THE UNIVERSITY OF CHICAGO

# MAPPING $N^6$-METHYLATION ON ADENINE BASE IN TRANSCRIPTOME AND GENOME AT HIGH RESOLUTION

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY

KAI CHEN

CHICAGO, ILLINOIS

AUGUST 2016

*For my family who has constantly supported me and encouraged me to follow my heart, who has provided me the best love in the world*

*For my girlfriend who stays with me whatever difficulties I am facing and shows me her love*

*For all my friends who always stand by my side and never fail me when I need their help*

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF SCHEMES

## LIST OF TABLES

## ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Professor Chuan He. I thank him for all the opportunities he provides. His scientific insight and guidance helps me with my graduate research on these challenging and exciting topics; his patience and encouragement lead me to overcome the difficulties in scientific exploration.

I would like to thank Professor Tao Pan for his instruction on how to address the RNA related problems. I benefit a lot from his expertise in RNA biology and thank him to serve on my thesis committee. I am grateful to Professor Joseph A. Piccirilli for being my committee member again. He served on my committee in the oral examination. I am happy that he would like to come again.

I would also like to thank my collaborators, Dr. Lijia Ma from Professor Kevin White's group, Mr. Zhike Lu, and Professor Xin He from the Department of Human Genetics for working together on data analysis and statistical modelling.

I thank the University of Chicago facilities, and Mass Spectrometry Facility at Fudan University (Shanghai, China). Their solid supports facilitated my graduate research.

All the work mentioned here cannot be done without other He group members, including but not limited to Dr. Guan-Zheng Luo and Mr. Hang Yin. I want to express my gratitude to all the current members and alumni in He group, particularly Dr. Ye Fu and Professor Xin Deng. I have learned a lot from them, and I have enjoyed all the fruitful discussion with them. I thank Sarah F. Reichard, MA for editing the manuscripts.

Finally, I would like to thank my family, my girlfriend and all my friends for all their love and company. It is them who make me stronger and support me to go through all the hard time in life.

## ABSTRACT

Both genome and transcriptome carry a wide range of different chemical modifications. Among these modifications, methylation is the most abundant, exerting important functions in multiple biological processes in both DNA and RNA related pathways in prokaryotes and eukaryotes. The recent discoveries of $N^6$-methyladenosine ($m^6A$) in transcriptome with reversible dynamics and regulatory roles and $N^6$-methyladenine (6mA) in eukaryotic genome as an epigenetic mark have drawn considerable attentions of the scientific community. The development of high-throughput sequencing (next generation sequencing) provides us a powerful tool to study this chemical modification in genome-wide and/or transcriptome-wide level. The high resolution map, hence, is of great necessity for further investigation in its biological meanings. To meet the requirement, two major strategies are designed and will be discussed in the thesis.

The photo-crosslinking-assisted strategy introduces covalent bond between antibody and nucleic acid and improves the efficiency and specificity of immunoprecipitation; the nuclease digestion further narrows down the detection region, thus significantly increasing the mapping resolution. The strategy was applied to the detection of $m^6A$ in bacterial mRNA, the study of relative positions of methylation and RNA binding proteins, and the first glimpse of 6mA pattern and methylation motif in *Chlamydomonas* genome.

The deamination based $m^6A$ sequencing approach, inspired by deamination based bisulfite sequencing to distinguish 5-methylcytosine from cytosine, is designed to achieve the single nucleotide resolution $m^6A$ map. The model study and pilot experiment have demonstrated NGS detectable A-to-G conversion is introduced by deamination treatment, which is dependent on methylation level, indicating the potential of this method to transcriptome-wide differentiate and quantify $m^6A$.

## LIST OF PUBLICATIONS BASED ON THE WORK PRESENTED IN THIS THESIS

1. Chen, K., Zhao, Boxuan S., & He, C. Nucleic Acid Modifications in Regulation of Gene Expression. *Cell Chem. Biol*. **23**, 74-85 (2016).
2. Chen, K.*, Lu, Z.*, Wang, X., Fu, Y., Luo, G.-Z., Liu, N., Han, D., Dominissini, D., Dai, Q., Pan, T. & He, C. High-Resolution $N^6$-Methyladenosine ($m^6A$) Map Using Photo-Crosslinking-Assisted $m^6A$ Sequencing. *Angew. Chem. Int. Ed*. **54**, 1587-1590 (2015).
3. Deng, X.*, Chen, K.*, Luo, G.-Z., Weng, X., Ji, Q., Zhou, T. & He, C. Widespread occurrence of $N^6$-methyladenosine in bacterial mRNA. *Nucleic Acids Res.* gkv596 (2015).
4. Fu, Y.*, Luo, G.-Z.*, Chen, K., Deng, X., Yu, M., Han, D., Hao, Z., Liu, J., Lu, X., Doré, Louis C., Weng, X., Ji, Q., Mets, L. & He, C. $N^6$-Methyldeoxyadenosine Marks Active Transcription Start Sites in *Chlamydomonas*. *Cell* **161**, 879-892 (2015).
5. Chen, K., Luo, G.-Z. & He, C. Chapter Nine - High-Resolution Mapping of $N^6$-Methyladenosine in Transcriptome and Genome Using a Photo-Crosslinking-Assisted Strategy in *Meth. Enzymol.* Vol. Volume 560 (ed He, Chuan) 161-185 (Academic Press, 2015)

*These authors contributed equally to the work.

# 1 Introduction

## 1.1 Nucleic acid modifications in "epigenome" and "epitranscriptome"

Epigenetics is defined as "the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence".[1] Chemical modifications on nucleic acid and histone, besides the Polycomb-trithorax group (Pc-G/trx) protein complexes, carry attributes of epigenetic processes and play key roles in gene expression regulation.[2]

Nucleic acids have a wide range of different chemical modifications. In contrast to previous views that these modifications are static and only play fine-tuning functions, recent research advances paint a much more dynamic picture. These diverse modifications in genome and transcriptome are employed as specific marks to exert essential or critical influences in a variety of cellular processes in eukaryotic organisms, which are composed of "epigenome" and "epitranscriptome", one of the current focuses in biological research.

## 1.2 Cytosine epigenetic marks in DNA

The existence of cytosine methylation (5mC) in genomic DNA was first reported by Wyatt in 1951.[3] More than two decades later, the regulatory maintenance of the 5mC pattern across cell divisions was proposed.[4,5] The activity of the speculated writer enzymes, mammalian methyltransferases, was detected in cellular extracts early on.[6,7] But it was not until in 1983 that the first DNA methyltransferase, Dnmt1, was purified by the Ingram group.[8] Dnmt1 was shown to preferentially methylate the hemimethylated DNA at CpG sites; and its loss in mouse embryonic stem cells (mESCs) leads to genome-wide depletion of the CpG methylation, indicating the methylation-maintaining role of Dnmt1 during DNA replication. Besides the maintenance of 5mC, *de novo* DNA methylation, i.e. the establishment of 5mC on unmethylated DNA, was detected in early pluripotent embryonic cells by the Jaenisch group in 1982.[9] Subsequent homology

studies in mouse carried out by the Li group led to the discovery of Dnmt3a and Dnmt3b, which are two enzymes responsible for *de novo* methylation of proviral DNA and repetitive sequences.[10,11] Later the same group showed that these methyltransferases are also required for the establishment of *de novo* methylation on maternal imprinted genes with the cooperation of Dnmt3L.[12]

The functional outcomes of DNA methylation are generally associated with the repression of gene expression. Early studies largely benefited from the inhibitory effects of 5-azacytidine on DNA methylation in living cells, in which the reactivation of silenced genes was shown to be achieved by the use of this nucleoside analog.[13-15] The later study using *dnmt1* knockout mice also revealed that loss of methylation led to the reactivation of several naturally inactive genes.[16] These findings suggested the repressive nature of DNA methylation.

A more direct approach for the functional investigation of 5mC in genomic DNA involves the discovery and characterization of proteins that recognize 5mC and carry out subsequent actions, i.e. 5mC effectors or readers (**Figure 1.1**). The first 5mC reader to be characterized was methyl-CpG binding protein complex MeCP1, which was identified by the Bird group.[17] The subsequent studies eventually revealed four 5mC readers comprising the methyl-CpG binding domain (MBD) family, including MeCP2, MBD1, MBD2, and MBD4 (MBD3 in this family is not a 5mC reader).[18] Among them, MeCP2, MBD1, and MBD2 have been shown to be involved in 5mC-dependent transcriptional repression.[19] An unrelated p120 catenin partner protein Kaiso was also found to be a specific 5mC reader and functions as a methylation-dependent transcriptional repressor.[20] The discovery and subsequent characterization of 5mC readers led to a more comprehensive mechanistic elucidation of DNA methylation in gene expression regulation. This

specific binding of reader proteins to methylated CpG results in repression of gene expression and represents a fundamental epigenetic mechanism of particular importance in higher organisms.

DNA methylation has long been associated with regulation of gene expression. Together with histone modifications, DNA methylation modulates the chromatin structure and affects cognate gene expression by maintaining various expression patterns across cell types.[21,22] The presence of DNA methylation in the promoter region is directly connected to repression of transcription. How the repression is induced by methylation is described by two possible models: DNA methylation may either recruit its reader proteins that act as transcription repressors, e.g. MeCP and Kaiso proteins, preventing transcriptional factors from accessing to the promoter region as an "indirect" model; or serve as a disruptor to interfere with the binding of certain transcription factors and thus prevent the activation of corresponding genes in a "direct" model.[23-27] In contrast, DNA methylation in the gene body shows positive correlation with gene expression.[28-35] Gene body DNA methylation may affect the transcription elongation process, regulate RNA splicing, and alter nucleosome-positioning, which further highlight diverse functions of DNA methylation in gene expression.[31,36-38] It should be noted that the transcription regulation roles of DNA methylation typically synergize with various histone marks as the methyltransferases, demethylases and readers of DNA methylation interact with various histone marks or histone modification enzymes.

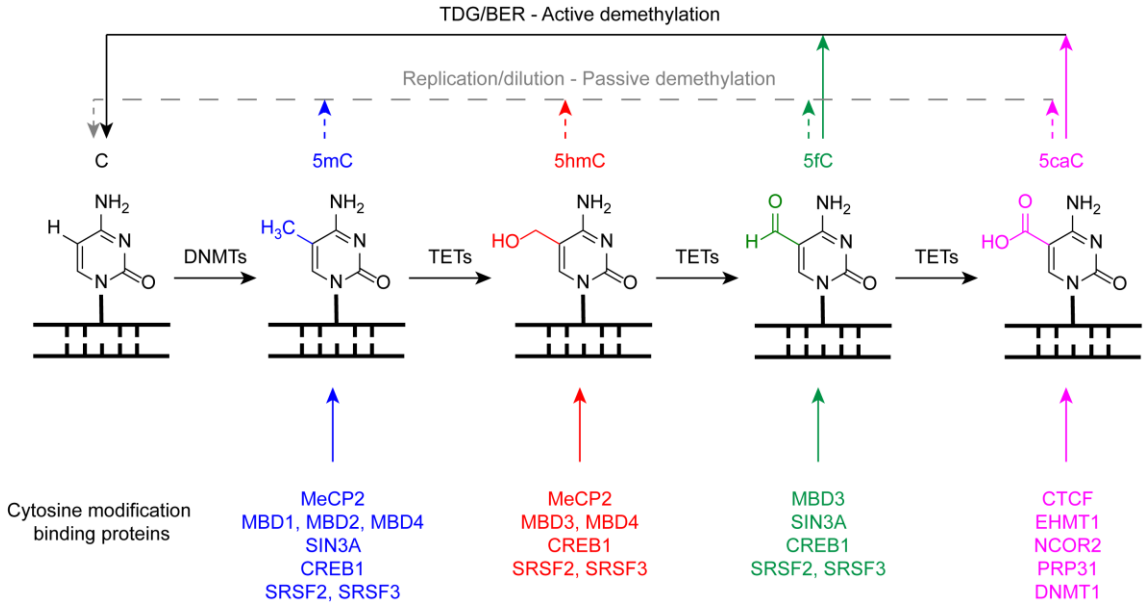



**Figure 1.1. Scheme of the reversible cytosine methylation in DNA, and binding proteins that are known to or proposed to bind modified cytosine derivatives.[39]**

Based on early observations, DNA 5mC methylation was considered to be dynamic and reversible.[40-42] However, although the writers and readers of 5mC were identified and characterized early on, identity of eraser, enzymes that remove 5mC methyl mark remained a mystery for several decades. This all changed with the discovery of 5-hydroxymethylcytosine (5hmC) in the mammalian genome and the identification of TET (ten-eleven translocation) proteins. TET proteins are methylcytosine dioxygenase that utilize dioxygen to oxidize 5mC to 5hmC.[43-45] Subsequent studies demonstrated that the TET enzymes can further oxidize 5hmC to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC).[46-48] Both 5fC and 5caC can be recognized and excised by human thymine DNA glycosylase (TDG), followed by base excision repair (BER) to replace the modified cytosine with a normal cytosine, completing the active demethylation process.[46,49] Additionally, the 5mC oxidation derivatives of 5hmC, 5fC, and 5caC may also be passively diluted to the unmethylated stage through cell division (**Figure 1.1**).[50]

The close relationship between cytosine methylation and levels of gene expression has resulted in intense research that focused on investigation of presence and patterns of cytosine methylation within promoter regions.[51] However, recent progress reveals that distal regulatory elements, genomic regions that are further away from the genes but are occupied by transcriptional factors and can loop back to interact and regulate transcription, may undergo much more dynamic methylation and demethylation.[52-58] Genome-wide studies support the hypothesis that active demethylation, associated with the presence of methylated cytosine oxidation derivatives, may play critical roles in cell development and stem cell maintenance.[59]

Furthermore, the process of active demethylation is known to occur in certain biological contexts. For example, during fertilization, the loss of 5mC in paternal chromosome and the appearance of 5hmC/5fC/5caC have been observed by immunostaining, suggesting the association between active DNA demethylation and embryonic development. The genome-wide distributions of all 5mC oxidation derivatives have also been mapped by recent advances in developing next-generation sequencing methods, which showed wide-spread active demethylation events at the genomic level along with the association of 5hmC/5fC/5caC with functional elements.[50,60-62] In addition, some human cancers have been associated with aberrant TET activity. Reduced 5hmC abundance and downregulation of TET activity were observed during tumor progression, including melanoma, hepatocellular carcinoma, and hematopoietic malignancies.[63-66] These findings indicate that 5mC oxidation derivatives could be used as markers in cancer diagnostics and prognostics. The DNA 5mC oxidation and demethylation pathway could also be targeted for therapeutic interventions.

## 1.3 A new eukaryotic DNA epigenetic mark: $N^6$-methyladenine (6mA)

Another methylation modification, $N^6$-methyladenine (6mA or $m^6dA$), exists in the genomic DNA of prokaryotes and plays critical roles (**Figure 1.2A**).[67] In bacteria, 6mA serves as an important marker participating in DNA repair, replication, and cell defense.[68-73] In particular, 6mA is a marker in restriction–modification (R-M) systems, in which 6mA can be recognized by corresponding restriction endonucleases as a label to prevent the host genome from restriction digestion and further enable the degradation of unmethylated foreign DNA.[74] The deletion of 6mA methyltransferase in pathogenic *Escherichia coli* leads to global transcription changes, indicating significant regulatory functions of 6mA besides a host genetic marker.[75] Intriguingly, although bacteria employ R-M systems to cleave foreign genomic DNA such as bacteriophage DNA, 6mA methyltransferases were found to be encoded by some viral DNA.[76-79]

It should be noted that besides 6mA, bacterial genomes also contain $N^4$-methylcytosine (4mC or $m^4dC$) and 5mC (**Figure 1.2A**).[80,81] These cytosine modifications are also used by bacterial restriction–modification (R-M) systems as defense mechanisms. These two cytosine modifications can be differentiated at base resolution using a revised TAB-seq protocol.[82,83]

In addition to prokaryotes, several eukaryotes have relatively abundant 6mA in genomic DNA.[84-87] However, functions of 6mA in eukaryotic systems remained unclear until very recently. The lack of R-M systems suggests that 6mA mainly exerts regulatory roles in these unicellular eukaryotes.[88] On the other hand, the existence of 6mA in higher eukaryotes, which was supported by indirect evidence, hinted that 6mA might be another potential epigenetic DNA mark in eukaryotes in addition to 5-methylcytosine.[67]

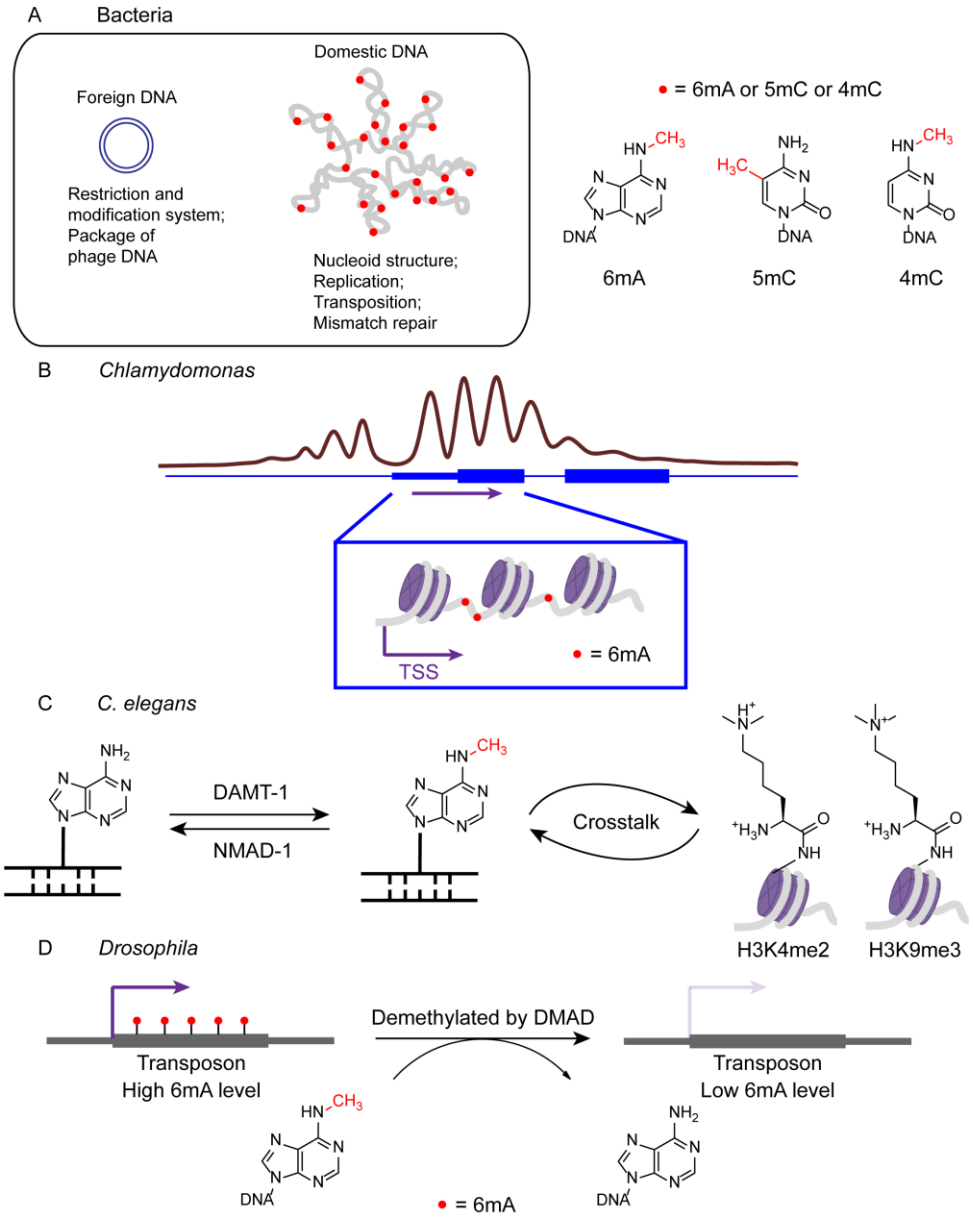



**Figure 1.2. $N^6$-methylation on adenine in genomic DNA.** (A) A brief overview of biological function of methyl groups in bacterial genomic DNA. (B) High-throughput mapping of $N^6$-methyladenine (6mA) in *Chlamydomonas reinhardtii* revealed a unique distribution pattern in the genome with complete depletion at transcription start sites (TSS) and high enrichment at the linker region between nucleosomes. (C) In *Caenorhabditis elegans* 6mA is installed by DAMT-1 and reversibly removed by NMAD-1. The "crosstalk" between 6mA and histone modification, particularly the histone H3 methylation, indicates critical roles that 6mA may play in gene expression regulation. (D) 6mA in *Drosophila melanogaster* could be converted back to A by Tet homolog DMAD. Intriguingly, the 6mA level is correlated with the expression level of transposon, supporting the regulatory significance of 6mA in eukaryotes.

In 2015, three groups reported the presence of 6mA in three different eukaryotes independently, shedding light on the function of this methylation modification in eukaryotes.[89-91] The existence of 6mA in alga genomic DNA was verified decades ago, but the methylation distribution and biological function were not addressed then. By employing and developing several high-throughput sequencing approaches to map 6mA in *Chlamydomonas* genomic DNA, it was revealed that 6mA is not only enriched at ApT dinucleotides around transcription start sites (TSS), but also labels the active transcribed genes, and marks the linker DNA regions between adjacent nucleosomes, indicating the potential gene activation function of adenine methylation in the *Chlamydomonas* genome (**Figure 1.2B**).[89]

The discovery of 6mA, its demethylase NMAD-1 and potential methyltransferase DAMT-1 in *C. elegans* changed the previous view that *C. elegans* lacks DNA methylation, raising an intriguing possibility that 6mA serves as DNA methylation mark instead of 5mC. The phenotypes of deletion of *nmad-1* and *damt-1* and the crosstalk between adenine methylation and histone modifications indicate a potential gene activation role of 6mA (**Figure 1.2C**).[90]

By knocking out demethylase candidates in the *Drosophila* genome and monitoring the 6mA level, Zhang, Huang *et al* found that the *Drosophila* Tet homolog is likely responsible for the demethylation of 6mA in the *Drosophila* genome. The identified DNA 6mA demethylase (DMAD) regulates the level of 6mA during embryogenesis and tissue homeostasis processes. Further sequencing analyses have revealed that the dynamic demethylation is correlated with transposon expression and plays a critical role in development (**Figure 1.2D**).[91]

## 1.4 Beyond DNA: $N^6$-methyladenosine ($m^6A$) methylation on messenger RNA

In addition to genomic DNA being regulated by different chemical modifications, RNA molecules are also decorated with similar modifications and some of them have been appreciated

for decades. For example, the existence of $N^6$-methyladenosine ($m^6A$) in mRNA was discovered in 1974 in both eukaryotic and viral mRNAs.[92-97] $m^6A$ is the most prevalent internal modification in mRNAs and long non-coding RNAs (lncRNAs) in higher eukaryotes.[98] It was revealed that, in the mammalian transcriptome, approximately 3 $m^6A$ marks exist per mRNA molecule and occur within a consensus motif of G($m^6A$)C (70%) or A($m^6A$)C (30%), but the methylation percentage at each site varies substantially.[99-105]

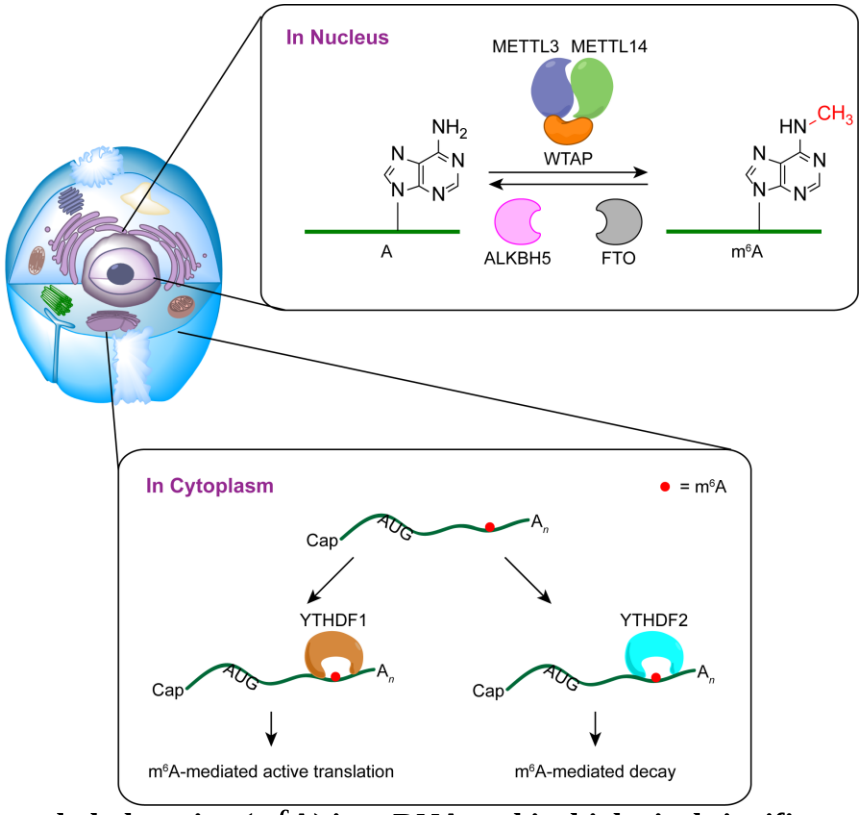



**Figure 1.3. $N^6$-methyladenosine ($m^6A$) in mRNA and its biological significance.** The reversible methylation and demethylation process occurs in the nucleus, catalyzed by methyltransferase complex and demethylases, respectively. The $m^6A$ modification has profound effects on mRNA fate: it switches mRNA to active translation mode, and also accelerates its decay rate.

The methylation of adenosine in mRNA could be dynamic and could introduce biological regulations.[106] The methyl group on $N^6$ position of adenosine is installed in the nucleus by $m^6A$ 'writer', a multicomponent complex that contains methyltransferase like 3 (METTL3), methyl-

transferase like 14 (METTL14), and Wilms' tumor 1-associating protein (WTAP) (**Figure 1.3**).[107-112] The deficiency in the methyltransferase complex leads to significant phenotypes, such as blocking the subsequent differentiation of mESCs, lethality at the early stage of mouse embryo development, developmental arrest or defects in gametogenesis in yeast, flies, and plants.[107,113-120] In zebrafish, the knockdown of METTL3 leads to smaller head, eyes, and brain ventricle, and curved notochord.[111] Moreover, the $m^6A$ level is thought to be regulated by miRNA through the modulated binding of METTL3 to mRNA.[121]

Two human AlkB family proteins, the fat mass and obesity-associated protein (FTO) and ALKBH5, serve as $m^6A$ 'eraser', exerting the function of RNA demethylases to remove $m^6A$ methylation in mammalian polyA-tailed RNA (**Figure 1.3**).[122,123] FTO has significant effects on development while ALKBH5 affects spermatogenesis, suggesting the effects of $m^6A$ on multiple biological phenomena.[122-126] The 'reader' proteins YTHDF1 and YTHDF2, specifically binding to $m^6A$ and shown to interact with thousands of mRNA targets, mediate methylation-dependent translation efficiency regulation and mRNA decay, respectively (**Figure 1.3**).[127,128] Besides the direct reader proteins, two groups have independently demonstrated that $m^6A$ methylation also modulates the mRNA and lncRNA structure transcriptome-wide, which dramatically affects protein-RNA interactions to impact mRNA abundance and the alternative splicing of the methylated RNA.[129,130] hnRNPA2B1 has also been shown to selectively recognize methylated pri-microRNA in order to promote microRNA maturation.[131,132] In a separate study, $m^6A$ depletion prolongs nuclear retention and delays the nuclear exit export of mature mRNAs of clock genes *per2* and *arntl*, which connect $m^6A$ to the pace of the circadian cycle and the clock speed and stability.[133] The discoveries of $m^6A$ also present in other organisms further indicate its critical roles in biological functions.[134,135]

Transcriptome-wide sequencing revealed that $m^6A$ shows distinct distribution patterns in eukaryotic transcriptomes. In mammals, $m^6A$ is enriched at 3'-UTR around stop codons and also marks long exons, which may be related to mRNA splicing.[116,136,137] In *Arabidopsis* mRNA, this mark is enriched not only at 3'-UTR but also 5'-UTR.[135] It has been suggested that the selectivity of methylation installation machinery is modulated by the cofactors in the $m^6A$ methyltransferase complex, yet the detailed mechanisms underlying the specific distribution remain to be unveiled.[137]

## 1.5 Scope of thesis

The recent progress in understanding the biological functions and reversible properties of nucleic acid modifications, particularly 6mA in genome and $m^6A$ in transcriptome, leads to a much more dynamic landscape on how the chemical diversity regulates and get regulated by gene expression. The scientific research on nucleic acid modification has specified the requirement of relatively macro-level tools to investigate the characteristics in a genome-wide or transcriptome-wide manner, which has been partially met by the rapid development of high-throughput sequencing (next generation sequencing or NGS). However, given that $N^6$-methylation on adenine base cannot be directly read by NGS, specific approaches/strategies aiming to facilitate the sequencing detection are still needed for high resolution genome-wide/transcriptome-wide mapping. The thesis will discuss some progress in 6mA/$m^6A$ sequencing method development.

**Chapter 2** presents a photo-crosslinking-assisted $m^6A$ sequencing strategy (PA-$m^6A$-seq) to more accurately define sites with $m^6A$ modifications. Using this strategy, we obtained a high resolution map of $m^6A$ in a human transcriptome.

**Chapter 3** presents the application of PA-m$^6$A-seq to transcriptome-wide profile m$^6$A in *Escherichia coli* and *Pseudomonas aeruginosa* to demonstrate the existence and distribution of m$^6$A in bacterial mRNA.

**Chapter 4** presents how the photo-crosslinking-assisted strategy followed by nuclease digestion assisted to identify 6mA containing regions in *Chlamydomonas* genomic DNA at higher resolution, resulting in the discovery of periodic pattern of 6mA distribution.

**Chapter 5** presents some preliminary results of a novel deamination based approach to distinguish m$^6$A from A, illustrating the potential to achieve single nucleotide resolution mapping.

**Chapter 6** summarizes the current results and discusses future directions in nucleic acid chemical modifications research.

## 1.6 References


1 Russo, V. E., Martienssen, R. A. & Riggs, A. D. *Epigenetic mechanisms of gene regulation*. (Cold Spring Harbor Laboratory Press, 1996).

2 Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6-21 (2002).

3 Wyatt, G. R. Recognition and Estimation of 5-Methylcytosine in Nucleic Acids. *Biochem. J.* **48**, 581-584 (1951).

4 Holliday, R. & Pugh, J. E. DNA Modification Mechanisms and Gene Activity during Development. *Science* **187**, 226-232 (1975).

5 Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell Genet.* **14**, 9-25 (1975).

6 Kalousek, F. & Morris, N. R. Deoxyribonucleic Acid Methylase Activity in Rat Spleen. *J. Biol. Chem.* **243**, 2440-2442 (1968).

7 Roy, P. H. & Weissbach, A. DNA methylase from Hela cell nuclei. *Nucleic Acids Res.* **2**, 1669-1684 (1975).

8 Bestor, T. H. & Ingram, V. M. Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. *Proc. Natl. Acad. Sci.-Biol.* **80**, 5559-5563 (1983).

9 Jahner, D. *et al. De novo* methylation and expression of retroviral genomes during mouse embryogenesis. *Nature* **298**, 623-628 (1982).

10 Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247-257 (1999).

11 Okano, M., Xie, S. P. & Li, E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat. Genet.* **19**, 219-220 (1998).

12 Hata, K., Okano, M., Lei, H. & Li, E. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development* **129**, 1983-1993 (2002).
13 Clough, D. W., Kunkel, L. M. & Davidson, R. L. 5-Azacytidine-induced reactivation of a herpes simplex thymidine kinase gene. *Science* **216**, 70-73 (1982).
14 Jones, P. A. & Taylor, S. M. Cellular differentiation, cytidine analogs and DNA methylation. *Cell* **20**, 85-93 (1980).
15 Mohandas, T., Sparkes, R. S. & Shapiro, L. J. Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science* **211**, 393-396 (1981).
16 Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366**, 362-365 (1993).
17 Meehan, R. R., Lewis, J. D., Mckay, S., Kleiner, E. L. & Bird, A. P. Identification of a Mammalian Protein That Binds Specifically to DNA Containing Methylated CpGs. *Cell* **58**, 499-507 (1989).
18 Hendrich, B. & Bird, A. Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol. Cell. Biol.* **18**, 6538-6547 (1998).
19 Bird, A. P. & Wolffe, A. P. Methylation-Induced Repression-Belts, Braces, and Chromatin. *Cell* **99**, 451-454 (1999).
20 Prokhortchouk, A. *et al.* The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. *Genes Dev.* **15**, 1613-1618 (2001).
21 Cheng, X. & Blumenthal, R. M. Coordinated Chromatin Control: Structural and Functional Linkage of DNA and Histone Methylation. *Biochemistry* **49**, 2999-3008 (2010).
22 De Carvalho, D. D., You, J. S. & Jones, P. A. DNA methylation and cellular reprogramming. *Trends Cell Biol.* **20**, 609-617 (2010).
23 Boyes, J. & Bird, A. DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. *Cell* **64**, 1123-1134 (1991).
24 Iguchi-Ariga, S. M. & Schaffner, W. CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. *Genes Dev.* **3**, 612-619 (1989).
25 Kovesdi, I., Reichel, R. & Nevins, J. R. Role of an adenovirus E2 promoter binding factor in E1A-mediated coordinate gene control. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 2180-2184 (1987).
26 Nan, X., Cross, S. & Bird, A. in *Novartis Foundation Symposium 214 - Epigenetics* 6-21 (John Wiley & Sons, Ltd., 2007).
27 Watt, F. & Molloy, P. L. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes Dev.* **2**, 1136-1143 (1988).
28 Aran, D., Toperoff, G., Rosenberg, M. & Hellman, A. Replication timing-related and gene body-specific methylation of active human genes. *Hum. Mol. Genet.* **20**, 670-680 (2011).
29 Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nat. Biotechnol.* **27**, 361-368 (2009).

30 Hellman, A. & Chess, A. Gene Body-Specific Methylation on the Active X Chromosome. *Science* **315**, 1141-1143 (2007).
31 Laurent, L. *et al.* Dynamic changes in the human methylome during differentiation. *Genome Res.* **20**, 320-331 (2010).
32 Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322 (2009).
33 Rauch, T. A., Wu, X., Zhong, X., Riggs, A. D. & Pfeifer, G. P. A human B cell methylome at 100−base pair resolution. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 671-678 (2009).
34 Shann, Y.-J. *et al.* Genome-wide mapping and characterization of hypomethylated sites in human tissues and breast cancer cell lines. *Genome Res.* **18**, 791-801 (2008).
35 Yang, X. *et al.* Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. *Cancer Cell* **26**, 577-590 (2014).
36 Chodavarapu, R. K. *et al.* Relationship between nucleosome positioning and DNA methylation. *Nature* **466**, 388-392 (2010).
37 Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484-492 (2012).
38 Lorincz, M. C., Dickerson, D. R., Schmitt, M. & Groudine, M. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat. Struct. Mol. Biol.* **11**, 1068-1075 (2004).
39 Liyanage, V. *et al.* DNA Modifications: Function and Applications in Normal and Disease States. *Biology* **3**, 670 (2014).
40 Mayer, W., Niveleau, A., Walter, J., Fundele, R. & Haaf, T. Embryogenesis: demethylation of the zygotic paternal genome. *Nature* **403**, 501-502 (2000).
41 Oswald, J. *et al.* Active demethylation of the paternal genome in the mouse zygote. *Curr. Biol.* **10**, 475-478 (2000).
42 Wu, S. C. & Zhang, Y. Active DNA demethylation: many roads lead to Rome. *Nat. Rev. Mol. Cell Biol.* **11**, 750-750 (2010).
43 Ito, S. *et al.* Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**, 1129-1133 (2010).
44 Kriaucionis, S. & Heintz, N. The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science* **324**, 929-930 (2009).
45 Tahiliani, M. *et al.* Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science* **324**, 930-935 (2009).
46 He, Y. F. *et al.* Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science* **333**, 1303-1307 (2011).
47 Ito, S. *et al.* Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science* **333**, 1300-1303 (2011).
48 Pfaffeneder, T. *et al.* The Discovery of 5-Formylcytosine in Embryonic Stem Cell DNA. *Angew. Chem. Int. Ed.* **50**, 7008-7012 (2011).
49 Maiti, A. & Drohat, A. C. Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine POTENTIAL IMPLICATIONS FOR ACTIVE DEMETHYLATION OF CpG SITES. *J. Biol. Chem.* **286**, 35334-35338 (2011).
50 Inoue, A. & Zhang, Y. Replication-Dependent Loss of 5-Hydroxymethylcytosine in Mouse Preimplantation Embryos. *Science* **334**, 194-194 (2011).

51 Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010-1022 (2011).

52 Booth, M. J. *et al.* Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. *Science* **336**, 934-937 (2012).

53 Lu, X. *et al.* Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. *Cell Res.* **25**, 386-389 (2015).

54 Shen, L. *et al.* Genome-wide Analysis Reveals TET- and TDG-Dependent 5-Methylcytosine Oxidation Dynamics. *Cell* **153**, 692-706 (2013).

55 Song, C.-X. *et al.* Genome-wide Profiling of 5-Formylcytosine Reveals Its Roles in Epigenetic Priming. *Cell* **153**, 678-691 (2013).

56 Wu, H., Wu, X., Shen, L. & Zhang, Y. Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. *Nat. Biotechnol.* **32**, 1231-1240 (2014).

57 Xia, B. *et al.* Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale. *Nat. Methods* **12**, 1047-1050 (2015).

58 Yu, M. *et al.* Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome. *Cell* **149**, 1368-1380 (2012).

59 Shen, L., Song, C.-X., He, C. & Zhang, Y. Mechanism and Function of Oxidative Reversal of DNA and RNA Methylation. *Annu. Rev. Biochem.* **83**, 585-614 (2014).

60 Inoue, A., Shen, L., Dai, Q., He, C. & Zhang, Y. Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. *Cell Res.* **21**, 1670-1676 (2011).

61 Lu, X., Zhao, B. S. & He, C. TET Family Proteins: Oxidation Activity, Interacting Molecules, and Functions in Diseases. *Chem. Rev.* **115**, 2225-2239 (2015).

62 Smith, Z. D. *et al.* A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature* **484**, 339-344 (2012).

63 James, C. *et al.* The hematopoietic stem cell compartment of JAK2V617F-positive myeloproliferative disorders is a reflection of disease heterogeneity. *Blood* **112**, 2429-2438 (2008).

64 Ko, M. *et al.* Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **468**, 839-843 (2010).

65 Lian, Christine G. *et al.* Loss of 5-Hydroxymethylcytosine Is an Epigenetic Hallmark of Melanoma. *Cell* **150**, 1135-1146 (2012).

66 Liu, C. *et al.* Decrease of 5-Hydroxymethylcytosine Is Associated with Progression of Hepatocellular Carcinoma through Downregulation of TET1. *PLoS ONE* **8**, e62828 (2013).

67 Ratel, D., Ravanat, J.-L., Berger, F. & Wion, D. $N^6$-methyladenine: the other methylated base of DNA. *BioEssays* **28**, 309-315 (2006).

68 Campbell, J. L. & Kleckner, N. E. coli *oriC* and the *dnaA* gene promoter are sequestered from *dam* methyltransferase following the passage of the chromosomal replication fork. *Cell* **62**, 967-979 (1990).

69 Collier, J., McAdams, H. H. & Shapiro, L. A DNA methylation ratchet governs progression through a bacterial cell cycle. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 17111-17116 (2007).

70 Low, D. A., Weyand, N. J. & Mahan, M. J. Roles of DNA adenine methylation in regulating bacterial gene expression and virulence. *Infect. Immu.* **69**, 7197-7204 (2001).
71 Lu, M., Campbell, J. L., Boye, E. & Kleckner, N. SeqA: a negative modulator of replication initiation in E. coli. *Cell* **77**, 413-426 (1994).
72 Messer, W. & Noyer-Weidner, M. Timing and targeting: the biological functions of Dam methylation in E. coli. *Cell* **54**, 735-737 (1988).
73 Ogden, G. B., Pratt, M. J. & Schaechter, M. The Replicative Origin of the *Escherichia Coli* Chromosome Binds to Cell Membranes Only When Hemimethylated. *Cell* **54**, 127-135 (1988).
74 Murray, N. E. 2001 Fred Griffith review lecture. Immigration control of DNA in bacteria: self versus non-self. *Microbiology* **148**, 3-20 (2002).
75 Fang, G. *et al.* Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat. Biotechnol.* **30**, 1232-1239 (2012).
76 Arnold, H. P., Ziese, U. & Zillig, W. SNDV, a Novel Virus of the Extremely Thermophilic and Acidophilic *Archaeon Sulfolobus*. *Virology* **272**, 409-416 (2000).
77 Baranyi, U., Klein, R., Lubitz, W., Krüger, D. H. & Witte, A. The archaeal halophilic virus-encoded Dam-like methyltransferase M.φCh1-I methylates adenine residues and complements *dam* mutants in the low salt environment of *Escherichia coli*. *Mol. Microbiol.* **35**, 1168-1179 (2000).
78 Magrini, V. *et al.* Temperate *Myxococcus xanthus* phage Mx8 encodes a DNA adenine methylase, Mox. *J. Bacteriol.* **179**, 4254-4263 (1997).
79 Schlagman, S. L. & Hattman, S. Molecular cloning of a functional *dam*$^+$ gene coding for phage T4 DNA adenine methylase. *Gene* **22**, 139-156 (1983).
80 Ehrlich, M., Wilson, G. G., Kuo, K. C. & Gehrke, C. W. $N^4$-methylcytosine as a minor base in bacterial DNA. *J. Bacteriol.* **169**, 939-943 (1987).
81 Vanyushin, B. F., Belozersky, A. N., Kokurina, N. A. & Kadirova, D. X. 5-Methylcytosine and 6-Methylaminopurine in Bacterial DNA. *Nature* **218**, 1066-1067 (1968).
82 Kahramanoglou, C. *et al.* Genomics of DNA cytosine methylation in *Escherichia coli* reveals its role in stationary phase transcription. *Nat Commun* **3**, 886 (2012).
83 Yu, M. *et al.* Base-resolution detection of N4-methylcytosine in genomic DNA using 4mC-Tet-assisted-bisulfite- sequencing. *Nucleic Acids Res.* gkv738 (2015).
84 Cummings, D. J., Tait, A. & Goddard, J. M. Methylated bases in DNA from *Paramecium aurelia*. *Biochim. Biophys. Acta* **374**, 1-11 (1974).
85 Harrison, G. S., Findly, R. C. & Karrer, K. M. Site-specific methylation of adenine in the nuclear genome of a eucaryote, *Tetrahymena thermophila*. *Mol. Cell. Biol.* **6**, 2364-2370 (1986).
86 Hattman, S., Kenny, C., Berger, L. & Pratt, K. Comparative study of DNA methylation in three unicellular eucaryotes. *J. Bacteriol.* **135**, 1156-1157 (1978).
87 Rae, P. M. & Spear, B. B. Macronuclear DNA of the hypotrichous ciliate *Oxytricha fallax*. *Proc. Natl. Acad. Sci. U.S.A.* **75**, 4992-4996 (1978).
88 Ehrlich, M. & Zhang, X.-Y. in *J. Chromatogr. Libr.* Vol. Volume 45, Part B (eds W. Gehrke Charles & C. T. Kuo Kenneth) B327-B362 (Elsevier, 1990).

89 Fu, Y. *et al.* $N^6$-Methyldeoxyadenosine Marks Active Transcription Start Sites in *Chlamydomonas*. *Cell* **161**, 879-892 (2015).
90 Greer, Eric L. *et al.* DNA Methylation on $N^6$-Adenine in *C. elegans*. *Cell* **161**, 868-878 (2015).
91 Zhang, G. *et al.* $N^6$-Methyladenine DNA Modification in *Drosophila*. *Cell* **161**, 893-906 (2015).
92 Desrosiers, R., Friderici, K. & Rottman, F. Identification of Methylated Nucleosides in Messenger RNA from Novikoff Hepatoma Cells. *Proc. Natl. Acad. Sci. U.S.A.* **71**, 3971-3975 (1974).
93 Desrosiers, R. C., Friderici, K. H. & Rottman, F. M. Characterization of Novikoff hepatoma mRNA methylation and heterogeneity in the methylated 5' terminus. *Biochemistry* **14**, 4367-4374 (1975).
94 Dubin, D. T. & Taylor, R. H. The methylation state of poly A-containing messenger RNA from cultured hamster cells. *Nucleic Acids Res.* **2**, 1653-1668 (1975).
95 Perry, R. P. & Kelley, D. E. Existence of methylated messenger RNA in mouse L cells. *Cell* **1**, 37-42 (1974).
96 Wei, C. M. & Moss, B. Methylation of Newly Synthesized Viral Messenger-Rna by an Enzyme in Vaccinia Virus. *Proc. Natl. Acad. Sci. U.S.A.* **71**, 3014-3018 (1974).
97 Perry, R. P., Kelley, D. E., Friderici, K. & Rottman, F. The methylated constituents of L cell messenger RNA: Evidence for an unusual cluster at the 5' terminus. *Cell* **4**, 387-394 (1975).
98 Wei, C. M., Gershowitz, A. & Moss, B. Methylated Nucleotides Block 5' Terminus of Hela Cell Messenger RNA. *Cell* **4**, 379-386 (1975).
99 Carroll, S. M., Narayan, P. & Rottman, F. M. N6-methyladenosine residues in an intron-specific region of prolactin pre-mRNA. *Mol. Cell. Biol.* **10**, 4456-4465 (1990).
100 Harper, J. E., Miceli, S. M., Roberts, R. J. & Manley, J. L. Sequence specificity of the human mRNA $N^6$-adenosine methylase *in vitro*. *Nucleic Acids Res.* **18**, 5735-5741 (1990).
101 Horowitz, S., Horowitz, A., Nilsen, T. W., Munns, T. W. & Rottman, F. M. Mapping of N6-methyladenosine residues in bovine prolactin mRNA. *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5667-5671 (1984).
102 Kane, S. E. & Beemon, K. Precise Localization of $m^6A$ in Rous Sarcoma Virus RNA Reveals Clustering of Methylation Sites: Implications for RNA Processing. *Mol. Cell. Biol.* **5**, 2298-2306 (1985).
103 Schibler, U., Kelley, D. E. & Perry, R. P. Comparison of methylated sequences in messenger RNA and heterogeneous nuclear RNA from mouse L cells. *J. Mol. Biol.* **115**, 695-714 (1977).
104 Wei, C., Gershowitz, A. & Moss, B. 5'-Terminal and internal methylated nucleotide sequences in HeLa cell mRNA. *Biochemistry* **15**, 397 - 401 (1976).
105 Wei, C.-M. & Moss, B. Nucleotide sequences at the $N^6$-methyladenosine sites of HeLa cell messenger ribonucleic acid. *Biochemistry* **16**, 1672 - 1676 (1977).
106 He, C. Grand challenge commentary: RNA epigenetics? *Nat. Chem. Biol.* **6**, 863-865 (2010).
107 Bokar, J. A. in *Fine-tuning of RNA functions by modification and editing* 141-177 (Springer, 2005).

108 Bokar, J. A., Rath-Shambaugh, M. E., Ludwiczak, R., Narayan, P. & Rottman, F. Characterization and partial purification of mRNA $N^6$-adenosine methyltransferase from HeLa cell nuclei. Internal mRNA methylation requires a multisubunit complex. *J. Biol. Chem.* **269**, 17697-17704 (1994).

109 Bokar, J. A., Shambaugh, M. E., Polayes, D., Matera, A. G. & Rottman, F. M. Purification and cDNA cloning of the AdoMet-binding subunit of the human mRNA ($N^6$-adenosine)-methyltransferase. *RNA* **3**, 1233-1247 (1997).

110 Liu, J. *et al.* A METTL3-METTL14 complex mediates mammalian nuclear RNA *N*$^6$-adenosine methylation. *Nat. Chem. Biol.* **10**, 93-95 (2014).

111 Ping, X.-L. *et al.* Mammalian WTAP is a regulatory subunit of the RNA $N^6$-methyladenosine methyltransferase. *Cell Res.* **24**, 177-189 (2014).

112 Wang, Y. *et al. N*$^6$-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat. Cell Biol.* **16**, 191-198 (2014).

113 Batista, Pedro J. *et al.* $m^6A$ RNA Modification Controls Cell Fate Transition in Mammalian Embryonic Stem Cells. *Cell Stem Cell* **15**, 707-719 (2014).

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

115 Clancy, M. J., Shambaugh, M. E., Timpte, C. S. & Bokar, J. A. Induction of sporulation in *Saccharomyces cerevisiae* leads to the formation of *N*$^6$-methyladenosine in mRNA: a potential mechanism for the activity of the IME4 gene. *Nucleic Acids Res.* **30**, 4509-4518 (2002).

116 Dominissini, D. *et al.* Topology of the human and mouse $m^6A$ RNA methylomes revealed by $m^6A$-seq. *Nature* **485**, 201-206 (2012).

117 Geula, S. *et al.* $m^6A$ mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science* **347**, 1002-1006 (2015).

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

119 Schwartz, S. *et al.* High-Resolution Mapping Reveals a Conserved, Widespread, Dynamic mRNA Methylation Program in Yeast Meiosis. *Cell* **155**, 1409-1421 (2013).

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

121 Chen, T. *et al.* $m^6A$ RNA Methylation Is Regulated by MicroRNAs and Promotes Reprogramming to Pluripotency. *Cell Stem Cell* **16**, 289-301 (2015).

122 Jia, G. *et al. N*$^6$-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat. Chem. Biol.* **7**, 885-887 (2011).

123 Zheng, G. *et al.* ALKBH5 Is a Mammalian RNA Demethylase that Impacts RNA Metabolism and Mouse Fertility. *Mol. Cell* **49**, 18-29 (2013).

124 Boissel, S. *et al.* Loss-of-Function Mutation in the Dioxygenase-Encoding FTO Gene Causes Severe Growth Retardation and Multiple Malformations. *Am. J. Hum. Genet.* **85**, 106-111 (2009).

125 Claussnitzer, M. *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N. Engl. J. Med.* **373**, 895-907 (2015).

126 Fischer, J. *et al.* Inactivation of the *Fto* gene protects from obesity. *Nature* **458**, 894-898 (2009).
127 Wang, X. *et al.* $N^6$-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505**, 117-120 (2014).
128 Wang, X. *et al.* $N^6$-methyladenosine Modulates Messenger RNA Translation Efficiency. *Cell* **161**, 1388-1399 (2015).
129 Liu, N. *et al.* $N^6$-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature* **518**, 560-564 (2015).
130 Spitale, R. C. *et al.* Structural imprints in vivo decode RNA regulatory mechanisms. *Nature* **519**, 486-490 (2015).
131 Alarcón, Claudio R. *et al.* HNRNPA2B1 Is a Mediator of $m^6A$-Dependent Nuclear RNA Processing Events. *Cell* **162**, 1299-1308 (2015).
132 Alarcón, C. R., Lee, H., Goodarzi, H., Halberg, N. & Tavazoie, S. F. $N^6$-methyladenosine marks primary microRNAs for processing. *Nature* **519**, 482-485 (2015).
133 Fustin, J.-M. *et al.* RNA-Methylation-Dependent RNA Processing Controls the Speed of the Circadian Clock. *Cell* **155**, 793-806 (2013).
134 Deng, X. *et al.* Widespread occurrence of $N^6$-methyladenosine in bacterial mRNA. *Nucleic Acids Res.* gkv596 (2015).
135 Luo, G.-Z. *et al.* Unique features of the $m^6A$ methylome in *Arabidopsis thaliana*. *Nat. Commun.* **5** (2014).
136 Meyer, K. *et al.* Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3' UTRs and near Stop Codons. *Cell* **149**, 1635-1646 (2012).
137 Schwartz, S. *et al.* Perturbation of m6A Writers Reveals Two Distinct Classes of mRNA Methylation at Internal and 5' Sites. *Cell Rep.* **8**, 284-296 (2014).

## 2 Photo-Crosslinking-Assisted m$^6$A Sequencing (PA-m$^6$A-seq)

### 2.1 Introduction

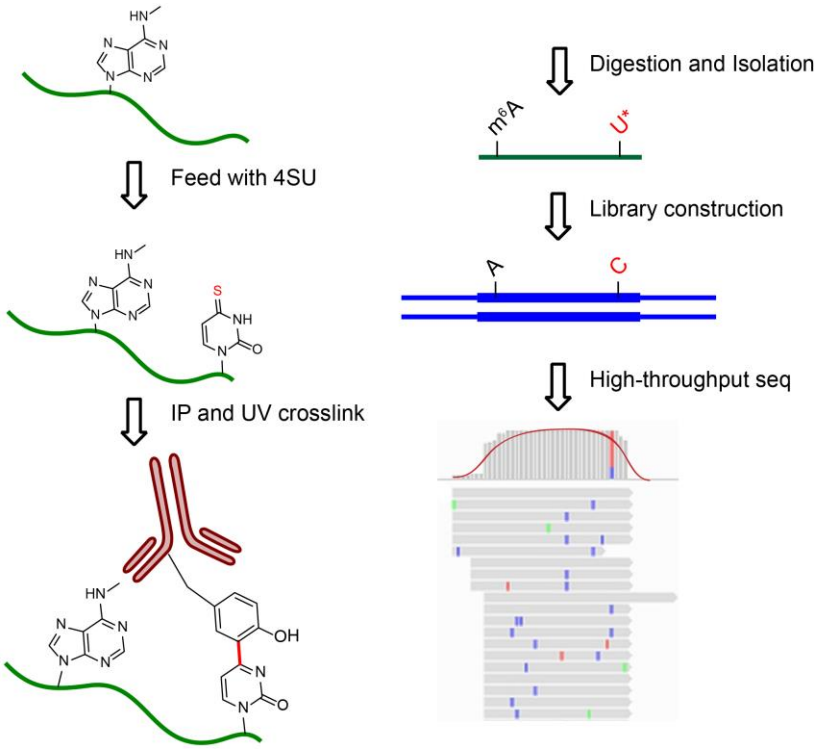
Precise knowledge of m$^6$A locations within the mammalian transcriptome is essential to understanding its biological function. The recently developed high-throughput method, termed m$^6$A-seq or MeRIP-seq (m$^6$A-specific methylated RNA immunoprecipitation with next-generation sequencing), utilizes anti-m$^6$A antibodies for the capture and enrichment of the m$^6$A-containing RNA fragments, followed by high-through sequencing to profile m$^6$A distributions in mammalian transcriptomes. This modification was shown to accumulate at 3'-UTR around stop codons and within exons. The resolution of these maps hovers around 200 nt and therefore cannot pinpoint the precise locations of the m$^6$A.[1,2] A higher resolution map of yeast m$^6$A methylome has been generated with an improved approach of m$^6$A-seq using shorter fragments to identify m$^6$A sites.[3] A ligation-based detection and SCARLET (site-specific cleavage and radioactive-labeling followed by ligation-assisted extraction and TLC) were also developed in order to precisely determine methylation sites with single-nucleotide resolution.[4,5] The SCARLET method, based on site-specific RNase H or DNAzyme cleavage, is effective but also time consuming, and is not yet feasible for high throughput applications.[6-8]

Photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) is a photo-crosslinking-based method to identify binding sites of RNA-binding proteins with high resolution.[9] A photoactivatable ribonucleoside, 4-thiouridine (4SU) or 6-thioguanosine (6SG), is incorporated into messenger RNA and covalently crosslinks with nearby aromatic amino acid residues in RNA-binding proteins upon 365 nm UV irradiation. Inspired by PAR-CLIP, we applied a similar approach, named photo-crosslinking-assisted m$^6$A-sequencing (PA-m$^6$A-

seq), which efficiently improves the accuracy of the methylation site assignments, and provides a high-resolution transcriptome-wide mammalian m$^6$A map (~23 nt).

## 2.2 Result and discussion

The photo-crosslinking-assisted m$^6$A-seq strategy is shown in **Scheme 2.1**.[1,9,10] 4SU, in which oxygen at 4' position is substituted by sulfur, forming thioketone structure. The effect of substitution of sulfur, similar to the effect of substitution of bromine in 5-bromouridine, significantly decreases the bond dissociation energy, facilitating the homolysis of carbon-sulfur bond and the formation of radical. The rearrangement and deprotonation of 4-thiouridine lead to the T-to-C transition then the base-pair reading change in PCR step.[11-13]




**Scheme 2.1. The strategy scheme of photo-crosslinking-assisted m$^6$A-seq (PA-m$^6$A-seq).** Covalently crosslinked 4SU is labelled as U*, which is read as C in RT-PCR. The example of high-throughput sequencing result is shown on the bottom right. Blue vertical bars represent T-to-C transition induced by 4SU and covalent crosslinking, compared to reference genome hg19.

### 2.2.1 Validation of PA-m$^6$A-seq strategy

The specificity and immunoprecipitation capability of anti-m$^6$A antibody have been well documented in previously published work.[14-21] However, it is still necessary to confirm that crosslinking comes from the specific recognition of m$^6$A by the antibody. Two parallel immunoprecipitation reactions were established, one with anti-m$^6$A polyclonal rabbit IgG, the other with normal rabbit IgG. With the same treatment, only the anti-m$^6$A antibody afforded visible radioactive signals, demonstrating specific m$^6$A recognition by the selected antibody is critical for crosslinking (**Figure 2.1A**).

To further confirm that 365 nm UV irradiation triggers 4SU-based crosslinking, a 21-mer RNA oligonucleotide containing 4SU and m$^6$A was synthesized as a model substrate. Along with the irradiation, the radioactive signal, representing the crosslinked complex, also appeared, confirming that the 4SU-mediated crosslinking is light-dependent (**Figure 2.1B**).

Furthermore, crosslinked 21-mer RNA oligonucleotide was isolated and inserted into vector for Sanger sequencing, showing that the covalent crosslinking between 4SU and antibody is indeed able to introduce a T-to-C transition nearby m$^6$A, as in the *in vivo* PAR-CLIP (see Experimental section). The T-to-C transition in model work ensures the reliability of this strategy and allows for the use of PAR-CLIP algorithms to analyze PA-m$^6$A-seq data (**Figure 2.1C**). With the model work above demonstrating the effectiveness of the strategy, we proceeded with two biological replicates of 4SU-incorporated HeLa cell mRNA for high-throughput sequencing.

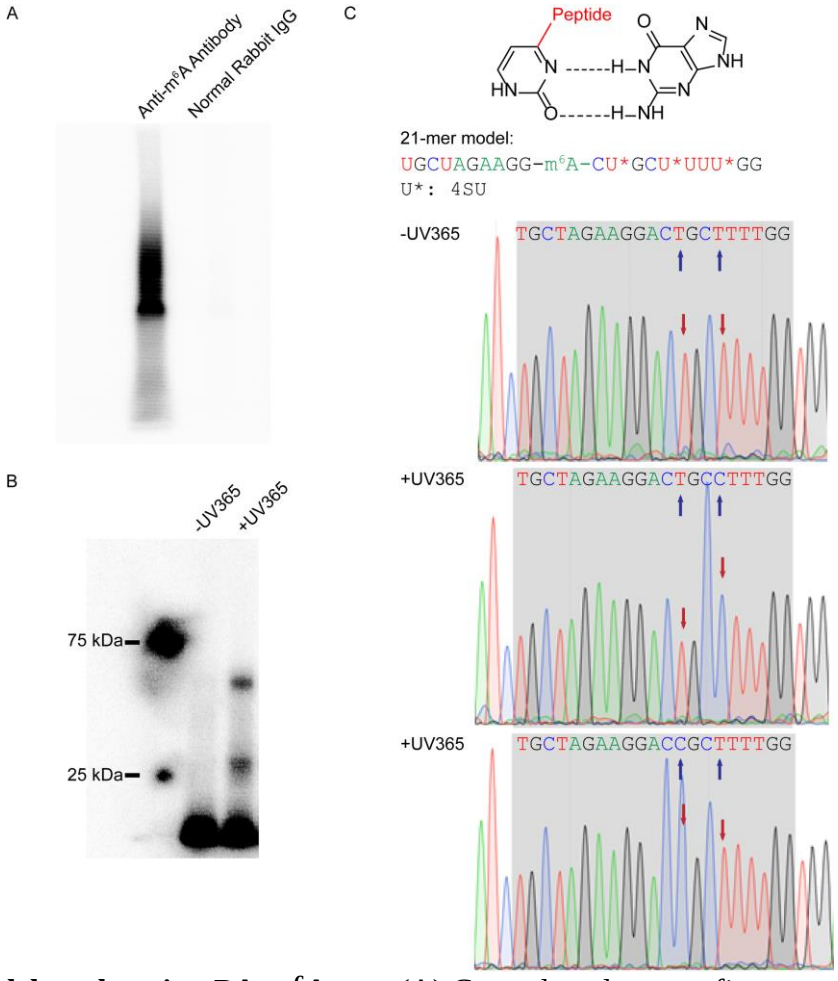



**Figure 2.1. Model study using PA-m$^6$A-seq.** (A) Control study to confirm crosslinking is based on recognition of antibody against m$^6$A-containing RNA. (B) Control study to prove UV365 is the trigger of covalent crosslinking by using a synthesized 21-mer RNA oligonucleotide. (C) T-to-C transition introduced by 4SU and UV irradiation. Proposed mechanism on how the 4SU crosslinked with protein changes Watson-Crick base pair (top). The sequence of 21-mer model is shown in the middle. Sanger sequencing results of 21-mer model with and without UV365 proved the crosslinked 4SU near m$^6$A was read as C. The blue arrows indicate the 4SU sites on the model, while red arrows point the chromatogram reads of corresponding 4SU with or without crosslinking.

### 2.2.2 HeLa transcriptome-wide mapping by using PA-m$^6$A-seq

The libraries were constructed following the procedure shown in **Scheme 2.1**, and were subjected to high-throughput sequencing. We identified 13,486 m$^6$A peaks within the human transcriptome, with an average length of 23 nt; much shorter than previous published results (**Figure**

**2.2**).[1,2] We further classified our reads into five different segments: 5' untranslated region (5'UTR), coding DNA sequence (CDS), 3'UTR, intergenic region, and intronic region. The distribution of our data confirmed that m$^6$A is significantly enriched near the stop codon and mainly localized in CDS and 3'UTR, consistent with previously published results (**Figure 2.3A**).[1,2]

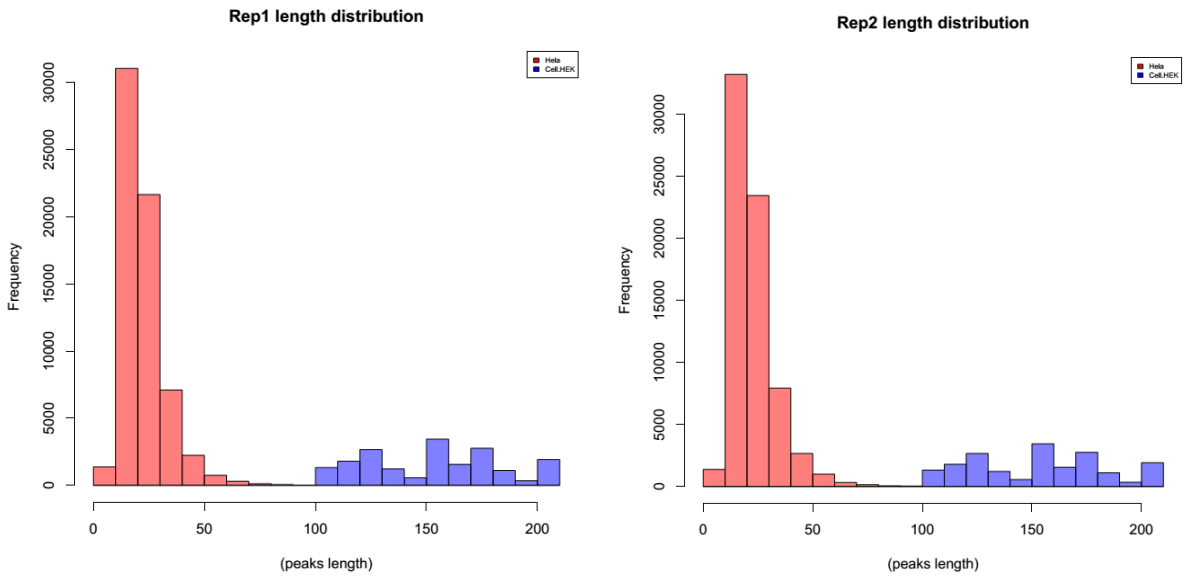



**Figure 2.2. The peak length distribution of PA-m$^6$A-seq vs MeRIP-seq.** Red histogram represents PA-m$^6$A-seq, blue representing MeRIP-seq.

The consensus sequence for the m$^6$A modification is known as $R_1R_2$-m$^6$A-CH (R = G or A; H = A or C or U; $R_2$: G > A).[16,22-27] The unbiased motif search based on two sets of high-throughput data uncovered a similar consensus. Here, we used the HOMER motif discovery tool to analyze the possible consensus sequence of m$^6$A.[28] Indeed, GGACU is the most enriched motif in our data (**Figure 2.3A**).

### 2.2.3 Comparison between PA-m$^6$A-seq and normal m$^6$A-seq

After validation, we performed further comparison and analyses with our higher resolution map. The methylation sites identified by PA-m$^6$A-seq can be confirmed by SCARLET and m$^6$A-seq/MeRIP-seq. Blue horizontal bars and blue peaks were chosen to represent PA-m$^6$A-seq and

original m$^6$A-seq peaks, respectively. The SCARLET-identified sites were emphasized in red, showing in the same figures. For example, methylation sites on β-actin mRNA (ACTB) and homo sapien basigin mRNA (BSG) were previously identified by m$^6$A-seq and precisely detected using SCARLET.[1,5] We uploaded our new high-resolution map and compared its contents of methylation sites in ACTB and BSG transcripts to the published results, which showed that these SCARLET-identified methylated regions exist in a 30 nt region in PA-m$^6$A-seq map, while in around 200-nt region using m$^6$A-seq/MeRIP-seq (**Figure 2.3B**).

The higher resolution map also suggests a 'clustering' property of m$^6$A deposition on transcripts, which is similar to methylation of cytosine on genomic DNA. Multiple methylation sites were identified in transcripts such as MALAT1 by PA-m$^6$A-seq (**Figure 2.3C**). These single sites previously confirmed by using SCARLET were also discovered by our PA-m$^6$A-seq strategy.[5] Intriguingly, the multicity of methylation or the 'clustering' property of m$^6$A on transcripts implies the likelihood that these transcripts are highly affected by m$^6$A reader proteins as recently suggested, resembling DNA 5-methylcytosine methylation.[29]

The power of PA-m$^6$A-seq lies in its ability to identify single consensus methylation sequences within a ~23 base region, enabling single-base resolution detection of the m$^6$A modification. Hence, we used SCARLET as an independent approach to validate new methylation sites found in PA-m$^6$A-seq, which confirmed the ability of PA-m$^6$A-seq to pinpoint methylation site in a transcriptome-wide manner (see Experimental section).

Next, we analysed the spatial relationship between the methylation sites and the binding sites of two RNA-binding proteins shown to recognize m$^6$A. Wang *et al* proved that YTHDF2 is a selective binder/reader of m$^6$A, and identified over 3,000 RNA targets using PAR-CLIP.[29] The high resolution methylation sites revealed by PA-m$^6$A-seq overlap very well with the binding

sites of YTHDF2. The high resolution map obtained in this study allowed us to conclude that most YTHDF2-binding sites are within 30-50 nt of the m$^6$A sites, strongly supporting direct interactions of YTHDF2 with m$^6$A and the regulatory role of m$^6$A in YTHDF2-mediated RNA decay (**Figure 2.3D**).

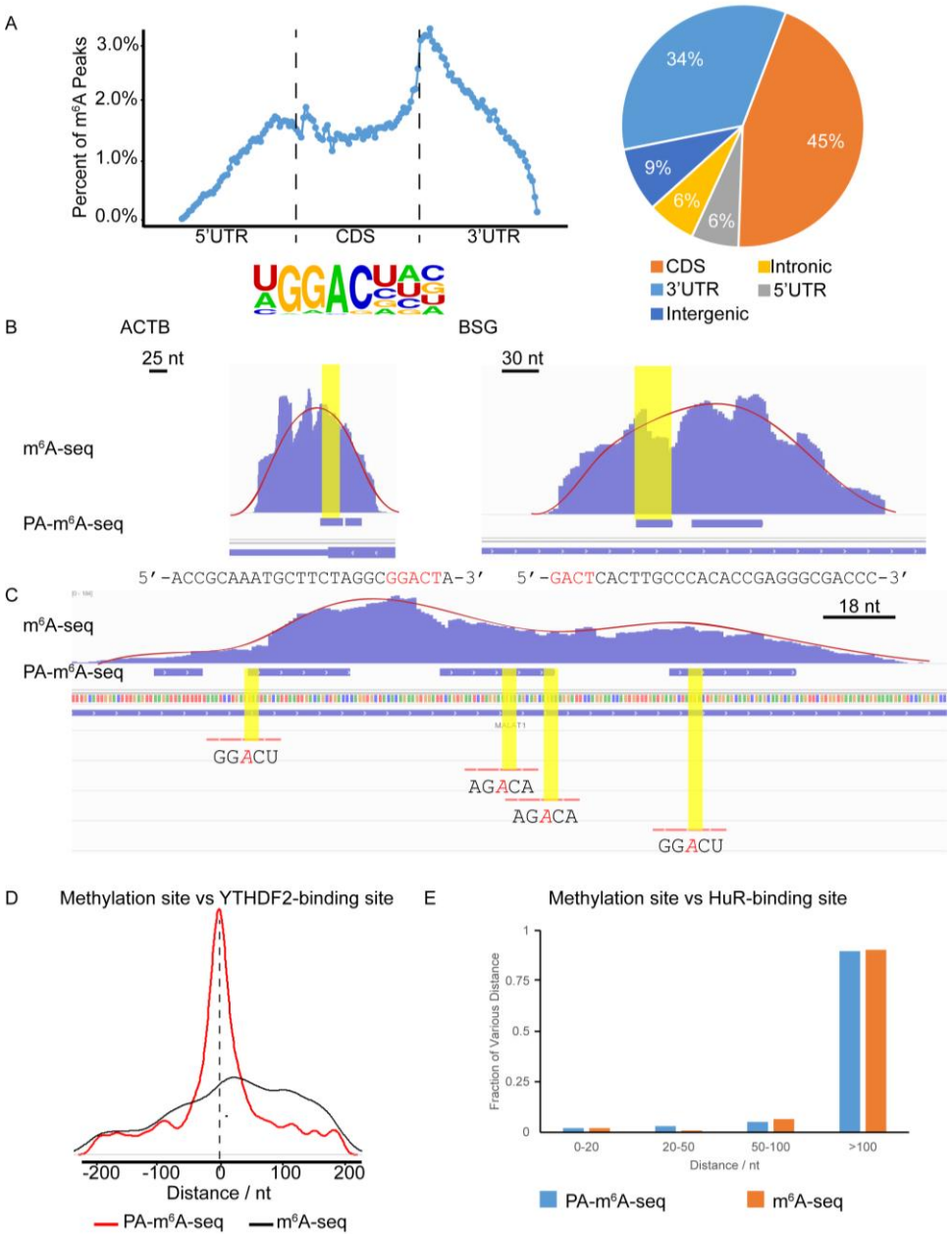



**Figure 2.3. PA-m$^6$A-seq applied to polyA-tailed RNA purified from HeLa cells.** In following figures, blue bars represent methylation sites identified by PA-m$^6$A-seq, while blue 'peaks' above those bars are from normal m$^6$A-seq. (A) Validation of PA-m$^6$A-seq strategy. Metagene

(*Continued*)
profile and pie-chart of the enrichment of RNA segments are consistent with previous reported distribution of $m^6A$, and the motif search yielded GGACU as the predominant one, which was the same as the result from normal $m^6A$-seq. (B) Comparison of predicted methylation sites in β-actin (ACTB) and homo sapien basigin (BSG) from PA-$m^6A$-seq with peaks from normal $m^6A$-seq and single sites by SCARLET. Sequences of predicted sites are shown below, consensus motif containing $m^6A$ in red. All input background of normal $m^6A$-seq has been subtracted. Both sites were verified by SCARLET. (C) Multiple methylation sites in MALAT1 transcript. The methylation sites confirmed by SCARLET, which were covered by peak from normal $m^6A$-seq, were also identified by PA-$m^6A$-seq with higher resolution. Yellow regions indicate the RRACH motif. Red lines are the probes used to verify these sites with SCARLET. (D) Distance of predicted methylation sites using PA-$m^6A$-seq vs peaks obtained from normal $m^6A$-seq to YTHDF2-binding sites. (E) Distance of predicted methylation sites obtained from PA-$m^6A$-seq vs peaks obtained from normal $m^6A$-seq to HuR-binding sites.

Together with YTHDF2, another well-studied RNA-binding protein, HuR (ELAVL1), was also pulled down using an $m^6A$-containing bait.[1] In contrast with YTHDF2, the consensus sequence recognized by HuR is very different from that of the $m^6A$ site.[30] A recent work evaluating binding of HuR to various probes containing $m^6A$ suggested interesting spatial constraints that affect potential interactions between HuR and $m^6A$.[31] We attempted to further probe the binding sites of HuR and the high resolution $m^6A$ sites on mRNA by applying the same analysis shown above for YTHDF2, and plotting the distribution of distances between the high-resolution methylation sites and HuR-binding sites. The results of both PA-$m^6A$-seq and normal $m^6A$-seq analyses indicate that the majority of the HuR-binding sites are further away (100 nt) from the $m^6A$ site (**Figure 2.3E**). This analysis suggests that HuR may "indirectly" (through other proteins or mRNA structure change) interact with $m^6A$ if it associates with $m^6A$.

### 2.2.4 Discussion and summary

In summary, we have established a photo-crosslinking-assisted strategy to improve the resolution of $m^6A$-seq/MeRIP-seq, and report a high-resolution methylation map of the mammalian transcriptome. The covalent crosslinking and effective RNase T1 digestion resulted in a signifi-

cant resolution improvement on localization of $m^6A$ on fragmented RNA. On the whole, the general analyses validated our strategy, and the higher resolution map showed the ability to reveal new methylation sites at base resolution. Furthermore, more precise methylation location made it possible to explore the distance between $m^6A$ peaks and the binding sites of RNA-binding proteins. These analyses confirm that YTHDF2 directly binds $m^6A$, whereas direct interaction of $m^6A$ by HuR is less likely, which suggests HuR may mediate the role of $m^6A$ indirectly. The PA-$m^6A$-seq approach presented here can be widely applied to study precise $m^6A$ sites in various organisms and the UV-crosslinking strategy is compatible with investigation on nucleic acid-protein interactions.

## 2.3 Experimental section

### 2.3.1 Preparation of 4-thiouridine incorporated polyA-tailed RNA

Culture cells in an appropriate growth medium supplemented with appropriate supplies, if needed. When the confluence of cells is around 80%, add 1 M 4-thiouridine DMSO stock to cell culture medium directly to a final concentration of 200 μM. Shake the plates gently in order to expand the chemical. Incubate for 16 hrs in a cell incubator. Add 5 mL of TRIzol reagent directly to the cells after removing culture media. Homogenize adhesive cells in TRIzol reagent and pool the lysate in a 50 mL tube. Extract total RNA following the TRIzol reagent protocol by adding chloroform for phase separation and then adding isopropanol for RNA precipitation. Dissolve total RNA in RNase-free water. Isolate polyA-tailed RNA from total RNA by using FastTrack mRNA isolation kit.

### 2.3.2 Anti-$m^6A$ immunoprecipitation and UV crosslinking

Set up 600 μL anti-$m^6A$ immunoprecipitation reaction as following:

| | |
|---|---|
| 4-thiouridine incorporated polyA-tailed RNA | 12 μg |
| Anti-$m^6A$ polyclonal antibody | 10 μg |
| 5X IP Buffer (50 mM Tris-HCl, pH 7.4, 750 mM NaCl, 0.5% Igepal CA-630) | 120 μL |
| RNase Inhibitor (40X) | 15 μL |
| RNase-free $H_2O$ | to 600 μL |

Mix thoroughly but gently. Transfer to head-over-tail rotating wheel to incubate for 2 hrs at 4 ℃. During the reaction incubation, pre-block Dynabeads protein A magnetic slurry with 1X IP Buffer supplement with 0.5 mg/mL BSA and 1 U/μL RNase inhibitor. After 2 hrs incubation, split the reaction mix into 8-10 wells in 96-well cell culture plate. Keep the plate on ice and expose in UV 365 nm three times with energy dosage of 0.15 J/$cm^2$.

### 2.3.3 Enzymatic treatment on beads

Pool the crosslinked fractions together, add RNase T1 to a final concentration of 0.2 U/μL for the 1st round RNase T1 digestion. Incubate the reaction at 22 ℃ for 10 min. Quench the reaction on ice for 5 min. Transfer each reaction to 80 μL pre-blocked protein A slurry. Mix gently, and then incubate on head-over-tail rotating wheel for 1.5 hrs at 4 ℃. Then place the tube on magnetic rack, discard the supernatant and wash the beads with 500 μL ice-cold IP Wash Buffer (50 mM HEPES-KOH, pH 7.4, 300 mM KCl, 0.05% Igepal CA-630) three times by gentle vortex. Resuspend the beads in one volume of IP Wash Buffer, then conduct 2nd round RNase-T1 digestion with a final concentration of 20 U/μL. Incubate for 15 min at 22 ℃. Quench the reaction on ice for 5 min. Wash the beads with 500 μL ice-cold High Salt Wash Buffer (50 mM HEPES-KOH, pH 7.4, 500 mM KCl, 0.05% Igepal CA-630) by gentle vortex. Resuspend the beads in one volume of 1X Antarctic Phosphatase Buffer (NEB) containing 1 U/μL Antarctic Phosphatase. Incubate for 20 min at 37 ℃ to dephosphorylate the 5' terminal of the RNA frag-

ments. Wash the beads with 500 μL ice-cold Phosphatase Wash Buffer (50 mM Tris-HCl, pH 7.4, 20 mM EGTA, 0.5% Igepal CA-630) twice, 500 μL ice-cold PNK Wash Buffer (50 mM Tris-HCl, pH 7.4, 50 mM NaCl, 10 mM $MgCl_2$) twice. Resuspend the beads in one volume of 1X T4 PNK Buffer (NEB) containing 1 U/μL T4 PNK. Incubate for 30 min at 37 ºC first to remove 3' terminal phosphor group then add ATP to a final concentration of 1 mM and incubate for another 20 min at 37 ℃. Wash the beads with 500 μL ice-cold PNK Wash Buffer three times. To release the RNA fragment bound by antibody, resuspend the beads in one volume of 1X Proteinase K Buffer (50 mM Tris-HCl, pH 7.4, 75 mM NaCl, 6.25 mM EDTA, 1% SDS) containing 2 mg/mL Proteinase K. Incubate for 30 min at 55 ℃. Then transfer the supernatant (which contains RNA fragments) in a new tube. Repeat the proteinase K digestion again and combine the supernatant together.

### 2.3.4 RNA isolation and purification

Use equal volume of phenol chloroform to extract RNA fragments in supernatant. Wash the aqueous phase once more with equal volume of chloroform. Add 1/10 volume of 3 M sodium acetate (pH 5.3) and 4 μL glycogen. Then precipitate RNA fragments with 2.5 volume of cold pure ethanol. Store the mixture at -80 ℃ overnight. Spin down the RNA fragments, wash the pellet with 75% (v/v) ethanol, and dissolve the pellet into 50 μL RNase-free $H_2O$. Clean up by Zymo Research RNA Clean & Concentrator and elute in 10 μL RNase-free $H_2O$.

### 2.3.5 Library construction, high-throughput sequencing and data analysis

Apply the purified RNA to small RNA library construction. The library is purified by agarose gel size selection and extraction. Library quality control by Bioanalyzer and qPCR. Adapters were trimmed by FASTX and also the first and last bases were removed because of the sequencing quality. Then the reads were mapped to the human genome (hg19) using Bowtie with 2

mismatch at most.[32] PARalyzer was used for peak calling.[33] 23,880,658 and 21,456,137 reads were mapped to the genome separately for the each replicates and 22,396 and 25,509 peaks were called by PARalyzer. We took the 13,486 peaks shared by two replicates as a reliable peak sets for the following analysis. Refseq genes were used for peaks annotation. These peaks were mapped to 6,176 genes (13,499 isoforms).

The overlap analysis results are shown in **Figure 2.4**.

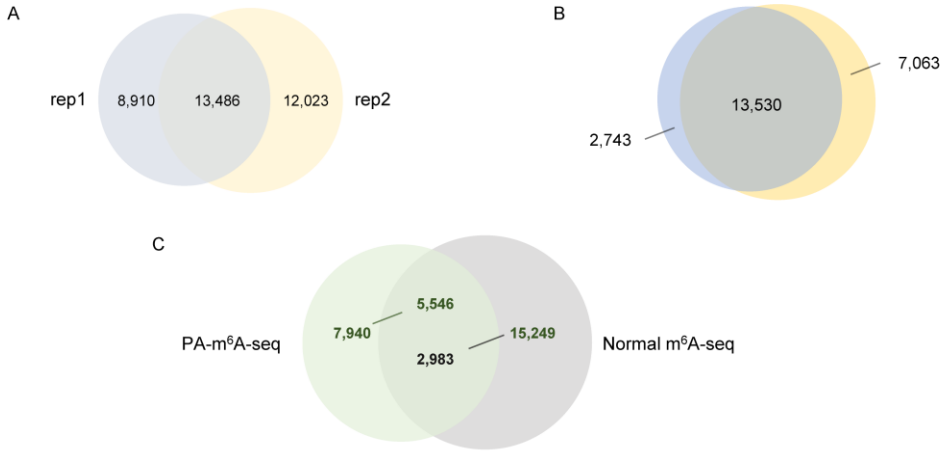



**Figure 2.4. The overlap analysis results.** (A) Venn diagram showing overlap of peaks between PA-m$^6$A-seq biological replicates. 22,396 and 25,509 peaks were called, respectively. Among them, 13,486 peaks were overlapped between two replicates. (B) Technical replicates of PA-m$^6$A-seq from same batch of HeLa 4SU-containing mRNA. 16,273 and 20,593 peaks were called, respectively. Among them, 13,530 peaks were overlapped between two replicates. (C) Venn diagram showing overlap of Peaks between PA-m$^6$A-seq in HeLa and normal m$^6$A-seq in HeLa. For PA-m$^6$A-seq, 5,546 peaks out of 13,486 were overlapped with normal m$^6$A-seq.

### 2.3.6 Model study

#### 2.3.6.1 Oligonucleotide synthesis

The oligonucleotides were synthesized on an Expedite nucleic acid synthesis system on a 1 μmole scale using 0.067 M acetonitrile solutions of phosphoramidites from Glen Research.

2.3.6.2 Oligonucleotide deprotection and purification

Half of the CPG-bound RNA oligonucleotide was transferred from the column to a screw cap glass vial, to which was added 2 mL of 1.0 M DBU, 1,8-Diazabicyclo[5.4.0]undec-7-ene, in anhydrous acetonitrile for 2 hours at RT. Acetonitrile was used to wash after DBU treatment to completely remove DBU and dry the support. Then CPG was treated by 1 mL of *tert*-butylamine: MeOH: $H_2O$ (1: 1: 2) containing 50 mM NaHS to deprotect for 3 hours at 55 ℃. The vial was cooled and the supernatant was collected and desalted on NAP-10 column to remove NaHS. The supernatant was then dried on a Speed-Vac concentrator. The dried material was dissolved in 115 µL DMSO first. Then 60 µL TEA was added to maintain a basic solution. 75 µL TEA•3HF was applied to the mixture to remove silyl protection groups. After incubation of 2.5 hours at 65 ℃, 1.75 mL quenching buffer (Tris buffer) was used to stop the deprotection reaction. The clear solution was applied to RNA purify cartridge (Glen Research) for further purification. The eluate was lyophilized and dissolved in RNase-free $H_2O$ and the oligonucleotide was qualified and analyzed by MALDI-TOF (**Figure 2.5**).

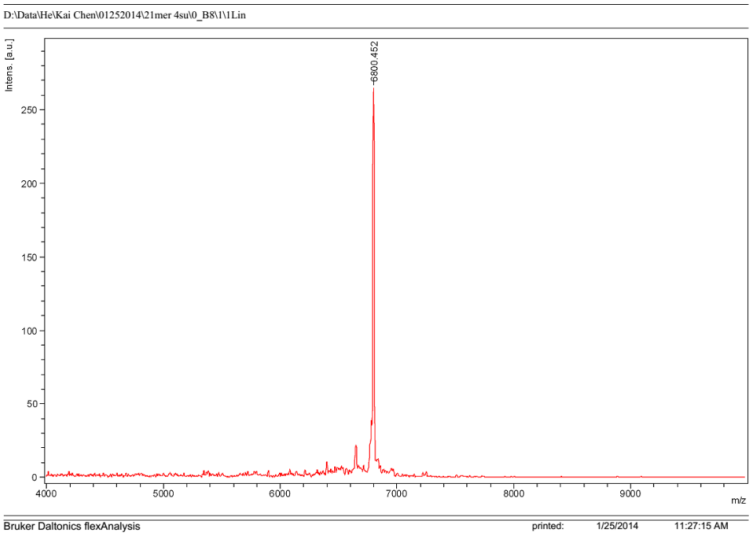



**Figure 2.5. MALDI-TOF of the synthesized 21-mer RNA oligonucleotide.** The sequence is: 5'-UGCUAGAAGG-$m^6$A-CU*GCU*UUU*GG-3'; U*: 4-thiouridine.

2.3.6.3 Insertion of oligonucleotide for Sanger sequencing

The model RNA IP assay was performed as described above. After isolated and purified by TRIzol extraction and ethanol precipitation, the pellet was dissolved in RNase-free $H_2O$. The 3' adapter and 5' adapter from Illumina TruSeq Small RNA Prep Kit were ligated to the 3' and 5' ends of the model using the Illumina protocol, respectively. Then, smallRNA-RT primer was used for reverse transcription catalyzed by SuperScript II reverse transcriptase (Life Technologies) to generate cDNA, which was amplified using Zymo *Taq* polymerase (Zymo Research) and a pair of smallRNA PCR primers. The sequences of smallRNA primers are shown below.

| | |
|---|---|
| KC_smallRNA-F | 5'-GAGTTCTACAGTCCGACGATC-3' |
| KC_smallRNA-RT | 5'-TTGGCACCCGAGAATTCCA-3' |
| KC_smallRNA-R | 5'-CCTTGGCACCCGAGAATTCCA-3' |

The PCR product was purified by QIAquick PCR purification kit and apply to TA TOPO-cloning kit (Life Technologies) following the manufacturer's instruction.

### 2.3.7 SCARLET assay

The SCARLET assay was performed exactly followed the published protocol.[5] The sequences of oligonucleotides used are shown below.

*PITX1-1 Chimera:* 5'-mGmUmGmGmGmGmUCCGGmAmAmAmAmGmCmA-3'

*PITX1-1 Splint:*

5'-TTTTGTCTTTTTGGAGGGCAGAGTGGGGTCTATTAACTCACAGGACCGGCGATGGCTG-3'

*YARS Chimera:* 5'-mAmAmAmGmAmGmUCCAAmAmCmCmCmGmCmU-3'

*YARS Splint:*

5'-CTTGCCGAGTTCTGCACCGAATAAAGAGTCTATTAACTCACAGGACCGGCGATGGCTG-3'

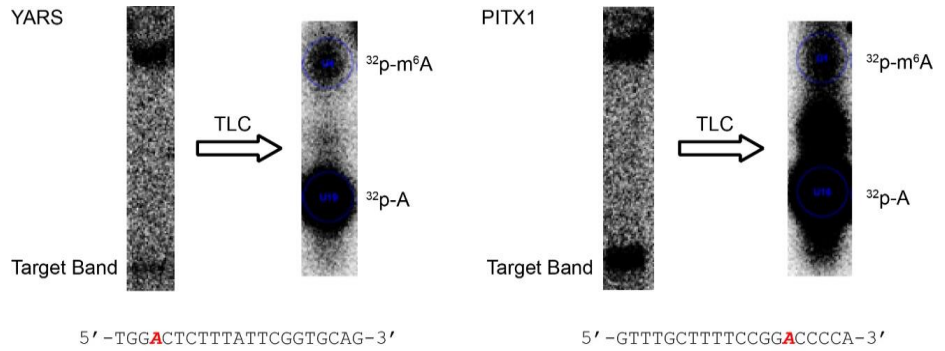



**Figure 2.6. The SCARLET results of methylation sites identified by PA-m$^6$A-seq.** The methylated A is shown in red.

## 2.4 References


1 Dominissini, D. *et al.* Topology of the human and mouse m$^6$A RNA methylomes revealed by m$^6$A-seq. *Nature* **485**, 201-206 (2012).

2 Meyer, K. *et al.* Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3' UTRs and near Stop Codons. *Cell* **149**, 1635-1646 (2012).

3 Schwartz, S. *et al.* High-Resolution Mapping Reveals a Conserved, Widespread, Dynamic mRNA Methylation Program in Yeast Meiosis. *Cell* **155**, 1409-1421 (2013).

4 Dai, Q. *et al.* Identification of recognition residues for ligation-based detection and quantitation of pseudouridine and *N*$^6$-methyladenosine. *Nucleic Acids Res.* **35**, 6322-6329 (2007).

5 Liu, N. *et al.* Probing *N*$^6$-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA. *RNA* **19**, 1848-1856 (2013).

6 Hengesbach, M., Meusburger, M., Lyko, F. & Helm, M. Use of DNAzymes for site-specific analysis of ribonucleotide modifications. *RNA* **14**, 180-187 (2008).

7 Yu, Y.-T., Shu, M. D. & Steitz, J. A. A new method for detecting sites of 2'-O-methylation in RNA molecules. *RNA* **3**, 324-331 (1997).

8 Zhao, X. & Yu, Y.-T. Detection and quantitation of RNA base modifications. *RNA* **10**, 996-1002 (2004).

9 Hafner, M. *et al.* Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell* **141**, 129-141 (2010).

10 Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N. & Rechavi, G. Transcriptome-wide mapping of *N*$^6$-methyladenosine by m$^6$A-seq based on immunocapturing and massively parallel sequencing. *Nat. Protoc.* **8**, 176-189 (2013).

11 Dietz, T. M. & Koch, T. H. Photochemical Coupling of 5-Bromouracil to Tryptophan, Tyrosine and Histidine, Peptide-Like Derivatives in Aqueous Fluid Solution. *Photochem. Photobiol.* **46**, 971-978 (1987).

12 Dietz, T. M., Vontrebra, R. J., Swanson, B. J. & Koch, T. H. Photochemical coupling of 5-bromouracil (BU) to a peptide linkage. A model for BU-DNA protein photocrosslinking. *J. Am. Chem. Soc.* **109**, 1793-1797 (1987).

13 Ito, S., Saito, I. & Matsuura, T. Acetone-sensitized photocoupling of 5-bromouridine to tryptophan derivatives via electron-transfer process. *J. Am. Chem. Soc.* **102**, 7535-7541 (1980).

14 Bringmann, P. & Lührmann, R. Antibodies specific for $N^6$-methyladenosine react with intact snRNPs U2 and U4/U6. *FEBS Lett.* **213**, 309-315 (1987).
15 Dante, R. & Niveleau, A. Inhibition of in vitro translation by antibodies directed against $N^6$-methyladenosine. *FEBS Lett.* **130**, 153-157 (1981).
16 Horowitz, S., Horowitz, A., Nilsen, T. W., Munns, T. W. & Rottman, F. M. Mapping of $N^6$-methyladenosine residues in bovine prolactin mRNA. *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5667-5671 (1984).
17 Jia, G. *et al.* $N^6$-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat. Chem. Biol.* **7**, 885-887 (2011).
18 Munns, T. W., Liszewski, M. K., Oberst, R. J. & Sims, H. F. Antibody Nucleic Acid Complexes. Immunospecific Retention of $N^6$-Methyladenosine-Containing Transfer Ribonucleic Acid. *Biochemistry* **17**, 2573-2578 (1978).
19 Munns, T. W., Liszewski, M. K. & Sims, H. F. Characterization of Antibodies Specific for $N^6$-Methyladenosine and for 7-Methylguanosine. *Biochemistry* **16**, 2163-2168 (1977).
20 Munns, T. W., Oberst, R. J., Sims, H. F. & Liszewski, M. K. Antibody-nucleic acid complexes. Immunospecific recognition of 7-methylguanine- and $N^6$-methyladenine-containing 5'-terminal oligonucleotides of mRNA. *J. Biol. Chem.* **254**, 4327-4330 (1979).
21 Munns, T. W., Sims, H. F. & Liszewski, M. K. Immunospecific Retention of Oligonucleotides Possessing $N^6$-Methyladenosine and 7-Methylguanosine. *J. Biol. Chem.* **252**, 3102-3104 (1977).
22 Canaani, D., Kahana, C., Lavi, S. & Groner, Y. Identification and mapping of $N^6$-methyladenosine containing sequences in Simian Virus 40 RNA. *Nucleic Acids Res.* **6**, 2879-2899 (1979).
23 Dimock, K. & Stoltzfus, C. M. Sequence specificity of internal methylation in B77 avian sarcoma virus RNA subunits. *Biochemistry* **16**, 471-478 (1977).
24 Harper, J. E., Miceli, S. M., Roberts, R. J. & Manley, J. L. Sequence specificity of the human mRNA N6-adenosine methylase *in vitro*. *Nucleic Acids Res.* **18**, 5735-5741 (1990).
25 Kane, S. E. & Beemon, K. Precise Localization of $m^6A$ in Rous Sarcoma Virus RNA Reveals Clustering of Methylation Sites: Implications for RNA Processing. *Mol. Cell. Biol.* **5**, 2298-2306 (1985).
26 Schibler, U., Kelley, D. E. & Perry, R. P. Comparison of methylated sequences in messenger RNA and heterogeneous nuclear RNA from mouse L cells. *J. Mol. Biol.* **115**, 695-714 (1977).
27 Wei, C.-M. & Moss, B. Nucleotide Sequences at the $N^6$-Methyladenosine Sites of HeLa Cell Messenger Ribonucleic Acid. *Biochemistry* **16**, 1672-1676 (1977).
28 Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime *cis*-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576-589 (2010).
29 Batista, Pedro J. *et al.* $m^6A$ RNA Modification Controls Cell Fate Transition in Mammalian Embryonic Stem Cells. *Cell Stem Cell* **15**, 707-719 (2014).
30 Lebedeva, S. *et al.* Transcriptome-wide Analysis of Regulatory Interactions of the RNA-Binding Protein HuR. *Mol. Cell* **43**, 340-352 (2011).
31 Wang, Y. *et al.* $N^6$-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat. Cell Biol.* **16**, 191-198 (2014).

32 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
33 Corcoran, D. L. *et al.* PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.* **12**, R79 (2011).

# 3 Mapping m$^6$A in bacterial mRNA by photo-crosslinking-assisted approach

## 3.1 Introduction

Although m$^6$A has been well documented in the rRNA in bacteria, its presence on mRNA is still elusive. In *Escherichia coli*, A1618 and A2030 of 23S rRNA are methylated by methyltransferases RlmF and RlmJ, respectively.[1,2] Both deletion and overexpression of *rlmF* result in a loss of cell fitness and growth defect, while an *rlmJ* mutant shows mild phenotypes under various growth conditions.[1,2] Interestingly, the modifications of m$^2$A or m$^8$A on A1607, A2503 and A2508 play important roles in antibiotic resistance, an extensively studied subject in microbiology during the last 10 years.[3]

In order to investigate the potential presence and functions of m$^6$A in bacterial mRNA, we calculated the m$^6$A/A ratio in mRNA from seven diverse bacterial species, which reveal that m$^6$A is an abundant mRNA modification in Gram-negative bacteria. High-resolution transcriptome wide m$^6$A profiling in two model bacteria *E. coli* and *Pseudomonas aeruginosa* reveal a conserved and distinct m$^6$A distribution pattern. Most m$^6$A-modified genes are involved in energy metabolism and small RNAs, suggesting potential functional roles of m$^6$A in these processes.

## 3.2 Result and discussion

### 3.2.1 m$^6$A is presented in mRNA of a wide range of bacterial species

Although m$^6$A is the most abundant internal mRNA modification in eukaryotes, its potential presence in the kingdom of bacteria has yet to be investigated. To this end, we selected seven diverse model bacterial species (*E. coli, P. aeruginosa, Pseudomonas syringae, Staphylococcus aureus, Bacillus subtilis, Anabaena* sp. PCC 7120 and *Synechocystis* sp. PCC 6803) to grow in a common laboratory environment and measured their m$^6$A/A ratios in purified mRNA. Unlike eukaryotes, bacterial mRNA lacks a polyA$^+$ tail, which makes it challenging to purify mRNA.

By following the protocols described in the experimental section, we were able to remove >90% rRNA in the purified mRNA sample (**Figure 3.1**).

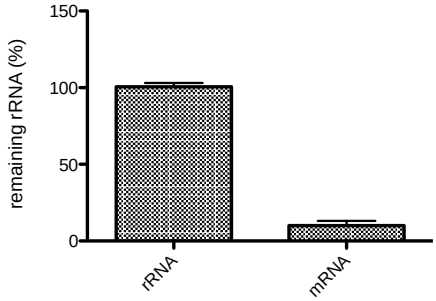



**Figure 3.1. qPCR verification of mRNA enrichment.** qPCR was performed against the rRNA (primers targeting 16S rRNA) background to check the relative enrichment level of the mRNA sample from the wild type *E. coli*.

We then tried to use an UHPLC-QQQ-MS/MS approach in order to quantify the $m^6A$/A level in the bacterial mRNA samples that contain residual rRNA (<10%). Given that two $m^6A$ (catalyzed by RlmF and RlmJ) and two $N^6$,$N^6$-dimethladenosine ($m^6{}_2A$, catalyzed by KsgA) are known to be present in rRNA of *E. coli* and other related bacterial species, the values of $m^6{}_2A$ levels can be used as an internal reference for the $m^6A$ level from the residue of rRNA presented in the purified mRNA. We first determined the $m^6{}_2A/m^6A$ ratio of rRNA as 1.30 in the wild-type strain and 2.04 in either an *rlmF* mutant or an *rlmJ* mutant.[1,2,4] As a negative control, the $m^6{}_2A$ modification was not detectable in a *ksgA* mutant (**Figure 3.2**). Based on the $m^6{}_2A/m^6A$ ratio in rRNA, we were able to accurately calculate the real $m^6A$/A level as ($m^6A$-$m^6{}_2A$/1.30)/A in the purified mRNA samples.

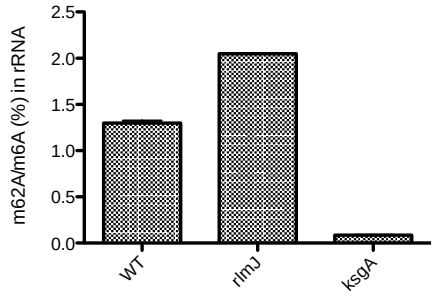



**Figure 3.2. The ratios of $m^6{}_2A/m^6A$ in rRNA from the wild type, the *rlmJ* mutant and the *ksgA* mutant of *E. coli.***

We observed the presence of $m^6A$ in all of the tested bacterial species, whose $m^6A/A$ ratio varied within the range of 0.02–0.28% (**Figure 3.3A**). We obtained the $m^6A/A$ ratios in mRNA from three Gram-negative bacteria (*E. coli*, *P. aeruginosa* and *P. syringae*) (>0.2%) and from two Gram-positive bacteria (*S. aureus* and *B. subtilis*) (<0.08%). Unlike *E. coli* and *Pseudomonas* spp., two other Gram-negative cyanobacteria (Anabaena sp. PCC 7120 and *Synechocystis* sp. PCC 6803) showed low $m^6A/A$ ratios (<0.04%). In order to test if $m^6A/A$ ratios also vary among different strains in the same species, three strains of *E. coli* (K-12, 5α and XL-blue), two strains of *P. aeruginosa* (MPAO1 and PA14) and six strains of *S. aureus* (Newman, USA100, USA400, USA700, RN4220 and COL) were tested, all of which revealed a constant ratio in the same species (**Figure 3.3B**). These results clearly demonstrate the widespread occurrence of $m^6A$ in bacterial mRNA. Gram-negative bacteria tend to have higher $m^6A/A$ ratios in mRNA than Gram-positive bacteria. The high $m^6A/A$ ratio (>0.2%) in mRNA from *E. coli* and *Pseudomonas* spp. is comparable to that from eukaryotes, suggesting that $m^6A$ could be an important mRNA modification playing functional roles in these and other bacteria.[5]

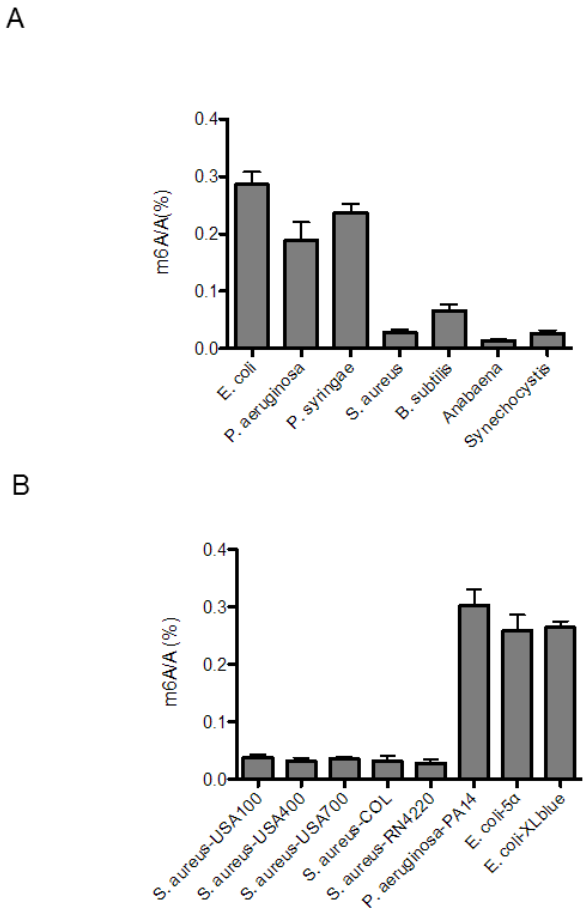



**Figure 3.3. Presence of m$^6$A in bacterial mRNA.** (A) The m$^6$A/A ratios of mRNA isolated from 7 bacterial species. (B) The m$^6$A/A ratios of mRNA isolated from different strains of *E. coli* and *P. aeruginosa*. Error bars represent standard deviations, which were calculated from three replicates.

### 3.2.2 m$^6$A distribution exhibits a distinct topology in *E. coli*

To obtain the transcriptome-wide m$^6$A map of *E. coli*, we employed an m$^6$A-specific antibody for pull-down coupled with high-throughput sequencing.[6,7] In order to obtain a high-resolution m$^6$A-map, bacterial mRNAs were subjected to a modified photo-crosslinking-assisted m$^6$A sequencing approach (PA-m$^6$A-seq), which significantly improves the m$^6$A peak resolution from ~200 nt to around 23 nt.[8] In total, we identified 265 m$^6$A peaks representing the transcripts of 213 genes in *E. coli*.

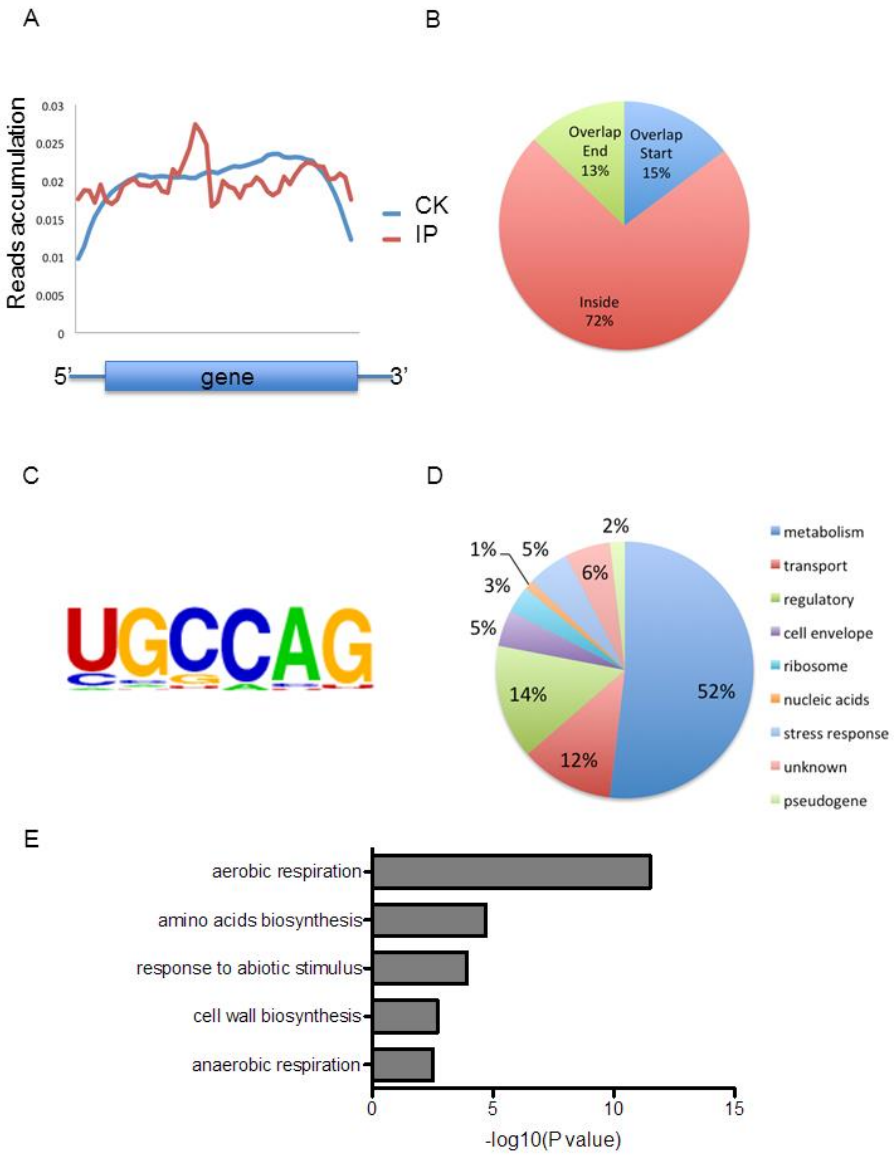



**Figure 3.4. Overview of m$^6$A methylome in *E. coli*.** (A) The m$^6$A peak distribution within different gene contexts. The *y*-axis represents (number of reads/length unit)/(number of total reads), which is an indicator of the extent to which sequencing reads are enriched in different segments across the entire transcript. (B) Accumulation of m$^6$A reads along transcripts. Each transcript is divided into three parts: Overlap Start, Inside, and Overlap End. (C) The UGCCAG conserved sequence motif for m$^6$A-containing peak regions. (D) Pie chart displaying the percentage of genes containing m$^6$A peaks with functional categories. (E) GO-enrichment analysis of all the genes with m$^6$A peaks. The effect size (number of enriched genes/total genes in the GO category) for each category is 15/45 (aerobic respiration), 15/122 (amino acids biosynthesis), 6/138 (response to abotic stimulus), 5/61 (cell wall biosynthesis), and 6/39 (anaerobic respiration), respectively. The statistical test (*p*-value) used by DAVID was the Fisher Exact test.

We next analyzed the distribution of $m^6A$ in the whole transcriptome of *E. coli*. We determined the distribution of $m^6A$ reads along transcripts in both the $m^6A$-IP and non-IP (input) samples. Intriguingly, we found that reads from $m^6A$-IP tend to be equally distributed throughout a gene, with a peak in the middle of open reading frames (ORFs) (**Figure 3.4A**). The prevalence is quite different from that observed in mammals, which accumulates around the stop codon and within 3' UTRs.[6,7] To further confirm the preferential locations of $m^6A$ on transcripts, we investigated the metagene profiles of $m^6A$ peaks. Consistent with the distribution of reads, $m^6A$ peaks are abundant inside ORF (72%), followed by regions at the start of gene (15%) and then the end of gene (13%) (**Figure 3.4B**).

We then used the HOMER tool to identify a leading $m^6A$ consensus sequence (UGCCAG, $P < 1e\text{-}14$), which could be found in more than 41.2% of all $m^6A$ peaks (**Figure 3.4C**).[9] This motif is different from the conserved one (RRACU, R = A/G) found in eukaryotes. The unique feature of the $m^6A$ distribution suggests a likely unique role of $m^6A$ in perhaps bacteria-specific pathways.

### 3.2.3 $m^6A$-containing mRNAs in important biological pathways in *E. coli*

Diverse functions are encoded by $m^6A$-containing genes, which include metabolism (52%), transportation (12%), gene regulation (11%), cell envelope (5%), ribosome (3%), nucleic acids (1%), stress response (5%), genes with unknown function/annotation (6%) and pseudogenes (2%) (**Figure 3.4D**). To further uncover potential functional insights on $m^6A$ in *E. coli*, we selected 213 $m^6A$-containing transcripts and identified the enriched gene ontology (GO) terms using the DAVID tool.[10] We found that these genes are highly enriched in aerobic respiration, amino acids biosynthesis, response to abiotic stresses, cell wall biosynthesis and anaerobic respiration (**Figure 3.4E**). The first two classes belong to housekeeping genes that are involved in central energy

production and metabolism, while the latter three classes are bacteria-specific categories. The $m^6A$ distribution pattern suggests that $m^6A$ may play roles in these important biological pathways. For instance, hydrogenase 1 mediates hydrogen uptake and transport in the process of anaerobic respiration. Four (*hyaA*, *hyaB*, *hyaC* and *hyaD*) genes of the six-gene-operon encoding hydrogenase 1 contain multiple $m^6A$ peaks inside the transcript, suggesting a clustering of $m^6A$ in this operon (**Figure 3.5A**).[11] We also observed concentrated $m^6A$ peaks in *gabD* and *gabT* genes, which encode succinate-semialdehyde dehydrogenase and 4-aminobutyrate aminotransferase in the pathway of amino acid metabolism (**Figure 3.5B**).[12]

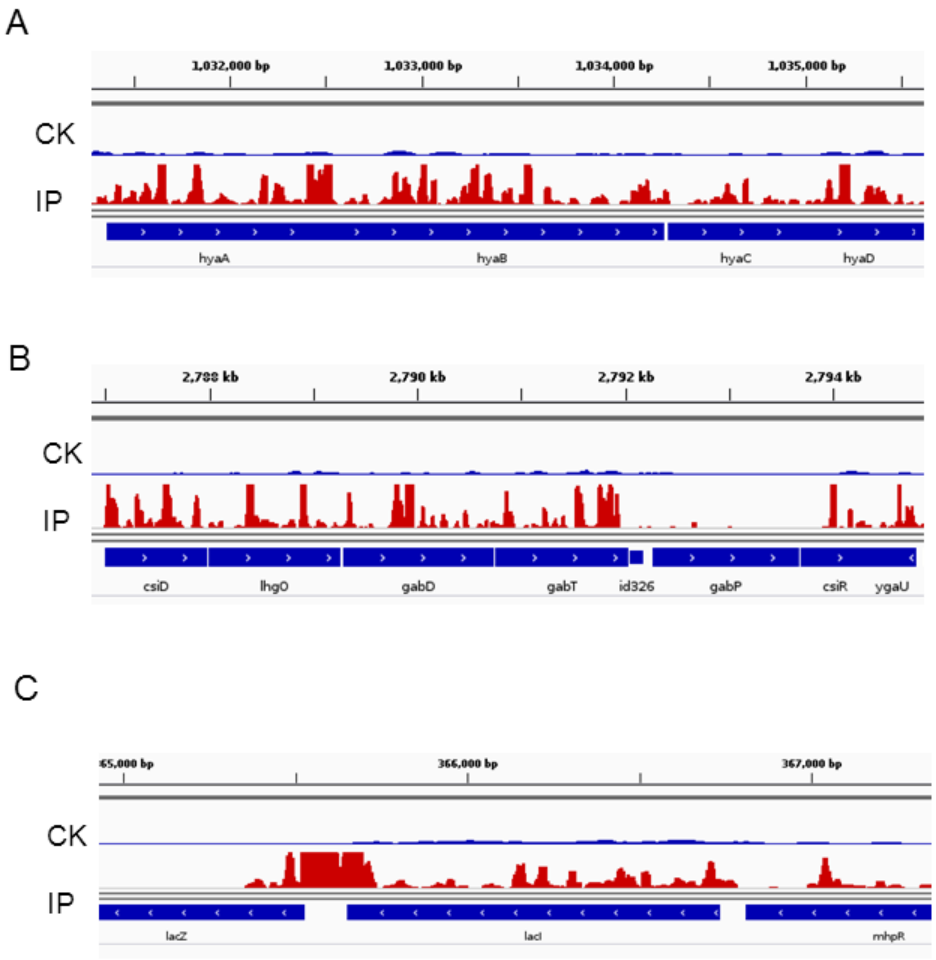



**Figure 3.5. Accumulation of $m^6A$ reads in *hyaABCD* genes (A), *gabDT* (B), and *lacZI* (C) in *E. coli* transcriptome.** CK represents the control sample, and IP represents $m^6A$ sample.

We next sought to determine if the unique $m^6A$ position patterns are related to bacteria-specific GO categories. As aforementioned, we classified genes into three subgroups according to the location of $m^6A$ peaks on a gene: Overlap Start ($m^6A$ peaks within 100-bp from the start codon), Overlap End ($m^6A$ peaks within 100-bp from the stop codon) and Inside ($m^6A$ peaks inside the coding region) (**Figure 3.4B**). We then performed GO-enrichment analysis for each subgroup. As expected, the same five GO categories were enriched in the Inside subgroup that consists of 72% of all $m^6A$ peaks. Two GO categories (aerobic respiration and stress responses) were enriched in the Overlap Start subgroup, while amino acids biosynthesis and response to stimulus were enriched in the Overlap End subgroup (**Figure 3.6**).

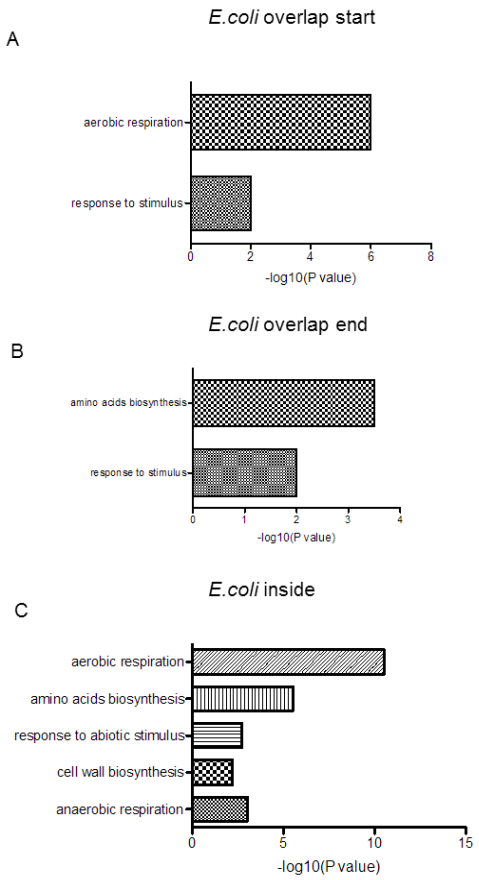



**Figure 3.6. GO-enrichment analysis of all *E. coli* genes with $m^6A$ peaks.** (A) Overlap Start; (B) Overlap End; (C) Inside.

Beside these five GO categories, we identified high $m^6A$ peaks in a group of functionally important genes, such as *lacZ* and *lacI* (**Figure 3.5C**). LacI negative regulates the classic lac operon (*lacZYA*) that is required for transport and metabolism of lactose.[13] Interestingly, we also noticed 15 small RNAs carrying $m^6A$ modification. These newly found $m^6A$ marks in these transcripts could open a new angle to study novel regulatory roles in well-established pathways.

### 3.2.4 Unique patterns of *P. aeruginosa* methylome

Given that *P. aeruginosa*, a widely-spread human opportunistic pathogen, also possesses a high $m^6A$/A ratio in mRNA, we applied the same modified photo-crosslinking-assisted $m^6A$-seq approach to obtain a high-resolution map of its $m^6A$ methylome. We identified 109 $m^6A$ peaks representing the transcripts of 68 genes in *P. aeruginosa*. The $m^6A$-modified transcripts identified are around half of those in *E. coli*.

We next determined the distribution of $m^6A$ reads along transcripts in both the $m^6A$-IP and non-IP (input) samples. Like in *E. coli*, we found that reads from $m^6A$-IP are equally distributed throughout a gene, with two peaks in the middle of ORFs as well as in the beginning of genes (**Figure 3.7A**). $m^6A$ peaks are abundant inside ORF (77%), followed by the start of gene regions (15%) and end of gene regions (8%) (**Figure 3.7B**). We were also able to identify an $m^6A$ consensus sequence (GGYCAG, Y = C/U, $P < 1e\text{-}16$), which were found in >70% of all $m^6A$ peaks (**Figure 3.7C**). This motif is almost identical to the one in *E. coli* (UGCCAG). A similar feature of the $m^6A$ distribution in both *E. coli* and *P. aeruginosa* indicates that $m^6A$ possesses functions unique to these bacteria.

The $m^6A$-containing genes cover different gene categories in *P. aeruginosa*, including metabolism (40%), gene regulation (8%), transportation (1%), virulence (4%), ribosome (12%), stress response (4%) and genes with unknown function/annotation (31%) (**Figure 3.7D**). DAVID

analysis revealed significant GO enrichments in amino acids metabolism, glycolysis, and tricarboxylic acid (TCA) cycle (**Figure 3.7E**), all of which belong to housekeeping genes that are involved in central energy production and metabolism. The $m^6A$ pattern suggests that $m^6A$ may play roles in these essential pathways in *P. aeruginosa*. For instance, three adjacent genes (PA3415-PA3417 encoding branched-chain alpha-keto acid dehydrogenase, pyruvate dehydrogenase β and α subunit, respectively) involved in glycolysis and TCA carry numerous $m^6A$ peaks (**Figure 3.8A**).[14] We also observed multiple $m^6A$ peaks in the next downstream gene *ldh*, which encodes leucine dehydrogenase in the pathway of amino acid metabolism (**Figure 3.8A**). Beside these housekeeping genes, we identified high $m^6A$ peaks in two important small RNAs, namely RsmY and RsmZ, as well as in two key virulence genes *rhlA* and *rhlB* (**Figure 3.8B**).

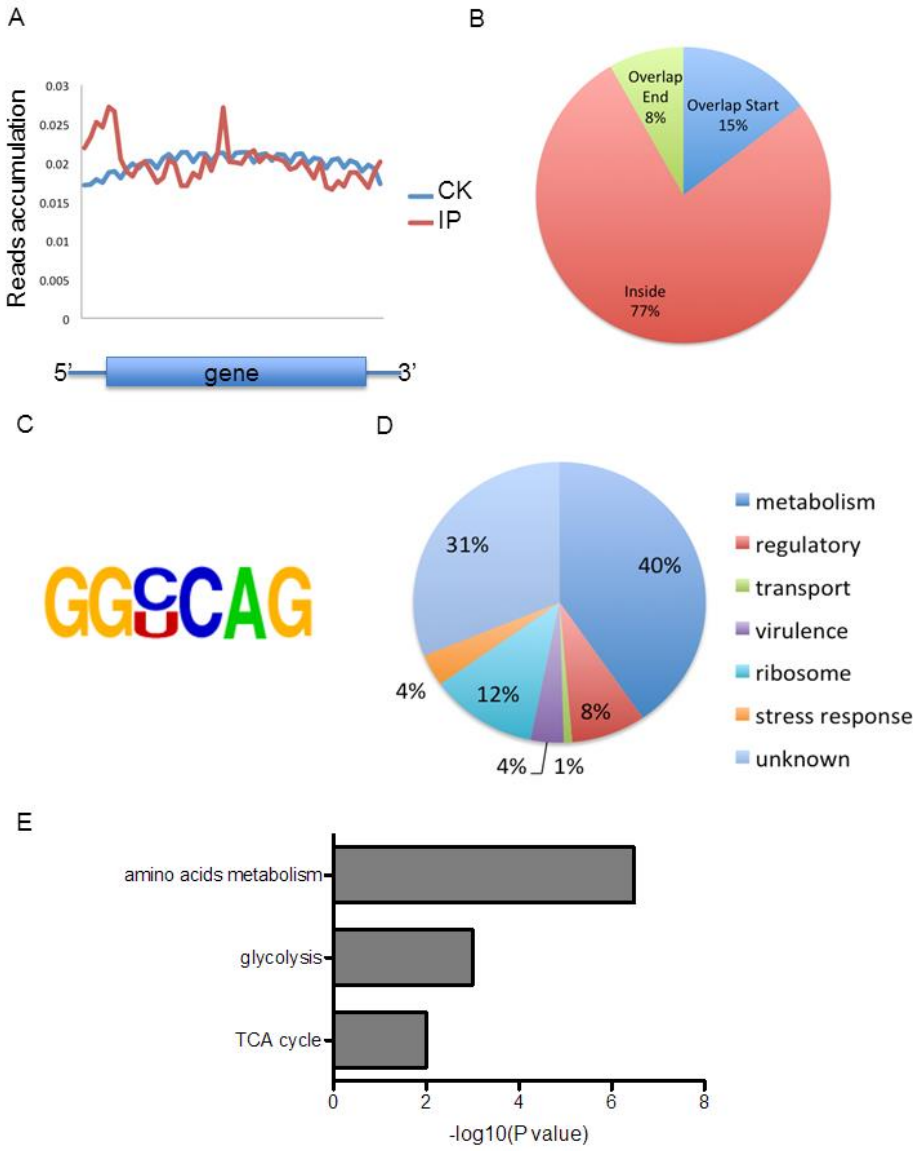



**Figure 3.7. Overview of m$^6$A methylome in *P. aeruginosa*.** (A) The m$^6$A peak distribution within different gene contexts. The y-axis represents (number of reads/length unit)/(number of total reads), which is an indicator of the extent to which sequencing reads are enriched in different segments across the entire transcript. (B) Accumulation of m$^6$A reads along transcripts. Each transcript is divided into three parts: Overlap Start, Inside, and Overlap End. (C) The GGCCAG conserved sequence motif for m$^6$A-containing peak regions. (D) Pie chart displaying the percentage of genes containing m$^6$A peaks with functional categories. (E) KEGG-enrichment analysis of all the genes with m$^6$A peaks. The effect size (number of enriched genes/total genes in the KEGG category) for each category is 7/148 (amino acids metabolism), 4/37 (glycolysis), and 3/56 (TCA cycle), respectively.

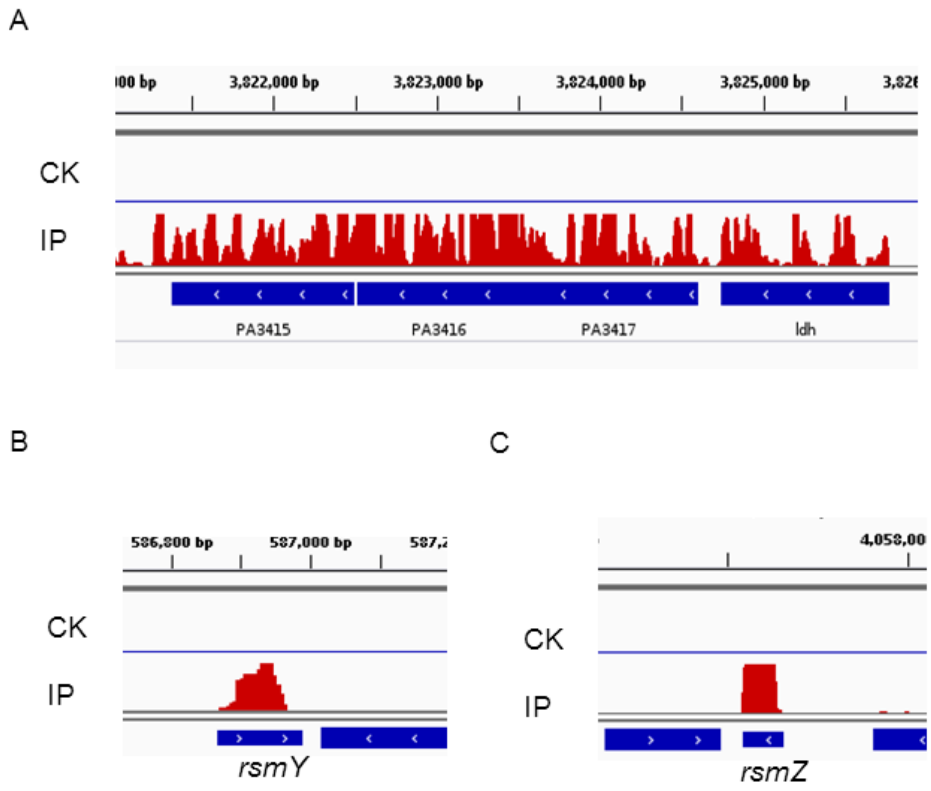



**Figure 3.8. Accumulation of m$^6$A reads in PA3415-3417 and *ldh* (A), *rsmY* (B), and *rsmZ* (C) in *P. aeruginosa* transcriptome.** CK represents control sample, and IP represents m$^6$A sample.

### 3.2.5 Temperature tunes m$^6$A level in *P. aeruginosa*

In humans and mice, dynamic changes of certain m$^6$A peaks have been observed under different stress conditions, indicating a link between m$^6$A and stress response.[6] In order to test if this trend also exists in bacteria, we measured the m$^6$A/A ratios under a variety of growth environments or stress conditions (such as varying temperatures, different growth media, exposure to different antibiotics and oxidative stresses) for both *E. coli* and *P. aeruginosa*. For most tested conditions, we did not observe a significant difference in the m$^6$A/A level compared to the normal condition (LB, mid-log phase, 37 ℃) for both bacteria. Interestingly, we noticed that increasing the culture temperature (from 37-45 ℃) led to a clear decrease in the m$^6$A/A ratio in *P.*

*aeruginosa* (**Figure 3.9**). Although *P. aeruginosa* still slowly grew, particularly at 45 ℃, the $m^6A$ modification was almost abolished.

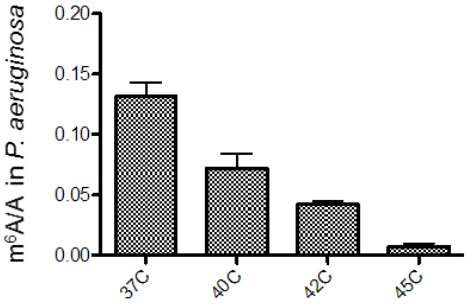



**Figure 3.9. Growth temperature significantly affects the $m^6A/A$ ratio in *P. aeruginosa*.** Error bars are calculated from three replicates.

### 3.2.6 Discussion and summary

Recent discoveries and characterization of $m^6A$ erasers (demethylase), binders ($m^6A$-specific binding protein) and writers (methyltransferase) as well as advances in profiling the $m^6A$ methylomes in eukaryotic systems reveal that $m^6A$ is a reversible and dynamic modification with important regulatory functions.[5] On the other hand, the $m^6A$ methylomes in bacterial mRNA remain poorly studied. Here, we report the presence of $m^6A$ in a wide range of bacterial species and the ratio of $m^6A/A$ in mRNA from diverse bacterial strains vary within the range of 0.02–0.28%. We noticed that *S. aureus* and *B. subtilis* showed very low $m^6A/A$ ratios, which suggests that they may not possess an $m^6A$ methylase that could be present in the Gram-negative bacteria. Based on the genome annotation in NCBI, there are at least 43 proteins containing an S-adenosyl methionine (SAM)-binding domain in *E. coli* K-12, but only 24 in *S. aureus* USA300. We further present the high-resolution, transcriptome-wide $m^6A$ distributions in *E. coli* and human pathogen *P. aeruginosa*, which contain 265 and 109 peaks, respectively.

In order to provide additional insights into the overall $m^6A$ patterns in the bacterial kingdom, we compared these two newly identified methylomes. *E. coli* and *P. aeruginosa* share many similarities in their $m^6A$ distributions that are distinct from those of mammals: (i) a similar motif GCCAG instead of RRACU motif in mammals; (ii) most peaks are in the middle of coding region, while mammalian $m^6A$ peaks enrich around the stop codon and at 3' UTRs; (iii) enrichment of GO categories of energy and amino acids metabolism; (iv) many small noncoding RNAs were found to carry $m^6A$ for both organisms. These shared characteristics suggest that other bacterial species, especially Gram-negative bacteria, may have similar $m^6A$ characteristics in mRNA. The new consensus sequence (GCCAG) is different from known rRNA methylation sites, including the two $m^6A$ sites on rRNA (CACA*GGU for RlmF and GUGA*AGA for RlmJ) and one on $tRNA^{val}$ (UACA*AGG for YfiC). Given that our recent study demonstrates that $m^6A$ and its specific binding protein, YTHDF2, affect the translation status and lifetime of mRNA in eukaryotes, $m^6A$ may play a similar role in bacteria.[15]

On the other hand, there are clear species-specific features of the $m^6A$ distribution between these two bacteria. Although the two species share major GO categories (energy and protein metabolism), we observed a very low rate of overlapping genes. Besides rRNA genes that are previously known to carry $m^6A$ modification, only one gene (*aceA* encoding isocitrate lyase) was shared between *E. coli* and *P. aeruginosa*. Each bacterium has distinct functional categories of mRNAs that carry $m^6A$. For example, genes involved in cell wall biosynthesis and anaerobic respiration are enriched in the *E. coli* methylome only, while a group of virulence genes (RsmYZ and *rhlAB*) are enriched in the *P. aeruginosa* methylome (**Figure 3.8B** and **3.8C**). RsmYZ binds to RsmA and dissociates RsmA away from its mRNA targets, which in turn tunes a group of important virulence pathways including Type III secretion system (T3SS) and biofilm

formation.[16,17] *rhlAB* encodes rhamnosyltransferase, producing rhamnolipid biosurfactants that are involved in uptake of hydrophobic substrates, virulence, biofilm and antibiotic resistance.[18] $m^6A$ marks in these virulence genes could connect RNA modification to bacterial pathogenesis. Our result also suggests a relationship between $m^6A$ and *P. aeruginosa* adaption to temperature changes. Alternatively, the putative $m^6A$ methylase in *P. aeruginosa* may be inactive at high temperature.

Given that multiple known rRNA or tRNA adenine methylases have been characterized in *E. coli*, we measured the $m^6A/A$ ratios in two methylase mutants, *rlmF* and *rlmJ*. However, the ratios were not significantly lower than the wild-type strain, suggesting that they are not mRNA methyltransferases (**Figure 3.10**). As a negative control, the *ksgA* mutant lost the $m^6{}_2A$ modification and showed a higher $m^6A/A$ ratio than the other strains, suggesting a small content of rRNA in the mRNA sample. We also tried to look for bacterial homologs of mammalian $m^6A$ methyltransferases (METTL3, METTL14 and WTAP), but could not identify one, reminiscent of a distinct bacterial $m^6A$ motif that is different from that of mammals. These results suggest that bacterial $m^6A$ modification in mRNA is possessing of a mechanism that differs from eukaryotes.

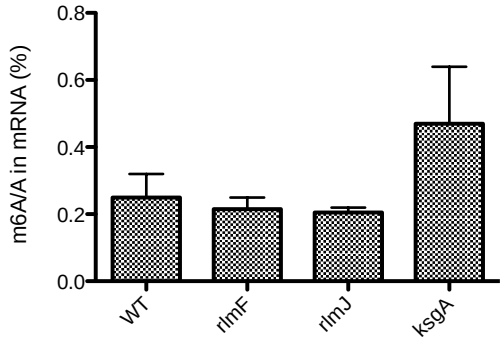



**Figure 3.10. The $m^6A/A$ levels in mRNA from the wild type ($m^6A$-$m^6{}_2A$/1.30)/A, the *rlmF* mutant ($m^6A$-$m^6{}_2A$/2.04)/A, the *rlmJ* mutant ($m^6A$-$m^6{}_2A$/2.04)/A and the *ksgA* mutant ($m^6A/A$) of *E. coli.***

Recent $m^6A$ profiling in yeast revealed dynamic changes in methylation during meiosis, which led us to test if bacterial $m^6A$ patterns also vary in different growth stage.[19] To this end, we measured the $m^6A/A$ ratios during lag phase, log phase, stationary phase and death phase for both *E. coli* and *P. aeruginosa*. No significant difference was recorded throughout the bacterial growth curve, which indicates that $m^6A$ is a stable modification during bacterial growth.

The first bacterial $m^6A$ maps presented here provide a starting roadmap for uncovering bacterial distinct $m^6A$ functions in the future. Recent breakthroughs in the characterization of $m^6A$-associated proteins as well as in the development of high-throughput assays in mammals present a very useful toolbox for us to study $m^6A$ in bacteria. Given the high abundance of $m^6A$ in numerous bacterial species, we foresee unique functions of $m^6A$ modification in mRNA in the wide bacteria kingdom.

## 3.3 Experimental section

### 3.3.1 Bacterial strains and mRNA purification

The strains and culture conditions used in this study are listed in **Table 3.1**. Total RNA was purified from bacterial pellets of 2 mL culture by using an RNeasy Mini Kit (Qiagen) that removes tRNA. Two micrograms of total RNA were applied to a MICROExpressTM Bacterial mRNA Enrichment Kit (Life technologies). A Ribo-Zero$^{TM}$ rRNA Removal Kit (Bacteria) (Epicentre) was used in order to further remove remaining rRNA. All procedures in the manufacturer's protocols were strictly followed. In order to verify the removal of rRNA, a qPCR (7300 Real-Time PCR System, Applied Biosystems) was done against the rRNA background in order to check relative enrichment levels. One nanogram of either total RNA or purified mRNA from *E. coli* was used per qPCR reaction (Power SYBR Real-Time PCR mater mix, Life technologies). The primers used were 5'-CTCCTACGGGAGGCAGCAG-3' and 5'-GTATTACCGCGGCGGCTG-3'.

**Table 3.1. Strains and growth conditions**

| Strain | Growth condition |
|---|---|
| *Escherichia coli* K-12 (CGSC) | LB, 37 ℃ overnight |
| *Escherichia coli* K-12 *rlmJ mutant* (CGSC) | LB, 37 ℃ overnight |
| *Escherichia coli* K-12 *rlmF mutant* (CGSC) | LB, 37 ℃ overnight |
| *Escherichia coli* K-12 *ksgA mutant* (CGSC) | LB, 37 ℃ overnight |
| *Escherichia coli* 5α | LB, 37 ℃ overnight |
| *Escherichia coli* XL-blue | LB, 37 ℃ overnight |
| *Pseudomonas aeruginosa* MPAO1 | LB, 37 ℃ overnight |
| *Pseudomonas aeruginosa* PA14 | LB, 37 ℃ overnight |
| *Pseudomonas syringae* pv. tomato DC3000 | King's B medium, 28°C for 2 d |
| *Staphylococcus aureus* Newman | TSB medium, 37 ℃ overnight |
| *Staphylococcus aureus* USA100 | TSB medium, 37 ℃ overnight |
| *Staphylococcus aureus* USA400 | TSB medium, 37 ℃ overnight |
| *Staphylococcus aureus* USA700 | TSB medium, 37 ℃ overnight |
| *Staphylococcus aureus* COL | TSB medium, 37 ℃ overnight |
| *Staphylococcus aureus* RN4220 | TSB medium, 37 ℃ overnight |
| *Bacillus subtilis* | LB, 37 ℃ overnight |
| *Anabaena* sp. PCC 7120 | Z8 medium, 25 ℃ overnight |
| *Synechocystis* sp. PCC 6803 | Z8 medium, 25 ℃ overnight |

The *P. aeruginosa* MPAO1 strain were cultured overnight at different temperatures (37, 40, 42 or 45 ℃) and then subjected to the mRNA purification protocol described above in the temperature variation studies.

### 3.3.2 Ultra-high pressure liquid chromatography coupled with triple-quadrupole tandem mass spectrometry (UHPLC-QQQ-MS/MS) analysis for $m^6A$/A ratio

The highly purified bacterial mRNA was subjected to an UHPLC-QQQ-MS/MS (Agilent) analysis. 200 ng of mRNA or rRNA (on the beads of the mRNA Enrichment Kit) were digested by nuclease P1 (2 U) in 40 μL of nuclease buffer (25 mM of NaCl and 2.5 mM of $ZnCl_2$) at 37 ℃ for 2 hrs, followed by the addition of $NH_4HCO_3$ (1 M, 2 μL) and alkaline phosphatase (0.5 U) at 37 ℃ for 2 hrs. The nucleosides were separated by reverse phase ultra-performance liquid chromatography by a C18 column on an Agilent 6410 QQQ triple-quadrupole LC mass spectrometer in positive electrospray ionization mode. The nucleosides were quantified using the nucleoside-to-base ion mass transitions of 282 to 150 ($m^6A$), 294 to 164 ($m^6{}_2A$) and 268 to 136 (A). Quantification was performed by comparison with the standard curve obtained from pure nucleoside standards. Three biological repeats have been performed for all bacterial strains.

### 3.3.3 High-throughput and high-resolution $m^6A$ sequencing

Procedures were slightly modified from previously described protocols.[8] In a 0.5 mL IP reaction, 5 μg purified bacterial mRNA and 15 μL of 0.5 mg/ml rabbit anti-$m^6A$ antibody (202003; Synaptic Systems) were incubated for 2 hrs at 4 ℃ in IPP buffer (150 mM NaCl, 0.1% NP-40, 10 mM Tris–HCl, pH 7.4, 1 U/μL RNasin). The mixture was exposed to UV irradiation at 254 nm 3× (90 s each time), before RNase T1 (0.1 U/μL) digestion for 15 min at 22 ℃. After the digestion reaction was quenched on ice for 5 min, 200 μL pre-blocked protein A bead slurry was added into reaction for 1 hr at 4 ℃. After washing thrice with IP wash buffer (50 mM HEPES-

KOH, pH 7.5, 300 mM KCl, 0.05% NP-40, with proteinase inhibitor and RNasin), the beads were treated by a second round of RNase T1 digestion (15 U/μL) at 22 ℃ for 15 min. The beads were cooled down on ice for 5 min and then thrice washed with high salt wash buffer (50 mM HEPES-KOH, pH 7.5, 500 mM KCl, 0.05% NP-40, with proteinase inhibitor and RNasin). The beads were then treated with Antarctic phosphatase (0.5 U/μL) for 20 min at 37 ℃. After dephosphorylation, beads were washed twice with phosphatase wash buffer (50 mM Tris–HCl, pH 7.5, 20 mM EGTA, 0.5% NP-40) and twice with PNK buffer without DTT (50 mM Tris–HCl, pH 7.5, 50 mM NaCl, 10 mM $MgCl_2$). Polynucleotide kinase (1 U/μL) and 200 μM adenosine triphosphate (ATP) was then added to the beads at 37 ℃ for 15 min. The RNA fragments were further purified by proteinase K digestion and TRIzol extraction. For IP samples, small RNA libraries were made by using NEBNext Small RNA Library Prep Set for Illumina (NEB). The input samples followed the above procedures without anti-$m^6A$ antibody pull-down and RNase T1 digestion. Libraries for input samples were made by using TruSeq RNA Sample Preparation Kits (Illumina, non-strand specific). Six libraries were constructed, containing one control sample and two duplicate IP samples for each strain.

### 3.3.4 Data analyses

All samples were sequenced using the HiSeq 2000 system (Illumina, with 50-bp and single end mode) at the Genomics Core Facility at the University of Chicago. FastQC was done to check the quality of each dataset; all datasets were obtained in high quality and can afford further reliable analyses. Two *E. coli* IP libraries obtained 1,718,364 mapped reads, and two *P. aeruginosa* libraries obtained 3,720,658 mapped reads. Sequence data were analyzed by following the procedures described previously.[20] Briefly, TopHat (version 2.0.0, with the parameter: `-p 8 –read-mismatches 2 –max-multihits 1`) with Bowtie was run in order to align the

input and IP-sequenced samples to the *E. coli* K-12 substr. MG1655 (ASM584v2, NC 000913.3) and *P. aeruginosa* PAO1 (ASM676v1, NC 002516.2) genomes and annotation files.[21,22] In TopHat each read was only mapped to the genome once. The enriched peaks were identified using MACS software (version 2.0.0, with the parameter: `callpeak -t ip.bam -c ck.bam -f BAM -g 6000000 –nomodel -n -p 1e-5`) (24).[23] Consensus sequence motifs were identified by using HOMER (version 4.7, with the parameter: `-p 3 -rna -len 6`, **Table 3.2**).[9]

**Table 3.2. Top m$^6$A motifs**

| Species | Motif | % of targets | *p*-value | Location |
|---|---|---|---|---|
| *E. coli* |  | 41.2% | 1e-14 | 68.1 +/- 79.4bp |
| |  | 25.8% | 1e-19 | 41.4 +/- 47.1bp |
| |  | 70.16% | 1e-16 | 702.7 +/- 823.5bp |
| *P. aeruginosa* |  | 56.85% | 1e-30 | 759.0 +/- 884.9bp |
| |  | 43.95% | 1e-13 | 795.2 +/- 910.8bp |

A scrambled sequence was used as the background. Gene function analysis (GO enrichment) was performed with the online DAVID (version 6.7) tool (http://david.abcc.ncifcrf.gov/ ).[10] The

$m^6A$ peaks were divided into three categories based on their relative positions in their corresponding genes: Overlap Start (±100 nucleotides around the start codon), Overlap End (±100 nucleotides around the stop codon) and Inside (other locations inside a coding region). The functional association with each gene was determined by NCBI annotation.

### 3.4 References


1 Golovina, A. Y. *et al.* The last rRNA methyltransferase of *E. coli* revealed: The *yhiR* gene encodes adenine-$N^6$ methyltransferase specific for modification of A2030 of 23S ribosomal RNA. *RNA* **18**, 1725-1734 (2012).

2 Sergiev, P. V., Serebryakova, M. V., Bogdanov, A. A. & Dontsova, O. A. The *ybiN* Gene of *Escherichia coli* Encodes Adenine-$N^6$ Methyltransferase Specific for Modification of A1618 of 23 S Ribosomal RNA, a Methylated Residue Located Close to the Ribosomal Exit Tunnel. *J. Mol. Biol.* **375**, 291-300 (2008).

3 Poehlsgaard, J. & Douthwaite, S. The bacterial ribosome as a target for antibiotics. *Nat. Rev. Microbiol.* **3**, 870-881 (2005).

4 Cunningham, P. R. *et al.* Site-specific mutation of the conserved $m^6{}_2A$ $m^6{}_2A$ residues of *E. coli* 16S ribosomal RNA. Effects on ribosome function and activity of the ksgA methyltransferase. *Biochimi. Biophys. Acta (BBA) - Gene Structure and Expression* **1050**, 18-26 (1990).

5 Fu, Y., Dominissini, D., Rechavi, G. & He, C. Gene expression regulation mediated through reversible $m^6A$ RNA methylation. *Nat. Rev. Genet.* **15**, 293-306 (2014).

6 Dominissini, D. *et al.* Topology of the human and mouse $m^6A$ RNA methylomes revealed by $m^6A$-seq. *Nature* **485**, 201-206 (2012).

7 Meyer, K. *et al.* Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3' UTRs and near Stop Codons. *Cell* **149**, 1635-1646 (2012).

8 Chen, K. *et al.* High-Resolution $N^6$-Methyladenosine ($m^6A$) Map Using Photo-Crosslinking-Assisted $m^6A$ Sequencing. *Angew. Chem. Int. Ed.* **54**, 1587-1590 (2015).

9 Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime *cis*-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576-589 (2010).

10 Alvord, G. *et al.* The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* **8**, 183 (2007).

11 Menon, N. K., Robbins, J., Wendt, J. C., Shanmugam, K. T. & Przybyla, A. E. Mutational analysis and characterization of the *Escherichia coli hya* operon, which encodes [NiFe] hydrogenase 1. *J. Bacteriol.* **173**, 4851-4861 (1991).

12 Bartsch, K., von Johnn-Marteville, A. & Schulz, A. Molecular analysis of two genes of the *Escherichia coli gab* cluster: nucleotide sequence of the glutamate:succinic semialdehyde transaminase gene (*gabT*) and characterization of the succinic semialdehyde dehydrogenase gene (*gabD*). *J. Bacteriol.* **172**, 7035-7042 (1990).

13 Beckwith, J. R. Regulation of the Lac Operon. *Science* **156**, 597-604 (1967).

14 Winsor, G. L. *et al.* Pseudomonas Genome Database: improved comparative analysis and population genomics capability for Pseudomonas genomes. *Nucleic Acids Res.* **39** gkq869 (2011).
15 Wang, X. *et al.* $N^6$-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505**, 117-120 (2014).
16 Intile, P. J., Diaz, M. R., Urbanowski, M. L., Wolfgang, M. C. & Yahr, T. L. The AlgZR Two-Component System Recalibrates the RsmAYZ Posttranscriptional Regulatory System To Inhibit Expression of the *Pseudomonas aeruginosa* Type III Secretion System. *J. Bacteriol.* **196**, 357-366 (2014).
17 Valverde, C., Heeb, S., Keel, C. & Haas, D. RsmY, a small regulatory RNA, is required in concert with RsmZ for GacA-dependent expression of biocontrol traits in *Pseudomonas fluorescens* CHA0. *Mol. Microbiol.* **50**, 1361-1379 (2003).
18 Ochsner, U. A., Fiechter, A. & Reiser, J. Isolation, characterization, and expression in *Escherichia coli* of the *Pseudomonas aeruginosa rhlAB* genes encoding a rhamnosyltransferase involved in rhamnolipid biosurfactant synthesis. *J. Biol. Chem.* **269**, 19787-19795 (1994).
19 Schwartz, S. *et al.* High-Resolution Mapping Reveals a Conserved, Widespread, Dynamic mRNA Methylation Program in Yeast Meiosis. *Cell* **155**, 1409-1421 (2013).
20 Luo, G.-Z. *et al.* Unique features of the $m^6A$ methylome in *Arabidopsis thaliana*. *Nat. Commun.* **5** (2014).
21 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359 (2012).
22 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562-578 (2012).
23 Zhang, Y. *et al.* Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, 1-9 (2008).

# 4 6mA-CLIP-exo sequencing to map 6mA in *Chlamydomonas* genomic DNA

## 4.1 Introduction

Covalent modifications of individual bases in DNA can encode inheritable genetic information beyond the four canonical DNA bases.[1] Methylations of DNA, including 5mC and 6mA, are the most abundant modifications in both prokaryotic and eukaryotic organisms.[2,3] The well-studied 5mC modification in multicellular eukaryotes regulates diverse cellular and developmental processes; however, the biological function of 6mA in eukaryotes is still unclear.[4,5]

6mA is known to be present in the genomic DNA of viruses, bacteria, protists, fungi, and algae, and has been detected in plant DNA and mosquito DNA.[6] In bacteria, 6mA plays crucial roles in the regulation of DNA mismatch repair, chromosome replication, cell defense, cell cycle regulation, transcription and virulence.[7-10] The maps of 6mA in several bacteria strains have been obtained by using single-molecule real-time (SMRT) sequencing.[11,12]

Besides bacteria, certain unicellular eukaryotes also contain 6mA in their genomes. For instance, the protozoan *Tetrahymena*, *Oxytricha fallox*, and *Paramecium aurelia* have relatively abundant 6mA but little 5mC.[13-15] On the other hand, green algae *Chlamydomonas reinhardtii* and *Volvox carteri* possess both 6mA and 5mC.[14,16] Although common in bacteria, no corresponding restriction endonucleases have been reported in these species. Therefore, 6mA in these unicellular eukaryotic genomes has long been suspected of possessing functions other than exclusion of foreign DNA or viruses. Additionally, evidences for the existence of 6mA in plants, insects, and mammals have also been reported.[17]

*Chlamydomonas reinhardtii* (referred to hereafter as *Chlamydomonas*) is a unicellular green alga that has been widely used as a model organism to study photosynthesis, eukaryotic flagella, and biomass production.[18] The high level (~0.3-0.5 mol%) of 6mA in the nuclear DNA of *Chla-*

*mydomonas* prompted us to study its distribution and function, which could help to decipher the long mystery of 6mA in eukaryotes and to develop bioengineering tools that may facilitate biomass and biofuel production.[14,19]

In this study, we employed/developed several methods for mapping 6mA sites in genomic DNA. We first applied 6mA immunoprecipitation sequencing, or 6mA-IP-seq, which is an antibody-based profiling method to obtain the genome-wide distribution of 6mA. We then developed a 6mA-CLIP-exo strategy of employing photo-crosslinking followed by exonuclease digestion to achieve a much higher resolution. Lastly, we developed a restriction enzyme based 6mA sequencing, or 6mA-RE-seq, to detect 6mA sites at single nucleotide resolution in genome-wide. Application of these three approaches to the *Chlamydomonas* genome revealed that 6mA marks more than 14,000 genes, accounting for 84% of all *Chlamydomonas* genes. This methylation is highly enriched around transcription start sites (TSS) with a bimodal distribution and significant local depletion at TSS. We used RNA-seq to quantify gene expression, and found that the presence of 6mA is correlated with actively expressed genes. This pattern is distinct from that of 5mC, which accumulates mostly in gene bodies in *Chlamydomonas*. At single nucleotide resolution, we also discovered that 6mA is enriched around TSS but exhibits an unexpected, strongly periodic pattern, suggesting controlled deposition of 6mA in association with nucleosome spacing. Nucleosome profiling revealed that 6mA around TSS occurs primarily within the linker DNA between nucleosomes. Our data show that 6mA is an abundant DNA mark associated with actively expressed genes in *Chlamydomonas*. These methods and results should stimulate future functional investigations of 6mA in *Chlamydomonas* and other eukaryotic organisms.

## 4.2 Result and discussion

### 4.2.1 6mA is a stable modification in *Chlamydomonas* genomic DNA

To accurately quantify the level of 6mA in genomic DNA, we applied an LC-MS/MS assay using pure 6mA nucleoside as an external standard (**Figure 4.1**).[20] In agreement with the previous data, we detected ~0.4 mol% of 6mA (6mA/A) in the genomic DNA isolated from *Chlamydomonas* cultured in mixotrophic conditions, i.e., Tris-Acetate-Phosphate (TAP) medium under constant light (**Figure 4.2A**).[14]

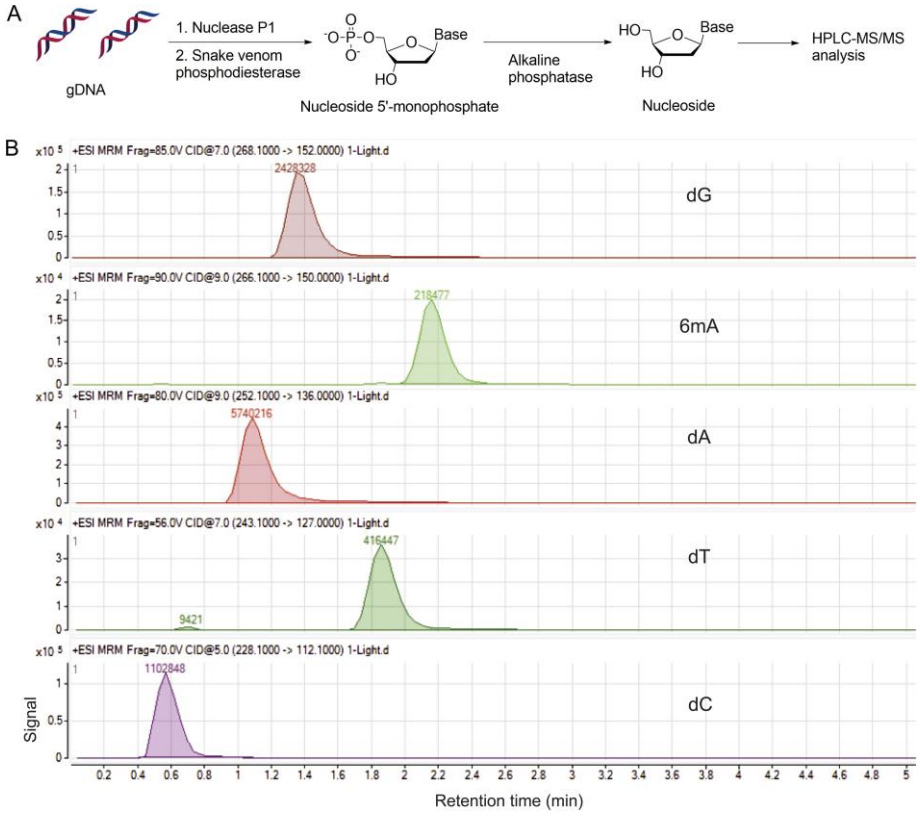



**Figure 4.1. Measuring the 6mA content in genomic DNA.** (A) Digestion and quantification of DNA bases. (B) Snapshot of UHPLC-QQQ-MS/MS for four unmodified bases and 6mA.

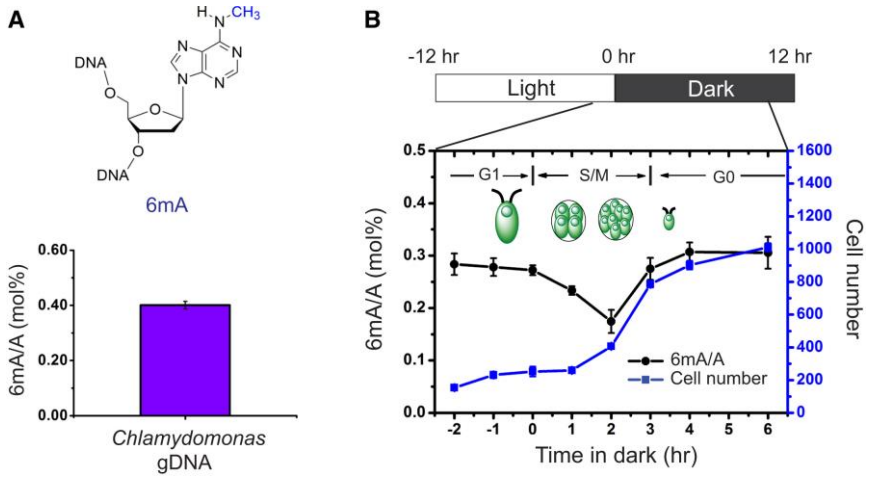



**Figure 4.2. The presence and conservation of 6mA in *Chlamydomonas* genomic DNA.** (A) The presence of 6mA in isolated *Chlamydomonas* genomic DNA as determined by UHPLC-QQQ-MS/MS. Ratio of 6mA/A are shown (n = 6, mean ± S.E.M.). (B) The level of 6mA decreases at the beginning of S/M phase during DNA replication, and increases back to the original level at the late stage of S/M phase. During a multiple fission cell cycle in naturally synchronized cells, cells begin to replicate its DNA 1 hr after dark and divide two times, leading to ~4-fold increase of the total cell number. Ratios of 6mA/A are shown on the left axis (n = 4, mean ± S.E.M.) with calibrated cell concentration (cells/μL) shown on the right axis (n = 2, mean ± S.E.M.).

To determine whether 6mA is stable during cell growth, we monitored the 6mA level during a multiple fission cell cycle in naturally synchronized cells induced by a 12 hrs/12 hrs light/dark cycle in minimal media cultures.[21] Under such growth conditions, cells grow in size during the light phase (G1 phase) and then undergo two to three rapid rounds of alternating DNA replications and cellular divisions (S/M phase) from 1 hr to 5 hrs after entering the dark phase. Cells were mostly synchronized and rapidly divided under this light-dark phase transition according to cell counting measured by flow cytometry (**Figure 4.2B**). The proportion of 6mA in genomic DNA was measured before and after the switch from light to dark. Our results showed that the overall 6mA level in genomic DNA decreased by ~40% in 2 hrs after dark, corresponding to the time period when DNA was replicated. This level then rose quickly back to the original level within 2 hrs. This result indicated that 6mA is installed on the newly synthesized DNA within a

short time period after DNA replication and is stably maintained during cell proliferation (**Figure 4.2B**).

### 4.2.2 Genome-wide mapping of 6mA with 6mA-IP-Seq

Although the existence of 6mA in *Chlamydomonas* has been known, its distribution/localizations are unclear. To generate a map of the genome-wide distribution of 6mA, we applied 6mA-IP-seq. Similar to the methylated DNA immunoprecipitation (MeDIP) that has been widely applied to enrich 5mC-containing DNA fragments, we sought to use a 6mA-specific antibody to enrich the 6mA-containing DNA fragments.[22] An antibody that recognizes the $N^6$-methyladenine base has recently been applied to genome-wide profiling of $m^6A$ sites in RNA.[23,24] By performing dot blotting assay on synthesized 6mA containing DNA oligonucleotide, we confirmed that this anti-6mA antibody can also specifically recognize 6mA in both single-stranded and double-stranded DNA (**Figure 4.3**).

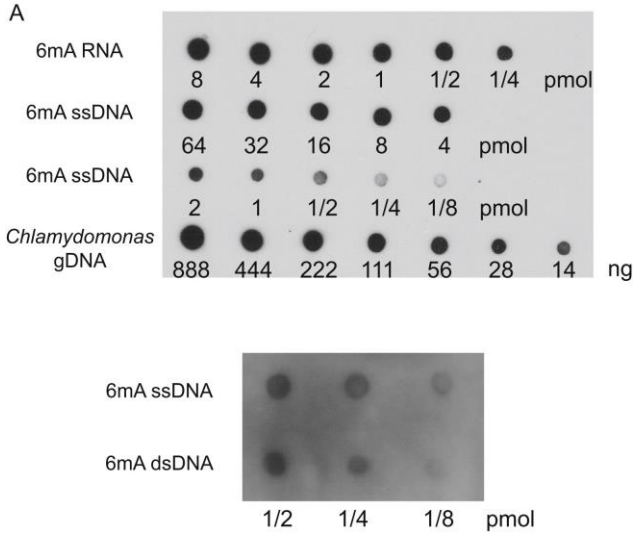
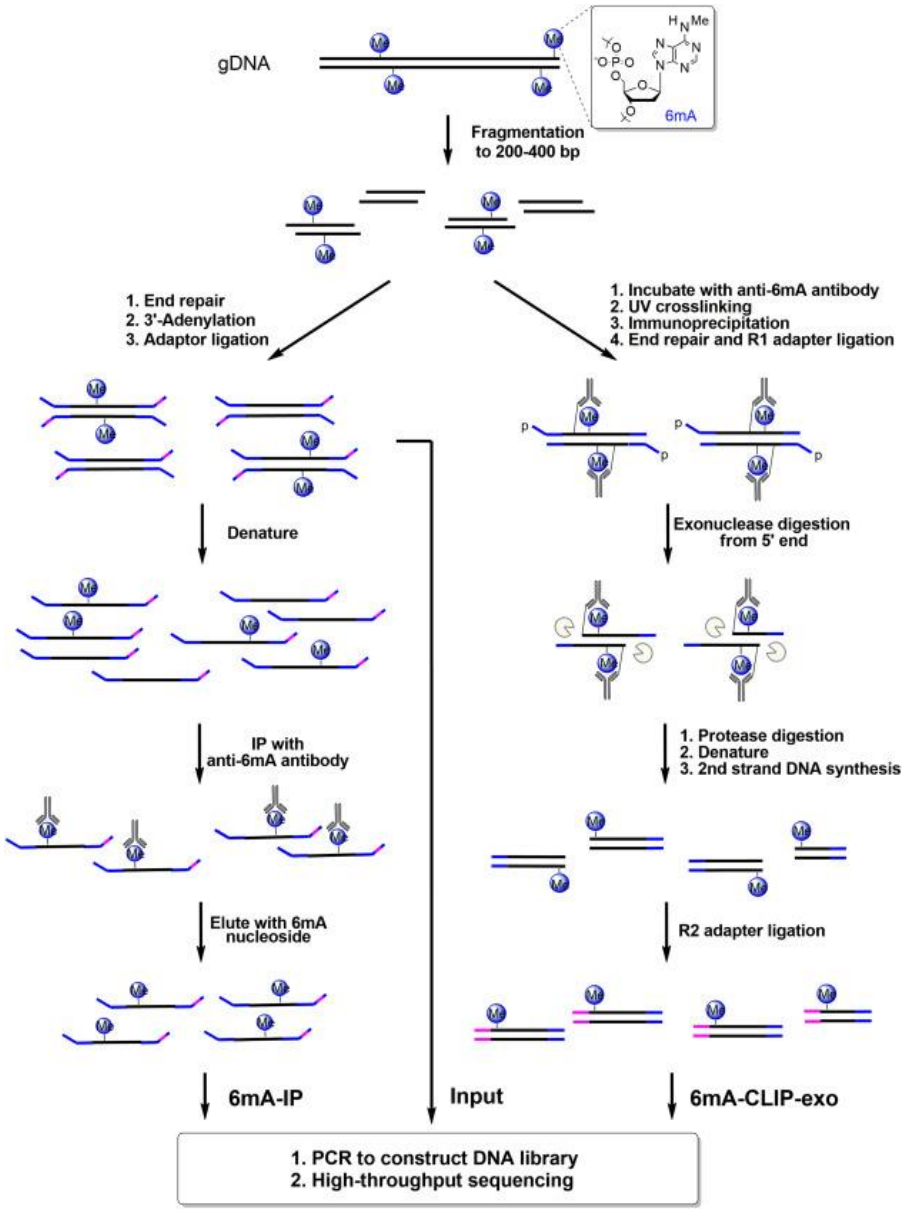



**Figure 4.3. Dot blotting assay of 6mA containing oligonucleotide.** 6mA dot blotting shows that the anti-6mA antibody can also recognize 6mA in single-stranded and double-stranded DNA.




**Scheme 4.1. Schematic diagram of 6mA-IP-seq and 6mA-CLIP-exo.** For 6mA-IP-seq (left), fragmented genomic DNA (gDNA) is ligated to a Y-shaped adapter with specific index sequence, denatured, and immunoprecipitated using anti-6mA antibody. The captured DNA is eluted with 6mA single nucleotide, and PCR amplified to construct the DNA library. Simultaneously, the input library was obtained from the ligated DNA before immunoprecipitation. For 6mA-CLIP-exo (right), fragmented gDNA is incubated with anti-6mA antibody, crosslinked by 254 nm UV irritation, and immunoprecipitated. The crosslinked DNA is ligated to adapter R1 on beads, followed by 5' to 3' exonuclease digestion. Antibody-protected DNA is preserved, and a 2nd-strand DNA synthesis is performed after protease digestion of the antibody. A second ligation to adapter R2 provides the template for PCR amplification to construct the library for high-throughput sequencing. Boundaries were determined by the sequencing ends of the 6mA-CLIP-exo-seq to provide a high resolution localization of 6mA.

We then isolated genomic DNA from *Chlamydomonas*, and fragmented it into 200-400 base pairs by sonication. The fragmented DNA was ligated to an adapter with specific index sequence (**Scheme 4.1**), which was then denatured to single-stranded DNA, and immunoprecipitated using the anti-6mA antibody. The captured DNA was eluted through the competition with 6mA single nucleotide, and PCR amplified to construct the DNA library (**Scheme 4.1**). Simultaneously, an input library was obtained by PCR amplification of the ligated DNA before immunoprecipitation. Both libraries were subjected to high-throughput sequencing. The obtained sequencing reads were mapped to a reference genome of *Chlamydomonas* (JGI version 9.1), and 6mA sites were identified using a peak-detection algorithm.[25] The false discovery rate (FDR) was estimated to be below 0.01.

### 4.2.3 6mA bases are highly enriched around TSS with a bimodal distribution

We performed 6mA-IP-seq on *Chlamydomonas* cultured under mixotrophic (constant light) or heterotrophic (constant dark) conditions in TAP medium during the pre-stationary phase. For each condition, we performed two biological replicates. After peak calling, we identified 25,803 and 28,982 high-confidence 6mA peaks in light samples and 22,005 and 21,016 peaks in dark samples (FDR < 0.01), respectively. Among them, more than 95% of the peaks mutually occur in both replicate samples, indicating the high reproducibility of our approach (**Figure 4.4A**). About 88% of the peaks are common under both light and dark conditions, suggesting a faithful instal-lation/maintenance mechanism of 6mA at specific genomic regions. Consistent with the previous measurements that 6mA was only detected in *Chlamydomonas* nuclear DNA but not chloroplast DNA, all the 6mA peaks were mapped to the nuclear genome but not the chloroplast genome. To our surprise, we observed that 6mA is highly enriched around the TSS of 14,868 genes, consti-tuting 84% of all the genes in the *Chlamydomonas* genome (**Figure 4.5A**). A closer examination

of the distribution revealed that the 6mA sites enriched around TSS (500 to +800 bp, ~91% of all 6mA peaks) exhibit a bimodal distribution with a significant local depletion at TSS. The summit of the peak tends to locate within 500 bp downstream of TSS (**Figure 4.5A**). The rest of the 6mA peaks (~9%) not associated with TSS do not show specific patterns and reside in both gene bodies and intergenic regions. The average peak width of the identified peaks is around 320 bp, which is consistent with the fragmentation size of our sequenced DNA (200-400 bp). We cannot quantify the number of methylation sites under each 6mA peak; however, some peaks are noticeably broader, with certain peaks containing multiple sub-peak summits, suggesting the presence of multiple methylation sites in these regions (**Figure 4.5B**). Thus, our observation revealed a region-specific bimodal methylation pattern of 6mA highly enriched around TSS.

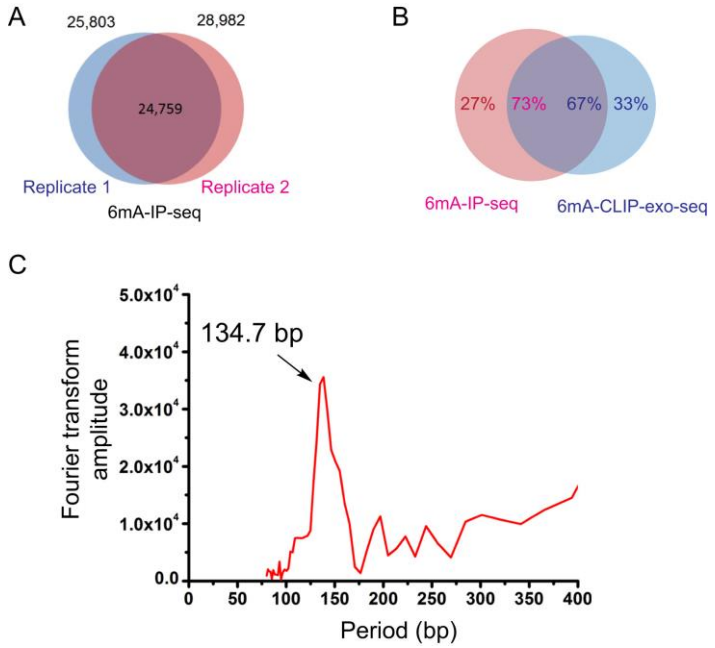



**Figure 4.4. Consistency of the 6mA-IP-seq and 6mA-CLIP-exo results.** (A) Overlap of results from two IP samples showing high consistency. (B) Overlap between 6mA-IP-seq and 6mA-CLIP-exo showing high consistency between these two methods. (C) Fourier transformation of 6mA localization profiles around TSS from 6mA-CLIP-exo results. A potential periodicity of 130-140 bp can be observed by performing Fourier transformation.

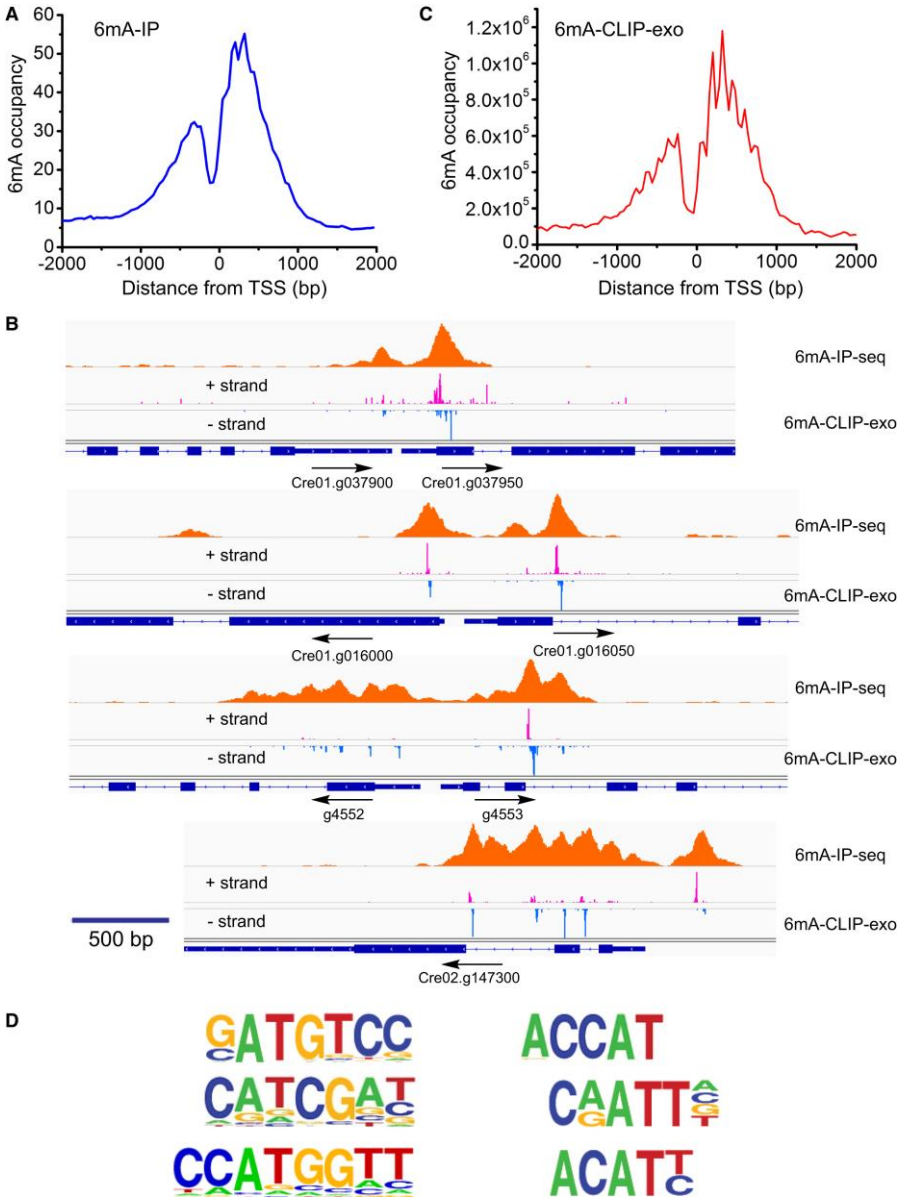



**Figure 4.5. A bimodal distribution of 6mA around transcription start sites.** (A) Distribution of 6mA peaks around TSS measured by 6mA-IP-seq. 6mA is enriched around TSS with a bimodal distribution and a local depletion at TSS. 6mA occupancy represents the reads coverage averaged by gene number in 6mA-IP-seq. (B) Snapshot of 6mA peak determined by both 6mA-IP-seq and 6mA-CLIP-exo in specific gene loci. 6mA peaks can be detected both upstream and downstream of TSS in single direction promoter region and bidirectional promoter region. Some enrichment peaks are located in the first and second introns. Boundaries of 6mA-CLIP-exo-seq on both DNA strands were marked by magenta and blue color. Regions between the two nearest boundaries were determined as a 6mA-containing sequence. Black arrows indicate the transcription direction. (C) Distribution of 6mA peaks around TSS measured by 6mA-CLIP-exo. The enrichment of 6mA near TSS shows a similar pattern as that obtained using 6mA-IP-seq. In addition, several spikes could be observed from the large peak. (D) The dinucleotide sequence ApT is enriched in 6mA-CLIP-exo peaks, including CATG.

### 4.2.4 6mA-CLIP-exo with immunoprecipitation, photo-crosslinking, and exonuclease digestion

Inspired by chromatin immunoprecipitation followed by exonuclease digestion (ChIP-exo), a method to map the locations at which a protein binds to the genome, we introduced photo-crosslinking after the antibody-based 6mA enrichment followed by exonuclease digestion in an attempt to identify 6mA peaks with higher resolution. DNA/antibody complexes were covalently crosslinked with UV irradiation before being captured by magnetic Protein A beads.[26,27] The crosslinked DNA was ligated to adapter R1 before being treated with two 5'-3' exonucleases, Lambda exonuclease and $RecJ_f$ exonuclease, to digest the DNA from the 50 end. The presence of crosslinked antibody stopped the exonuclease digestion before the crosslinking site. Antibody was then removed by proteinase K digestion, and DNA fragments were recovered for primer extension. The double stranded DNA (dsDNA) product was ligated to adapter R2 and sequenced (**Scheme 4.1**). By mapping the read ends to the *Chlamydomonas* genome, we determined the boundary sites of antibody-protected regions, which contain one or more 6mA sites. As expected, we successfully improved the resolution to ~33 bp (**Figure 4.5B**) and identified 30,899 6mA-containing sequences with 67% overlapping with 6mA peaks identified from 6mA-IP-seq. Meanwhile, 73% of 6mA peaks from 6mA-IP-seq contain at least one 6mA-containing sequence identified from 6mA-CLIP-exo (**Figure 4.4B**). These higher-resolution 6mA peaks showed the same enrichment around TSS with a bimodal distribution, a local depletion at TSS, and a potential periodic pattern (**Figure 4.4C**, **Figure 4.5C**). A motif search revealed multiple high-frequency sequences (**Figure 4.5D**), most of which contain an ApT dinucleotide motif (Figure 3D), reminiscent of the CpG methylation in most eukaryotic organisms and suggesting ApT as the general consensus sequence.

### 4.2.5 Validation of individual methylation sites

The methylation status of 6mA in specific motif sites can be validated by digestion with restriction enzymes originating from bacteria and viruses that are sensitive to 6mA methylation. For instance, CviAII is sensitive to 6mA and only digests the unmethylated CATG sequence, whereas DpnII cuts only the unmethylated GATC sequence.[28,29] We then applied the restriction-enzyme-digestion assay followed by quantitative PCR (6mA-RE-qPCR) to quantitatively evaluate the methylation status on specific motif sequences (**Figure 4.6A**). In this assay, we treated the isolated genomic DNA with CviAII or DpnII overnight to fully digest the unmethylated recognition motifs. We then designed PCR primers to specifically amplify the region flanked by the candidate 6mA site. In principle, the percentage of 6mA in the target 6mA site could be determined by quantitative PCR (qPCR) amplification of the restriction enzyme digested genomic DNA using undigested genomic DNA as a control, given that 6mA hinders digestion. This strategy was tested by analyzing nine specific CATG sites and five specific GATC sites within identified 6mA peaks from 6mA-IP-seq and 6mA-CLIP-exo, along with two CATG sites and two GATC sites in regions that are not methylated based on 6mA-IP-seq results. The 6mA-RE-qPCR assay identified 8/9 of these CATG sites and 3/5 GATC sites within 6mA peaks to be completely or partially (a lower methylation frequency in a population of DNA molecules) methylated. The control sites not identified by 6mA-IP-seq were not methylated by this assay (**Figure 4.6B** and **4.6C**). Therefore, this assay provides locus-specific validation of the 6mA-IP-seq results.

### 4.2.6 Genome-wide identification of single 6mA sites using 6mA-RE-Seq

The *Chlamydomonas* genome is GC rich (G+C content 64%) and around 120 million base pairs.[18] The ~0.4 mol% 6mA/A ratio corresponds to ~85,000 fully methylated 6mA sites. Our 6mA-IP-seq identified roughly 25,000 peaks, with each peak potentially covering multiple 6mA

sites (most of them are fully methylated at ~100%, see below), consistent with 6mA-IP-seq results showing that most 6mA peaks in the *Chlamydomonas* genome cluster around TSS sites. The 6mA-CLIP-exo results revealed several high-frequency sequences that include CATG and GATC. After we validated these two sequences as genuine 6mA methylation sites that mark TSS regions in *Chlamydomonas*, we sought to develop a high-throughput assay to map 6mA methylation in these selected sequences in genome wide at single-base resolution and to quantitatively determine the modification percentage at each site.

Genomic DNA was isolated and treated with CviAII or DpnII and then sonicated to around 300 base pair fragments, end-repaired by T4 DNA polymerase, 3'-adenylated, and ligated to DNA adapters. The unmethylated CATG or GATC motifs would be digested and should be enriched at the end of the DNA fragments. The methylated motifs should resist restriction enzyme mediated digestion and be present in the internal locations of DNA fragments. After PCR amplification of the fragments, a DNA library can be prepared for high-throughput sequencing. The ratio of a specific CATG or GATC sites with sequence reads internal versus at the end represents the relative methylation to unmethylation ratio. An input sample from genomic DNA without enzyme digestion serves as a control. Through mapping sequencing reads to the reference genome, we can identify the methylation status for every CATG or GATC motif in genome wide. We named this approach-as diagrammed in **Figure 4.6D**, 6mA-RE-seq and applied it to *Chlamydomonas* genomic DNA. While the specificity of DpnII to non-methylated DNA has been well characterized, the specificity of CviAII in cutting only non-methylated but not hemi- or fully methylated sequences was further confirmed using synthetic DNA probes (**Figure 4.7A**).[28]

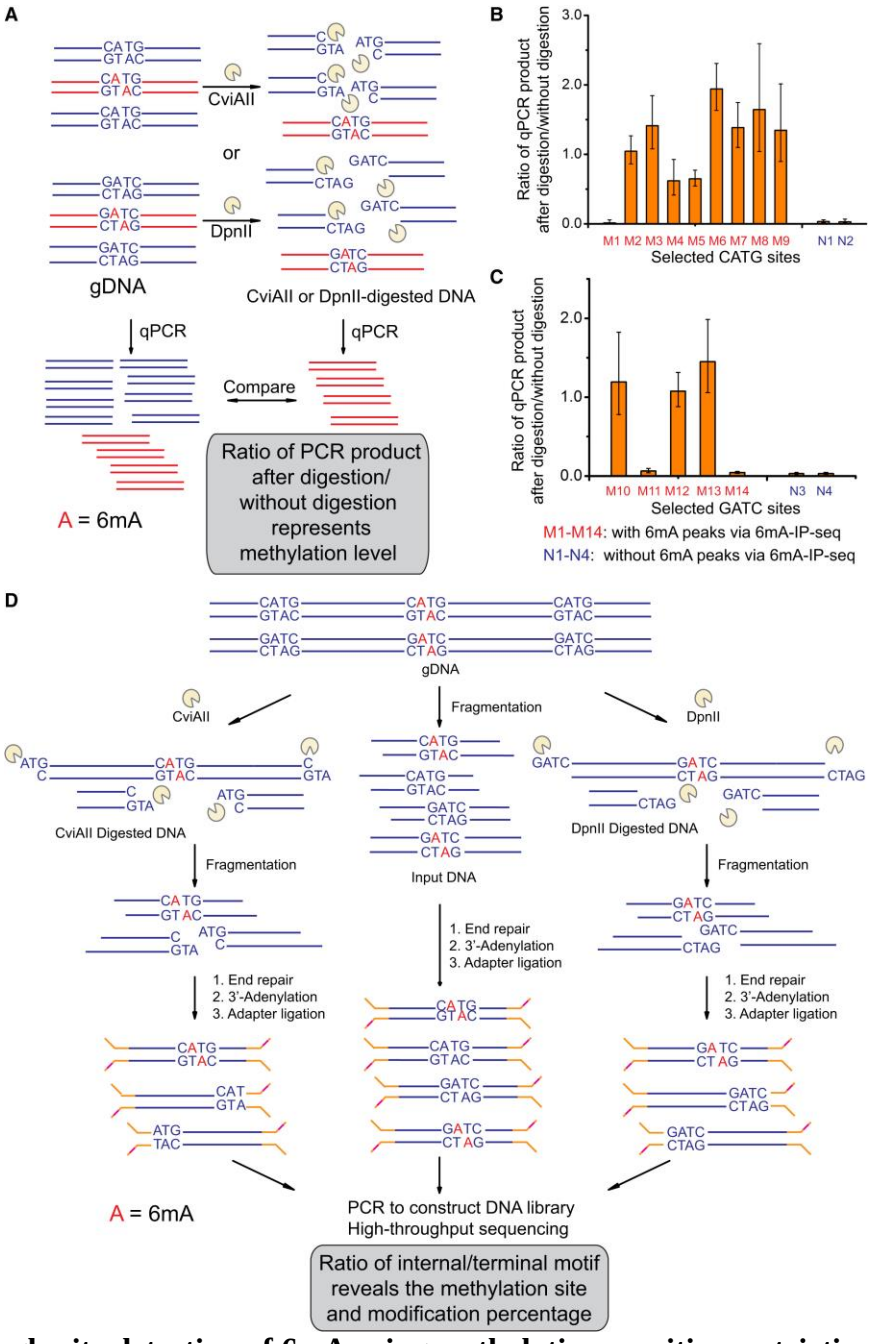



**Figure 4.6. Single site detection of 6mA using methylation sensitive restriction enzymes.** (A) Schematic diagram of 6mA-RE-qPCR for validation of specific 6mA. Restriction enzymes CviAII or DpnII that are sensitive to 6mA methylation in CATG or GATC were used to digest the unmethylated CATG or GATC sites in genomic DNA, respectively. The undigested CATG or GATC sites represent the methylated fraction, and can be PCR-amplified by using primers that cover these sites. (B, C) qPCR results of 11 selected CATG sites and 7 GATC sites validated the accuracy of 6mA-IP-seq. After CviAII- or DpnII-mediated digestion, qPCR was performed using specific primers covering these sites. Relative abundances of undigested CATG or GATC sites were calculated from the ΔCt value between digested and undigested DNA samples

(*Continued*)
(n = 3, mean ± S.E.M.). (D) Schematic diagram of 6mA-RE-seq. gDNA is digested with CviAII or DpnII, sonicated to small fragments around 100 base pair, and constructed into sequencing libraries. The ratio for CATG or GATC internal of sequence reads versus at the end of sequence reads of a specific genomic site represents the relative methylation to unmethylation ratio. An input sample from gDNA without CviAII- or DpnII-based digestion serves as a control.

By applying 6mA-RE-seq to two biologically independent samples of *Chlamydomonas* grown under constant light or dark conditions, we obtained a high-resolution 6mA map of all CATG and GATC motifs in the *Chlamydomonas* genome. As expected, most of the sequencing reads were initiated with ATG or GATC for samples digested by CviAII or DpnII, which resulted from the digestion of unmethylated CATG or GATC sites, respectively (**Figure 4.7B** and **4.7C**). Meanwhile, the intact CATG or GATC motifs that appear internal to the sequencing reads were counted as specific 6mA sites. We developed a bioinformatics algorithm with which to calculate the methylation level of individual 6mA sites within corresponding genomic sequences by calculating the ratio of reads obtained from fragment terminals to total reads of each site. We successfully identified 24,970 and 19,778 C6mATG sites with high confidence (FDR < 0.01) in light and dark samples, respectively. 4,967 and 4,174 high-confidence G6mATC sites were found in the same samples. Among the methylated sites discovered, 15,883 C6mATG sites and 3,337 G6mATC sites were identified from both light and dark samples, showing consistency of the method and reinforcing 6mA as a persistent DNA modification in *Chlamydomonas* (**Figure 4.8A**). These single 6mA sites include methylation sites that we have also validated using 6mA-RE-qPCR (**Figure 4.6B** and **4.6C**, primer sequences shown in **Table 4.1**). The sites without methylation based on 6mA-IP-seq and 6mA-RE-qPCR results were determined to be unmethylated by 6mA-RE-seq as well (**Figure 4.6B** and **4.6C**, primer sequences shown in **Table 4.1**). Approximately 78% (13,076/15,883 for C6mATG and 2,069/3,337 for G6mATC) of the total

detected sites overlap with 6mA peaks identified by 6mA-IP-seq (**Figure 4.8B**). We plotted base-resolution 6mA sites that overlap with corresponding 6mA peaks as identified from 6mA-IP-seq. The 6mA peaks are highly enriched around the identified single 6mA sites, with peak summits right on top of the single 6mA sites (**Figure 4.8C**). In addition, most of these methylation sites are close to 100% methylated, as indicated by the ratio of internal versus terminal sequencing reads (**Figure 4.9A and 4.9B**).

**Table 4.1. Primer sequences for qPCR-verified methylated sites in 6mA-RE-seq**

| Site | Forward primer | Reverse primer |
|---|---|---|
| 1 (Non-methylated) | GCGTGTCCAATCACACAATC | CACATTCGCATAGCTCAGGA |
| 2 (Methylated) | ACTGGGGGAAGTCGTAGACC | GCGTGTCGATGTGAATAACTG |
| 3 (Methylated) | ATTGCACTGGGCAGAAAAAC | ACGCTCCGGACATAACTACG |
| 4 (Partial) | CGACAACCGCGATGTAACT | CGATGTGCTGGAGCATCTAA |
| 5 (Methylated) | GGCACGCGCCAGTTATAGTA | CACTGGGTCTTGCAACGATA |
| 6 (Methylated) | GCAAGTATGTCCGACGCTCT | GCAAATGAGCAACACCACAT |
| 7 (Methylated) | TAGCGTTGTGTGACCTCCTG | GGCTGCTTTAGCTGCGTACT |
| 8 (Methylated) | GCCTCATTAAAGGCATTGGA | GTCACAAGTAGCGGGACCAT |
| 9 (Methylated) | GAACAAGGCCTGTTTTGGAC | GATGCGCATTGCCTGATAGT |
| 10 (Methylated) | TCCCTGGTTATGAGGTGAGC | GCTCGGAGTCTGAGAACAGC |
| 11 (Non-methylated) | TGCCCTGGCTTCTGTAACTC | CGATGAGCGATGATGTCGTA |
| 12 (Methylated) | CGTCGAAGTCCTCCTGTTGT | ATATGAGGAGGCCCCTGAAC |
| 13 (Partial) | TCCGAAGGCTAGGTTGAAGA | GACCTGTTACCCGCCATTTA |
| 14 (Non-methylated) | ACATCCTGCAAATGGAAACG | CAAGTGAGTCGACGAGCAAG |

(*Continued*)

| | | |
|---|---|---|
| 15 (Non-methylated) | GGTTCGTCACGTGTTTGTTG | CTAACTGCAACCGGCACAC |
| 16 (Non-methylated) | CCCGTGCCATATCCCTCT | CGATCTCGACTTCGCTGAC |
| 17 (Non-methylated) | AAGCTGTCAACTCACATGCAA | AGAGCGGTGGTGGCAGTA |
| 18 (Non-methylated) | CTGCTGTTCCAGGCATGATA | CTGCCCAAGAACAGAAGGTG |

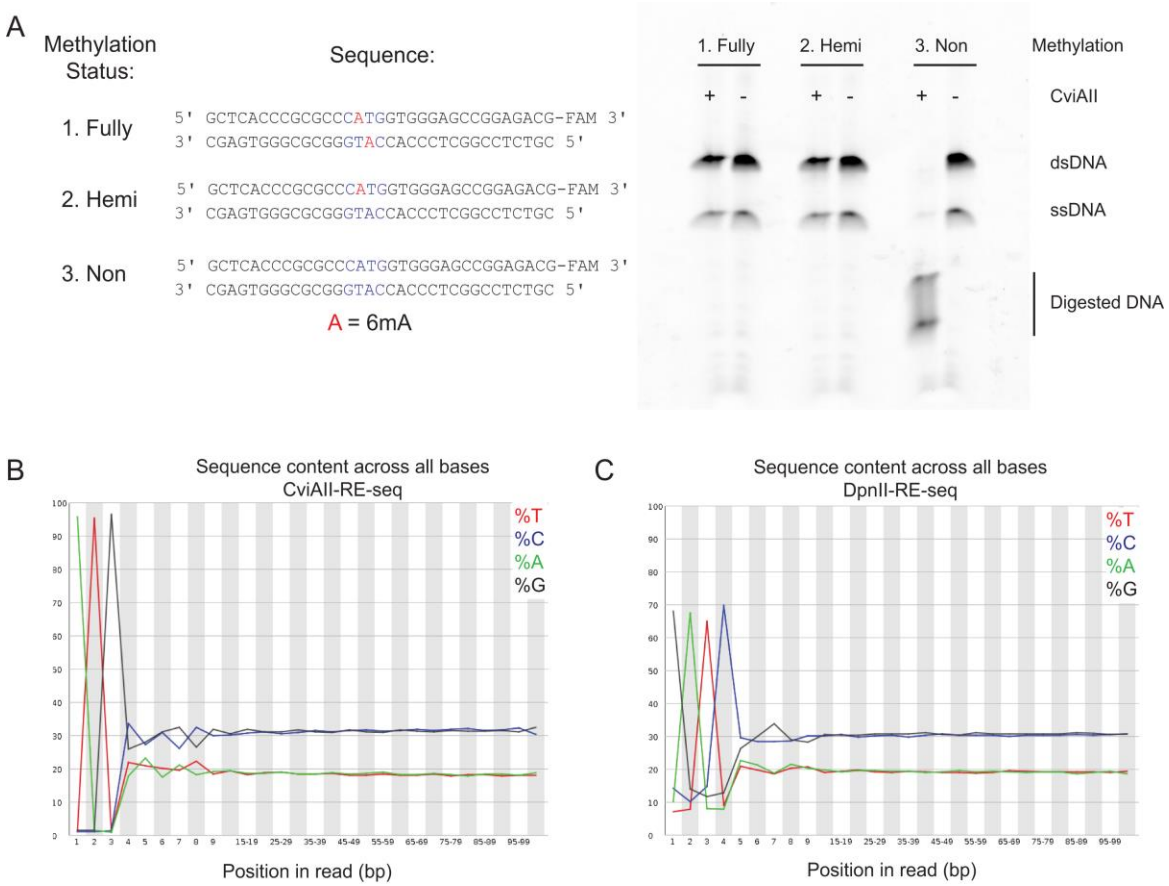



**Figure 4.7. Restriction enzyme digestion to identify 6mA at single-nucleotide resolution.** (A) Specificity of CviAII-mediated digestion on non-methylated DNA. Three DNA probes containing non-methylated, hemi-methylated, and fully-methylated DNA were tested for CviAII-based digestion. Only non-methylated DNA was digested. (B) Sequence content across all bases in the CviAII-digested 6mA-RE-seq showing an enrichment of ATG as the start of sequence reads. (C) Sequence content across all bases in the DpnII-digested 6mA-RE-seq showing an enrichment of GATC as the start of sequence reads.

We performed an extended motif search based on the newly identified sites to examine whether there is any additional preference of nucleotides flanking the CATG or GATC sequence; however, no additional consensus nucleotides were observed (**Figure 4.9C**). Considering the high frequency of CATG and GATC all over the genome (588,209 CATGs and 144,087 GATCs), the methylated sites occupy only 3%-4% of all available motifs. However, the identified CATG and GATC methylations represent ~30% (24,970/85,000) and ~6% (4,967/85,000) of all genomic 6mA sites, respectively. On the other hand, there are ~28% of the 6mA-IP-seq peaks that do not contain any CATG or GATC sequences along the entire genomic regions, indicating the presence of other 6mA sites in distinct sequence contexts besides these two motifs. Interestingly, individual 6mA sites located at these two different sequence contexts tend to cluster in short regions (**Figure 4.9D**). We also observed multiple CATG and GATC motifs in a single peak identified from 6mA-IP-seq, and the peak length linearly correlates with the number of CATG or GATC motifs present in the region (**Figure 4.9E**). Taken together, these results indicate that 6mA methylation occurs mainly to ApT in multiple sequence motifs that tend to cluster together.

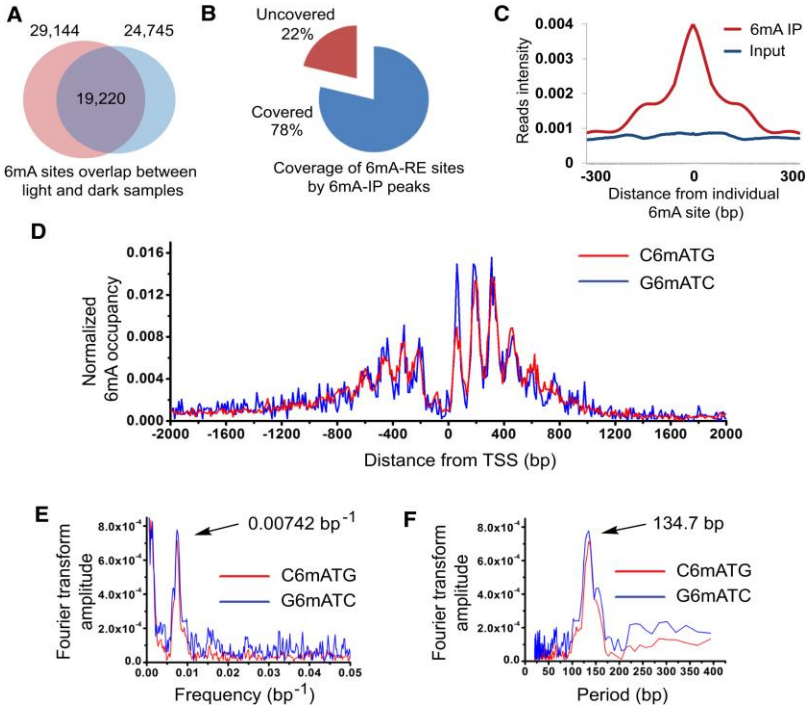



**Figure 4.8. Single-nucleotide resolution map of 6mA.** (A) Overlap of two 6mA-RE-seq samples under light and dark growth conditions. The majority of methylation sites were detected in both samples, indicating the consistency of this method. (B) A majority of the detected single 6mA sites by 6mA-RE-seq are covered by 6mA peaks identified by 6mA-IP-seq. (C) Overlap of 6mA sites identified by 6mA-RE-seq with the 6mA peak identified by 6mA-IP-seq. (D) 6mA occupancy around TSS normalized to the CATG and GATC distribution. A periodic pattern of 6mA around TSS could be observed for both C6mATG and G6mATC motifs. (E) Fourier transformation of 6mA distribution peaks. (F) Periods of the corresponding frequency in Fourier transformation. The dominant period length is 134.7 bp.

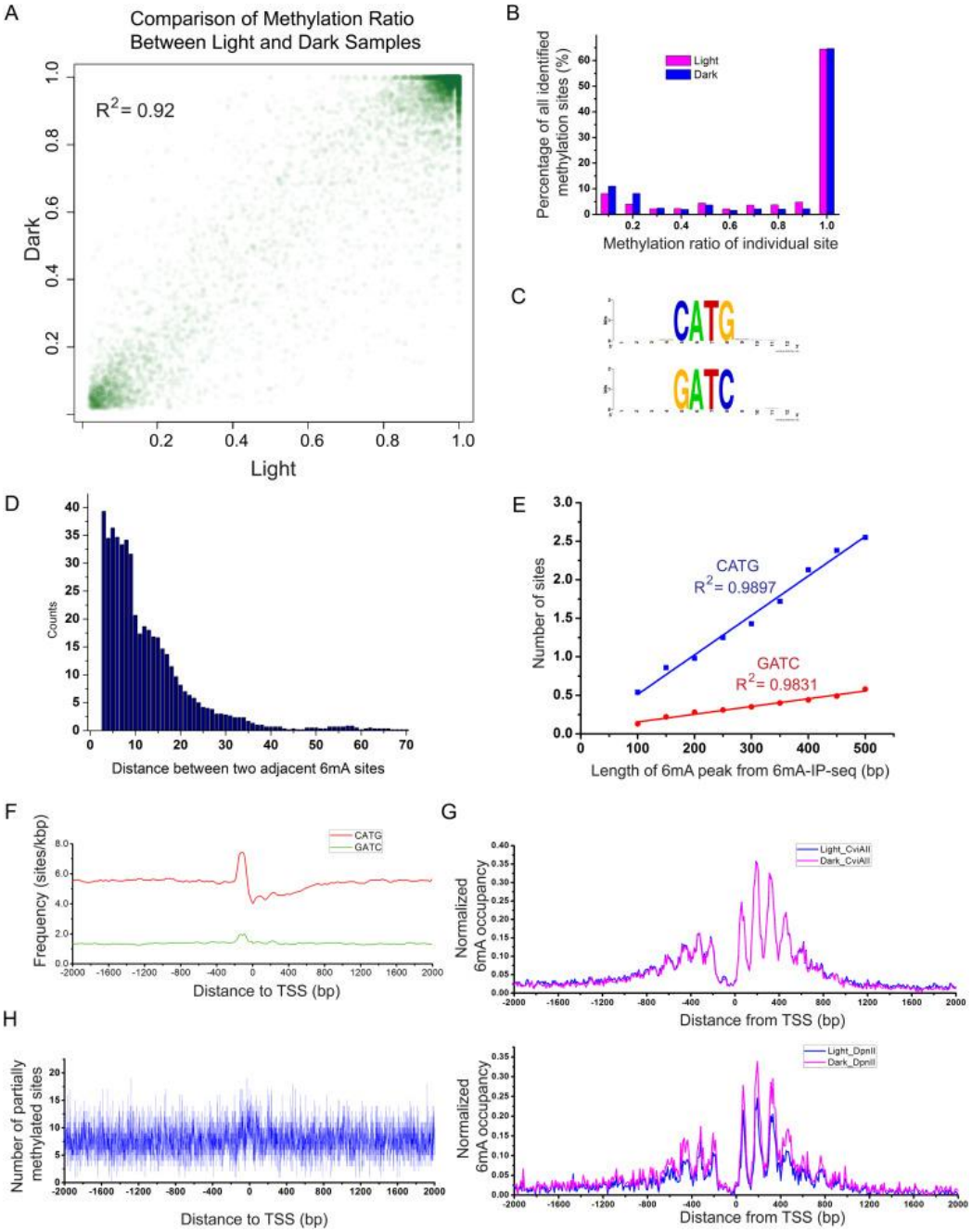



**Figure 4.9. Analysis of 6mA-RE-seq results.** (A) Comparison of methylation sites identified by 6mA-RE-seq in *Chlamydomonas* grown under light and dark conditions in TAP medium. The majority of the identified sites are fully methylated in both samples. Percentage of sites with different levels of methylation (0 = non-methylated, 1 = fully methylated). Density of the spot represents the number of sites in a non-linear scale. (B) A majority (> 90%) of the identified methylation sites by 6mA-RE-seq are fully methylated under both light and dark growth conditions. (C) Extended motif search was performed using 6mA sites identified from 6mA-RE-seq. Only CATG and GATC were detected as a consensus motif. (D) The distance between two closest 6mA sites located at CATG and GATC showing the high tendency for these sites to form clusters mostly within 30 bp to each other. (E) Length of 6mA-IP-seq peak is linearly correlated with

(*Continued*)
the number of CATG or GATC motifs within the peak. (F) Only a slight enrichment of the CATG motif was observed upstream of TSS. Lower frequency of GATC is presented in a similar pattern. This distribution was used to normalize the counts of methylated motifs. (G) The periodic pattern of 6mA around TSS is conserved for both GATC and CATG motifs in the *Chlamydomonas* samples grown under both light and dark conditions. (H) Partially methylated methylation sites determined by 6mA-RE-seq are randomly distributed near TSS.

### 4.2.7 Periodic distribution of 6mA near TSS sites

To further understand the methylation specificity, we calculated the density of individual fully methylated 6mA sites around TSS (over 90% 6mA sites are close to fully methylated). Strikingly, we observed an apparent periodic pattern of 6mA distribution near the TSS region (**Figure 4.8D**). To rule out the possibility that a biased distribution of the CATG or GATC sequences caused the periodic distribution pattern, we normalized the 6mA site frequency according to motif occurrence within each region (**Figure 4.9F**). Of particular note is an obvious discontinuity between peaks upstream and downstream of TSS, which corresponds to a local depletion at TSS (**Figure 4.8D**). Fourier analysis of the periodic profile showed that the frequency is one per 130-140 bp for both downstream and upstream 6mA peaks (**Figure 4.8E** and **4.8F**). The observed periodic pattern is similar to the one observed in the 6mA-CLIP-exo result, which is independent of sequence bias (**Figure 4.4C**). The pattern is also conserved in both biologically independent samples and is independent of culture conditions. Both motif sequences show exactly the same pattern (**Figure 4.9G**). For comparison with the fully methylated sites, we also analyzed the distribution of partially methylated sites (< 60% methylated measured by 6mA-RE-seq, corresponding to less than 10% of all 6mA sites). These partially methylated sites are evenly distributed without any obvious pattern or periodicity (**Figure 4.9H**). It is possible that the occur-

rence of these sites is governed by different mechanisms than those associated with the periodic, peri-TSS sites in *Chlamydomonas*.

### 4.2.8 6mA preferentially locates at linker DNA between two adjacent nucleosomes

The periodic distribution pattern of 6mA around TSS prompted us to study its correlation with nucleosome positioning. We performed nucleosome foot printing followed by high-throughput sequencing to reveal the exact position of each nucleosome in the *Chlamydomonas* genome.[30] Briefly, micrococcal nuclease (MNase) was used to digest unprotected DNA between nucleosomes while leaving the nucleosome-occupied DNA intact; the intact DNA was then subjected to library preparation and high-throughput sequencing. After MNase digestion, the purified DNA showed a clear band with ~150 bp length; the DNA is composed of the nucleosome-protected segments (**Figure 4.10A**). These DNA segments were fully sequenced by paired-end sequencing. The length distribution is enriched around 147 bp (**Figure 4.10B**), which perfectly matches the reported value for *Chlamydomonas*.[31] When we mapped the nucleosomes and 6mA locations to the *Chlamydomonas* genome, we found that most of the 6mA sites locate between two adjacent nucleosomes (**Figure 4.11A**). We then analyzed the statistical distribution of nucleosomes relative to individual 6mA sites, which revealed that the peaks of the closest nucleosomes are enriched ~75 bp upstream and ~78 bp downstream of the 6mA sites (**Figure 4.11B**). This pattern further supports that 6mA is mostly present in regions corresponding to the linker DNA between two adjacent nucleosomes (**Figure 4.11C**). The analysis of nucleosome-6mA correlation also showed that the downstream nucleosomes possess a progression with a steady phase of 170-180 bp periodicity (**Figure 4.10C**), while the upstream nucleosomes are relatively loosely phased, and this tight periodicity disappears around 2 to 3 nucleosomes away from the 6mA site.

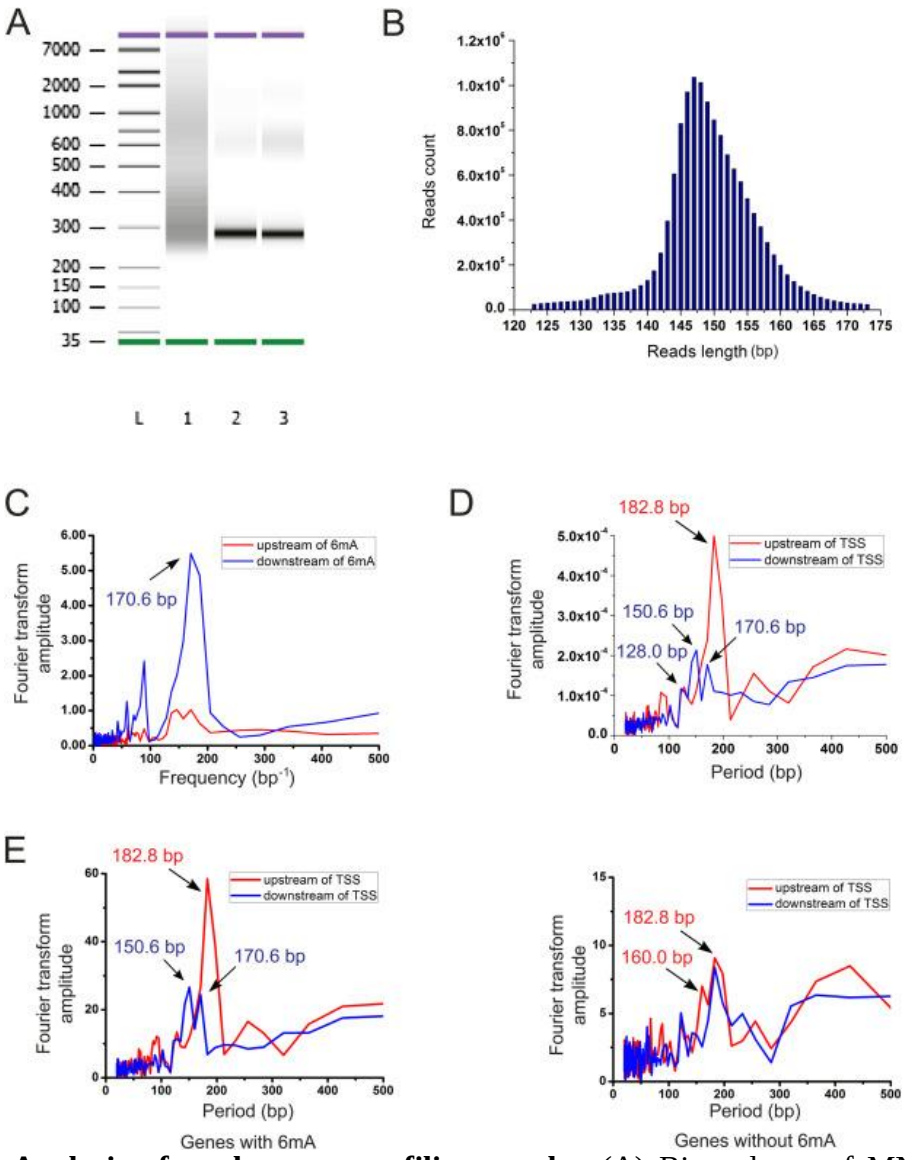



**Figure 4.10. Analysis of nucleosome profiling results.** (A) Bioanalyzer of MNase-digested chromatin DNA which had been ligated with 132 bp adaptor showing an average size of ~280 bp. L represents ladder. Lane 1 is the randomly fragmented input sample. Lane 2 and 3 are nucleosome profiling samples. (B) Length distribution of nucleosome-wrapped reads centering around 147 bp. (C) Fourier transformation of nucleosome profiles up- and down-stream of 6mA sites. The downstream nucleosomes are positioned with a strong periodicity of 170-180 bp. (D) Fourier transformation of nucleosome profiles up- and down-stream of TSS. The upstream nucleosomes show a constant period of ~183 bp, while the downstream nucleosomes exhibit multiple periods which are shorter than the regular period. (E) Fourier transformation of nucleosome profiles up- and down-stream of TSS for genes marked with and without 6mA. Genes without 6mA show lower amplitude of periodicity with a ~183 bp period for both upstream and downstream nucleosomes.

### 4.2.9 6mA may contribute to the positioning of nucleosomes in *Chlamydomonas*

To further understand the relationship between nucleosome distribution and 6mA, we plotted their density around TSS. We found that the periodic pattern of average nucleosome occupancy around TSS in *Chlamydomonas* has distinct features compared to other species (**Figure 4.11D**): first, the density of nucleosomes around TSS is much lower than that in gene body regions and upstream promoter regions; second, the periodicity between two nucleosomes is centered at 183 bp upstream of TSS, but has multiple period values downstream of TSS, including 171, 151, and 128 bp (**Figures 4.10D**). Previous studies of nucleosome distribution in *Chlamydomonas* and other organisms revealed that nucleosome-depleted regions (NDRs) are on average ~155-160 bp around TSS, and nucleosomes downstream of the NDRs are strictly phased in a 165-185 bp period, depending on the length of linker region between two adjacent nucleosomes.[31,32] The multiple periodic values we observed could be a result of convolution between the regular nucleosome periodicity of ~170 bp and the 6mA-influenced periodicity of 130-140 bp downstream of TSS on DNA. Nonetheless, when we compared the nucleosome distribution with the 6mA distribution around TSS, we found that they correlates with each other with ~180 degree phase shift, consistent with our finding that 6mA preferentially locates at linker regions. To probe the relationship of 6mA distribution and nucleosome positioning in detail, we divided all the genes into two groups: with or without 6mA around TSS. Interestingly, nucleosomes phase well for genes that contain 6mA around TSS, whereas the nucleosome phase pattern was weak for genes without 6mA (**Figure 4.10E** and **4.11E**). Taking these results together, we propose a model in which the DNA 6mA modification either restricts or marks the positions of nucleosomes near TSS in *Chlamydomonas* (**Figure 4.11F**). The 130-140 periodic pattern of 6mA leads to out-of-phase distribution and partial occupancy of nucleosomes around TSS. For exam-

ple, if the distance between two adjacent 6mA sites is larger than the length of a nucleosome, such as 270 bp, one nucleosome may reside between two adjacent 6mA, in place depending on the sequence content. If the distance between two adjacent 6mA sites is shorter than 150 bp, such as 135 bp; nucleosome will be missing, leaving a nucleosome-free region between them (**Figure 4.11A** and **4.11F**). The distribution pattern of 6mA may restrict the pattern of nucleosome positioning for each gene, such that the genome-wide pattern of nucleosome is correlated with 6mA distribution pattern.

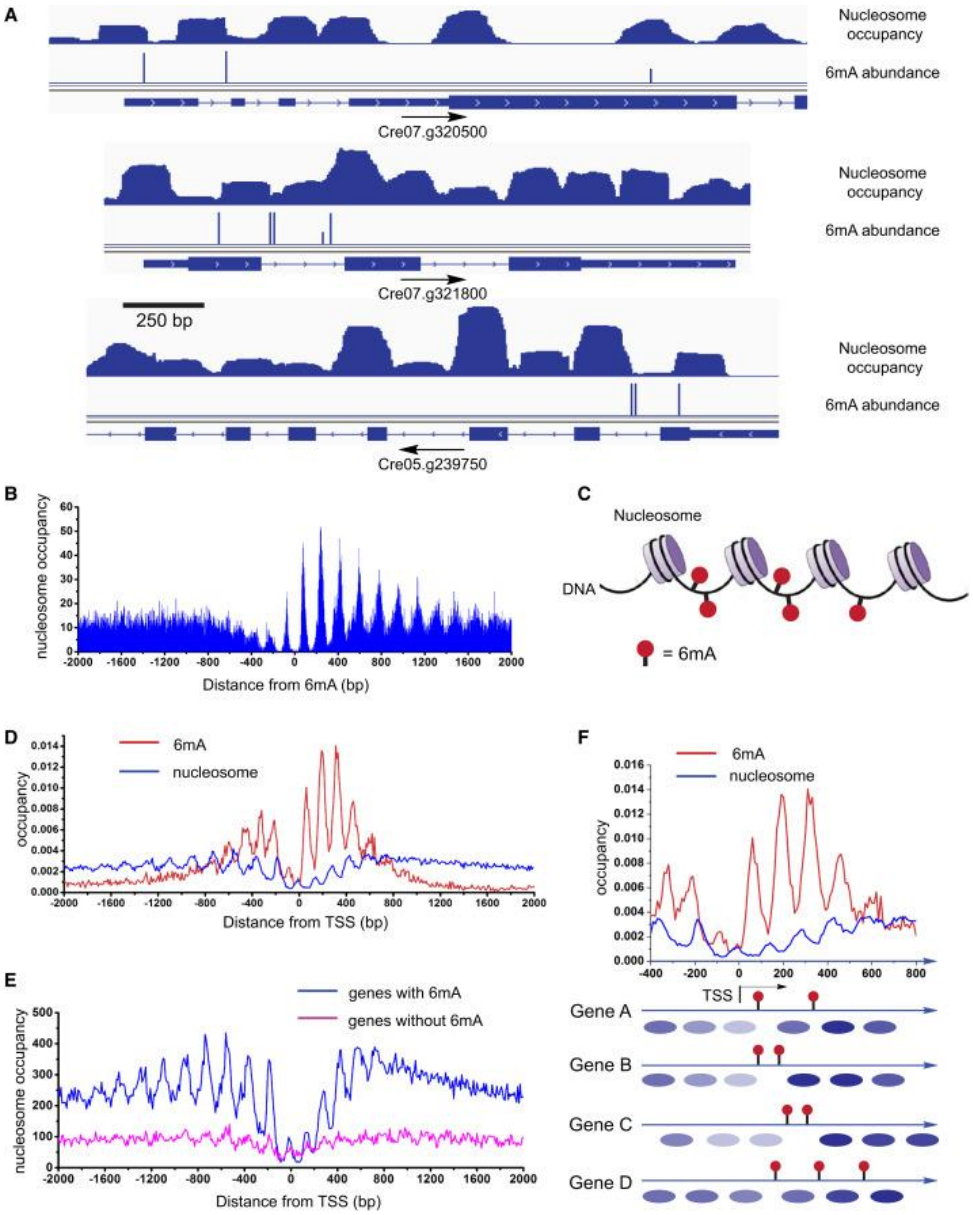



**Figure 4.11. 6mA resides at the DNA linker region between adjacent nucleosomes.** (A) Distribution of nucleosome and 6mA in selected genes. 6mA mainly lies at the boundary region of nucleosomes. Nucleosome occupancy is shown on the first line, 6mA sites identified from 6mA-RE-seq are shown on the second line. Genome annotations are shown on the bottom line. (B) Nucleosome occupancy around 6mA sites. 0 defines the 6mA site, with downstream noted as positive. Nucleosomes reside adjacent to but not on the 6mA site. Nucleosomes downstream of 6mA sites show a constant period of ~170-180 bp. (C) Schematic models of the relationship between nucleosome distribution and 6mA in genomic DNA showing that 6mA mainly distributes in the linker DNA between two adjacent nucleosomes. (D) Distribution profiles of 6mA and nucleosome around TSS showing they are mostly inversely correlated. (E) Nucleosomes exhibit a more consistent phase in relation to TSS in genes marked with 6mA than genes without 6mA. (F) Schematic illustration of the relationship between nucleosome positioning and 6mA location in individual genes. 6mA does not reside on nucleosome-wrapped DNA.

### 4.2.10 6mA marks the TSS regions of actively transcribed genes

The bimodal localization of 6mA around transcription start sites prompted us to investigate its relationship with gene expression. We used RNA-seq to analyze the expression of individual genes. We divided genes into two groups: high expression (80% of all genes) and low expression (20%), and plotted their 6mA peak abundances obtained from 6mA-CLIP-exo experiments (**Figure 4.12A**). We found a general trend that genes with lower expression tend to have low occupancies of 6mA around TSS regions. Specifically, among the 16% of genes without 6mA, ~64% are categorized as low expression or non-active genes. Correspondingly, on a genome-wide level, genes with 6mA around TSS express significantly higher than genes without 6mA (**Figure 4.13A**). The widely-studied 5mC methylation typically plays repressive roles in the regulation of gene expression. However, our results reveal that 6mA marks the TSS regions of actively transcribed genes in *Chlamydomonas*. Studies have shown that 6mA can reduce the stability of the DNA duplex due to the requirement of unfavorable *trans*- configuration for base-pairing. The presence of 6mA may lower the energy required for opening up the DNA duplex.[33] Based on the observed periodic distribution pattern, the tightly controlled deposition of 6mA is associated with nucleosome phasing around TSS. These 6mA modifications could affect nucleosome positioning, or recruit protein factors analogous to methyl-CpG-binding proteins as potential "readers" to impact transcription initiation.[34] Indeed, barley nuclear extract has been shown to contain specific 6mA-binding proteins, and 6mA embedded within GATC at the promoter region can increase the transcription activity of a transfected plasmid.[35]

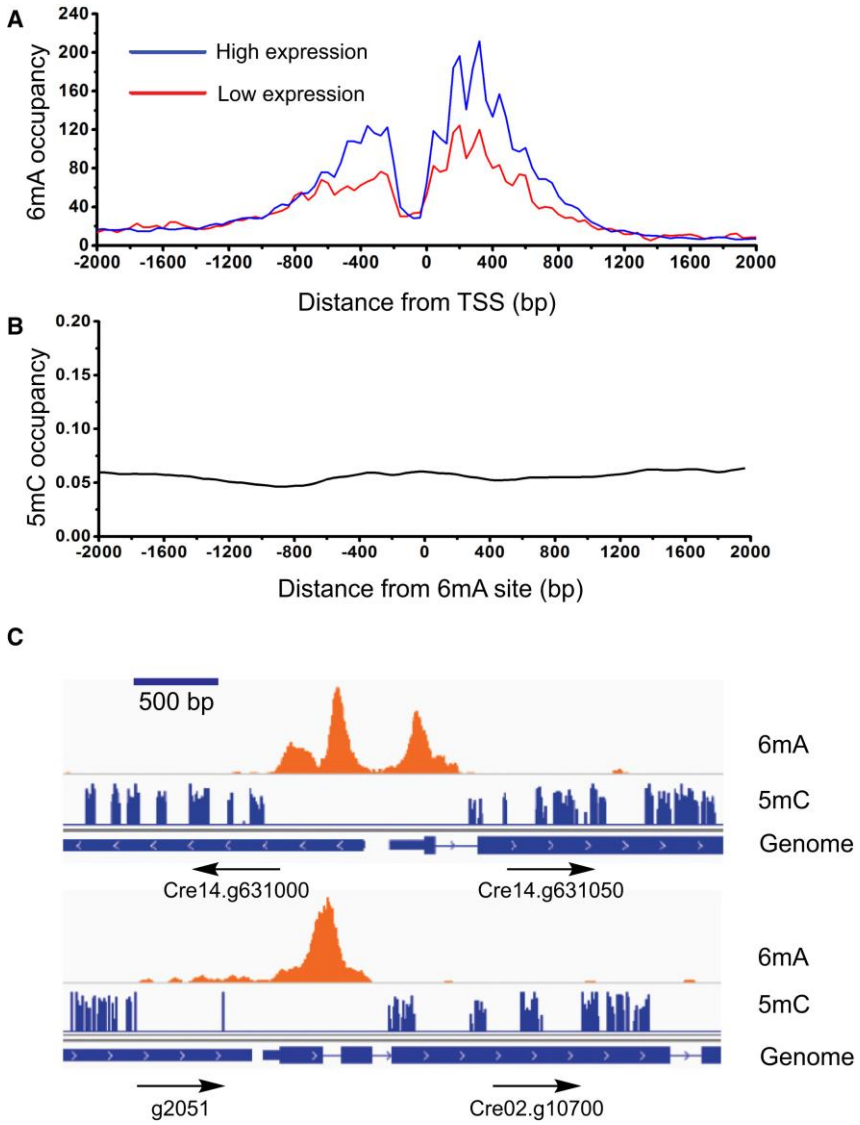



**Figure 4.12. Correlation of 6mA with active genes.** (A) The 6mA methylation is correlated with active genes. Two groups of genes with high (FPKM ≥ 1) and low (FPKM <1) expression levels are plotted with its methylation level determined from 6mA-CLIP-exo-seq. 6mA occupancy represents the reads coverage that are normalized to gene counts of each category in 6mA-CLIP-exo. FPKM stands for Fragments Per Kilobase Of Exon Per Million Fragments Mapped. (B) No correlation was observed between the distributions of 5mC and 6mA. Distance between 5mC and 6mA was plotted, showing no correlation between the two. (C) Selected examples showing that 5mC mainly appears in the gene body, while 6mA mainly resides near TSS region. 6mA peaks identified from 6mA-IP-seq are shown on the first line, 5mC sites identified from previous results are shown on the second line. Genome annotations are shown on the bottom line.

To study potential effects of 6mA on gene regulation, we profiled the mRNA transcriptome of algae cultured under constant light and dark conditions, and found 4,866 differentially expressed genes. In parallel, we used the restriction enzyme based method to quantify the methylation level of individual 6mA site under light and dark conditions. 6mA levels in most genes were similar under both light and dark conditions (**Figure 4.13B**). These results suggest that 6mA is a general mark of TSS regions that could be actively transcribed. Transcription factors and other factors may play more direct roles in determining the exact expression levels of individual genes.

### 4.2.11 6mA and 5mC mark distinct regions in the *Chlamydomonas* genome

As 5mC is also present in high abundance in the *Chlamydomonas* genome, we wondered if any relationship exists between these two DNA base modifications. *Chlamydomonas* has an unusual pattern of 5mC methylation: overall it has less CpG methylation compared to multicellular eukaryotes, but possesses all three types of methylation of CpG, CHG, and CHH enriched in exons of genes, and has only CpG methylation enriched in repeats and transposons.[36] We compared bisulfite sequencing data of 5mC with the 6mA distribution that we generated. There is no specific enrichment pattern of 5mC distribution around TSS regions (**Figure 4.13C**), and 5mC generally do not co-localize with 6mA (**Figure 4.12B**). 5mC appears mostly in gene bodies with a much broader distribution, and is absent near TSS regions (**Figure 4.12C** and **4.13C**). In addition, 5mC has been proposed to be negatively correlated with gene expression in general; we did not observe a strong correlation between the gene expression and 5mC occupancy around TSS region (**Figure 4.13D**).[37] This analysis indicates that 6mA and 5mC are two distinct marks in *Chlamydomonas* genome: 6mA may contribute to chromatin structures that enable initiation of gene transcription, while 5mC may contribute to transposon silencing, imprinting, exon definition and affect transcription elongation.[38]

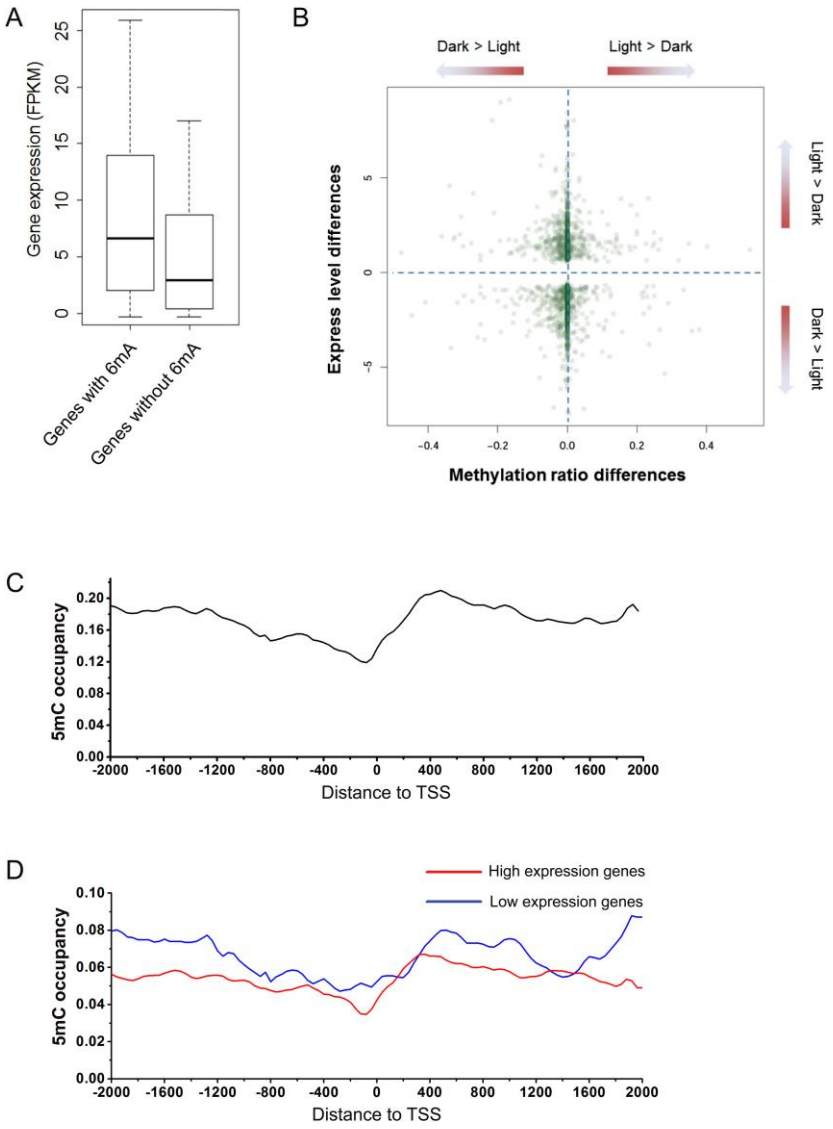



**Figure 4.13. Relationship between 6mA, 5mC, and gene expression in *Chlamydomonas*.** (A) Genes marked with 6mA around TSS exhibit higher expression than genes without 6mA. FPKM stands for Fragments Per Kilobase Of Exon Per Million Fragments Mapped. Box, the first and third quartiles; line, the median; whisker, max/min. (B) RNA sequencing results for both light and dark samples show no significant correlation between methylation level change and RNA level. (C) No enrichment but a slight depletion at TSS region was observed for 5mC. (D) Correlation between 5mC methylation and gene expression level. Two groups of genes with high and low expression levels are plotted with its 5mC level determined from a previous study.[36]

### 4.2.12 Discussion and summary

The 6mA and 5mC modifications are both abundant in the genome of the green algae *Chlamydomonas reinhardtii*. We showed that the total 6mA level is robustly maintained during cell

proliferation. We applied 6mA-IP-seq and further developed 6mA-CLIP-exo to profile 6mA in genome wide using antibodies that specifically recognize and enrich $N^6$-methylated adenine. We found that 6mA mainly resides around TSS with a bimodal distribution. The results from 6mA-CLIP-exo at higher resolution revealed that 6mA deposition occurs mainly at ApT dinucleotides within multiple sequence contexts. At least two sequence motifs, CATG and GATC, are confirmed by a restriction enzyme digestion assay using CviAII and DpnII that are sensitive to 6mA. We then applied this restriction-enzyme-based 6mA-RE-seq strategy to *Chlamydomonas* genomic DNA and obtained genome-wide 6mA maps at single-nucleotide resolution. The identified 6mA sites within these two specific sequences account for ~1/3 of the total 6mA in genomic DNA. 6mA sites within other sequence contexts likely show similar distribution patterns (**Figure 4.4C** and **4.5D**).

The results from the high-resolution maps of 6mA in two specific sequences not only validate the IP-based profiling data but also uncover a periodic pattern of 6mA. This periodicity may mark special features of transcription initiation in *Chlamydomonas* and could be related to nucleosome positioning around TSS. Indeed, we performed nucleosome footprinting coupled with high-throughput sequencing, and the results revealed a periodic pattern of nucleosome occupancy that correlates with the periodicity of 6mA distribution but is ~180 degrees out of phase around the TSS region. The individual 6mA sites exclusively mark the linker DNA between two adjacent nucleosomes. We propose two possible interpretations for this exclusive behavior. One possibility is that, unlike the nucleosome-wrapped DNA, the linker DNA is exposed and can thus be accessed for methylation. The other possibility is that the locations of 6mA sites contribute to the precise positioning of nucleosomes. Our results favor the latter hypothesis for the following reasons: first, we have shown that nucleosomes around TSS sites exhibit very low densities.

Low-occupancy nucleosomes unlikely serve as determining factors for 6mA deposition because it occurs at almost 100% at most of these sites. On the other hand, a high density of 6mA might act to reprogram the positioning of nucleosomes around TSS regions. Second, nucleosomes are likely more dynamic than the covalent 6mA mark on DNA in the TSS regions during transcription initiation. Precedence for a role of base methylation in affecting chromatin structure exists: 5mC has been shown to contribute to nucleosome positioning in other eukaryotes.[32] Additionally, 6mA may mark the TSS region for more efficient transcription initiation. Although it has been well known that the first intron is always important for transgenic gene expression in *Chlamydomonas*, the mechanism was unclear. We provide evidence that 6mA can reside in the first intron (examples shown in **Figure 4.5B**).[39] The periodic distribution, its specific location on the linker DNA between two adjacent nucleosomes at TSS, and its marking of gene activation all suggest that this unique DNA mark contributes to nucleosome positioning and transcription initiation.

We have shown that 6mA shares little correlation with 5mC in the *Chlamydomonas* genome, indicating that they are controlled through different pathways and likely exhibit distinct functions. Our transcriptome analysis found an association of 6mA with gene activation; whereas, 5mC appears to negatively correlate with gene expression. Studies of 5mC have dominated notions of DNA epigenetics in eukaryotes, in particular in vertebrates, because of the critical roles played by 5mC. As shown here, 6mA can also be an important mark that could mark/affect gene activation in eukaryotes. Analogous to 5mC recognition by methyl-CpG-binding proteins, proteins that specifically recognize 6mA at TSS may exist; these proteins could interact with or be part of transcription initiation complexes that contribute to gene activation. It is also possible that 6mA may coordinate with other epigenetic factors such as histone modifications that are also enriched

around the TSS region. Highly dense and narrow distributions of modifications such as H3K9 acetylation (H3K9ac) and H3K4 trimethylation (H3K4me3) near transcription start sites have been associated with constitutive expression of genes involved in translation in *Arabidopsis*.[40] Cooperative interactions among 6mA, histone modification, and transcriptional factors could serve as a general mechanism for transcription activation in *Chlamydomonas* and possibly other eukaryotic organisms.

The *E. coli* Dam DNA methyltransferase methylates the $N^6$ position of adenine at GATC sites. Compared to prokaryotic 6mA modification in genomic DNA, the 6mA methylation in the *Chlamydomonas* genome exists in a more complex manner with multiple potential sequences mainly centered on ApT, resembling eukaryotic 5mC methylation of CpG. The methyltransferases that are involved in establishing or maintaining the patterns of 6mA sites remain to be determined.[41] It should be noted that Greer and Shi *et al* have recently discovered two enzymes, MAD-1 and DMT-1, which can install or remove 6mA in the genome of *Caenorhabditis elegans*, respectively.[42]

In summary, our study has demonstrated that 6mA is an abundant DNA modification in the *Chlamydomonas* genome. It is enriched specifically around TSS and preferentially marks actively transcribed genes. A periodic distribution pattern with depletion at the TSS coupled with an almost exclusive marking of the linker DNA between adjacent nucleosomes indicates a process of controlled deposition, as well as functional roles in nucleosome positioning and transcriptional initiation. Although 5mC is well known to mark gene repression at promoter and enhancer sites in vertebrates, we show in this work that a different DNA base modification, 6mA, flanks TSS and marks actively transcribed genes. The ribose version of 6mA modification (with 2'-OH) exists as the most abundant internal mRNA modification in almost all eukaryotes. It has recently

been shown to be reversible and plays important regulatory functions.[43] We suspect that 6mA could be widely present in eukaryotic genomes as well; in certain species, 6mA may carry important roles in regulating gene expression; in other organisms, 6mA may play complementary roles to 5mC at different stages of development.

## 4.3 Experimental section.

### 4.3.1 Preparation of *Chlamydomonas* genomic DNA

Culture the algae under constant light to exponential growth stage in Erlenmeyer flask in 22 ℃ shaker. Collect the algae and extract genomic DNA (gDNA) by Zymo Quick-gDNA Micro Prep Kit. Sonicate gDNA into ~150 bp fragment by using Bioruptor Pico (following user manual)

### 4.3.2 Anti-6mA immunoprecipitation and UV crosslinking

Set up 500 μL anti-6mA immunoprecipitation reaction as following:

| | |
|---|---|
| Algae gDNA fragment | 10 μg |
| Anti-$m^6A$ polyclonal antibody | 8 μg |
| 5X IP Buffer (50 mM Tris-HCl, pH 7.4, 750 mM NaCl, 0.5% Igepal CA-630 | 100 μL |
| DNase-free $H_2O$ | to 500 μL |

Mix thoroughly but gently. Transfer to head-over-tail rotating wheel to incubate at 4 ℃ overnight. Pre-block Dynabeads protein A magnetic slurry with 1X IP Buffer supplement with 0.5 mg/mL BSA for around 1 hr. Split the reaction mix into 8-10 wells in 96-well cell culture plate. Keep the plate on ice and expose in UV 254 nm six times with energy dosage of 0.15 $J/cm^2$.

### 4.3.3 Enzymatic treatment on beads

Pool the crosslinked fractions together, then transfer each reaction to 80 μL pre-blocked protein A slurry. Mix gently then incubate on head-over-tail rotating wheel for 1.5 hrs at 4 ℃. Wash

the beads with 500 μL ice-cold following buffers once as order: 1X IP Buffer, Exo High Salt Wash Buffer (50 mM HEPES-KOH, pH 7.4, 1 M NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate), Exo Wash Buffer 2 (50 mM HEPES-KOH, pH 7.4, 0.5 M NaCl, 2 mM EDTA, 1% Igepal CA-630, 0.1% sodium deoxycholate), Exo Wash Buffer 3 (10 mM Tris-HCl, pH 7.9, 25 mM LiCl, 2 mM EDTA, 1% Triton X-100, 1% sodium deoxycholate), TE Buffer and 10 mM Tris-HCl, pH 7.9 by gentle vortex. Then, perform the following enzymatic treatment on beads step by step.

Polishing reaction: resuspend the beads in 80 μL 10 mM Tris-HCl, pH 8.0, 10 μL 10X NEBuffer 2, 5 μL 20X BSA (2 mg/mL), 3 μL 10 mM dNTP Mix and 2 μL T4 DNA polymerase; incubate for 30 min at 25 ℃ and then wash the beads as previously described.

Kinase reaction: resuspend the beads in 87 μL 10 mM Tris-HCl, pH 7.5, 10 μL 10X T4 DNA ligase buffer and 3 μL T4 PNK; incubate for 30 min at 37 ℃ and then wash the beads as previously described.

R1 adapter ligation reaction: resuspend the beads in 82 μL 10 mM Tris-HCl, pH 7.5, 10 μL 10X T4 DNA ligase buffer, 5 μL R1 Adapter (40 μM) and 3 μL T4 DNA polymerase; incubate for 2 hrs at 25 ℃ and then wash the beads as previously described. The sequence of R1 adapter is shown below:

R1 adapter upper: 5’-OH-TGGAATTCTCGGGTGCC-OH-3’

R1 adapter lower: 5’-Phos-CCTTGGCACCCGAGAATTCCA-OH-3’

Filling-in reaction: resuspend the beads in 75 μL 10 mM Tris-HCl, pH 7.5, 10 μL 10X Phi29 DNA polymerase Buffer, 5 μL 20X BSA (2 mg/mL), 3 μL 10 mM dNTP Mix and 2 μL Phi29 DNA polymerase; incubate for 20 min at 30 ℃ and then wash the beads as previously described.

Phosphorylation reaction (optional): resuspend the beads in 87 μL 10 mM Tris-HCl, pH 7.5, 10 μL 10X T4 DNA ligase buffer and 3 μL T4 PNK; incubate for 30 min at 37 ℃ and then wash the beads as previously described.

Lambda exonuclease digestion: resuspend the beads in 86 μL 10 mM Tris-HCl, pH 7.5, 10 μL 10X Lambda exonuclease buffer and 4 μL Lambda exonuclease; incubate for 30 min at 37 ℃ and then wash the beads as previously described.

$RecJ_f$ exonuclease digestion: resuspend the beads in 87 μL 10 mM Tris-HCl, pH 7.9, 10 μL 10X $RecJ_f$ exonuclease buffer and 3 μL $RecJ_f$ exonuclease; incubate for 30 min at 37 ℃ and then wash the beads as previously described.

After finishing the enzymatic treatment on beads, add 150 μL ChIP elution buffer (25 mM Tris-HCl, pH 8.0, 200 mM NaCl, 0.5% SDS), mix thoroughly, and incubate for 20 min at 65 ℃ on Heat Shake with 1000 rpm speed. Then, spin down the beads briefly, place the tube on magnetic rack for 30 sec until beads are captured and transfer the supernatant (which contains ChIP complex) in a new tube. Repeat elution step once, and combine the supernatant together. Then, reverse crosslinking by adding 10 μL Proteinase K to the supernatant and incubate overnight at 65 ℃ on Heat Shake with 600 rpm speed.

Next, extract DNA fragment by adding equal volume 25: 24: 1 phenol/chloroform/isoamyl alcohol. Then wash aqueous phase once with chloroform and isolate the DNA fragment by ethanol precipitation. Store the mixture at -80 ℃ for 3 hrs. Spin down DNA fragment precipitation and wash the pellet with 75% (v/v) ethanol, dissolve the pellet into 50 μL DNase-free $H_2O$. Clean up by Zymo Research DNA Clean & Concentrator and elute in 12 μL TE Buffer.

### 4.3.4 Enzymatic treatment in solution

After isolating the DNA fragment from beads, perform the following enzymatic treatment in solution step by step.

Denaturing and primer extension: add 2 μL 10X Phi29 DNA polymerase Buffer, 2 μL 20X BSA (2 mg/mL), 1 μL 10 mM dNTP Mix and 1 μL R1 primer extension (20 μM) to the isolated DNA fragment in TE buffer. The R1 primer extension sequence is:

5’ OH-GCCTTGGCACCCGAGAATTCCA-OH 3’

Incubate the reaction mixture at 95 ℃ for 5 min for denaturing and then 60 ℃ for 5 min for primer annealing. Leave the reaction to cool down to room temperature. Then add 2 μL Phi29 DNA polymerase (total volume 20 μL) and continue to incubate at 30 ℃ for 30 min for primer extension and double stranded cDNA formation, followed by a 10 min 65 ℃ deactivation step.

R2 Adapter ligation reaction: add 3 μL 10X T4 DNA ligase Buffer, 1 μL R2 Adapter (30 μM), 4 μL TE Buffer and 2 μL T4 DNA ligase to double stranded cDNA (total volume 30 μL). Incubate for 2 hrs at 25 ℃

### 4.3.5 dsDNA isolation and purification

Purify the R2 adapter ligated dsDNA by using 52 μL well-mixed AMPure beads. Elute the double stranded cDNA fragment in 30 μL TE Buffer. The sequence of R2 adapter is shown below:

R2 adapter upper: 5’-OH-GTTCAGAGTTCTACAGTCCGACGATC-OH-3’

R2 adapter lower: 5’-OH-GATCGTCGGACTGTAGAACTCTGAACCACCA-OH-3’

### 4.3.6 Library construction, high-throughput sequencing and data analysis

Prepare the following PCR amplification reaction:

| | |
|---|---|
| Purified double stranded cDNA | 15 μL |
| 2X Phusion High-Fidelity PCR Master Mix | 25μL |
| PCR Primer (from Illumina Small RNA Kit) | 2 μL |
| PCR Index Primer (from Illumina Small RNA Kit) | 2 μL |
| $H_2O$ | 6 μL |

Then purify library by using 50 μL well-mixed AMPure beads or agarose gel size selection and purification. The library is applied to Illumina HiSeq single end 50 bp sequencing. Align single-end 50 bp reads to reference genome by Bowtie.[44] Call peaks by using MACE (Model based Analysis of ChIP-exo).[45]

### 4.3.7 Model study

#### 4.3.7.1 6mA incorporated model DNA preparation

Before applying ChIP-exo protocol to 6mA-CLIP-exo, whether anti-6mA antibody is able to recognize and pulldown 6mA-containing dsDNA in immunoprecipitation must be verified. To confirm the IP efficiency and effectiveness, a pair of dsDNA models with or without 6mA incorporated was prepared by PCR amplification. The sequences were chosen from lambda DNA. The primers and sequence of each model are shown below:

Lambda-1 (6mA-containing, using $N^6$-methyl-2'-deoxyadenosine-5'-triphosphate)

Lambda-1F: 5'-CCTGGGCCATGTAAGCTGAC-3'

Lambda-1R: 5'-CCACACCCTGCTTGCTGAG-3'

Lambda-4 (unmodified control)

Lambda-4F: 5'-GCGAGAATTTTTAGCCCAAGC-3'

Lambda-4R: 5'-TCAGCATCTAGCATGCAACC-3'

4.3.7.2 Antibody recognition against 6mA in denatured ssDNA and dsDNA

Lambda-1 and lambda-4 models (each one 4 ng) were spiked in 4 μg fragmented *Chlamydomonas* genomic DNA. Prior to immunoprecipitation reaction, the mixture was split into two halves, one heated at 95 ℃ for 5 min and quickly cooled down on ice to denature the double stranded structure, the other kept on ice. A small portion of each of the two mixtures was kept as qPCR input. Then, two IP reactions were set up as previously described, using the denatured mixture and the untreated one and incubated at 4 ℃ overnight in 1 X IP buffer (150 mM NaCl, 0.1% Igepal CA-630, 10 mM Tris-HCl, pH 7.4) on a head-over-tail rotator. The antibody-DNA complex was captured at 4 ℃ for 2 hrs by 20 μL of Dynabeads® Protein A slurry, followed by 1 X IP buffer wash step. The pulled down DNA was released by proteinase K digestion and purified by phenol chloroform isoamyl alcohol extraction and ethanol precipitation. The pulled down DNA and left over input were applied to qPCR using specific primers to determine the enrichment folds (**Figure 4.14**).

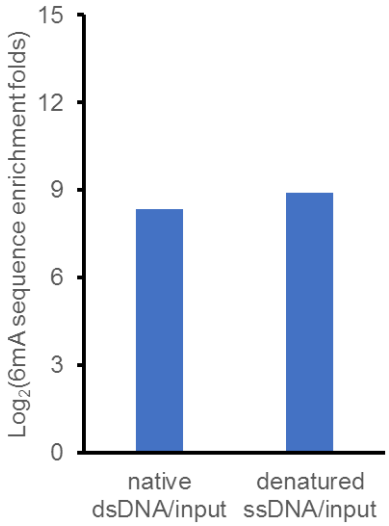



**Figure 4.14. qPCR analysis on antibody recognition against 6mA in denatured ssDNA and native dsDNA.** The 6mA sequence enrichment folds confirm that anti-6mA antibody is able to interact with 6mA in both native dsDNA and denatured ssDNA.

4.3.7.3 6mA containing synthesized DNA oligonucleotide dot blotting assay

The 6mA containing DNA oligonucleotides were synthesized on an Expedite nucleic acid synthesis system on a 1 μmole scale using 0.067 M acetonitrile solutions of phosphoramidites from Glen Research. The sequences are shown below:

Probe 1 (121-1): 5'-TTTTAAATG(6mA)TCAATATCAT-3'

Probe 2 (121-2): 5'-ATGATATTG(6mA)TCATTTAAAA-3'

The oligonucleotides were synthesized, deprotected and purified followed manufacturer's protocol, then quality checked by using MALDI-TOF.

The dsDNA probe was slow annealed on thermocycler in a 50 μL reaction containing 10 mM HEPES-KOH, pH 7.5, 50 mM KCl, 40 μM probe 1 and probe 2. Both ssDNA probe and dsDNA probe were serial diluted and dotted onto Amersham Hybond-$N^+$ membrane (GE Healthcare), subjected to UV 254 nm crosslinking. Anti-6mA antibody was applied to the membrane with 1000-fold dilution in 3% BSA and incubated with membrane at 4 ℃ overnight. The HRP conjugated anti-rabbit secondary antibody was used to afford the chemiluminescence signal, which proved that anti-6mA antibody is able to recognized 6mA in both ssDNA and dsDNA with similar affinity, consistent with previous described qPCR result.

## 4.4 References


1 Bird, A. Perceptions of epigenetics. *Nature* **447**, 396-398 (2007).

2 Sasaki, H. & Matsui, Y. Epigenetic events in mammalian germ-cell development: reprogramming and beyond. *Nat. Rev. Genet.* **9**, 129-140 (2008).

3 Wion, D. & Casadesus, J. $N^6$-methyl-adenine: an epigenetic signal for DNA-protein interactions. *Nat. Rev. Microbiol.* **4**, 183-192 (2006).

4 Law, J. A. & Jacobsen, S. E. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet.* **11**, 204-220 (2010).

5 Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204-220 (2013).

6 Ratel, D., Ravanat, J.-L., Berger, F. & Wion, D. $N^6$-methyladenine: the other methylated base of DNA. *BioEssays* **28**, 309-315 (2006).

7 Collier, J., McAdams, H. H. & Shapiro, L. A DNA methylation ratchet governs progression through a bacterial cell cycle. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 17111-17116 (2007).
8 Low, D. A., Weyand, N. J. & Mahan, M. J. Roles of DNA adenine methylation in regulating bacterial gene expression and virulence. *Infect. Immu.* **69**, 7197-7204 (2001).
9 Lu, M., Campbell, J. L., Boye, E. & Kleckner, N. SeqA: a negative modulator of replication initiation in E. coli. *Cell* **77**, 413-426 (1994).
10 Messer, W. & Noyer-Weidner, M. Timing and targeting: the biological functions of Dam methylation in E. coli. *Cell* **54**, 735-737 (1988).
11 Fang, G. *et al.* Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat. Biotechnol.* **30**, 1232-1239 (2012).
12 Murray, I. A. *et al.* The methylomes of six bacteria. *Nucleic Acids Res.* **40**, 11450-11462 (2012).
13 Cummings, D. J., Tait, A. & Goddard, J. M. Methylated bases in DNA from *Paramecium aurelia*. *Biochim. Biophys. Acta* **374**, 1-11 (1974).
14 Hattman, S., Kenny, C., Berger, L. & Pratt, K. Comparative study of DNA methylation in three unicellular eucaryotes. *J. Bacteriol.* **135**, 1156-1157 (1978).
15 Rae, P. M. & Spear, B. B. Macronuclear DNA of the hypotrichous ciliate *Oxytricha fallax*. *Proc. Natl. Acad. Sci. U.S.A.* **75**, 4992-4996 (1978).
16 Babinger, P., Kobl, I., Mages, W. & Schmitt, R. A link between DNA methylation and epigenetic silencing in transgenic Volvox carteri. *Nucleic Acids Res.* **29**, 1261-1271 (2001).
17 Ehrlich, M. & Zhang, X.-Y. in *J. Chromatogr. Libr.* Vol. Volume 45, Part B (eds W. Gehrke Charles & C. T. Kuo Kenneth) B327-B362 (Elsevier, 1990).
18 Merchant, S. S. *et al.* The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**, 245-250 (2007).
19 Radakovits, R., Jinkerson, R. E., Darzins, A. & Posewitz, M. C. Genetic engineering of algae for enhanced biofuel production. *Eukaryotic cell* **9**, 486-501 (2010).
20 Jia, G. *et al.* $N^6$-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat. Chem. Biol.* **7**, 885-887 (2011).
21 Bisova, K., Krylov, D. M. & Umen, J. G. Genome-wide annotation and expression profiling of cell cycle regulatory genes in *Chlamydomonas reinhardtii*. *Plant Physiol.* **137**, 475-491 (2005).
22 Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* **37**, 853-862 (2005).
23 Dominissini, D. *et al.* Topology of the human and mouse $m^6A$ RNA methylomes revealed by $m^6A$-seq. *Nature* **485**, 201-206 (2012).
24 Meyer, K. *et al.* Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3' UTRs and near Stop Codons. *Cell* **149**, 1635-1646 (2012).
25 Zhang, Y. *et al.* Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, 1-9 (2008).
26 Chen, K. *et al.* High-Resolution $N^6$-Methyladenosine ($m^6A$) Map Using Photo-Crosslinking-Assisted $m^6A$ Sequencing. *Angew. Chem. Int. Ed.* **54**, 1587-1590 (2015).

27 Rhee, H. S. & Pugh, B. F. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr. Protoc. Mol. Biol. / edited by Frederick M. Ausubel ... [et al.]* **Chapter 21**, Unit 21 24 (2012).
28 Vovis, G. F. & Lacks, S. Complementary action of restriction enzymes endo R-DpnI and Endo R-DpnII on bacteriophage f1 DNA. *J. Mol. Biol.* **115**, 525-538 (1977).
29 Zhang, Y., Nelson, M., Nietfeldt, J. W., Burbank, D. E. & Van Etten, J. L. Characterization of Chlorella virus PBCV-1 CviAII restriction and modification system. *Nucleic Acids Res.* **20**, 5351-5356 (1992).
30 Chodavarapu, R. K. *et al.* Relationship between nucleosome positioning and DNA methylation. *Nature* **466**, 388-392 (2010).
31 Lodha, M. & Schroda, M. Analysis of chromatin structure in the control regions of the chlamydomonas HSP70A and RBCS2 genes. *Plant Mol. Biol.* **59**, 501-513 (2005).
32 Huff, J. T. & Zilberman, D. Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. *Cell* **156**, 1286-1297 (2014).
33 Engel, J. D. & von Hippel, P. H. Effects of methylation on the stability of nucleic acid conformations. Studies at the polymer level. *J. Biol. Chem.* **253**, 927-934 (1978).
34 Sternberg, N. Evidence that adenine methylation influences DNA-protein interactions in *Escherichia coli*. *J. Bacteriol.* **164**, 490-493 (1985).
35 Rogers, J. C. & Rogers, S. W. Comparison of the effects of $N^6$-methyldeoxyadenosine and $N^5$-methyldeoxycytosine on transcription from nuclear gene promoters in barley. *Plant J.* **7**, 221-233 (1995).
36 Feng, S. H. *et al.* Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 8689-8694 (2010).
37 Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484-492 (2012).
38 Cerutti, H., Johnson, A. M., Gillham, N. W. & Boynton, J. E. Epigenetic Silencing of a Foreign Gene in Nuclear Transformants of Chlamydomonas. *Plant Cell* **9**, 925-945 (1997).
39 Eichler-Stahlberg, A., Weisheit, W., Ruecker, O. & Heitzer, M. Strategies to facilitate transgene expression in *Chlamydomonas reinhardtii*. *Planta* **229**, 873-883 (2009).
40 Ha, M., Ng, D. W., Li, W. H. & Chen, Z. J. Coordinated histone modifications are associated with gene expression variation within and between species. *Genome Res.* **21**, 590-598 (2011).
41 Iyer, L. M., Abhiman, S. & Aravind, L. Natural history of eukaryotic DNA methylation systems. *Prog. Mol. Biol. Transl. Sci.* **101**, 25-104 (2011).
42 Greer, Eric L. *et al.* DNA Methylation on $N^6$-Adenine in *C. elegans*. *Cell* **161**, 868-878 (2015).
43 Fu, Y., Dominissini, D., Rechavi, G. & He, C. Gene expression regulation mediated through reversible $m^6A$ RNA methylation. *Nat. Rev. Genet.* **15**, 293-306 (2014).
44 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
45 Wang, L. *et al.* MACE: model based analysis of ChIP-exo. *Nucleic Acids Res.* **42**, e156 (2014).

# 5 Deamination based approach to map $m^6A$ at single nucleotide resolution

## 5.1 Introduction

The photo-crosslinking-assisted strategy was also recently employed by other groups to transcriptome-wide map $m^6A$ at high resolution.[1,2] Particularly, Linder *et al* reported an approach to map $m^6A$ in -AC- context at single nucleotide resolution by using monoclonal anti-$m^6A$ antibody which can covalently UV crosslink to cytosine beside $m^6A$ and specifically induce C-to-T mutational signature in high-throughput sequencing.[2] All the crosslinking based methods indeed increase the mapping resolution, providing more insights into the distribution properties of $m^6A$ in transcriptome. On the other hand, however, none of these strategies are independent of immunoprecipitation, nor do they distinguish unmethylated adenosine from $m^6A$ directly. Given the specificity issue and unexpected biases of immunoprecipitation and crosslinking, introducing an antibody free single nucleotide resolution approach to "read out" methylation directly is still an unmet requirement.

In terms of differentiating methylated nucleotide from unmethylated base, bisulfite-sequencing (BS-seq) is one of the most widely used golden standard to map 5-methylcytosine, affording the single nucleotide resolution and quantitative DNA methylation information.[3-6] Bisulfite treatment is believed to have an external nucleophile, $HSO_3^-$, to attack the cytosine ring system at the 5,6-double bond, forming a transient adduct (structure **2** in **Figure 5.1A**). The bisulfite does not further react with the amino group of this cytosine derivative; rather, structure **2** proceeds a deamination reaction in aqueous solution, forming a uracil derivative (structure **3** in **Figure 5.1A**). The basic solution then converts structure **3** back to uracil.[7,8] In contrary, the rate of deamination for 5-methylcytosine (5mC) mediated by bisulfite is much slower than for cyto-

sine, making bisulfite treatment a powerful chemical tool to analyze methylation status in the genome (**Figure 5.1B**).[9,10]

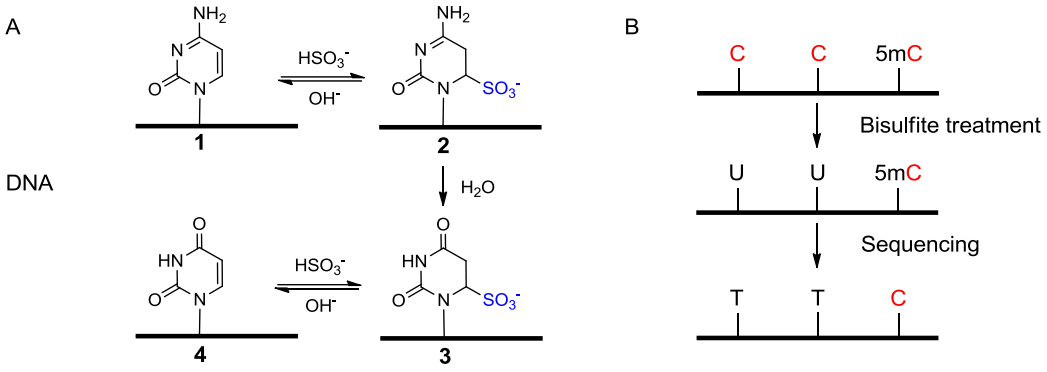



**Figure 5.1. Mechanism of the bisulfite mediated cytosine conversion to uracil (A) and a simplified scheme of bisulfite sequencing (BS-seq) (B).**

Inspired by bisulfite sequencing, we proposed a deamination based approach to distinguish methylated adenosine from unmethylated. To achieve the proposed result, instead of a chemical compound in bisulfite treatment, adenosine deaminase that acting on RNA (ADAR) is employed to deaminate unmethylated adenosine into inosine (I) at a much higher rate than $m^6A$, introducing high-throughput sequencing readable A-to-G transition (Deam-seq). Published studies showing that the $m^6A$ methylation significantly decreases the rate of deamination reaction catalyzed by ADAR theoretically support the proposed approach (**Figure 5.2** and **Scheme 5.1**).[11] Given that ADAR can only work on double stranded RNA (dsRNA) and the A-to-I conversion introduces I-to-T mismatches destabilizing the dsRNA, the deamination efficiency of one-round treatment is found to be very limited. To overcome the drawback to make the strategy suitable for transcriptome-wide $m^6A$ mapping, an iterative deamination based $m^6A$-seq (iterative Deam-seq) is designed and under optimization.

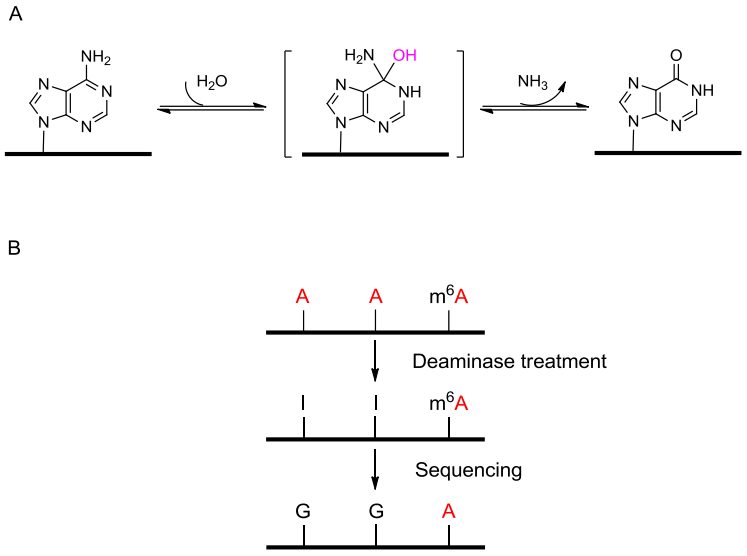
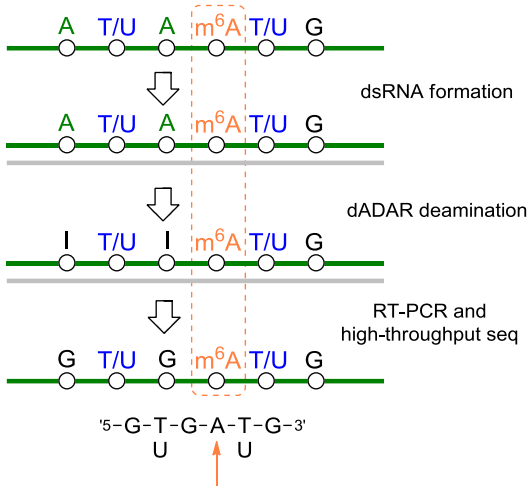



**Figure 5.2. Adenosine conversion to inosine (I).** (A) ADAR mediated, zinc dependent deamination reaction and the proposed "tetravalent" intermediate. (B) A simplified scheme of deamination based $m^6A$ sequencing, comparable with Figure 5.1(B).




**Scheme 5.1. Proposed scheme of Deam-seq.** Given that ADAR can only convert adenosine in dsRNA to inosine, the first step is to form dsRNA using transcriptome as "on-transcript-strand". Then, dADAR is employed to perform deamination on unmethylated adenosine, while $m^6A$ keeps unchanged. Finally, the dADAR treated RNA is subjected to RT-PCR and high-throughput sequencing. The A-to-G transition in sequencing data indicates unmethylated adenosine.

## 5.2 Result and discussion

### 5.2.1 ADAR introduces detectable A-to-G conversion

Adenosine deamination, leading to adenosine-to-inosine (A-to-I) RNA editing, plays key regulatory roles in post-transcriptional process, including mRNA re-coding and alternative splicing. ADAR is well known as an RNA editing enzyme converting adenosine to inosine to introduce codon change.[12-15] The deficit of ADAR leads to severe phenotype.[16-21] The deamination catalyzed by ADAR proceeds a zinc-dependent substitution mechanism which has water as the nucleophile and goes through a "tetravalent" intermediate.[12,22]

The *drosophila* ADAR (dADAR) is chosen as the deamination enzyme to perform transcriptome-wide *in vitro* deamination conversion. Before applying the concept to biological sample, it is necessary to validate the random deamination activity of dADAR on dsRNA. The Sanger sequencing result showed multiple detectable A-to-G transitions compared with original sequence, clearly demonstrating the deamination occurs across the entire transcribed dsRNA model and conceptually confirming the plausibility of the deamination based approach (**Figure 5.3A**).

### 5.2.2 A-to-G conversion ratio depends on $m^6A$ level

To test whether A-to-G mutation ratio is affected by methylation, both methylated model and unmethylated model were subjected to deamination treatment and Sanger sequencing. A-containing target strand and $m^6A$-containing target strand were *in vitro* transcribed by using normal adenosine triphosphate and $N^6$-methyladenosine triphosphate, respectively, and pooled together in a 1:1 ratio as the methylated control group. Both unmethylated group and methylated group were first annealed to complementary strand, and then subjected to ADAR mediated deamination reaction. Afterwards, RNA was purified, and the target strand was selectively converted to cDNA. The cDNA was further amplified by PCR and subjected to TA TOPO-cloning kit.

Single colonies were picked up for sequencing. The results, shown in **Figure 5.3B**, revealed significant differences between the samples with or without methylation. The deamination efficiency (indicated by A-to-G frequency) ranged around 5-15% for the methylated control group with a 50% $m^6A$ content, while the deamination efficiency of the unmethylated control group without $m^6A$ increased up to around 20-35%. The comparison of the two results verified feasibility of deamination based approach to differentiate methylated adenosine from unmethylated adenosine according to the different A-to-G transition frequency.

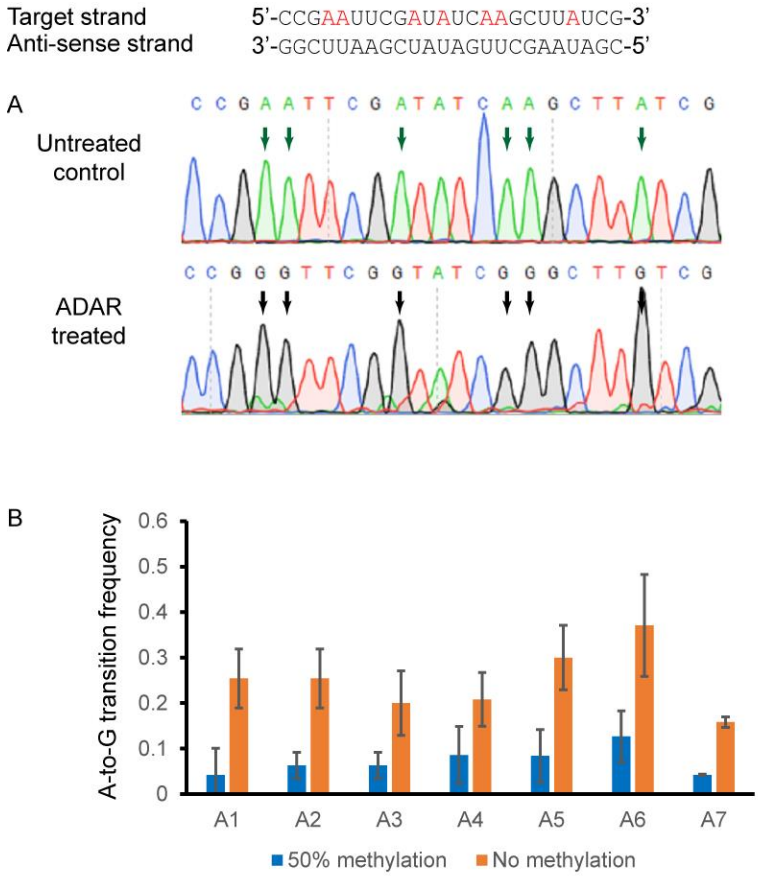



**Figure 5.3. ADAR introduces widely present deamination on dsRNA which is affected by methylation.** The model used is shown on the top. (A) The Sanger sequencing chromatogram of the target strand. dsRNA was first treated by ADAR, then reversely transcribed into cDNA and amplified using sequence specific primers. Sanger sequencing results clearly validate that ADAR introduces multiple sequencing detectable A-to-G mutations. (B) The comparison between 50% methylation model and no methylation model. It is shown that methylation significantly decreases the A-to-G transition frequency, indicating ADAR mediated deamination is hindered by $m^6A$.

### 5.2.3 Iterative deamination treatment increases conversion ratio

The ADAR-mediated adenosine deamination strategy (Deam-seq) has been illustrated by using model RNA (as described above), showing that ADAR is able to widely introduce predictable A-to-G transition in sequencing and distinguish $N^6$-methyladenosine ($m^6A$) from unmethylated adenosine (A). Then, we applied Deam-seq to HeLa total polyA-tailed RNA, testing how the strategy worked in a transcriptome-wide manner. In this pilot experiment, we decided to process two samples in parallel to further verify methylation effect on deamination reaction: HeLa total polyA-tailed RNA (Deam total) vs $m^6A$ antibody enriched HeLa polyA-tailed RNA (Deam $m^6A$). The experiment was conducted as following:

HeLa polyA-tailed RNA was purified by poly-dT magnetic beads and fragmented into ~200-nt length, then applied to T4 polynucleotide kinase mediated end repair. The majority of RNA fragment was used for $m^6A$ antibody enrichment. Deam total and Deam $m^6A$ RNA samples then served as target strand pools to form double-stranded RNA (dsRNA) by using homemade Phi6 RNA replicase, followed by recombinant *drosophila* ADAR treatment to introduce A-to-G transition on unmethylated adenosine. Treated samples were subjected to NGS library preparation and submitted to single-end 50 bp high-throughput sequencing.

The expected A-to-G transition could be observed directly by visualizing the raw data (**Figure 5.4A**) and analyzed by an adapted bisulfite sequencing analysis tool (**Figure 5.4B**, details in 5.3.5 Data analysis). We used the published $m^6A$-seq data from the same cell line to demonstrate the specificity of Deam-seq. In Deam $m^6A$ data, unchanged A sites covered 80% of $m^6A$-seq peaks based on IP experiment, which confirmed the high consistency of the two $m^6A$ identification technologies (**Figure 5.4C**). On the other side, only 3% of A-to-G sites (data not shown), which represent non-methylated A, overlapped with $m^6A$-seq peaks.

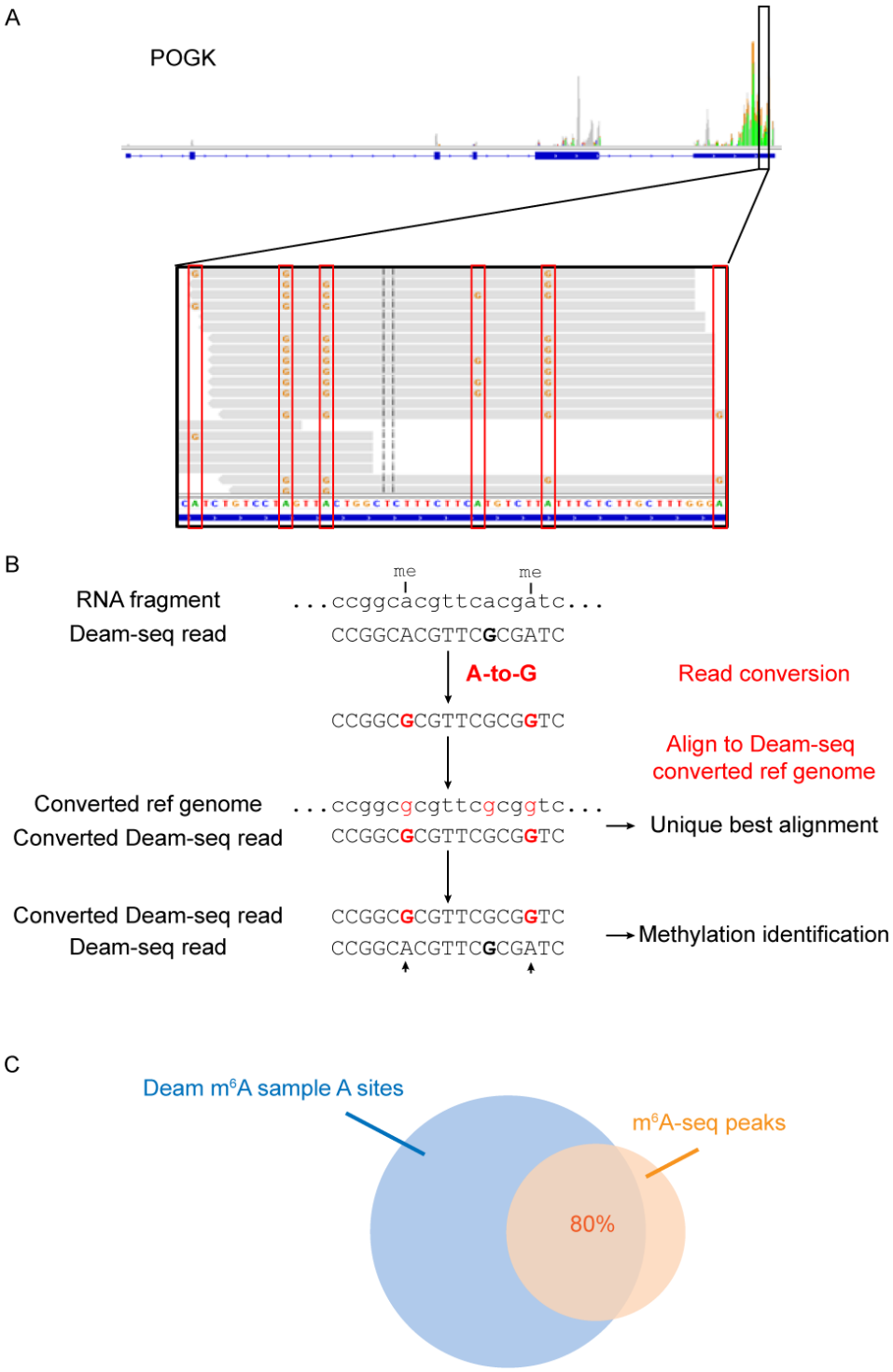
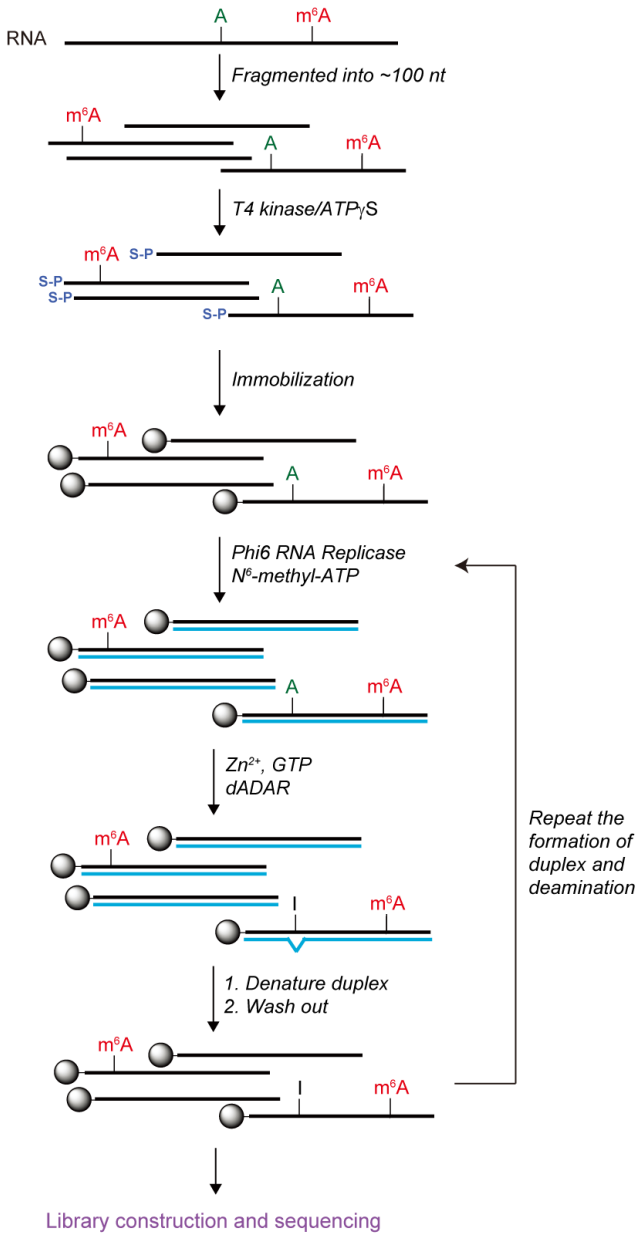



**Figure 5.4. Pilot experiment of Deam-seq.** (A) Visualization of Deam-seq aligned reads on IGV. POGK reads were further zoomed in and all adenosine sites shown were converted to guanosines in raw data. (B) Adapted Bismark analysis workflow demonstrated how A-to-G transition containing read is precisely mapped back to the reference genome. (C) The Venn diagram of overlapping analysis between unchanged A sites in Deam-seq on $m^6A$ enriched polyA-tailed RNA and published $m^6A$-seq peaks.

The pilot experiment validated Deam-seq in high-throughput data and indicated its potential to achieve single nucleotide resolution. However, even transition occurred on around 60% of adenosine sites covered by high-throughput sequencing, the A-to-G transition frequency ratio was quite low, significantly diminishing the power of this strategy. The major effect reducing the deamination ratio is that mismatch introduced destabilized the dsRNA substrate and eventually led dsRNA to falling apart into ssRNA as the enzyme works on it. It is very likely that dsRNA has been separated before ADAR is able to convert all unmethylated adenosine, preventing the enzyme from recognizing RNA molecule as its natural substrate. To overcome this problem, we came up with the iterative Deam-seq approach (iterative Deam-seq, **Scheme 5.2**). In this approach, the target strand is first labelled with a 5'-phosphothioate group by using T4 polynucleotide kinase and ATPγS for further biotinylation and selective capture (**Figure 5.4A**). The dsRNA formation and deamination reactions then are performed on labelled RNA in the first round treatment. After deamination reaction, artificial anti-sense ssRNA is destabilized and can be easily removed, while the target strand is freed for following multiple rounds of treatment. By using this strategy, the target strand is proposed to be deaminated more completely. We then applied this approach to HeLa total polyA-tailed RNA to test our design. The data analysis showed that adding one more round could significantly increase the A-to-G transition frequencies compared with one-round treatment, suggesting the feasibility of iterative Deam-seq approach (**Figure 5.4B**).




**Scheme 5.2. Proposed scheme of iterative Deam-seq.** The fragmented RNA is phosphorylated with phosphothioate group and labeled with biotin. Then, the fragment can be iteratively subjected to dsRNA formation and deamination in order to achieve maximum conversion ratio. Afterwards, biotinylated RNA fragment is captured and purified by streptavidin beads and applied to high-throughput sequencing.

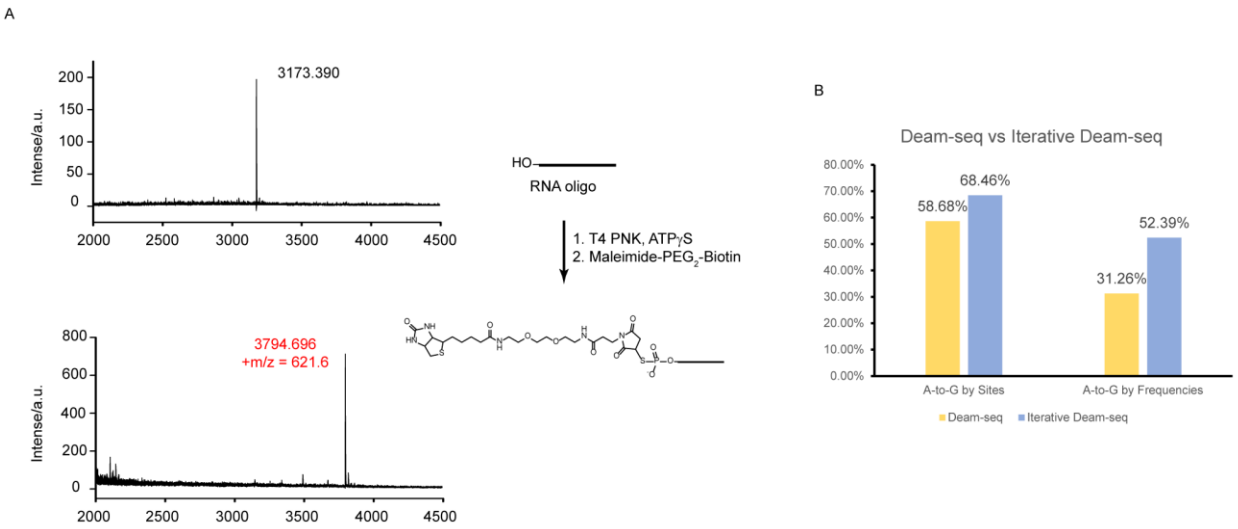



**Figure 5.5. The iterative Deam-seq approach.** (A) MALDI-TOF results validated the phosphothioate transfer and further sulfur-maleimide addition reaction. (B) Iterative Deam-seq pilot experiment on HeLa total polyA-tailed RNA. Adding one more round of treatment significantly increased the transition ratio of both transition site and transition frequency.

### 5.2.4 A-to-G transition is a reliable estimated measure of methylation level

We next wish to verify there is an inverse relationship between A-to-G transition ratio and methylation level, with which the sequencing detectable A-to-G conversion ratio statistic can be interpreted as unmethylation level, leading to quantitative measure of $m^6A$ abundance at single nucleotide resolution.

Given the imperfect A-to-G conversion and the complexity of transcriptome, simple descriptive statistics is not a useful tool to bridge the transition ratio to methylation level as we expected; however, it is good enough to roughly check the trend of the relationship in order to validate the design first. A group of spike-in controls with different $m^6A$ incorporation ratios (0%, 40% and 80% methylation level, respectively) were prepared by *in vitro* transcription and mixed together with same amount. The pooled controls were then subjected to the iterative deamination treatment as described in **5.2.3**; the library was prepared by TruSeq stranded mRNA library prepara-

tion kit and submitted to HiSeq 4000; the data were analyzed by using the pipeline described above.

The average A-to-G transition frequency, which was calculated as the ratio of the number of A-to-G events to the total number of all covered A sites, was used to investigate the relationship with average known unmethylation level. The direct trend relationship between average A-to-G transition frequency (*y*-axis) and average known unmethylation level (*x*-axis), shown as a linear regression in **Figure 5.5**, clearly supported the hypothesis that A-to-G transition is inversely affected by the methylation, suggesting that A-to-G transition is a reliable estimated measure of methylation level.

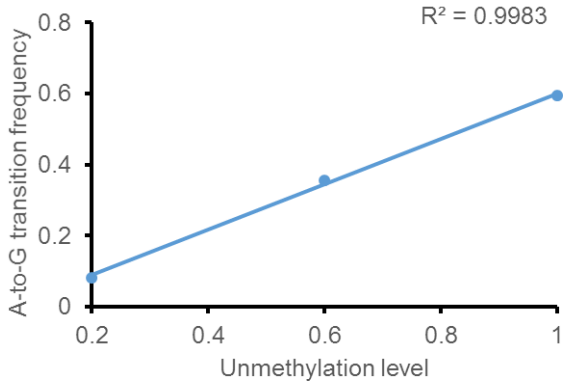



**Figure 5.6. The direct trend relationship between average A-to-G transition frequency (*y*-axis) and average known unmethylation level (*x*-axis).** The high $R^2$ suggests a strong positive correlation.

### 5.2.5 Discussion and summary

The deamination reaction, which is able to convert cytosine to uracil or to convert adenosine to guanosine, introduces readable mutations in high-throughput sequencing data. The Deam-seq is proposed to achieve the single nucleotide resolution $m^6A$ map by differing unmethylated adenosine from $m^6A$ using deaminase ADAR mediated deamination.

The dsRNA model confirmed the random deamination activity of ADAR, and the pilot experiment on HeLa polyA-tailed RNA and $m^6A$ enriched RNA demonstrated the transcriptome-wide plausibility of Deam-seq. The imperfect deamination led us to re-design the procedure, incorporating an iterative treatment strategy to achieve complete A-to-I conversion. The iterative Deam-seq showed significant improvement on deamination efficiency, suggesting the feasibility of the new design and the potential to quantitatively map $m^6A$ at single nucleotide resolution.

It has been proved that A-to-G transition frequency is a reliable measure to estimate the methylation level, indicating the Deam-seq is also able to reveal the methylation level quantitatively. The basic binomial model (under analysis, not shown in thesis) also strongly supports that methylation significantly affects the transition ratio, suggesting that Deam-seq is indeed likely to work on transcriptome as the way bisulfite sequencing works on genome and the current design is on the correct track. It is reasonable to expect deamination based approach is able to map $m^6A$ at single nucleotide resolution transcriptome-wide.

## 5.3 Experimental section

### 5.3.1 Recombinant *drosophila* ADAR expression and purification

The *drosophila* ADAR isoform N gene was a kind gift from Dr. O'Connell's lab. The dADAR gene was amplified, then cloned into a modified pFastBac-Dual vector with a FLAG-tag fused to the N terminus and a $His_6$-tag fused to the C terminus. The construct was verified by Sanger sequencing before used for following steps. Bacmids were generated in DH10Bac cells following the manufacturer's instruction for the Bac-to-Bac baculovirus expression system (Invitrogen), and baculovirus was generated and amplified in adhesive Sf-9 insect cells.

For protein expression and purification, the third passage of baculovirus (P3) was generally used to infect High Five (*Trichoplusia ni*) insect cells grown in SIM HF medium supplemented

with L-glutamine and anti-bacterial and anti-fungal agent. The infected High Five insect cells were incubated at 27 °C for 72 hrs for dADAR expression. Cells were harvested by centrifugation at 2,000 g for 10 min and homogenized in ice-cold lysis buffer containing 50 mM Tris-HCl, pH 8.0, 200 mM KCl and 1X protease inhibitor cocktail using a cell homogenizer. The cell lysate was cleared by centrifugation (13,000 rpm) at 4 °C for 1 hr. The supernatant was then incubated with the pre-equilibrated anti-FLAG M2 affinity gel at 4 °C for 2 hrs. The affinity gel was then washed with lysis buffer. Elution was performed twice by incubating the gel in buffer containing 50 mM Tris-HCl, pH 8.0, 200 mM KCl, and 1 mg/mL FLAG peptide at 4 ℃ for 1 hr. The protein was further purified using Heparin column (GE Healthcare) with a 150 mM to 1 M KCl gradient. The purified dADAR was concentrated and then subjected to gel filtration Superdex 200 with a buffer containing 50 mM Tris-HCl, pH 8.0, 200 mM KCl, 1μM $ZnCl_2$ and 1 mM DTT. The peak fractions were pooled together and concentrated to 1.2-1.5 mg/mL (**Figure 5.6A**).

### 5.3.2 Recombinant Phi6 RNA replicase expression and purification

The Phi6 RNA replicase gene was synthesized at GeneArt (Thermo Fisher). The gene was directly cloned into pMCSG19 vector by using Gibson assembly method. The construct was verified by Sanger sequencing then transformed to PRK1037 competent cell.

For protein expression and purification, the transformant was grown at 37 °C overnight as a starter culture. In the next day it was used to inoculate LB media and grown at 37 °C to an absorbance at 600 nm of 0.8. Then, the media culture was cooled down to 16 ℃ and induced by adding 1 mM isopropyl-β-D-thiogalactopyranoside at 16 ℃ for 18 hrs. The bacterial cells were pelleted by centrifugation and homogenized in lysis buffer containing 20 mM Tris-HCl pH 8.0, 200 mM NaCl and 1 mM phenylmethanesulfonyl-fluoride (PMSF) by a cell homogenizer. The supernatant was subjected to Ni-NTA columns for affinity purification. The protein was eluted in

20 mM Tris-HCl, pH 8.0, 200 mM NaCl and 500 mM imidazole, then directly subjected to Heparin column (GE Healthcare). Elution of the bound protein was performed with a 150 mM to 1 M NaCl gradient buffered with 50 mM Tris-HCl, pH 8.0 and 1 mM EDTA. Fractions containing Phi6 RNA replicase was further purified by Source Q column (GE Healthcare) with a 100 mM to 1 M NaCl gradient. The fraction was pooled together and concentrated for storage (**Figure 5.6B**).

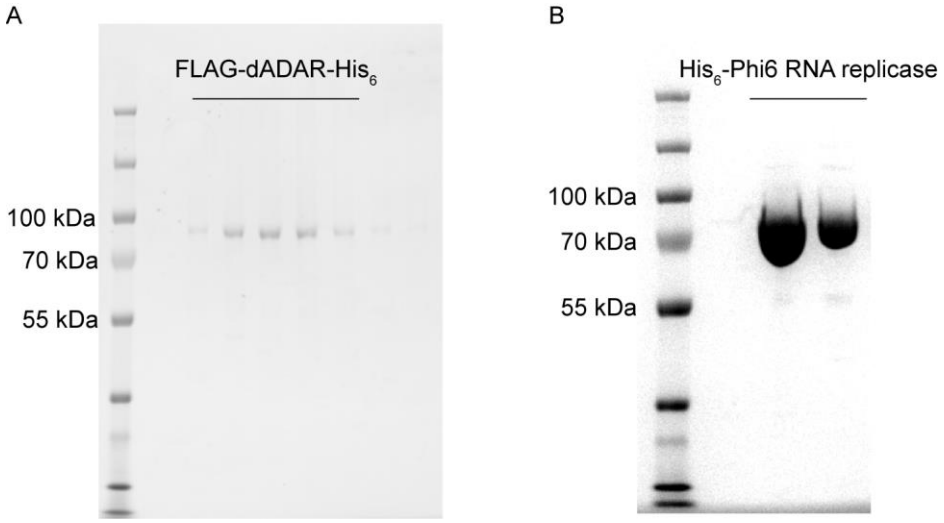



**Figure 5.7. Coomassie brilliant blue staining of protein SDS-PAGE.** (A) Recombinant FLAG-dADAR-$His_6$ purified from insect cell, after gel filtration. (B) Recombinant $His_6$-Phi6 RNA replicase purified from *E. coli*, after source Q ion exchange.

### 5.3.3 Iterative deamination treatment

HeLa polyA-tailed RNA was fragmented into 100-200 nt length with Diagenode bioruptor, then was subjected to Antarctic phosphatase mediated dephosphorylation reaction (NEB) at 37 ℃ for 1.5 hrs to remove 5’ phosphate group. The dephosphorylated fragment was isolated and purified with AMPure beads and re-leased in RNase-free water. T4 polynucleotide kinase (T4 PNK) was used to repair 3’ end of the RNA fragment to 3’-OH group and transfer phosphothioate group to 5’ end. The reaction was performed at 37 ℃ for 2.5 hrs with a supply of 1 mM ATPγS (Enzo Life Sciences).

After AMPure beads purification, the phosphothioate modified RNA fragment is first conjugated to maleimide biotin, then serves as template for Phi6 RNA replicase mediated dsRNA formation reaction. The RNA was first denatured at 80 ℃ for 3 min, cooled down on ice immediately. The reaction was performed in the system containing 1 U RNase inhibitor, 0.2 mM ATP, 0.2 mM CTP, 0.2 mM UTP, 0.6 mM GTP and 1.5 mM $MnCl_2$ buffered with 50 mM Tris-acetate, pH 8.75 and 50 mM $NH_4OAc$ at 32 ℃ for 4 hrs. The RNA was cleaned up with AMPure beads capture and release, followed by dADAR mediated deamination reaction performed in the system containing 1 U RNase inhibitor, 25 mM Tris-HCl, pH 8.0, 100 mM KCl, 10 μM $MgCl_2$, 20 μM $ZnCl_2$ and 1 mM DTT at 30 ℃ for 4 hrs.

The purified RNA was then subjected to dsRNA formation and deamination treatment iteratively in order to increase the conversion ratio.

Besides maleimide-sulfur addition, an alternative biotinylation is also tested (**Figure 5.7)**.

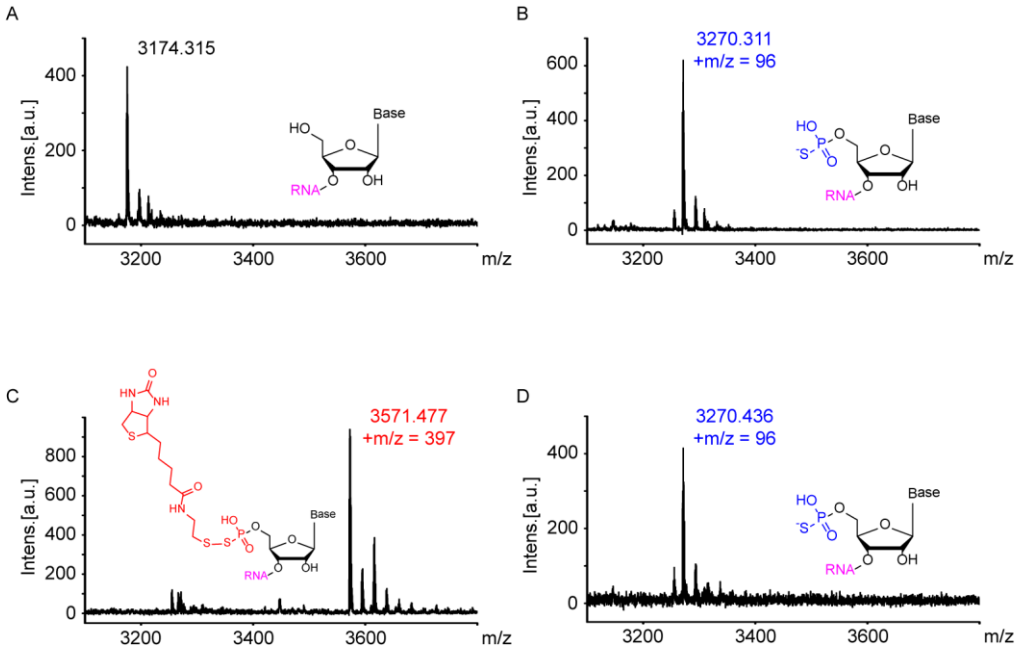



**Figure 5.8. MTSEA-biotin phosphothioate labeling.** The reactions were monitored by MALDI-TOF. (A) The RNA probe control. (B) The phosphothioate added probe. (C) The biotinylated probe. (D) The DTT mediated cleaved probe.

### 5.3.4 RNA selective isolation and purification, library construction

When the iterative treatment is done, biotin conjugated RNA is pulled down by streptavidin C1 beads (Thermo Fisher) at room temperature. The beads are washed followed the manufacturer's protocol. The bound RNA is eluted twice in elution buffer (50 mM Tris-HCl, pH 7.5, 75 mM NaCl, 6.25 mM EDTA, 1% (w/v) SDS containing 1 mM biotin, 2 mg/mL proteinase K) at 55 ℃ for 30 min on heat shake at 1,500 rpm. The eluted RNA is cleaned and released by AM-Pure beads, then subjected to stranded mRNA library preparation.

### 5.3.5 Data analysis

Regarding the fact that Deam-seq original reads contain multiple specific mutations which is very similar to bisulfite sequencing, we chose to adapt Bismark, a widely used bisulfite sequencing analysis tool, to our Deam-seq analysis by performing A-to-G transition instead of its original C-to-T transition.[23] The modified Bismark script is able to uniquely map the raw data back to human transcriptome and report both the transition sites and transition frequency of each site as analysis output, which could be used for further statistical test.

With the adapted Bismark tool, the pipeline to analyze Deam-seq and translate the raw data into frequency is described as following:

Step 1: use a homemade script to convert the raw data `.fastq` file to its reverse complementary sequence `.fastq.rc`. The Illumina TruSeq stranded mRNA library preparation kit specifically labels the transcriptome to effectively distinguish the first-strand cDNA from the second strand, keeping the stranded information in raw data. Given the great likelihood that Deam-seq libraries contain both "sense" (the transcriptome) and "anti-sense" (the artificial "reverse complement-ome") reads, it is of great necessity to process all reads to make sure that the reads are in their original directions as how they are annotated in genome.

Step 2: apply the adapted Bismark tool to map the converted raw data `.fastq.rc` file back to reference genome. In this step, both the raw reads from Step 1 and the reference genome are first transformed to replace all A sites with G sites, then two parallel alignment instances between original and converted versions of raw data and reference enable us to precisely and uniquely map the reads; meanwhile, conversion status of each A site is also recorded based on the parallel alignment, which is further reported as the A-to-G transition frequency.

Step 3: employ Bedtools intersect to extract the transcriptome oriented A-to-G transition frequency for further statistical analysis. Even the reverse complementary strand in dsRNA could be deaminated in the treatment, only the strand with the same direction as the transcriptome is the target for $m^6A$ analysis. Using the transcriptome based reference, Bedtools intersect tool efficiently extracts the "on-transcript-strand" part and reports in a separate file recording the conversion status of each A sites covered by raw data. An example report is shown in **Table 5.1**.

**Table 5.1. Transcriptome direction A-to-G report "on-transcript-strand"**

| Categories of A sites on-transcript-strand | Counts |
|---|---|
| A->A only Events [a] | 8,576,879 |
| A->A only Sites [b] | 2,217,080 |
| A->G only Events [c] | 9,234,159 |
| A->G only Sites [d] | 2,110,312 |
| A&G both A->A Events [e] | 25,384,512 |
| A&G both A->G Events [f] | 28,141,722 |
| A&G both Sites [g] | 2,702,153 |

*Notes:*
[a] A->A only Events: on A sites which do not have A-to-G mutation detected, the covered A reads counted as "events".

[b] A->A only Sites: the A sites which do not have A-to-G mutation detected.
[c] A->G only Events: on A sites which have all A read as G, the covered G reads counted as "events".
[d] A->G only Sites: the A sites which have all A read as G.
[e] A&G both A->A Events: on A sites which have both A and G detected, the covered A reads counted as "events".
[f] A&G both A->G Events: on A sites which have both A and G detected, the covered G reads counted as "events".
[g] A&G both Sites: the A sites which have both A and G detected.

### 5.3.6 Model study

#### 5.3.6.1 Phosphothioate transfer and maleimide biotin addition

To test the efficiency of phosphothioate transfer and maleimide biotin addition reaction, a 10-mer synthesized RNA oligonucleotide (purchased from IDT) was used as the model. A 20 μL phosphothioate transfer reaction containing 1X T4 PNK reaction buffer, 1 U RNase inhibitor, 50 μM RNA oligonucleotide, 1 U T4 PNK and 1 mM ATPgammaS was incubated at 37 ℃ for 2 hrs, then cleaned up by spin column. The maleimide biotin addition was performed at 37 ℃ for 1.5 hrs with 2 mM maleimide-$PEG_2$-biotin (final concentration). The RNA oligonucleotide was purified by spin column and the molecular weight change was monitored by MALDI-TOF. The result clearly demonstrated that both the phosphothioate transfer reaction and maleimide biotin addition reaction were complete and efficient.

#### 5.3.6.2 Preparation of spike-in controls with different methylation levels

We used lambda DNA to prepare *in vitro* transcription templates. The primers were designed to be compatible with T7 promoter based on the manufacturer's protocol (Thermo Fisher). Each template had around 1 kilo-base length. The primers are shown below:

T7-1 (0% $m^6A$ spike-in control)

T7-1F: 5'-TAATACGACTCACTATAGGGAGAGGTTTTTCGTCATGTTTTGAGTCT-3'

T7-1R: 5’-TCAGCATCTAGCATGCAACC-3’

T7-2 (40% $m^6A$ spike-in control)

T7-2F: 5’-TAATACGACTCACTATAGGGAGACCCTGAACTGTTGGTTAATACGCTTGAG-3’

T7-2R: 5’-CCACACCCTGCTTGCTGAG-3’

T7-3 (80% $m^6A$ spike-in control)

T7-3F: 5’-TAATACGACTCACTATAGGGAGATACCCGTCGTGGCTCTAATTCCGA-3’

T7-3R: 5’-AGATGACAACTTCCGCCATC-3’

The *in vitro* transcription reaction was performed as described in protocol; the $m^6A$ level in total rATP was adjusted by combining 100 mM $N^6$-methyladenosine-5’-triphosphate (TriLink Biotechnologies) with normal rATP. The reaction was incubated at 37 ℃ overnight, then added 1 μL Turbo DNase to chop off DNA template. The transcribed RNA was cleaned up by MEGA-clear Transcription Clean-Up Kit (Thermo Fisher). The difference in methylation level was confirmed by dot blotting (**Figure 5.8**).


**Figure 5.9. Dot blotting of spike-in controls.** The *in vitro* transcribed spike-in controls were detected by anti-$m^6A$ antibody. The chemiluminescence signals confirm the methylation levels.

The three RNA spike-in controls were pooled together in equal quantity, then fragmented into 100-200 nt length with Diagenode bioruptor for further iterative deamination treatment and high-throughput sequencing. The data was analyzed by using the pipeline described previously.

## 5.4 References


1 Ke, S. *et al.* A majority of $m^6A$ residues are in the last exons, allowing the potential for 3' UTR regulation. *Genes Dev.* **29**, 2037-2053 (2015).

2 Linder, B. *et al.* Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat. Methods* **12**, 767-772 (2015).

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

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

5 Lister, R. *et al.* Highly Integrated Single-Base Resolution Maps of the Epigenome in *Arabidopsis*. *Cell* **133**, 523-536 (2008).

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

7 Hayatsu, H., Wataya, Y. & Kai, K. The Addition of Sodium Bisulfite to Uracil and to Cytosine. *J. Am. Chem. Soc.* **92**, 724-726 (1970).

8 Shapiro, R., Servis, R. E. & Welcher, M. Reactions of Uracil and Cytosine Derivatives with Sodium Bisulfite. *J. Am. Chem. Soc.* **92**, 422-424 (1970).

9 Hayatsu, H. & Shiragami, M. Reaction of bisulfite with the 5-hydroxymethyl group in pyrimidines and in phage DNAs. *Biochemistry* **18**, 632-637 (1979).

10 Hayatsu, H., Wataya, Y., Kai, K. & Iida, S. Reaction of sodium bisulfite with uracil, cytosine, and their derivatives. *Biochemistry* **9**, 2858-2865 (1970).

11 Véliz, E. A., Easterwood, L. M. & Beal, P. A. Substrate Analogues for an RNA-Editing Adenosine Deaminase: Mechanistic Investigation and Inhibitor Design. *J. Am. Chem. Soc.* **125**, 10867-10876 (2003).

12 Bass, B. L. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* **71**, 817-846 (2002).

13 Jin, Y., Zhang, W. & Li, Q. Origins and evolution of ADAR-mediated RNA editing. *IUBMB Life* **61**, 572-578 (2009).

14 Nishikura, K. Functions and Regulation of RNA Editing by ADAR Deaminases. *Annu. Rev. Biochem.* **79**, 321-349 (2010).

15 Savva, Y. A., Rieder, L. E. & Reenan, R. A. The ADAR protein family. *Genome Biol.* **13**, 1-10 (2012).

16 Brusa, R. *et al.* Early-Onset Epilepsy and Postnatal Lethality Associated with an Editing-Deficient *GluR-B* Allele in Mice. *Science* **270**, 1677-1680 (1995).

17 Higuchi, M. *et al.* Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature* **406**, 78-81 (2000).

18 Keegan, L. P., Gallo, A. & O'Connell, M. A. The many roles of an RNA editor. *Nat. Rev. Genet.* **2**, 869-878 (2001).

19 Palladino, M. J., Keegan, L. P., O'Connell, M. A. & Reenan, R. A. A-to-I Pre-mRNA Editing in *Drosophila* Is Primarily Involved in Adult Nervous System Function and Integrity. *Cell* **102**, 437-449 (2000).

20 Tonkin, L. A. *et al.* RNA editing by ADARs is important for normal behavior in *Caenorhabditis elegans*. *EMBO J.* **21**, 6025-6035 (2002).

21 Wang, Q., Khillan, J., Gadue, P. & Nishikura, K. Requirement of the RNA Editing Deaminase ADAR1 Gene for Embryonic Erythropoiesis. *Science* **290**, 1765-1768 (2000).
22 Polson, A. G., Crain, P. F., Pomerantz, S. C., McCloskey, J. A. & Bass, B. L. The mechanism of adenosine to inosine conversion by the double-stranded RNA unwinding/modifying activity: a high-performance liquid chromatography-mass spectrometry analysis. *Biochemistry* **30**, 11507-11514 (1991).
23 Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572 (2011).

## 6 Concluding remarks and future outlooks

Our understanding of nucleic acid modifications has evolved over the past several decades. We now appreciate that DNA methylation, as a bona fide epigenetic marker, is not only inheritable and dynamic, but also involved in diverse regulatory processes. Aberrant DNA methylation patterns are known to be associated with many types of human diseases including cancer.[1,2] The imbalance of the DNA methylation, which was previously thought to be caused by the dysfunction of methylation machinery, now is also considered to be induced by the abnormal status of demethylation machinery through TET-mediated active 5mC demethylation. In human cancer cells, 5hmC is largely depleted and the expression levels of TET genes tend to be reduced.[3,4] Both 5mC and 5hmC could serve as disease markers for early diagnosis and prognosis. The recent discoveries of 6mA as a functional DNA mark in eukaryotic genomic DNA raise the possibility that 6mA plays regulatory roles complementary to 5mC. With a better understanding of 6mA methyltransferase and 6mA demethylase and discovery of potential reader proteins, DNA methylation looks to be a ubiquitous epigenetic marker in almost all kingdoms of life. The interplay between 5mC and 6mA in species that contain both marks, and the transition from 6mA to 5mC in regulating certain biological processes are fascinating subjects that remain to be investigated not only on the functional aspects but also with the perspective of evolution.

Unlike genomic DNA, RNA has more complicated post-transcriptional processing: RNA splicing significantly increase the complex of gene expression by alternatively joining exons and removing introns; RNA editing alters the nucleoside sequence of specific transcript, which may or may not change protein coding regions or potential splicing sites to further diversify the transcriptome; RNA chemical modifications, most of which do not affect nucleotide sequence, are much more diverse and functionally versatile, suggesting broader functional impacts (**Figure**

**6.1**).[5,6] While tRNA and rRNA modifications as well as mRNA cap methylations have been studied in the past, it was only recently that the internal $m^6A$ methylation in mRNA was shown to be reversible.[5,7-10] Recent studies have uncovered this mRNA methylation as a new realm of biological regulation at the post-transcriptional level. It is now believed that internal modifications, such as $m^6A$, are distributed in unique patterns and affect multiple RNA metabolic processes in order to impact gene expression. While some modifications may not be reversed in RNA, other methylations on heteroatoms resembling $m^6A$ could be more broadly spread and be dynamically/reversibly regulated by specific enzyme systems. They could affect RNA metabolism and function via RNA structure alteration or recognition by specific reader proteins. As new modifications and new functions continue to emerge, these chemical marks on RNA may collectively provide additional tuning that affect biological outcomes at the post-transcriptional level. With the development of new approaches to quantitatively analyze RNA modifications in a transcriptome-wide manner, a quantitative picture of how chemical modifications affect gene expression regulation and their effects in various human diseases will emerge. RNA modifications may very likely mirror histone modifications: multiple chemical marks on biomacromolecules that dynamically controlled by multiple enzymes and proteins to enable synergistic regulation of the metabolism, processing and function of the target RNA.

The extraordinary progress in high-throughput sequencing technology provides us a powerful tool to investigate the characteristics of chemical modifications on nucleic acids. Most chemical modifications do not disturb classical Watson-Crick base pairing, impeding the direct "readout" in sequencing data. Thus, a number of techniques and approaches have been developed to make the high-throughput sequencing applicable to the modifications research, including immunoprecipitation which employs the specific antibody to enrich modification containing frag-

ments and chemical and/or biochemical treatment which is designed to widely convert either modified nucleotide or unmodified nucleotide to another base, introducing "sequencable" transition/mutation. Both of these strategies have achieved great success, leading us to genome-wide or transcriptome-wide understanding of chemical modifications.

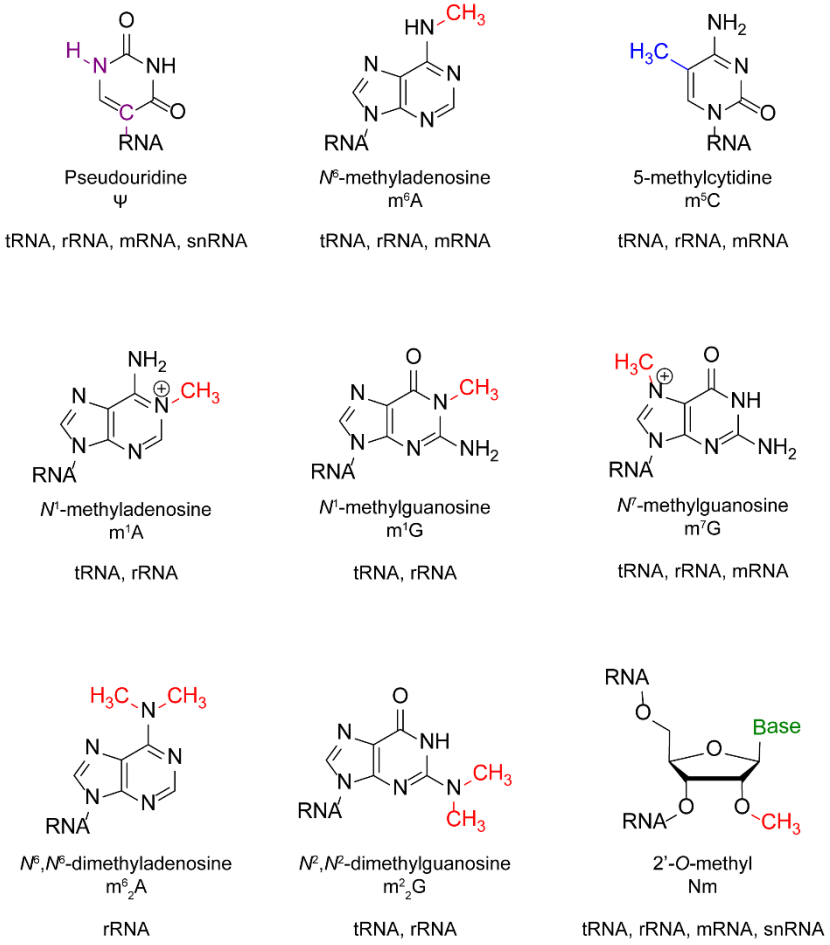



**Figure 6.1. A partial spectrum of diverse RNA chemical modifications.**

Recent research progress has revealed that $N^6$-methylation on adenine base in transcriptome and genome is dynamically tuned by methyltransferases, demethylases and specific binding proteins, getting involved in various regulatory and metabolic pathways. To provide transcriptome-wide or genome-wide pictures at high resolution, both immunoprecipitation based and deamination based approaches were developed or are under optimization.

The photo-crosslinking-assisted strategy coupled with nuclease digestion significantly improve the specificity and efficiency of immunoprecipitation, generating narrowed down detection regions which enable the discovery of modification containing consensus sequence and increase the sensitivity and accuracy. This strategy is compatible with and can be easily adopted to other nucleic acid modification studies.

The deamination based approach, on the other hand, is proposed to directly convert unmethylated adenosine in transcriptome to inosine and leave methylated adenosine unchanged, resulting in a bisulfite treatment-alike sequencing method. Currently available data validated the approach and demonstrated the likelihood of mapping $m^6A$ at single nucleotide resolution, and the recent iterative treatment has significantly increased the A-to-G conversion ratio.

Overall, our understanding on nucleic acid modifications still remains limited. The requirement to have more efficient and effective tools to study these modifications is unmet. This area provides numerous new opportunities for chemical biologists, including but not limited to introducing chemical strategies to map the modifications coupled with high-throughput sequencing, investigating the underlying installation and uninstallation mechanisms, manipulating the modification status to affect gene expression, and developing small molecules or other means to tune these pathways for fundamental research and therapeutic purposes in the future.

**References**


1 Feinberg, A. P., Ohlsson, R. & Henikoff, S. The epigenetic progenitor origin of human cancer. *Nat. Rev. Genet.* **7**, 21-33 (2006).

2 Jones, P. A. & Baylin, S. B. The Epigenomics of Cancer. *Cell* **128**, 683-692 (2007).

3 Jin, S.-G. *et al.* 5-Hydroxymethylcytosine Is Strongly Depleted in Human Cancers but Its Levels Do Not Correlate with IDH1 Mutations. *Cancer Research* **71**, 7360-7365 (2011).

4 Yang, H. *et al.* Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. *Oncogene* **32**, 663-669 (2013).

5 He, C. Grand challenge commentary: RNA epigenetics? *Nat. Chem. Biol.* **6**, 863-865 (2010).

6 Slotkin, W. & Nishikura, K. Adenosine-to-inosine RNA editing and human disease. *Genome Med.* **5**, 105 (2013).
7 Fu, Y., Dominissini, D., Rechavi, G. & He, C. Gene expression regulation mediated through reversible $m^6A$ RNA methylation. *Nat. Rev. Genet.* **15**, 293-306 (2014).
8 Pan, T. *N*6-methyl-adenosine modification in messenger and long non-coding RNA. *Trends Biochem. Sci.* **38**, 204-209 (2013).
9 Jia, G. *et al.* $N^6$-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat. Chem. Biol.* **7**, 885-887 (2011).
10 Zheng, G. *et al.* ALKBH5 Is a Mammalian RNA Demethylase that Impacts RNA Metabolism and Mouse Fertility. *Mol. Cell* **49**, 18-29 (2013).