a expression and reduce seizures and SUDEP incidence in a mouse model of Dravet syndrome. Sci Transl Med *12*. 10.1126/scitranslmed.aaz6100.

115. Kole, R., Krainer, A.R., and Altman, S. (2012). RNA therapeutics: beyond RNA interference and antisense oligonucleotides. Nat Rev Drug Discov *11*, 125–140. 10.1038/nrd3625.

116. Lim, K.H., Han, Z., Jeon, H.Y., Kach, J., Jing, E., Weyn-Vanhentenryck, S., Downs, M., Corrionero, A., Oh, R., Scharner, J., et al. (2020). Antisense oligonucleotide modulation of non-productive alternative splicing upregulates gene expression. Nat Commun *11*, 3501. 10.1038/s41467-020-17093-9.

117. Yoon, S.-J., Elahi, L.S., Pașca, A.M., Marton, R.M., Gordon, A., Revah, O., Miura, Y., Walczak, E.M., Holdgate, G.M., Fan, H.C., et al. (2019). Reliability of human cortical organoid generation. Nat Methods *16*, 75–78. 10.1038/s41592-018-0255-0.

118. McMahon, A.C., Barnett, M.W., O'Leary, T.S., Stoney, P.N., Collins, M.O., Papadia, S., Choudhary, J.S., Komiyama, N.H., Grant, S.G.N., Hardingham, G.E., et al. (2012). SynGAP isoforms exert opposing effects on synaptic strength. Nat Commun *3*, 900. 10.1038/ncomms1900.

119. Zeng, M., Shang, Y., Araki, Y., Guo, T., Huganir, R.L., and Zhang, M. (2016). Phase Transition in Postsynaptic Densities Underlies Formation of Synaptic Complexes and Synaptic Plasticity. Cell *166*, 1163-1175.e12. 10.1016/j.cell.2016.07.008.

120. Vlaskamp, D.R.M., Shaw, B.J., Burgess, R., Mei, D., Montomoli, M., Xie, H., Myers, C.T., Bennett, M.F., XiangWei, W., Williams, D., et al. (2019). SYNGAP1 encephalopathy. Neurology *92*, e96–e107. 10.1212/WNL.0000000000006729.

121. Yan, Q., Weyn-Vanhentenryck, S.M., Wu, J., Sloan, S.A., Zhang, Y., Chen, K., Wu, J.Q., Barres, B.A., and Zhang, C. (2015). Systematic discovery of regulated and conserved alternative exons in the mammalian brain reveals NMD modulating chromatin regulators. Proceedings of the National Academy of Sciences *112*, 3445–3450. 10.1073/pnas.1502849112.

122. Carvill, G.L., Engel, K.L., Ramamurthy, A., Cochran, J.N., Roovers, J., Stamberger, H., Lim, N., Schneider, A.L., Hollingsworth, G., Holder, D.H., et al. (2018). Aberrant Inclusion of a Poison

Exon Causes Dravet Syndrome and Related SCN1A-Associated Genetic Epilepsies. The American Journal of Human Genetics *103*, 1022–1029. 10.1016/j.ajhg.2018.10.023.

123. Raj, B., and Blencowe, B.J. (2015). Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. Neuron *87*, 14–27. 10.1016/j.neuron.2015.05.004.

124. Liyanage, V., Jarmasz, J., Murugeshan, N., Del Bigio, M., Rastegar, M., and Davie, J. (2014). DNA Modifications: Function and Applications in Normal and Disease States. Biology (Basel) *3*, 670–723. 10.3390/biology3040670.

125. Schiffers, S., Ebert, C., Rahimoff, R., Kosmatchev, O., Steinbacher, J., Bohne, A.-V., Spada, F., Michalakis, S., Nickelsen, J., Müller, M., et al. (2017). Quantitative LC-MS Provides No Evidence for m6dA or m4dC in the Genome of Mouse Embryonic Stem Cells and Tissues. Angewandte Chemie International Edition *56*, 11268–11271. 10.1002/anie.201700424.

126. Lentini, A., Lagerwall, C., Vikingsson, S., Mjoseng, H.K., Douvlataniotis, K., Vogt, H., Green, H., Meehan, R.R., Benson, M., and Nestor, C.E. (2018). A reassessment of DNA-immunoprecipitation-based genomic profiling. Nat Methods *15*, 499–504. 10.1038/s41592-018-0038-7.

127. Bird, A.P., and Southern, E.M. (1978). Use of restriction enzymes to study eukaryotic DNA methylation. J Mol Biol *118*, 27–47. 10.1016/0022-2836(78)90242-5.

128. Mahdavi-Amiri, Y., Chung Kim Chung, K., and Hili, R. (2021). Single-nucleotide resolution of N6-adenine methylation sites in DNA and RNA by nitrite sequencing. Chem Sci *12*, 606–612. 10.1039/D0SC03509B.

129. Schadt, E.E., Banerjee, O., Fang, G., Feng, Z., Wong, W.H., Zhang, X., Kislyuk, A., Clark, T.A., Luong, K., Keren-Paz, A., et al. (2013). Modeling kinetic rate variation in third generation DNA sequencing data to detect putative modifications to DNA bases. Genome Res *23*, 129–141. 10.1101/gr.136739.111.

130. Zhu, S., Beaulaurier, J., Deikus, G., Wu, T.P., Strahl, M., Hao, Z., Luo, G., Gregory, J.A., Chess, A., He, C., et al. (2018). Mapping and characterizing N6-methyladenine in eukaryotic genomes using single-molecule real-time sequencing. Genome Res *28*, 1067–1078. 10.1101/gr.231068.117.

131. Engel, J.D., and Von Hippel, P.H. (1974). Effects of methylation on the stability of nucleic acid conformations. Monomer level. Biochemistry *13*, 4143–4158. 10.1021/bi00717a013.

132. Engel, J.D., and von Hippel, P.H. (1978). Effects of methylation on the stability of nucleic acid conformations. Studies at the polymer level. Journal of Biological Chemistry *253*, 927–934. 10.1016/S0021-9258(17)38193-0.

133. Harcourt, E.M., Ehrenschwender, T., Batista, P.J., Chang, H.Y., and Kool, E.T. (2013). Identification of a Selective Polymerase Enables Detection of N6-Methyladenosine in RNA. J Am Chem Soc *135*, 19079–19082. 10.1021/ja4105792.

134. Sismour, A.M. (2005). The use of thymidine analogs to improve the replication of an extra DNA base pair: a synthetic biological system. Nucleic Acids Res *33*, 5640–5646. 10.1093/nar/gki873.

135. Pugliese, K.M., Gul, O.T., Choi, Y., Olsen, T.J., Sims, P.C., Collins, P.G., and Weiss, G.A. (2015). Processive Incorporation of Deoxynucleoside Triphosphate Analogs by Single-Molecule DNA Polymerase I (Klenow Fragment) Nanocircuits. J Am Chem Soc *137*, 9587–9594. 10.1021/jacs.5b02074.

136. Koerber, J.T., Maheshri, N., Kaspar, B.K., and Schaffer, D. V (2006). Construction of diverse adeno-associated viral libraries for directed evolution of enhanced gene delivery vehicles. Nat Protoc *1*, 701–706. 10.1038/nprot.2006.93.

137. Fromant, M., Blanquet, S., and Plateau, P. (1995). Direct Random Mutagenesis of Gene-Sized DNA Fragments Using Polymerase Chain Reaction. Anal Biochem *224*, 347–353. 10.1006/abio.1995.1050.

138. Zhang, M., Yang, S., Nelakanti, R., Zhao, W., Liu, G., Li, Z., Liu, X., Wu, T., Xiao, A., and Li, H. (2020). Mammalian ALKBH1 serves as an N6-mA demethylase of unpairing DNA. Cell Res *30*, 197–210. 10.1038/s41422-019-0237-5.

139. McIntyre, A.B.R., Alexander, N., Grigorev, K., Bezdan, D., Sichtig, H., Chiu, C.Y., and Mason, C.E. (2019). Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. Nat Commun *10*, 579. 10.1038/s41467-019-08289-9.

140. O'Brown, Z.K., Boulias, K., Wang, J., Wang, S.Y., O'Brown, N.M., Hao, Z., Shibuya, H., Fady, P.-E., Shi, Y., He, C., et al. (2019). Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. BMC Genomics *20*, 445. 10.1186/s12864-019-5754-6.

141. Beaulaurier, J., Zhang, X.-S., Zhu, S., Sebra, R., Rosenbluh, C., Deikus, G., Shen, N., Munera, D., Waldor, M.K., Chess, A., et al. (2015). Single molecule-level detection and long read-based phasing of epigenetic variations in bacterial methylomes. Nat Commun *6*, 7438. 10.1038/ncomms8438.

142. Powell, L.M., Lejeune, E., Hussain, F.S., Cronshaw, A.D., Kelly, S.M., Price, N.C., and Dryden, D.T.F. (2003). Assembly of EcoKI DNA methyltransferase requires the C-terminal region of the HsdM modification subunit. Biophys Chem *103*, 129–137. 10.1016/S0301-4622(02)00251-X.

143. Kurylo, C.M., Alexander, N., Dass, R.A., Parks, M.M., Altman, R.A., Vincent, C.T., Mason, C.E., and Blanchard, S.C. (2016). Genome Sequence and Analysis of Escherichia coli MRE600, a Colicinogenic, Nonmotile Strain that Lacks RNase I and the Type I Methyltransferase, EcoKI. Genome Biol Evol *8*, 742–752. 10.1093/gbe/evw008.

144. Anton, B.P., Mongodin, E.F., Agrawal, S., Fomenkov, A., Byrd, D.R., Roberts, R.J., and Raleigh, E.A. (2015). Complete Genome Sequence of ER2796, a DNA Methyltransferase-Deficient Strain of Escherichia coli K-12. PLoS One *10*, e0127446. 10.1371/journal.pone.0127446.

145. Tapella, R., Ashby, M., Sethuraman, A., and Rhall, P.B. Methylome analysis technical note. https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/Methylome-Analysis-Technical-Note.

146. Suen, T.C., and Hung, M.C. (1991). c-myc reverses neu-induced transformed morphology by transcriptional repression. Mol Cell Biol *11*, 354–362. 10.1128/mcb.11.1.354-362.1991.

147. Schechter, A.L., Stern, D.F., Vaidyanathan, L., Decker, S.J., Drebin, J.A., Greene, M.I., and Weinberg, R.A. (1984). The neu oncogene: an erb-B-related gene encoding a 185,000-Mr tumour antigen. Nature *312*, 513–516. 10.1038/312513a0.

148. Drebin, J.A., Stern, D.F., Link, V.C., Weinberg, R.A., and Greene, M.I. (1984). Monoclonal antibodies identify a cell-surface antigen associated with an activated cellular oncogene. Nature *312*, 545–548. 10.1038/312545a0.

149. Wang, Y., Huang, N., Li, H., Liu, S., Chen, X., Yu, S., Wu, N., Bian, X.-W., Shen, H.-Y., Li, C., et al. (2017). Promoting oligodendroglial-oriented differentiation of glioma stem cell: a repurposing of quetiapine for the treatment of malignant glioma. Oncotarget *8*, 37511–37524. 10.18632/oncotarget.16400.

150. Reindl, J., Shevtsov, M., Dollinger, G., Stangl, S., and Multhoff, G. (2019). Membrane Hsp70-supported cell-to-cell connections via tunneling nanotubes revealed by live-cell STED nanoscopy. Cell Stress Chaperones *24*, 213–221. 10.1007/s12192-018-00958-w.

151. Wheeler, M.A., Jaronen, M., Covacu, R., Zandee, S.E.J., Scalisi, G., Rothhammer, V., Tjon, E.C., Chao, C.-C., Kenison, J.E., Blain, M., et al. (2019). Environmental Control of Astrocyte Pathogenic Activities in CNS Inflammation. Cell *176*, 581-596.e18. 10.1016/j.cell.2018.12.012.

152. Wan, Q., Ni, L., Wu, L., Zhang, L., Liu, M., and Jiang, X. (2015). The determination of sex type of the cultured murine cell with quantitative PCR technique. Hum Cell *28*, 154–157. 10.1007/s13577-015-0109-3.

153. VAISSIERE, T., SAWAN, C., and HERCEG, Z. (2008). Epigenetic interplay between histone modifications and DNA methylation in gene silencing. Mutation Research/Reviews in Mutation Research *659*, 40–48. 10.1016/j.mrrev.2008.02.004.

154. Liu, S., Yin, F., Fan, W., Wang, S., Guo, X., Zhang, J., Tian, Z., and Fan, M. (2012). Over-expression of BMPR-IB reduces the malignancy of glioblastoma cells by upregulation of p21 and p27Kip1. Journal of Experimental & Clinical Cancer Research *31*, 52. 10.1186/1756-9966-31-52.

155. Lee, J., Son, M.J., Woolard, K., Donin, N.M., Li, A., Cheng, C.H., Kotliarova, S., Kotliarov, Y., Walling, J., Ahn, S., et al. (2008). Epigenetic-Mediated Dysfunction of the Bone Morphogenetic Protein Pathway Inhibits Differentiation of Glioblastoma-Initiating Cells. Cancer Cell *13*, 69–80. 10.1016/j.ccr.2007.12.005.

156. Hong, T., Yuan, Y., Wang, T., Ma, J., Yao, Q., Hua, X., Xia, Y., and Zhou, X. (2017). Selective detection of N6-methyladenine in DNA via metal ion-mediated replication and rolling circle amplification. Chem Sci *8*, 200–205. 10.1039/C6SC02271E.

157. Beh, L.Y., Debelouchina, G.T., Clay, D.M., Thompson, R.E., Lindblad, K.A., Hutton, E.R., Bracht, J.R., Sebra, R.P., Muir, T.W., and Landweber, L.F. (2019). Identification of a DNA N6-Adenine Methyltransferase Complex and Its Impact on Chromatin Organization. Cell *177*, 1781-1796.e25. 10.1016/j.cell.2019.04.028.

158. Mondo, S.J., Dannebaum, R.O., Kuo, R.C., Louie, K.B., Bewick, A.J., LaButti, K., Haridas, S., Kuo, A., Salamov, A., Ahrendt, S.R., et al. (2017). Widespread adenine N6-methylation of active genes in fungi. Nat Genet *49*, 964–968. 10.1038/ng.3859.

159. He, S., Zhang, G., Wang, J., Gao, Y., Sun, R., Cao, Z., Chen, Z., Zheng, X., Yuan, J., Luo, Y., et al. (2019). 6mA-DNA-binding factor Jumu controls maternal-to-zygotic transition upstream of Zelda. Nat Commun *10*, 2219. 10.1038/s41467-019-10202-3.

160. Zhang, G., Huang, H., Liu, D., Cheng, Y., Liu, X., Zhang, W., Yin, R., Zhang, D., Zhang, P., Liu, J., et al. (2015). N6-Methyladenine DNA Modification in Drosophila. Cell *161*, 893–906. 10.1016/j.cell.2015.04.018.

161. Greer, E.L., Blanco, M.A., Gu, L., Sendinc, E., Liu, J., Aristizábal-Corrales, D., Hsu, C.-H., Aravind, L., He, C., and Shi, Y. (2015). DNA Methylation on N6-Adenine in C. elegans. Cell *161*, 868–878. 10.1016/j.cell.2015.04.005.

162. Thijs, S., Op De Beeck, M., Beckers, B., Truyens, S., Stevens, V., Van Hamme, J.D., Weyens, N., and Vangronsveld, J. (2017). Comparative Evaluation of Four Bacteria-Specific Primer Pairs for 16S rRNA Gene Surveys. Front Microbiol *8*. 10.3389/fmicb.2017.00494.

163. Uchida, S., Hara, K., Kobayashi, A., Funato, H., Hobara, T., Otsuki, K., Yamagata, H., McEwen, B.S., and Watanabe, Y. (2010). Early Life Stress Enhances Behavioral Vulnerability to Stress through the Activation of REST4-Mediated Gene Transcription in the Medial Prefrontal Cortex of Rodents. The Journal of Neuroscience *30*, 15007–15018. 10.1523/JNEUROSCI.1436-10.2010.

164. Valori, V., Tus, K., Laukaitis, C., Harris, D.T., LeBeau, L., and Maggert, K.A. (2020). Human rDNA copy number is unstable in metastatic breast cancers. Epigenetics *15*, 85–106. 10.1080/15592294.2019.1649930.

165. Tunster, S.J. (2017). Genetic sex determination of mice by simplex PCR. Biol Sex Differ *8*, 31. 10.1186/s13293-017-0154-6.

166. Johnson, M.B., Kawasawa, Y.I., Mason, C.E., Krsnik, Ž., Coppola, G., Bogdanović, D., Geschwind, D.H., Mane, S.M., State, M.W., and Šestan, N. (2009). Functional and Evolutionary Insights into Human Brain Development through Global Transcriptome Analysis. Neuron *62*, 494–509. 10.1016/j.neuron.2009.03.027.

167. Hamdan, F.F., Daoud, H., Piton, A., Gauthier, J., Dobrzeniecka, S., Krebs, M.-O., Joober, R., Lacaille, J.-C., Nadeau, A., Milunsky, J.M., et al. (2011). De Novo SYNGAP1 Mutations in Nonsyndromic Intellectual Disability and Autism. Biol Psychiatry *69*, 898–901. 10.1016/j.biopsych.2010.11.015.

168. Berryer, M.H., Hamdan, F.F., Klitten, L.L., Møller, R.S., Carmant, L., Schwartzentruber, J., Patry, L., Dobrzeniecka, S., Rochefort, D., Neugnot-Cerioli, M., et al. (2013). Mutations in SYNGAP1 Cause Intellectual Disability, Autism, and a Specific Form of Epilepsy by Inducing Haploinsufficiency. Hum Mutat *34*, 385–394. 10.1002/humu.22248.

169. Roe, A.M., and Shur, N. (2007). From new screens to discovered genes: The successful past and promising present of single gene disorders. Am J Med Genet C Semin Med Genet *145C*, 77–86. 10.1002/ajmg.c.30121.

170. Dang, V.T., Kassahn, K.S., Marcos, A.E., and Ragan, M.A. (2008). Identification of human haploinsufficient genes and their genomic proximity to segmental duplications. European Journal of Human Genetics *16*, 1350–1357. 10.1038/ejhg.2008.111.

171. Miller, A.C., Comellas, A.P., Hornick, D.B., Stoltz, D.A., Cavanaugh, J.E., Gerke, A.K., Welsh, M.J., Zabner, J., and Polgreen, P.M. (2020). Cystic fibrosis carriers are at increased risk for a wide range of cystic fibrosis-related conditions. Proceedings of the National Academy of Sciences *117*, 1621–1627. 10.1073/pnas.1914912117.

172. Zschocke, J., Byers, P.H., and Wilkie, A.O.M. (2023). Mendelian inheritance revisited: dominance and recessiveness in medical genetics. Nat Rev Genet *24*, 442–463. 10.1038/s41576-023-00574-0.

173. Bochtler, M., and Fernandes, H. (2021). DNA adenine methylation in eukaryotes: Enzymatic mark or a form of DNA damage? BioEssays *43*, 2000243. 10.1002/bies.202000243.

174. Bradford, J.A., and Clarke, S.T. (2011). Dual-Pulse Labeling Using 5-Ethynyl-2′-Deoxyuridine (EdU) and 5-Bromo-2′-Deoxyuridine (BrdU) in Flow Cytometry. Curr Protoc Cytom *55*. 10.1002/0471142956.cy0738s55.

175. Crossley, B.M., Bai, J., Glaser, A., Maes, R., Porter, E., Killian, M.L., Clement, T., and Toohey-Kurth, K. (2020). Guidelines for Sanger sequencing and molecular assay monitoring. Journal of Veterinary Diagnostic Investigation *32*, 767–775. 10.1177/1040638720905833.

176. Fagan, P.A., Fàbrega, C., Eritja, R., Goodman, M.F., and Wemmer, D.E. (1996). NMR Study of the Conformation of the 2-Aminopurine:Cytosine Mismatch in DNA. Biochemistry *35*, 4026–4033. 10.1021/bi952657g.

177. Ge, R., Ye, C., Peng, Y., Dai, Q., Zhao, Y., Liu, S., Wang, P., Hu, L., and He, C. (2023). m6A-SAC-seq for quantitative whole transcriptome m6A profiling. Nat Protoc *18*, 626–657. 10.1038/s41596-022-00765-9.

178. Zhang, L.-S., Dai, Q., and He, C. (2023). BID-seq: The Quantitative and Base-Resolution Sequencing Method for RNA Pseudouridine. ACS Chem Biol *18*, 4–6. 10.1021/acschembio.2c00881.

179. Dai, Q., Zhang, L.-S., Sun, H.-L., Pajdzik, K., Yang, L., Ye, C., Ju, C.-W., Liu, S., Wang, Y., Zheng, Z., et al. (2022). Quantitative sequencing using BID-seq uncovers abundant pseudouridines in mammalian mRNA at base resolution. Nat Biotechnol. 10.1038/s41587-022-01505-w.

180. Guarné, A. (2012). The Functions of MutL in Mismatch Repair. In, pp. 41–70. 10.1016/B978-0-12-387665-2.00003-1.

181. Chen, H., Gu, L., Orellana, E.A., Wang, Y., Guo, J., Liu, Q., Wang, L., Shen, Z., Wu, H., Gregory, R.I., et al. (2020). METTL4 is an snRNA m6Am methyltransferase that regulates RNA splicing. Cell Res *30*, 544–547. 10.1038/s41422-019-0270-4.

182. Haag, S., Sloan, K.E., Ranjan, N., Warda, A.S., Kretschmer, J., Blessing, C., Hübner, B., Seikowski, J., Dennerlein, S., Rehling, P., et al. (2016). NSUN3 and ABH1 modify the wobble position of mt-tRNA-Met to expand codon recognition in mitochondrial translation. EMBO J *35*, 2104–2119. 10.15252/embj.201694885.

183. Li, M.-M., Nilsen, A., Shi, Y., Fusser, M., Ding, Y.-H., Fu, Y., Liu, B., Niu, Y., Wu, Y.-S., Huang, C.-M., et al. (2013). ALKBH4-dependent demethylation of actin regulates actomyosin dynamics. Nat Commun *4*, 1832. 10.1038/ncomms2863.

184. Woodcock, C.B., Yu, D., Zhang, X., and Cheng, X. (2019). Human HemK2/KMT9/N6AMT1 is an active protein methyltransferase, but does not act on DNA in vitro, in the presence of Trm112. Cell Discov *5*, 50. 10.1038/s41421-019-0119-5.

185. Zhu, C., Gao, Y., Guo, H., Xia, B., Song, J., Wu, X., Zeng, H., Kee, K., Tang, F., and Yi, C. (2017). Single-Cell 5-Formylcytosine Landscapes of Mammalian Early Embryos and ESCs at Single-Base Resolution. Cell Stem Cell *20*, 720-731.e5. 10.1016/j.stem.2017.02.013.

186. Safra, M., Sas-Chen, A., Nir, R., Winkler, R., Nachshon, A., Bar-Yaacov, D., Erlacher, M., Rossmanith, W., Stern-Ginossar, N., and Schwartz, S. (2017). The m1A landscape on cytosolic and mitochondrial mRNA at single-base resolution. Nature *551*, 251–255. 10.1038/nature24456.

187. Malbec, L., Zhang, T., Chen, Y.-S., Zhang, Y., Sun, B.-F., Shi, B.-Y., Zhao, Y.-L., Yang, Y., and Yang, Y.-G. (2019). Dynamic methylome of internal mRNA N7-methylguanosine and its regulatory role in translation. Cell Res *29*, 927–941. 10.1038/s41422-019-0230-z.