THE UNIVERSITY OF CHICAGO

NOVEL APPROACHES IN DNA METHYLATION ANALYSIS: SEQUENCING STRATEGY DEVELOPMENT AND APPLICATION TO CANCER METASTASIS STUDIES

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

QINZHE LIU

CHICAGO, ILLINOIS

MARCH 2024

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEGEMENT

First and foremost, I want to extend my deepest gratitude to my advisor, Professor Chuan He. From our first meeting back in 2017, I have been continually inspired by his exceptional intelligence and his genuinely supportive nature. Transitioning from a chemistry background to a bio-focused lab was challenging, but Professor He's guidance and encouragement were pivotal in helping me adapt and grow. His willingness to engage in thoughtful discussion at any hour can always bring me with new insights and motivation. When I struggled with depression and thought about the shift from scientific research to consulting, Professor He offered not just advisement but also compassionate support. His empathy, understanding, and inspirational conversations laid the foundation for my entire graduate study and the successful transition to a new career path.

I also wish to express my appreciation to my thesis committee members, Professor Weixin Tang and Professor Yamuna Krishnan. Professor Tang's collaboration on one of my main projects provided not only scientific insights but also shaped my attitude and approach as a scientist. Professor Krishnan, who supported me through both my candidacy exam and dissertation defense, has been a constant source of encouragement and guidance throughout my journey at UChicago as a graduate student.

I want to send thanks to my collaborator, Professor Feng Yue from Northwestern University and Professor Yiliang Ding from John Innes Centre. Our collaborative efforts have not only enriched my research but have also provided me with broader perspectives and innovative approaches to scientific inquiry.

I am also thankful to all my lab colleagues. Thank Dr. Jun Liu, Dr. Tong Wu, and Dr. Qiancheng You, for their patient mentorship during my early days in the lab and in the scientific research. Their dedication and exemplary work ethic have been inspiring, and their guidance invaluable in teaching me invaluable lesson on managing scientific projects.

Beyond their roles as mentors and teachers, they have also been true friends and supportive colleagues.

The initiation of the projects mentioned in my thesis coincided with the quarantine period, which brings unique challenges. Prof. Qing Dai, Dr. Chang Ye, Dr. Lisheng Zhang, Dr. Pingluan Wang and Dr. Xiaolong Cui were instrumental in guiding me on how to initiate, lead, and manage these projects under such constraints. Their advice, despite the physical separation imposed by the quarantine, was critical in the development of these project.

I am also grateful to the other members of our lab who, while not directly involved in my projects, provided support in both my research and personal life. A special mention to Mr. Zhongyu Zou, a friend of mine since our undergraduate days, whose mutual support in academic and everyday life has been a constant source of strength. Further, my sincere thanks go to Mr. Yuhao Zhong, Dr. Yu Xiao, and Dr. Yun Gao for their insightful experimental discussions and the enriching daily conversations that have greatly enhanced my research experience. I am also deeply grateful to our lab manager, Dr. Jordi Tauler, whose assistance in handling numerous operational aspects allowed me to focus intently on my research. I always believe the power of a community where members freely share their thoughts and offer mutual support. I have been fortunate to be part of such an incredible lab environment, learning and growing amongst a group of exceptionally talented and kind colleagues.

Thank the faculty member of Chemistry department, especially Ms. Melinda Moore and Ms. Vera Dragisich. They have been provided me with unlimited resources and support throughout my entire Ph.D. process. I could not make it this far without their help.

Last but not least, my heartfelt thanks to my family and friends. The unconditional love and support from my parents have been the bedrock of my achievements. Their belief in my potential and aspirations has been a constant source of encouragement and motivation.

My friends have provided companionship and support whenever needed. Pursuing a degree in a new country is challenging, and I could not have made it without their encouragement.

ABSTRACT

The study of epigenetics, particularly DNA methylation, provides crucial insights into gene regulation and its aberrations in diseases such as cancer. This dissertation is anchored in the exploration of DNA methylation, a fundamental epigenetic modification critically involved in cellular differentiation, developmental processes, and tumorigenesis. To advance the exploration of methylation landscapes, we have developed a novel enzymatic-based, bisulfite-free sequencing strategy, TT-5mC-seq. This method achieves single-base resolution in the detection of 5-methylcytosine (5mC) modifications, markedly reducing background noise without causing damage to the DNA. On the other hand, we have characterized the patterns of 5-hydroxymethylcytosine (5hmC) across various human tissues, both normal and cancerous, revealing the subtle yet consequential shifts in the epigenetics associated with disease states. This research provides not only new methodologies for probing DNA methylation but also enhances our understanding of 5hmC's potential as a biomarker for cancer metastasis.

1 Introduction

1.1 Epigenetics: beyond the nucleotide sequence

The remarkable diversity of life on earth is a direct result of the vast gene diversity found within populations of organisms. Nucleic acids — DNA and RNA — are the cornerstone of genetic inheritance, encoding the blueprints of life within their sequences of four canonical bases. DNA harbors genes, the fundamental units of heredity, which serve as templates for RNA molecules. RNA, in turn, guide the synthesis of proteins, which perform biological functions and allow organisms to interact with their environment.

Despite the significance of the nucleotide sequences of A, T, C, and G in DNA, and A, U, C, and G in RNA, these sequences alone do not fully account for the elaborate regulatory processes observed within cells. Phenotype changes were observed even without changes in the primary nucleic acid sequence. Given that, an additional tier of control is necessary—one that refines the genomic instructions without altering the DNA sequence itself. This tier of regulation, termed "epigenetics," was first brought out by C.H. Waddington in 1942[1]. The concept of epigenetics has evolved significantly over the decades. Originally used to describe the complex interactions between genes and their developmental manifestations, epigenetics now encompasses a broader spectrum of biological regulation with a revised definition as' *stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence*'[2]. This field fills the gap between the static nature of the genetic code and the dynamic nature of gene expression, influenced by both external environmental factors and intricate cellular mechanisms.

The core concept of epigenetics involves the study of molecular processes that regulate the accessibility and readability of genetic information, without changing the DNA sequence itself. This includes a wide array of mechanisms, including histone modifications, DNA modifications, and RNA regulatory factors (including non-coding RNA regulation). These

processes regulate gene expression collaboratively, combining in a wide regulatory network[3, 4]. Histone modifications, for instance, play a pivotal role in chromatin remodeling[5], while DNA modifications like methylation can directly influence gene transcription[6]. RNA-based mechanisms add another layer of regulation, impacting gene expression post-transcriptionally[7].

In development, epigenetic mechanisms are crucial for cellular differentiation, enabling cells with identical DNA to develop into diverse cell types[8]. In the context of disease, aberrant epigenetic modifications, particularly in the context of cancer, have been implicated in tumor genesis and progression[9]. Furthermore, epigenetics provides a framework for understanding how environmental factors such as diet can leave lasting impressions on gene expression[10]. This dynamic interplay between genetics and epigenetics is key to understanding not just the static blueprint of life but also the fluid and adaptable nature of gene regulation.

In this thesis, I will delve deeper into the realm of a crucial subfield of epigenetics-DNA modifications, exploring their distribution, biological functions, and dynamics. The following sections will highlight the role of DNA methylation in oncogenesis and its implications in epigenetic regulation, providing a comprehensive view of this crucial aspect of epigenetics.

1.2 The spectrum of DNA modifications in epigenetic regulation

Epigenetic modifications are integral to the dynamic regulation of genetic information, transcending the static DNA sequence to control gene activity and cellular function. Among the three primary arrays of epigenetics, histone modification, DNA modification, and RNA-based regulation, DNA modifications is a critical player with a widespread influence on gene function.

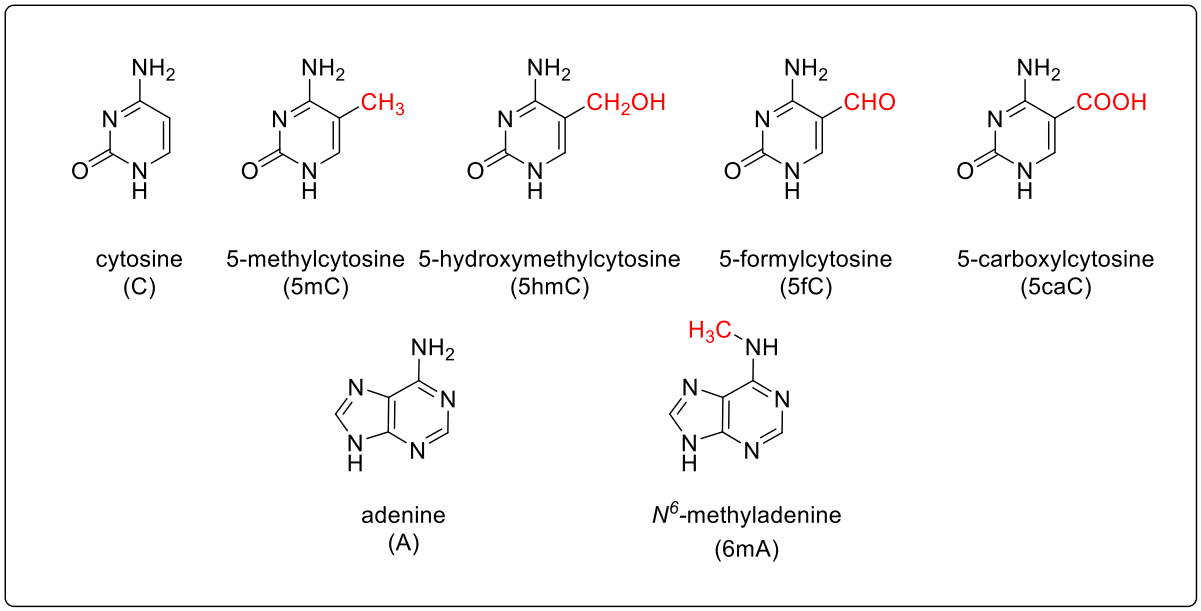


Figure 1.1 Chemical modifications of DNA
Cytosine modifications (5-methylcytosine with its oxidative derivatives) and adenine modification (N^6-methyladenine/6mA) in DNA are the well-studied examples of these DNA modifications.

Cytosine modifications include 5-methylcytosine (5mC), known for its role in gene silencing, and its oxidized derivatives such as 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC), which are all involved in DNA demethylation and gene regulation[3, 11]. 5mC in genomic DNA was initially identified by Wyatt in 1951[12]. In the next few decades, the idea of regulating and maintaining 5mC patterns throughout cell divisions was proposed[13, 14] and the DNMT1(DNA methyltransferases 1), the first DNA methyltransferase was identified[15, 16] and purified[17]. DNMT1 was found to preferentially methylate the hemimethylated DNA at CpG sites and the loss of DNMT1 in mouse embryonic stem cells (mESCs) resulted in a genome-wide depletion of CpG methylation[18], highlighting its role in maintaining methylation. The subsequent identification of other DNA methyltransferases (DNMTs) further consolidates our understanding of this modification.

Parallel to cytosine methylation, adenine is modified through methylation at the N^6 position to form N^6-methyladenine (6mA), a modification plays critical roles in the genomic DNA of prokaryotes, implicated in a range of cellular processes from DNA repair, replication, and cell defense[19-22]. Although prevalent in prokaryotes, the physiological functions and potential regulation of DNA 6mA modification in eukaryotes have yet to be fully explored. Since 2015, high levels of 6mA have been detected in the genomes of several lower eukaryotes with unique distribution patterns, suggesting a regulatory role for 6mA as an additional DNA marker in these systems[23]. In mammals, despite its lower abundance in most cases, 6mA is believed to play various roles in disease, animal development, and stress conditions[24].

The complex system of DNA modifications is essential for the fine-tuning of genomic regulation. While the methylated cytosines are well-studied transcriptional regulation, the emerging significance of adenine methylation in eukaryotes opens new avenues for understanding the epigenetic mechanisms at play. As we keep uncovering the roles of 5mC and 6mA, it will become even clearer how they contribute to the complexity of gene regulation, and the development of diseases.

1.3 Cytosine methylation: distribution and biological functions

5-methylcytosine (5mC), often termed the 'fifth base' of the human genome due to its pervasive influence on genomic regulation, has been extensively studied over the past decades. Representing more than 4% of all cytosines in human DNA, 5mC is deposited by DNA methyltransferases (DNMTs), which plays a pivotal role in maintaining methylation patterns during DNA replication.[25]. The dynamics of DNA methylation were further elucidated with the discovery of *de novo* DNA methylation in early pluripotent embryonic cells by the Jaenisch group in 1982[26]. This was a significant leap in understanding how methylation patterns are established and maintained. The subsequent identification of

DNMT3A and DNMT3B, responsible for *de novo* methylation of proviral DNA and repetitive sequences, and their role in establishing methylation on maternal imprinted genes, highlighted the dynamic nature of DNA methylation[27, 28].

However, methylation levels vary significantly among different species, indicating that its biological significance extends beyond mere abundance. The positioning of 5mC, whether in symmetric or asymmetric contexts, and its localization, particularly in CpG islands (CGIs) found in over 50% of vertebrate genes, are key determinants of its regulatory impact. While 70% - 80% of CpG sites in mammalian somatic tissues are methylated[29], most CGIs in somatic cells remain unmethylated, or only partially methylated. Understanding the interplay of 5mC with other epigenetic mechanisms, including histone modifications, is essential for a comprehensive view of genomic regulation.

The functional consequences of DNA methylation are closely associated with gene expression repression. Early studies took advantage of 5-azacytidine, an inhibitor of DNA methylation, demonstrated the reactivation of silenced genes in living cells, providing strong evidence of the repressive nature of DNA methylation[30-32]. This was further corroborated by studies on DNMT1 knockout mice, where loss of methylation led to the reactivation of several inactive genes[33].

The significance of 5mC also showcased by its recognition by specific proteins, termed '5mC readers.' The discovery of the methyl-CpG binding protein complex MeCP1 shed light on the molecular mechanisms by which DNA methylation influences gene expression[34]. Subsequent identification of other readers in the MBD family (MeCP2, MBD1, MBD2, MBD4) has further advanced our understanding of methylation-mediated gene regulation[35]. Further studies revealed MeCP2, MBD1, and MBD2 were involved in 5mC-transcriptional repression[36]. The binding of reader proteins to methylated CpG sites

specifically leads to the suppression of gene expression, serving as a fundamental epigenetic mechanism of significant importance in more complex organisms.

DNA methylation, alongside histone modifications, also modulates chromatin structure and regulates gene expression across different cell types in a different pathway. The presence of DNA methylation in promoter regions typically correlates with transcriptional repression in two different pathways. Early genome-wide analyses of DNA methylation patterns uncovered DNA methylation shows a robust inverse correlation with H3K4me2/3 because DNMT3A, DNMT3B, and DNMT3L cannot bind to the H3 tail di- or tri-methylated at H3K4 [37, 38],which indicated a ‘direct disruption model’. This repression can also occur indirectly through the recruitment of reader proteins like MeCP and Kaiso or directly by hindering the binding of transcription factors to DNA[39, 40].

Conversely, DNA methylation within gene bodies generally shows a positive correlation with gene expression. It impacts transcription elongation, RNA splicing, and nucleosome positioning, underscoring the multifaceted roles of DNA methylation in gene regulation[41-43]. Extensive positive correlations between active transcription and gene body methylation have been confirmed on the active X chromosome[44]. The interactions of methyltransferases, demethylases, and readers of DNA methylation with various histone marks further emphasize the integrated nature of epigenetic control over transcription.

1.4 DNA methylation dynamics: maintenance, active demethylation, and error correction

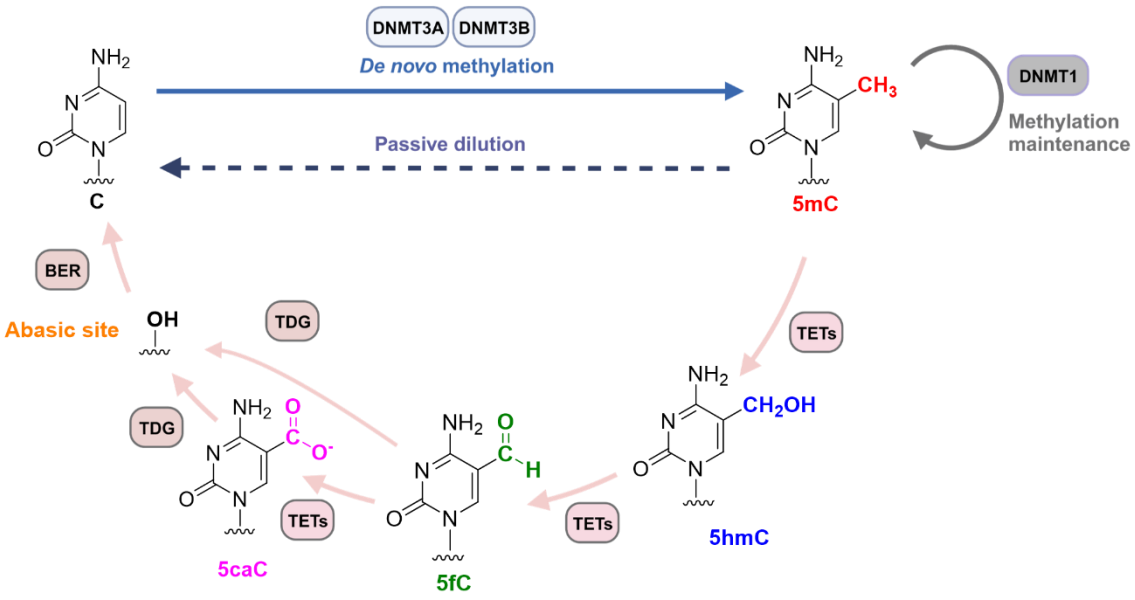


Figure 1.2 Scheme of reversible pathway for dynamic modifications of cytosine methylation
This scheme elucidates reversible pathway for modifying C within DNA. 5mC bases can be oxidized iteratively to 5hmC, 5fC and 5caC by TET enzymes. The removal of highly oxidized 5fC and 5caC is achieved through a process termed active restoration (AR). In this pathway, 5fC or 5caC is excised by TDG, resulting in the generation of an abasic site. This abasic site is subsequently repaired as part of the base excision repair (BER) process, ultimately restoring the original unmodified C. In the passive dilution pathway, 5hmC/5fC/5caC are gradually diluted during DNA replication in a replication-dependent manner to the restore unmodified C.

The revelation that DNA methylation is both dynamic and reversible changed the landscape of epigenetic research (**Figure 1.2**). The discovery of TET (ten-eleven translocation) proteins, which oxidize 5mC to 5-hydroxymethylcytosine (5hmC), marked a major breakthrough in our understanding of active demethylation[45-47]. Further studies showed that TET enzymes could oxidize 5hmC to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC)[48]. The TET family of enzymes, including TET1, TET2, and TET3, all use Fe (II) and α-ketoglutarate (α-KG) as co-substrates to catalyze the oxidation of 5mC all the way to 5caC, and marks the methylated cytosine for further processing.

The oxidized forms of 5mC, particularly 5fC and 5caC, are then recognized and excised by human thymine DNA glycosylase (TDG). TDG specifically binds to these modified bases and catalyzes their removal from the DNA, creating an abasic site. [49, 50], and prepares the DNA for the final restoration step, ensuring the fidelity of the genetic code is maintained.

Once abasic site is generated, base excision repair (BER) machinery steps in to complete the demethylation process. The BER pathway involves several steps, beginning with the recognition and removal of the abasic site. This is followed by the action of a DNA polymerase to inserts an unmodified cytosine (dCTP) at the abasic site. Finally, DNA ligase seals the nick in the DNA strand, restoring the DNA to its unmodified state. The entire BER process not only reinstates the original DNA sequence but also ensures that the epigenetic mark of methylation is accurately and efficiently removed[51].

In addition to the active demethylation pathway, DNA methylation dynamics also involve a mechanism known as passive dilution. In this pathway, 5mC was actively oxidized into other form and followed by passive dilution of the oxidized base to regenerate unmodified C in a replication-dependent manner.

The interplay between TET-mediated oxidation and DNMT-mediated methylation exemplifies the dynamic nature of the epigenome. DNMT3A and DNMT3B are responsible for *de novo* synthesize of methyl groups, and DNMT1 is involved in maintaining DNA methylation by copying methylation patterns onto the newly replicated DNA strand, while TET enzymes facilitate their removal, allowing for epigenetic flexibility[25].

The dynamics of DNA methylation are intricately linked to various biological processes. The process of active demethylation plays a crucial role in embryonic development, as evidenced by the loss of 5mC and the appearance of 5hmC/5fC/5caC during fertilization[52]. Genome-wide mapping of 5mC oxidation derivatives has revealed

widespread active demethylation events, associated with functional genomic elements, crucial for cell development and stem cell maintenance[53, 54].

1.5 DNA methylation in oncogenesis

The dynamic processes of methylation and demethylation, particularly involving 5mC and its oxidation derivative 5hmC, play critical roles in the regulation of gene expression and genomic stability, both of which are crucial in cancer development and progression. These processes are integral to cancer development and progression, where aberrations can lead to dysregulated gene expression patterns characteristic of various cancers.[55-57].

Aberrant patterns of DNA methylation have been identified as early events in tumor development. These include global hypomethylation leading to genomic instability and focal hypermethylation causing silencing of tumor suppressor genes. For example, DNA hypermethylation in cancer has been observed in CpG islands marked by transcriptionally active histone modifications and polycomb-mediated repression marks, often affecting genes involved in apoptosis and DNA repair[58, 59].

The establishment of cancer-specific epigenetic traits is closely linked to genetic aberrations of epigenetic modifiers. Mutations in enzymes like DNA methyltransferase family (DNMTs) and isocitrate dehydrogenase (IDH1/2) have been implicated in the altered methylation landscape of cancers[60-62]. These mutations can lead to impaired cellular differentiation, increased proliferation, or resistance to apoptosis, driving the progression of hematopoietic malignancies and gliomas.

As elucidated in recent studies, epigenetic alterations, triggered by environmental stimuli such as aging, chronic inflammation, and smoking, play a crucial role in cancer development. Remarkably, these changes are stably inherited by daughter cells even after the cessation of the initial stimulus, making them valuable biological markers for predicting cancer risk. When combined with traditional risk prediction approaches, DNA methylation

analysis enhances the predictive accuracy for cancers such as esophageal carcinoma [63]. Beyond risk assessment, the analysis of epigenetic characteristics has also improved diagnostic precision. Tumors like diffuse gliomas, known for their high interobserver variability in histopathological diagnosis, benefit from molecular analyses, such as unit-point DNA methylation testing, leading to improved diagnostic and treatment strategies[64, 65]. Epigenetic changes are prevalent across all cancer types, making them attractive biomarkers for early disease detection and potential targets for epigenetic drug interventions. These insights into epigenetic modifications underscore the importance of developing novel approaches in DNA methylation analysis and their application in understanding and combating cancer metastasis.

1.6 Scope of this thesis

The ongoing advancements in epigenetic research have significantly enhanced our understanding of the complex relationship between DNA methylation patterns and cancer development. In this context, the study of DNA methylation, particularly 5mC and 5hmC, has emerged as a critical area of investigation. These modifications are pivotal in the regulation of gene expression and genomic integrity, influencing everything from cell differentiation to tumor genesis. This thesis aims to contribute to this dynamic field by focusing on two key aspects: the development of an innovative sequencing method and the construction of a comprehensive human 5hmC tumor tissue map. These studies not only provide valuable insights into the mechanistic underpinnings of cancer but also pave the way for novel diagnostic and therapeutic strategies targeting epigenetic alterations.

Chapter 2 will detail the development TET/TDG-mediated 5mC labelling and sequencing strategy (TT-5mC-seq), a novel bisulfite-free method for high-resolution mapping and sequencing of 5mC in DNA.

Chapter 3 will describe the development of a comprehensive tumor tissue map, analyzing 5hmC distributions across a wide range of human cancer types and tissues.

2 TET/TDG-mediated 5mC labeling and sequencing strategy (TT-5mC-seq)

2.1 Introduction: DNA methylation detection

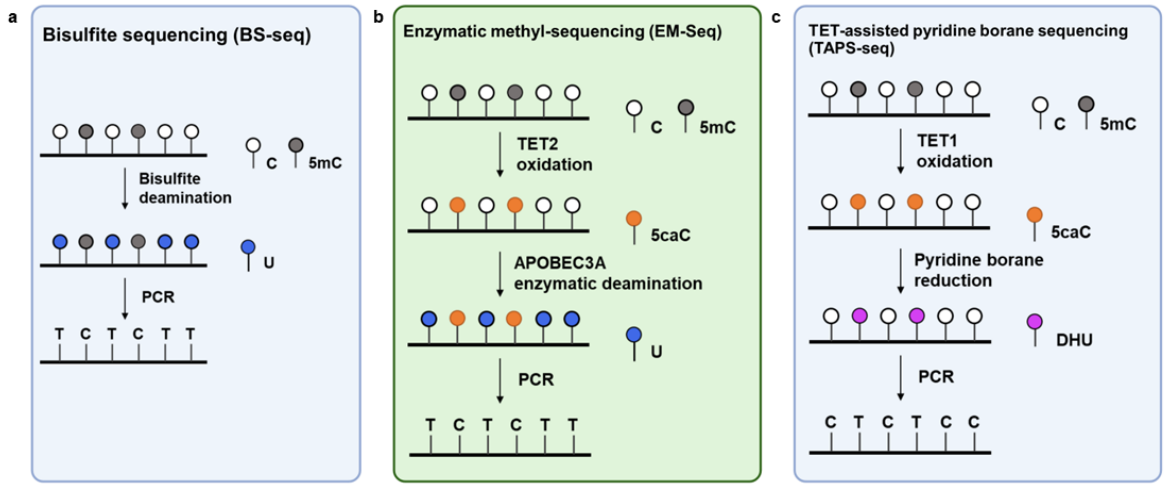


Figure 2.1 Scheme of three representative DNA 5mC modification detection sequencing technologies
(a)Bisulfite sequencing (BS-Seq) (b)Enzymatic Methyl-Seq (EM-Seq), and (c) TET-assisted pyridine borane sequencing (TAPS)

Since its development in 1992 by Frommer et al.[66], bisulfite sequencing (BS-Seq) has been the gold standard for DNA methylation analysis(**Figure 2.1a**). The advent of whole-genome bisulfite sequencing (WGBS) in 2009[67], which achieves over 99% conversion efficiency, further cemented BS-Seq's role in accurately mapping 5mC and 5hmC. Bisulfite methods employs sodium bisulfite, which specifically deaminates unmodified cytosine to uracil (U) while preserving 5mC and 5hmC. During subsequent PCR and sequencing, U is read as thymine (T), allowing for the differentiation of 5mC and 5hmC from unmodified cytosine when compared with the reference genome.

However, WGBS's approach, which involves harsh treatment conditions, results in significant DNA degradation, particularly problematic for low-input samples such as cell-free DNA (cfDNA). Additionally, the conversion of unmodified cytosine, which constitutes around 95% of all genomic cytosine, to uracil drastically reduces the DNA sequence complexity. This simplification leads to an increased need for deeper sequencing to achieve

accurate results, posing a challenge in terms of both efficiency and cost. While various improvements, such as post-bisulfite adaptor tagging (PBAT), have been developed to mitigate these issues and facilitate single-cell WGBS protocols[68-70], challenges in sequencing efficiency and genomic coverage bias remain.

In response to the limitations of WGBS, recent advancements have led to the development of bisulfite-free DNA methylation sequencing methods such as Enzymatic Methyl-Seq (EM-Seq) and TET-assisted pyridine borane sequencing (TAPS) (**Figure 2.1b, c**) [71, 72]. EM-seq first utilizes Tet methylcytosine dioxygenase 2 (TET2) to oxidize 5mC to 5hmC, 5fC and 5caC. Concurrently, β-glucosyltransferase (βGT) acts to glucosylate both TET2-derived and genomic 5hmC, forming 5-(β-glucosyloxymethyl) cytosine (5gmC). This process shields these modifications from subsequent deamination by APOBEC3A, which converts unmodified C to U. EM-Seq offers a protection rate of over 96% on 5mC and a low non-conversion rate of unmodified cytosine, making it compatible with DNA quantities as low as 100 pg. However, like WGBS, EM-Seq still indirectly maps 5mC and 5hmC modifications, leading to reduced genomic complexity.

In contrast, TET-assisted pyridine borane sequencing (TAPS), developed by Liu et al. in 2019[71], offers a direct detection method for 5mC and 5hmC, free from the harsh conditions of bisulfite treatment. TAPS uses a borane reduction chemistry to convert 5caC to dihydrouracil (DHU), a reaction that is milder and causes less DNA damage compared to bisulfite treatments. TAPS's method of directly detecting modified cytosines while preserving more genomic information marks a significant step forward, achieving higher mapping rates and sequencing quality at substantially reduced costs compared to WGBS. However, despite its advantages, TAPS is not without its limitations. The pico-borane reduction process, conducted at 70°C for 3 hours, can inadvertently convert some unmodified cytosine to DHU as well, resulting in a lower signal-to-noise ratio. This issue can impact the accuracy of

methylation detection. Furthermore, the use of pico-borane, not being a common reagent in many laboratories and clinics, poses additional practical challenges for the widespread adoption of TAPS.

In this chapter, we will introduce TET/TDG-mediated 5mC labeling and sequencing (TT-5mC-seq), a groundbreaking approach designed to overcome the challenges of existing DNA methylation sequencing methods. The TT-5mC-seq procedure is notably mild, getting rid of the use of harsh chemicals, thereby preserving DNA integrity. The new method detects 5mC to single-base resolution with significantly lower background noises, allowing us to sequence 5mC in DNA starting from low-input biological DNA samples. TT-5mC-seq also has an option to add an enrichment step to further increase signal to noise and reduce sequencing costs, enhancing both the efficiency and cost-effectiveness of the sequencing process.

2.2 Result and discussion

2.2.1 General design of TET/TDG-mediated 5mC sequencing (TT-5mC-seq)

To establish an approach that effectively addresses the limitations inherent in existing DNA methylation sequencing methods, we developed the TET/TDG-mediated 5mC Sequencing (TT-5mC-seq) **(Figure 2.2a)**. This method aims to provide a gentler, yet precise, strategy for detecting 5mC modifications without the typical drawbacks of DNA degradation and complexity reduction seen in traditional bisulfite sequencing (BS-Seq) method and the low signal-to-noise ratio issue in TAPS.

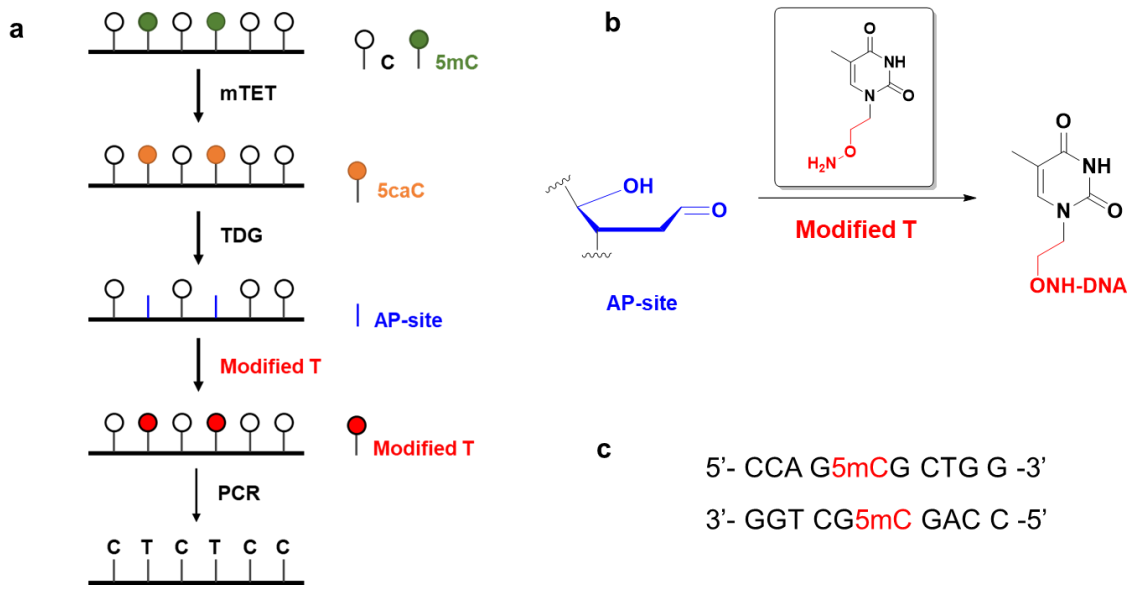


Figure 2.2 Schematic illustration of TT-5mC-seq
(a) Schematic diagram of TT-5mC-seq. 5mCs in genomic DNA are converted to 5caCs by TET-mediated oxidation. After TDG excision, abasic site (AP-site) is created at the original 5mC. N_3-T can specifically react with AP-site and leads to a 5mC-to-T mutation can be used to identify 5mC sites genome-wide at single-base substitution with or without enrichment. (b) The structure of modified T and the reaction between abasic site (AP-site) and modified T (c) The structure of 10-mer double strand model DNA with 5mC modification on both sides.

TT-5mC-seq was conceptualized to harness the natural active DNA demethylation process. It involves the oxidation of 5mC to 5caC by the TET1 enzyme, followed by excision through Thymine DNA Glycosylase (TDG), resulting in an abasic site. This site is then modified using a thymine (T) analog, ensuring that 5mC is read as thymine in subsequent

high-throughput sequencing, while unmodified cytosine is accurately read as cytosine. This innovative approach, which is chemical-free and preserves DNA integrity, offers a significant advancement in methylation sequencing, potentially enabling more precise and less invasive studies in areas where traditional methods have shown limitations.

2.2.2 Design and synthesis of nucleotide alternative

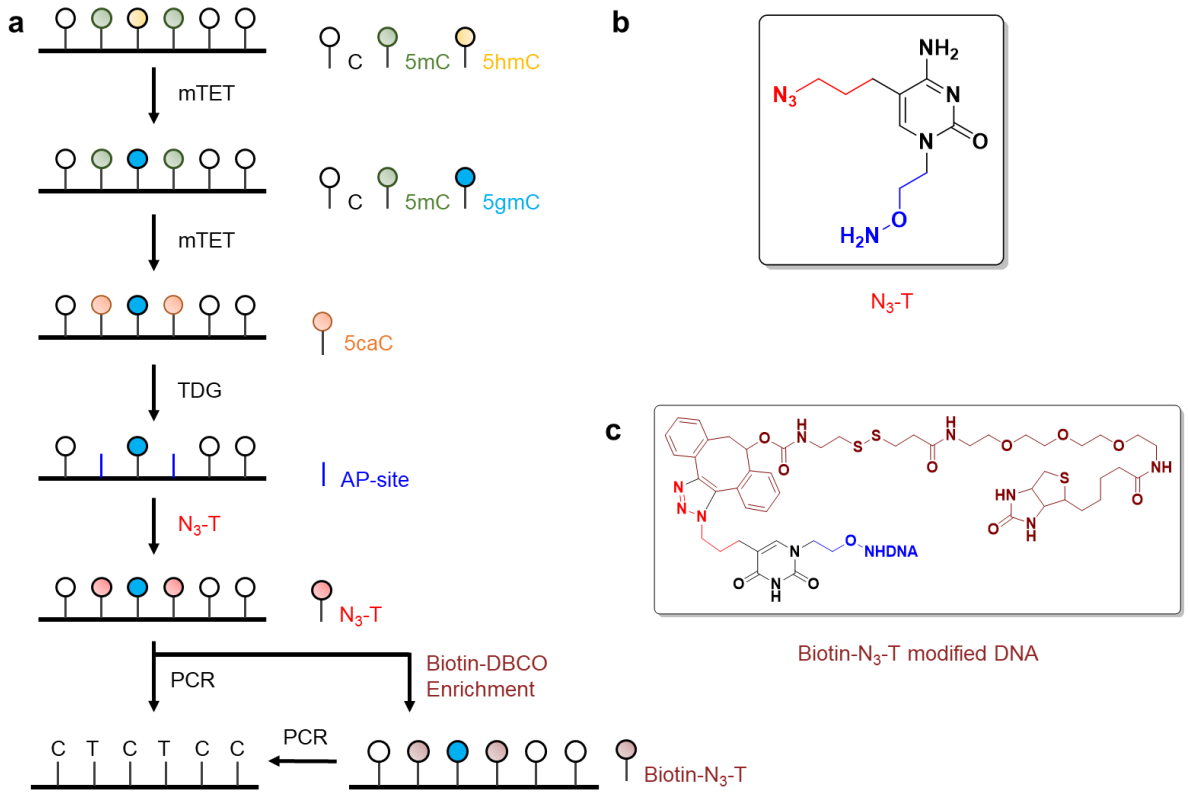


Figure 2.3 Schematic illustration of TT-5mC-seq with azide-T (N_3-T)
(a) Schematic diagram of modified TT-5mC-seq. 5mCs in genomic DNA are converted to 5caCs by TET-mediated oxidation. After TDG excision, abasic site (AP-site) is created at the original 5mC. N3-T can specifically react with AP-site and leads to a 5mC-to-T mutation can be used to identify 5mC sites genome-wide at single-base substitution with or without enrichment. (b) The structure of N_3-T (c) The structure of Biotin-N_3-T modified DNA.

Advancing the capabilities of TT-5mC-seq, we have developed and incorporated a chemically synthesized thymine mimic, known as N_3-T(**Figure 2.3b**), synthesized from commercial reagents in six steps (**Figure 2.4**). This novel compound comprises three key components. The first is a hydroxylamine functional group, specifically designed to react

with the aldehyde group at the abasic site, allowing for selective labeling. Once attached, the thymine nucleobase segment of N_3-T is read as T during PCR amplification, translating into a C-to-T mutation and thus enabling base-resolution sequencing. Additionally, N_3-T features an azide tether, which acts as a bioorthogonal handle. This feature offers versatility in application: we can either proceed directly to sequencing the labeled DNA, or we can utilize dibenzocyclooctyne-modified biotin (DBCO-biotin). In the latter approach, the methylated DNA fragment undergoes biotinylation via click chemistry (**Figure 2.3a**), facilitating the enrichment of specific DNA sequences for subsequent amplification and sequencing. This unique aspect of N_3-T integrates seamlessly into the sequencing workflow, enhancing both the accuracy and functionality of TT-5mC-seq.

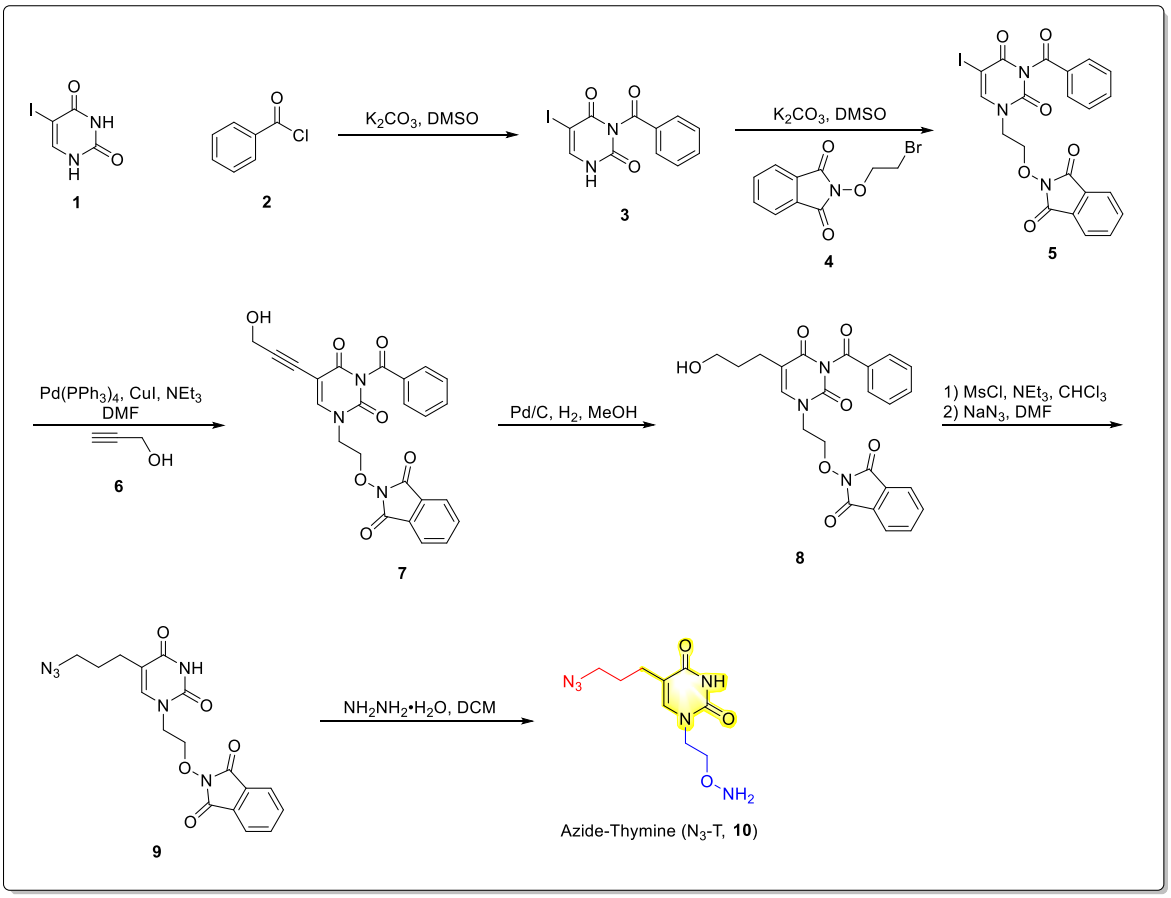


Figure 2.4 Design and synthetic scheme of N_3-T.
Illustration of the synthetic scheme for N_3-T, presenting each chemical reaction and intermediate compound leading to the final product.

2.2.3 Validation of TT-5mC-seq strategy

To validate the TT-5mC-seq strategy, we performed a comprehensive molecular characterization of each intermediate state during the DNA enzymatic reactions and the chemical addition of N_3-T. **Figure 2.5a** provides an overview of these steps as evidenced by MALDI-TOF mass spectrometry. Starting with a 10-mer model DNA containing 5-methylcytosine (**Figure 2.2c**), we applied the entire process to generate and subsequently identify the intermediates and final product (**Figure 2.5b**).

Here we present the MALDI-TOF mass spectra of the resulting 10-mer DNA at each stage of these reactions, confirming the molecular weights of 5mC, 5caC, the AP site, and the final N_3-T labeled product. The calculated and observed molecular weights closely align, validating the specificity of our labeling strategy and the efficiency of both enzymatic reaction and chemical addition throughout the process.

We then utilized an 82-mer synthetic DNA oligonucleotide, which comprised both cytosine and 5-methylcytosine bases. Post TT-5mC-seq treatment, Sanger sequencing analyses revealed a high conversion rate of 5mC to thymine (T), specifically 87.4% (**Figure 2.6a**). This result is indicative of the optimized conditions of our method being highly effective for the 5mC-to-T conversion. Crucially, cytosines (C) remained unaltered post-treatment, thus demonstrating the selectivity of the TT-5mC-seq and its ability to avoid undesired C-to-U conversions.

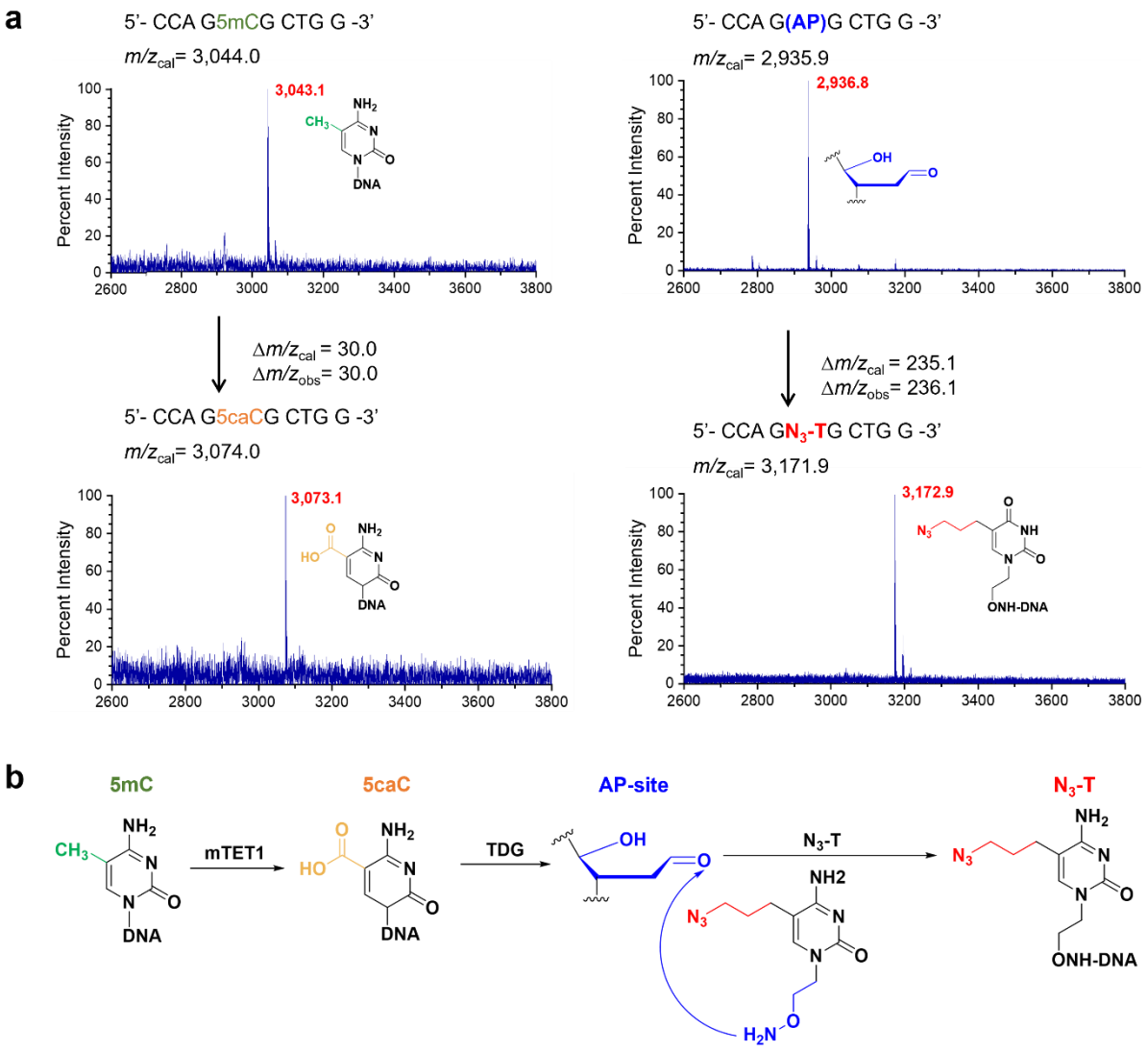


Figure 2.5 MALDI-TOF MS characterization of 5mC, 5caC, AP site, and N3-T containing 10-mer DNA in a model experiment.
(a) MALDI-TOF mass spectrum of TT-5mC-seq intermediates starting with a 10-mer model DNA, respectively, with the calculated molecular weight and observed molecular weight indicated. (b) Corresponding reactions of the mTET oxidation, TDG base excision and the subsequent reaction with N_3-T. Reactions were performed in duplex DNA with the complementary strand.

To further assess the practical applicability of TT-5mC-seq, particularly its potential for enrichment through click chemistry, a dot blot assay was employed. The assay confirmed the robustness and repeatability of the method, as evidenced by the clear and consistent signal observed for the DNA oligo labeled with N_3-T and subsequently treated with DBCO-S-S-

PEG4-biotin, as opposed to the untreated control (**Figure 2.6b**). These findings underscore the method's suitability for downstream applications requiring selective enrichment of methylated DNA fragments. We also examined the final product using MALDI-TOF mass spectrometry (**Figure 2.6c**), and presents the mass spectrum of the 10-mer model DNA following conjugation with biotin-N_3-T. The peak observed corresponds to the expected mass of the labeled DNA, confirming not only the precise addition of the biotinylated thymine analog but also the method's consistency and reliability. This mass spectral analysis provides robust evidence for the successful application of click chemistry in our sequencing strategy, and it underscores the potential of TT-5mC-seq for integrating into workflows that require the enrichment and detailed analysis of methylated DNA sequences.

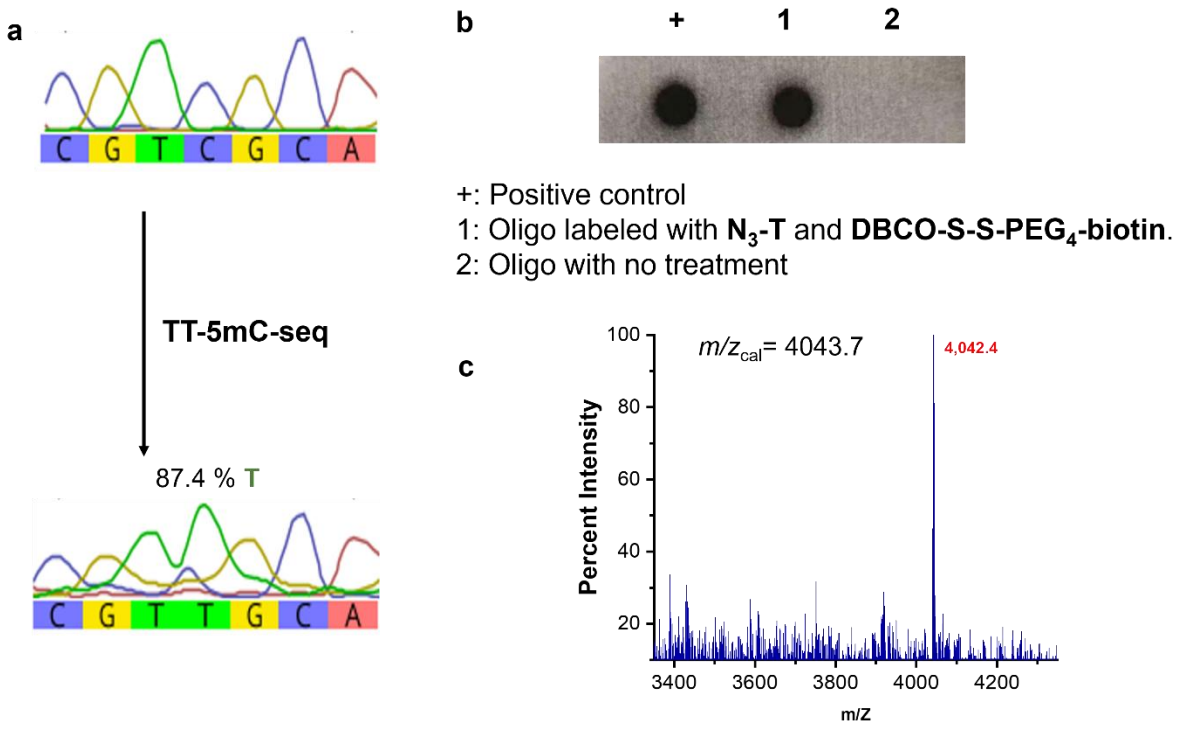


Figure 2.6 Mutation validation and biotin compatibility of TT-5mC-seq
(a) Sanger sequencing results for a model DNA containing fully methylated CpG sites before (top) and after (bottom) TT-5mC-seq. 5mC is converted to T after TT-5mC-seq. (b) Dot blot assay of TT-5mC-seq. Dot 1: Model DNA oligo labeled with N3-T and then further labeled with DBCO-S-S-PEG$_4$-biotin.; dot 2: Model DNA oligo with no treatment. (c) MALDI-TOF mass spectrum of 10-mer model DNA with biotin-N_3-T

2.2.4 Optimization of incorporation of a synthesized nucleotide alternative

Given the well-established foundation of mouse TET1 (mTET1), our methodology required minimal modifications from the widely accepted protocols in this initial step. For TDG excision, turnover rate of the excision reaction has long been a concern. To achieve complete transformation from 5caC to the abasic site, the 10-mer model 5caC modified DNA was mixed with TDG in different concentrations. It was found that found that a 10-fold molar ratio of TDG (100 nM TDG for 10 nM DNA substrate) could completely excise 5caC to afford the AP site. The reactions were performed at 22 °C for 60 min in reaction buffer containing 25 mM HEPES, pH 7.4, 0.5 mM EDTA, 0.5 mg/mL BSA, and 0.5 mM DTT. These two enzymatic processes provided a solid basis for the subsequent stage of our strategy: the incorporation of a synthesized nucleotide alternative at the abasic site.

To ensure a comprehensive validation of our optimization efforts, we expanded our model DNA oligo beyond the initial 10-mer containing symmetrical 5mC modifications. We introduced a 12-mer double-stranded DNA oligonucleotide, designed with 5mC modification on only one strand. This addition allowed us to better reflect the sparse and asymmetric methylation patterns found in the genome.

Optimization of the nucleotide alternative incorporation commenced with an investigation into the effect of pH on the reaction efficiency. MALDI-MS analysis revealed that a pH of 6 facilitated the reaction most effectively, achieving the highest rate of conversion (**Figure 2.7**). When the pH was increased to 7, the transformation rate notably decreased to only two-thirds of the optimal rate, suggesting a less favorable environment for the reaction at this higher pH level. Conversely, at a pH of 5, no starting reagents remained; however, MALDI-MS data indicated the presence of degradation products. This unexpected outcome was attributed to the more acidic conditions, which likely led to the instability and breakdown of the reagents.

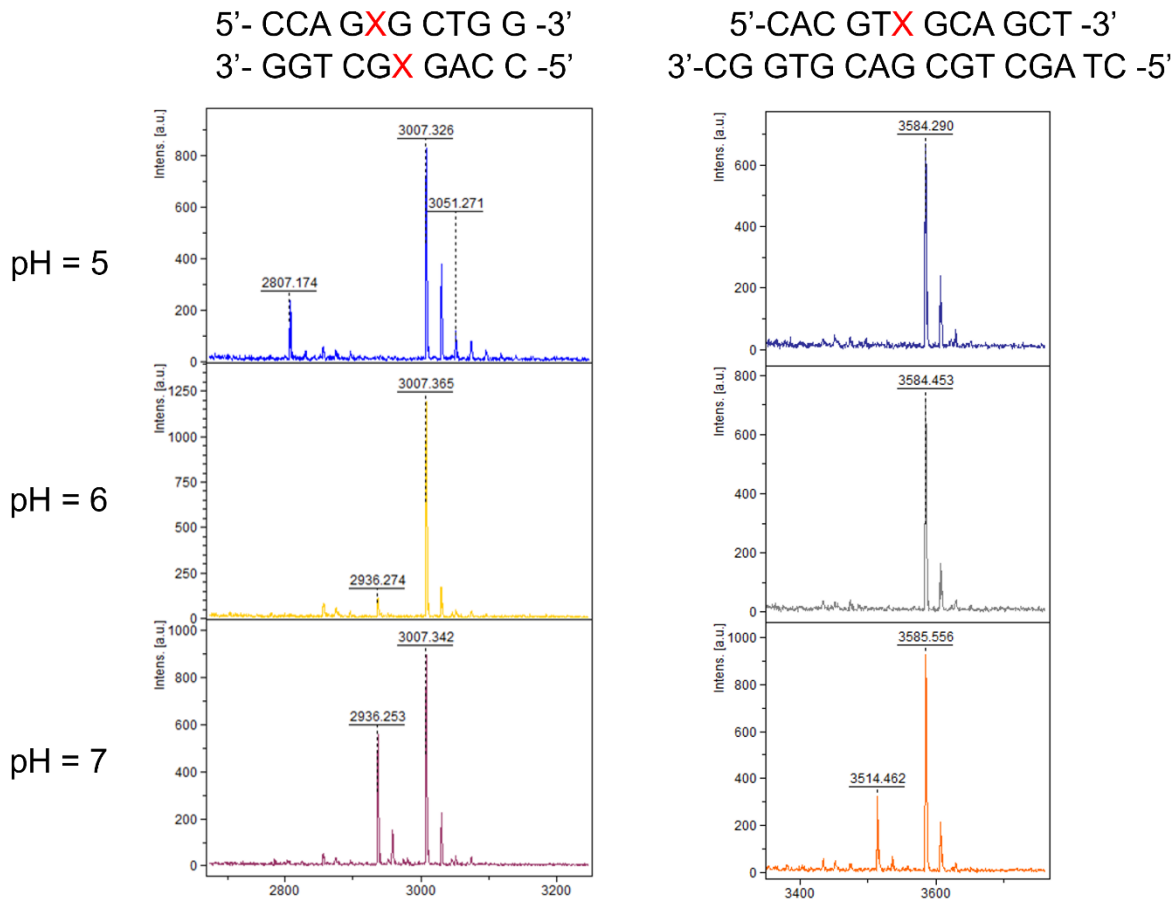


Figure 2.7 MALDI-TOF MS analysis of nucleotide alternative incorporation at different pH

In parallel to our pH optimization, we conducted a series of experiments to determine the optimal concentration of the nucleotide alternative necessary for efficient labeling of a given amount of substrate. Our objective was to identify the minimum effective concentration that would ensure complete transformation of the substrate into the final product without excess reagent, as indicated by MALDI-MS analysis (**Figure 2.8**). At the 2 mM, MALDI-MS revealing minimal to no conversion of the substrate. When the concentration was increased to 10 mM, more than half of the substrate transformed. Further increasing the concentration to 50 mM resulted in nearly complete conversion of the substrate. MALDI-MS analysis at this concentration showed an almost complete absence of the starting substrate, demonstrating that 50 micromolar was sufficient to drive the reaction to near-completion. At 100 mM, no

additional benefits were observed, indicating that increasing the concentration beyond this point would not enhance the reaction efficiency.

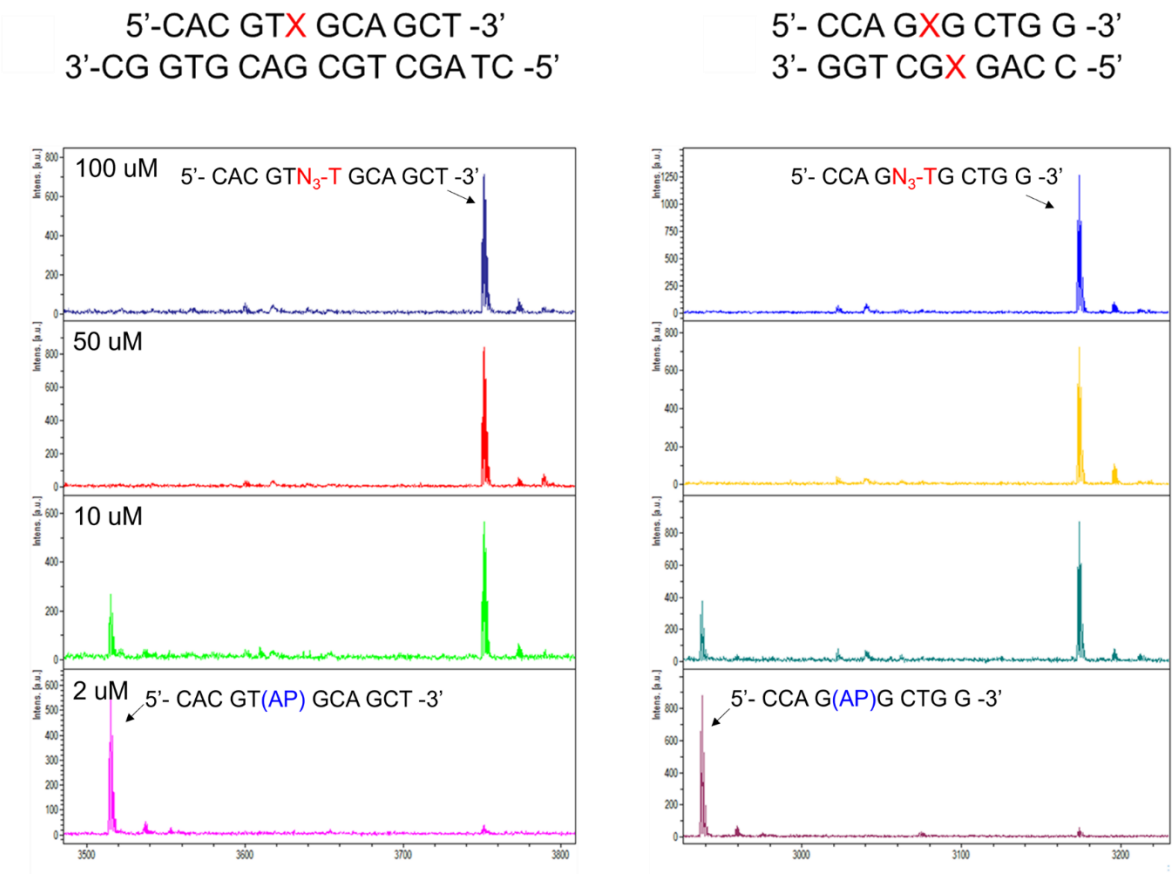


Figure 2.8 MALDI-TOF MS analysis of nucleotide alternative incorporation at different substrate concentrations

2.2.5 DNA polymerase selection for TT-5mC-seq

In the next phase of optimizing the TT-5mC-seq method, a range of high-fidelity DNA polymerases were assessed for their compatibility with our sequencing strategy. Notably, we experimented with Taq Platinum DNA Polymerase, EpiMark Hot Start Taq DNA Polymerase, Q5U Hot Start High-Fidelity DNA Polymerase, Bst DNA Polymerase, and Phusion High-Fidelity DNA Polymerase. These enzymes, while reliable for standard PCR, exhibited low conversion rates for 5mC during Sanger sequencing (**Figure 2.10a**), suggesting a limitation in their ability to process modified nucleotides. Suspecting that this inefficiency

might be due to insufficient read-through of N_3-T sites, we conducted primer extension experiments to further investigate (**Figure 2.10b**).

Given that our sequencing strategy requires incorporation of a synthetic nucleotide analog, the high-fidelity of common polymerases hindered their utility. Consequently, we shifted our approach towards polymerases with inherently lower fidelity that could accommodate modified nucleotides more readily. This led us to experiment with reverse transcriptase, which are known for their ability to transcribe RNA templates into complementary DNA and generally exhibit lower fidelity than DNA-directed DNA polymerases.

The primer extension results underscore the limited read-through capability of these high-fidelity polymerases. In contrast, reverse transcriptase, which naturally possess a lower fidelity than DNA polymerases, emerged as a more viable option for incorporating modified nucleotides. HIV Reverse Transcriptase (HIV RT), in particular, demonstrated a marked increase in the read-through rate and high mutation rate in the final sequencing result.

HIV RT was identified as a promising candidate. Upon integrating HIV RT into the TT-5mC-Seq protocol, we observed a significant improvement in the mutation rate, suggesting enhanced incorporation of the nucleotide analog.

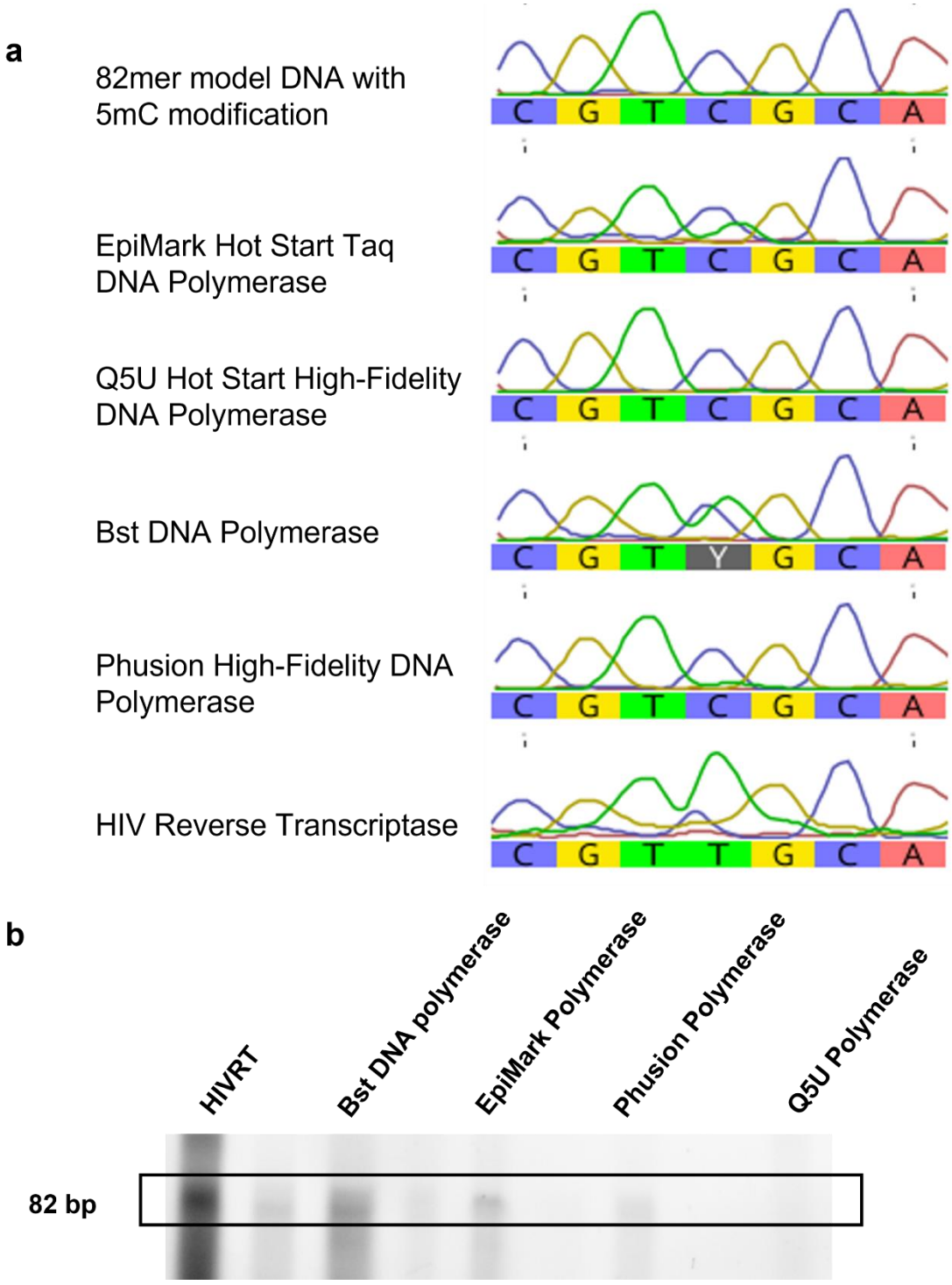


Figure 2.9 Performance evaluation of polymerases and reverse transcriptase.
(a) Sanger sequencing result for an 82-mer model DNA oligonucleotide with 5mC modifications processed by various high-fidelity DNA polymerases and HIV RT. The high-fidelity enzymes display precise but low mutation rates, while HIV RT demonstrates a markedly higher mutation rate, indicating better incorporation of the modified nucleotide. (b) Primer extension assay visualizing the read-through capability at mutation sites. HIV RT exhibits a higher read-through rate compared to the other tested polymerases, as evidenced by the more intense band at the expected 82 bp size.

2.2.6 Quality control and comparative analysis of TT-5mC-seq

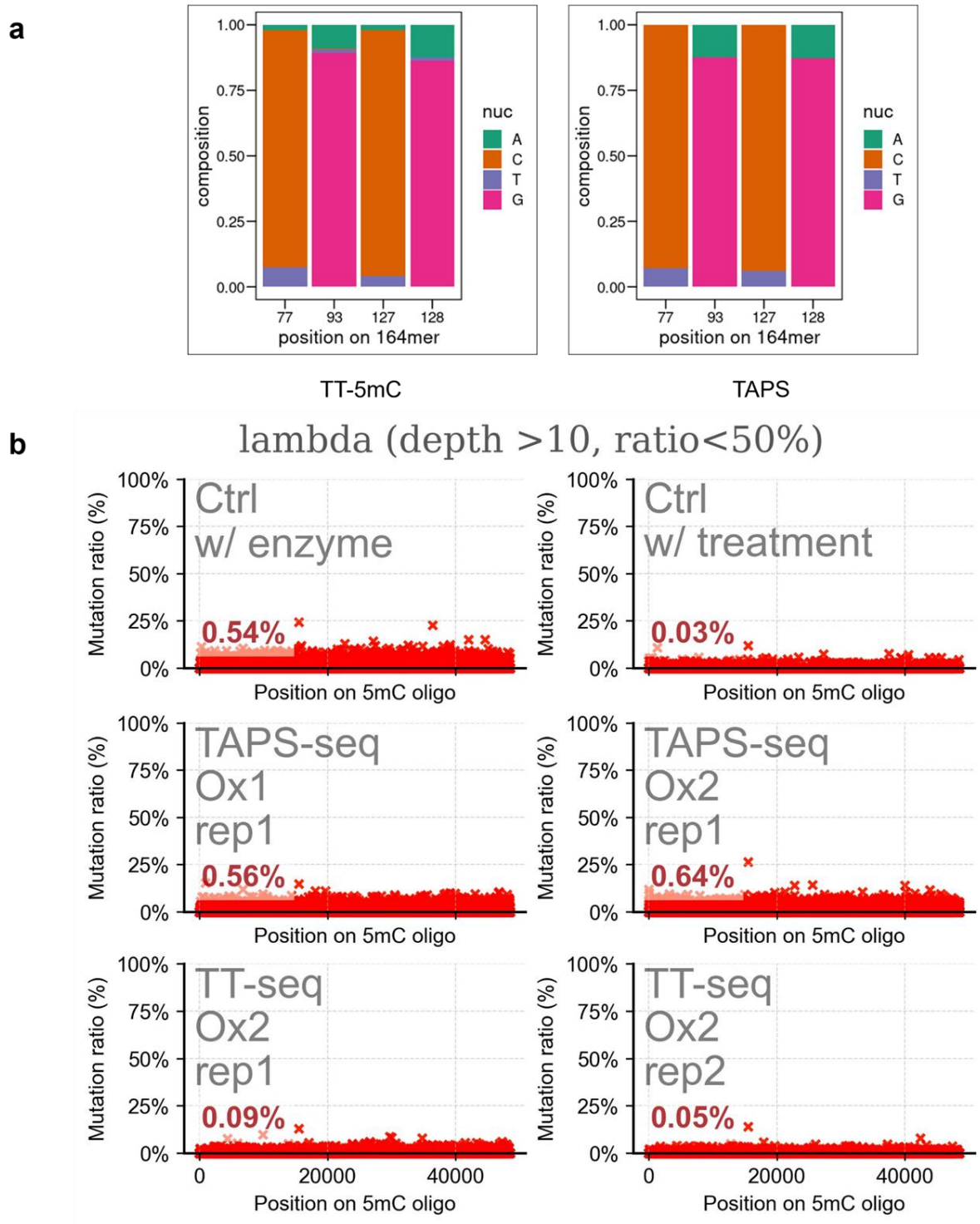


Figure 2.10 Comparative analysis of mutation ratios in spike-in and lambda DNA negative controls.
(a) Mutation ratio profiles of 164-mer spike in with four 5mC sites (b) Mutation ratio profiles from sequencing of lambda DNA without methylation. Top left: Control with enzyme treatment, top right: Control without any treatment, middle: TAPS-seq replicates, bottom: TT-5mC-seq replicates. The average mutation ratios are indicated for each condition.

For the purpose of evaluating the performance of our TET/TDG-mediated 5mC sequencing (TT-5mC-seq) strategy, we conducted a comparative analysis using a 164-mer 5mC modified DNA as spike in to measure the conversion rate, and lambda DNA as a negative control to measure background noise levels, alongside a comparison with TAPS-seq.

When compared to the established TAPS-seq method, TT-5mC-Seq exhibits comparable 5mC conversion rates, ensuring the reliability of methylation detection (**Figure 2.10a**). Our QC results reveal significant differences in background noise levels (**Figure 2.10b**). The control without any treatment demonstrates an expectedly low mutation ratio (0.03%), aligning with the natural fidelity of genomic DNA. Interestingly, the control that underwent 2-picoline-borane reduction but without the enzymatic treatment exhibited a higher mutation ratio (0.54%), similar to the noise level observed in TAPS-seq samples. These higher ratios indicate the potential for background noise introduced by the 2-picoline-borane reduction step itself.

Conversely, TT-5mC-seq showcases its robustness with mutation ratios (0.09% and 0.05%) that are markedly lower and comparable to the untreated control. This suggests that TT-5mC-seq maintains the integrity of the genomic DNA, avoiding the introduction of artifactual mutations that could compromise the accuracy of methylation detection.

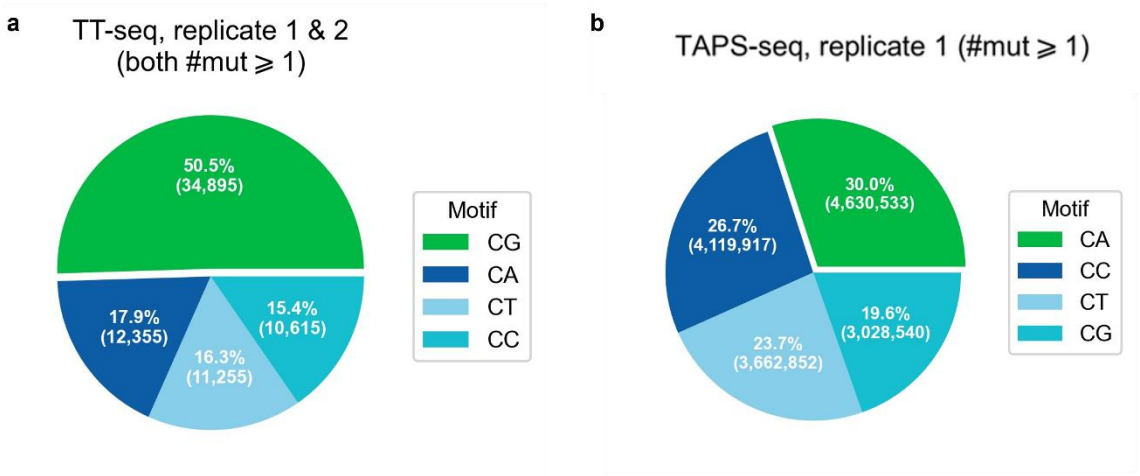


Figure 2.11 Mutation specificity comparison of TT-5mC-seq and TAPS-seq
Pie charts show the mutation ratio of nucleotide motifs in TT-5mC-seq replicates (a) and TAPS-seq (b), highlighting the higher specificity for CG motifs by TT-5mC-seq.

In the data analysis of whole-genome sequencing data from mouse embryonic stem cells (E14), TT-5mC-seq demonstrates a superior specificity for CG motifs, the primary site of methylation in mammalian genomes, with CG motifs accounting for 50.5% of mutations detected (**Figure 2.11a**). This high fidelity of CpG motif is important for accurate methylation mapping and is consistent with the biological emphasis on CpG methylation in gene regulation. In contrast, TAPS-seq displays a more dispersed mutation pattern among CA, CT, CC, and CG motifs (**Figure 2.11b**), suggesting a lower specificity for CpG methylation, which may not align as closely with the expected methylation patterns, potentially increasing the false-positive rate of results.

2.2.7 TT-5mC-seq is compatible with other alternative base substitutions.

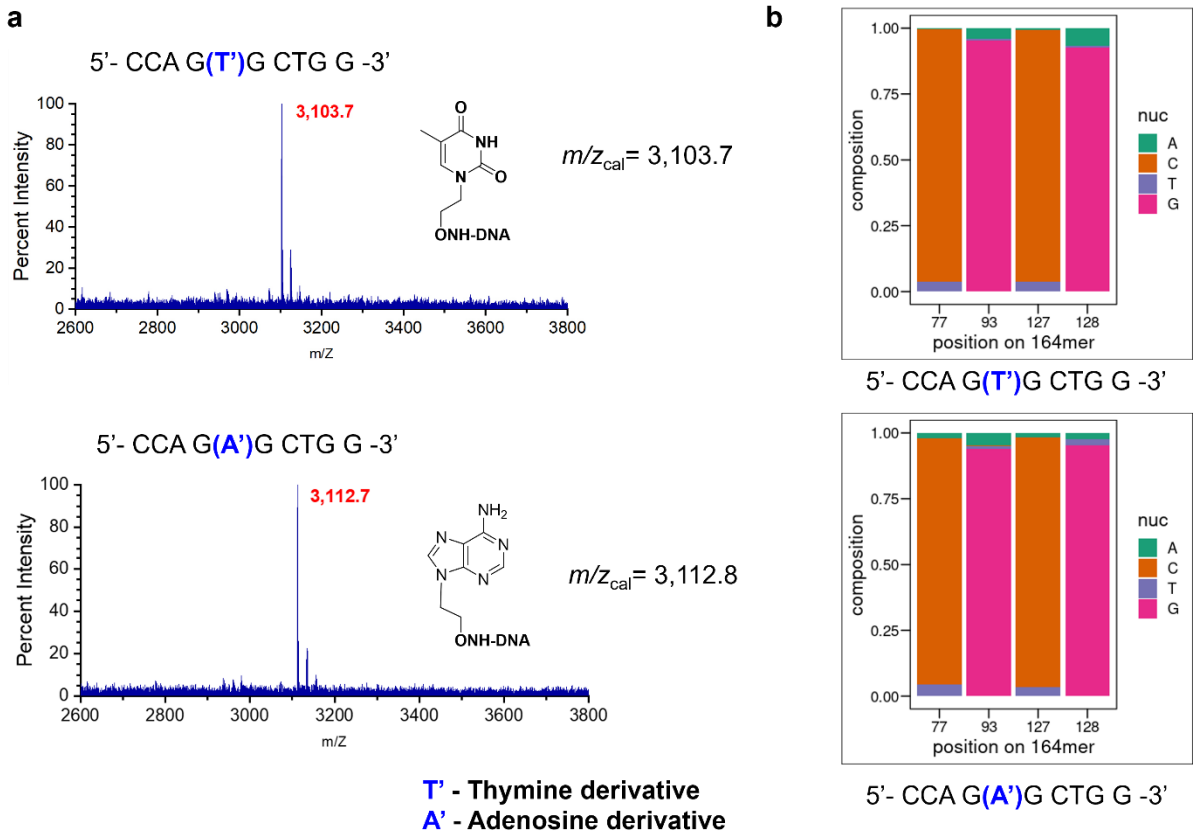


Figure 2.12 Characterization of alternative base substitutions

(a) MALDI-TOF MS of thymine derivative and adenosine derivative containing 10-mer DNA, respectively, with the calculated molecular weight and observed molecular weight indicated. (b) NGS result on 164mer spike-in probe treated with alternative base substitutions.

To demonstrate the versatility of TT-5mC-seq, we extended its application to the detection of alternative base substitutions. The method's adaptability was validated through the incorporation and subsequent sequencing of thymine and adenosine analogs within a 10-mer model DNA. The MALDI-TOF MS confirm the base analogs within the DNA oligo. The molecular weights were calculated theoretically and matched with the observed values to ensure accurate incorporation of the analogs (**Figure 2.12a**).

Next-generation sequencing was utilized to evaluate the mutation rates induced by these substitutions in a 164-mer spike-in probe with four 5mC modifications. The probe,

treated with the base analogs, was sequenced. While the mutation rate for 5mC was lower than that for N_3-methylthymine (N_3-T), the results indicated that the TT-5mC-Seq protocol could be effectively optimized for these alternative substitutions (**Figure 2.12b**).

The potential for further optimization was highlighted by the relative mutation rates, suggesting that TT-5mC-seq can be fine-tuned for enhanced detection in different situation. This adaptability paves the way for a broader application of the technique in epigenetic and genetic research.

2.3 Discussion and future perspective

In this work, we have successfully developed TT-5mC-seq, an enzymatic, bisulfite-free sequencing technique that achieves single-base resolution with as little as 5 ng of genomic DNA input. This method distinguishes itself by its precision in identifying 5-methylcytosine modifications, displaying significantly reduced background noise in comparison to existing methods. The successful integration of nucleotide analogs into TT-5mC-seq further demonstrates its potential for broad application in epigenetic research.

TT-5mC-seq presents a balance of specificity and sensitivity, outperforming established methods such as TAPS by reducing the background noise and incidence of false positives. Future efforts will focus on refining the method to enhance its detection capabilities and expand its scope of applications.

The proposed enrichment of our method through click chemistry presents an exciting future for enhancing the specificity of 5mC detection in complex genomic samples. To translate this into practice, *in vivo* testing will be required, followed by a validation step using clinical samples. Also, cfDNA analysis for non-invasive diagnostics offers a promising landscape for the application of TT-5mC-seq. The method's refinement could influence early detection and monitoring of pathological conditions.

As we advance, TT-5mC-seq aims to contribute significantly to our understanding of epigenetic mechanisms and to the field of personalized medicine. Through continued development and application, this method will enhance our ability to decipher complex methylation patterns, offering insights into disease pathogenesis and informing therapeutic strategies.

2.4 Experimental section

2.4.1 Expression and purification of recombinant human TDG enzyme

The PMCGS19 plasmid, encoding the catalytic domain of human thymine DNA glycosylase (hTDG) was provided by Dr. Liang Zhang (Shanghai Jiao Tong University School of Medicine). The enzyme was expressed in BL21(DE3) cells containing vector pRK1037.

Plasmid harboring the hTDG gene was transformed into BL21 (DE3) Escherichia coli (Invitrogen). Thaw one vial of competent BL21(DE3) cells on ice for 5 min. Add 1 μL of 50 ng/μL hTDG plasmid. The cells were gently mixed and incubated on ice for 30 minutes, then subjected to a heat shock at 42°C for 45 seconds. Post heat shock, the cells were cultured in 1 mL of SOC medium at 37°C for 1 hour. Subsequently, 50 μL of the culture was spread on a Kanamycin-supplemented agar plate (50 mg/mL) and incubated overnight at 37°C. A single colony was selected the following day and expanded in 20 mL LB broth with 50 mg/mL Kanamycin. This culture was further diluted into 4 L LB medium. When OD_{600} reached 0.6, isopropyl β-d-1- thiogalactopyranoside (IPTG) was added to a concentration of 0.5 mM to induce target protein expression., which was subsequently grown overnight at 25 °C. Bacteria were harvested and lysed in 60 mL lysis buffer (20 mM Tris-HCl, 500 mM NaCl, pH 7.5) by sonication for 15 min (10 sec on and 20 sec off, 30 cycles) on ice.

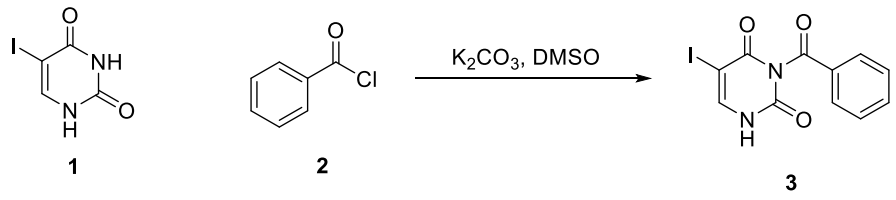
His-tagged hTDG was purified from the cleared lysate with the Histrap HP column (GE healthcare). The protein was eluted in elution buffer (20 mM Tris-HCl, 500 mM NaCl, 300 mM imidazole, pH 7.5) and further purified by gel filtration chromatography using a Hi-Load Superdex 200 16/60 column (GE Healthcare) in an FPLC (GE healthcare). The peak fraction was collected and concentrated using Amicon Ultra 10K centrifugal filter Device (Millipore).

2.4.2 Western blot

Western blot was used to validate the purified enzyme. The concentration of proteins was determined using a NanoDrop 8000 Spectrophotometer (Thermo Scientific). Subsequently, the lysates were heated to 95°C in the presence of 4× loading buffer (Biorad) for 5 minutes. The denatured proteins were loaded onto a 4-12% NuPAGE Bis-Tris gel and subsequently transferred to PVDF membranes (Life Technologies). The membranes were then blocked in PBST containing 5% milk for 30 minutes at room temperature. Next, they were incubated in a diluted primary antibody solution at 4°C overnight, followed by thorough washing and subsequent incubation with a secondary antibody conjugated to HRP for 1 hour at room temperature. Protein bands were visualized using the SuperSignal West Dura Extended Duration Substrate kit (Thermo) with a FluroChem R system (Proteinsimple).

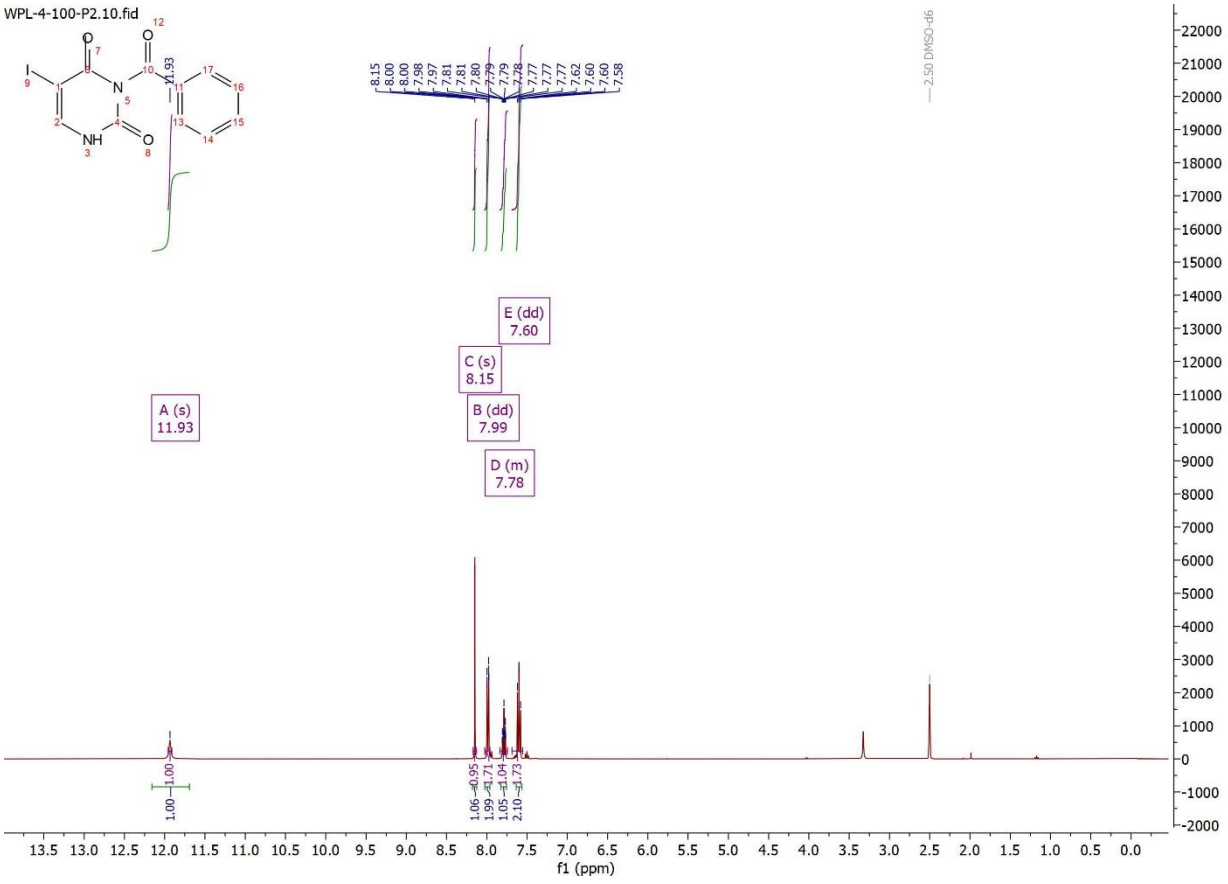
2.4.3 Synthesize of nucleotide derivatives

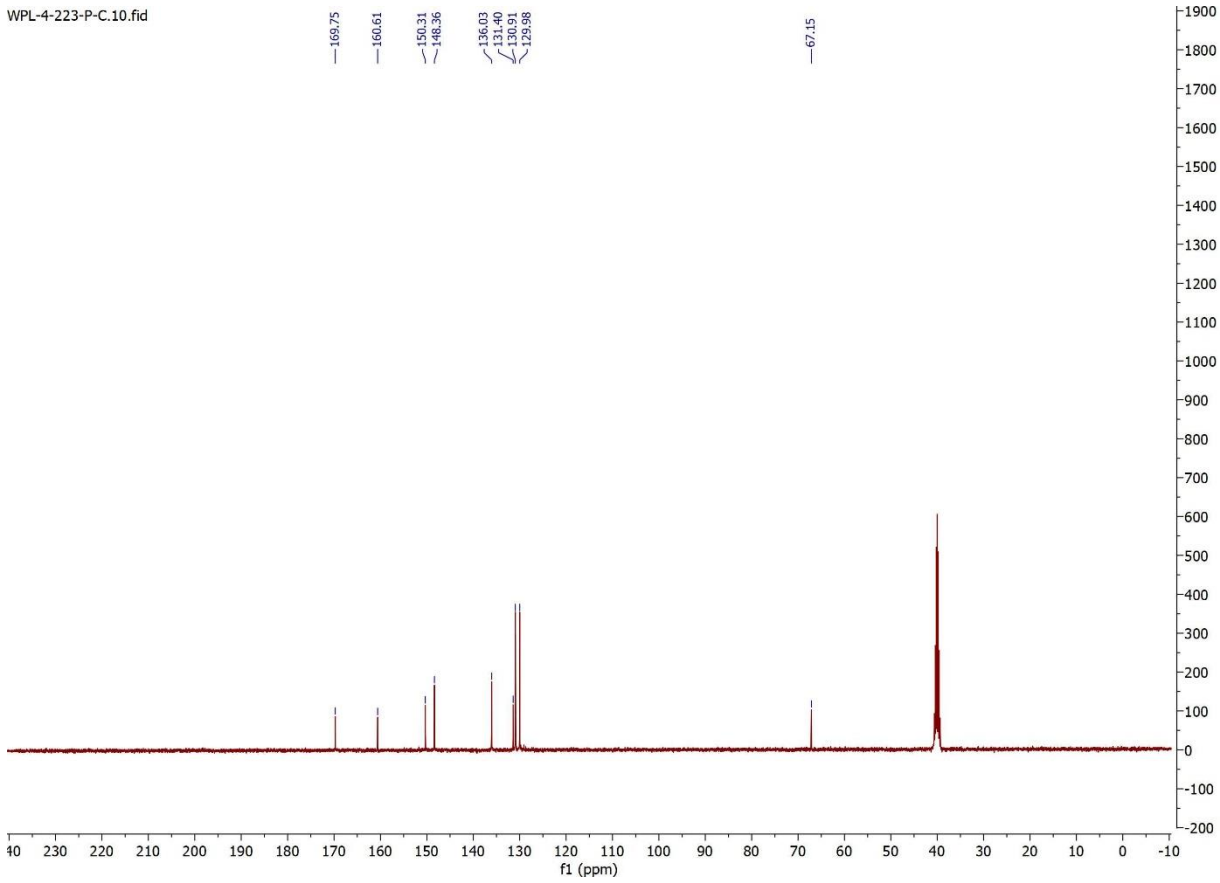
2.4.3.1 Synthesize of Azide-Thymine

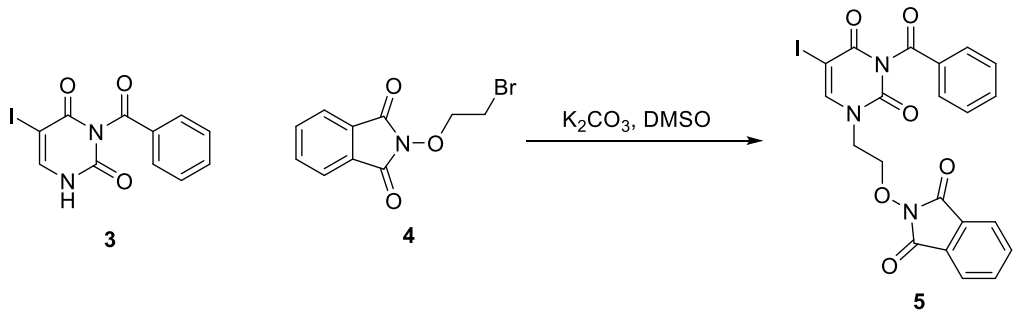


Synthesis of compound **3**: To a solution of 5-iodouracil (**1**, 1.0 g, 4.2020 mmol) in a mixture of pyridine and acetonitrile (2/5, v/v, 21 mL) was added Benzoyl chloride (**2**, 1.5 mL, 12.606 mmol). The reaction mixture was stirred for 24 h at room temperature. After evaporation of all the volatiles, the residue was purified by silica gel column chromatography (eluting with 1:1 hexanes/ethyl acetate) to give compound **3** (1.440 g, quant.) as a white foam.

^{1}H NMR (400 MHz, DMSO) δ 11.93 (s, 1H), 8.15 (s, 1H), 7.99 (dd, J = 8.4, 1.3 Hz, 2H), 7.84 – 7.74 (m, 1H), 7.60 (dd, J = 8.3, 7.4 Hz, 2H). ^{13}C NMR (101 MHz, DMSO) δ 169.75, 160.61, 150.31, 148.36, 136.03, 131.40, 130.91, 129.98, 67.15. HRMS $C_{11}H_7IN_2NaO_3^+$ $[M+Na]^+$ calculated 364.9394, found 364.9389.

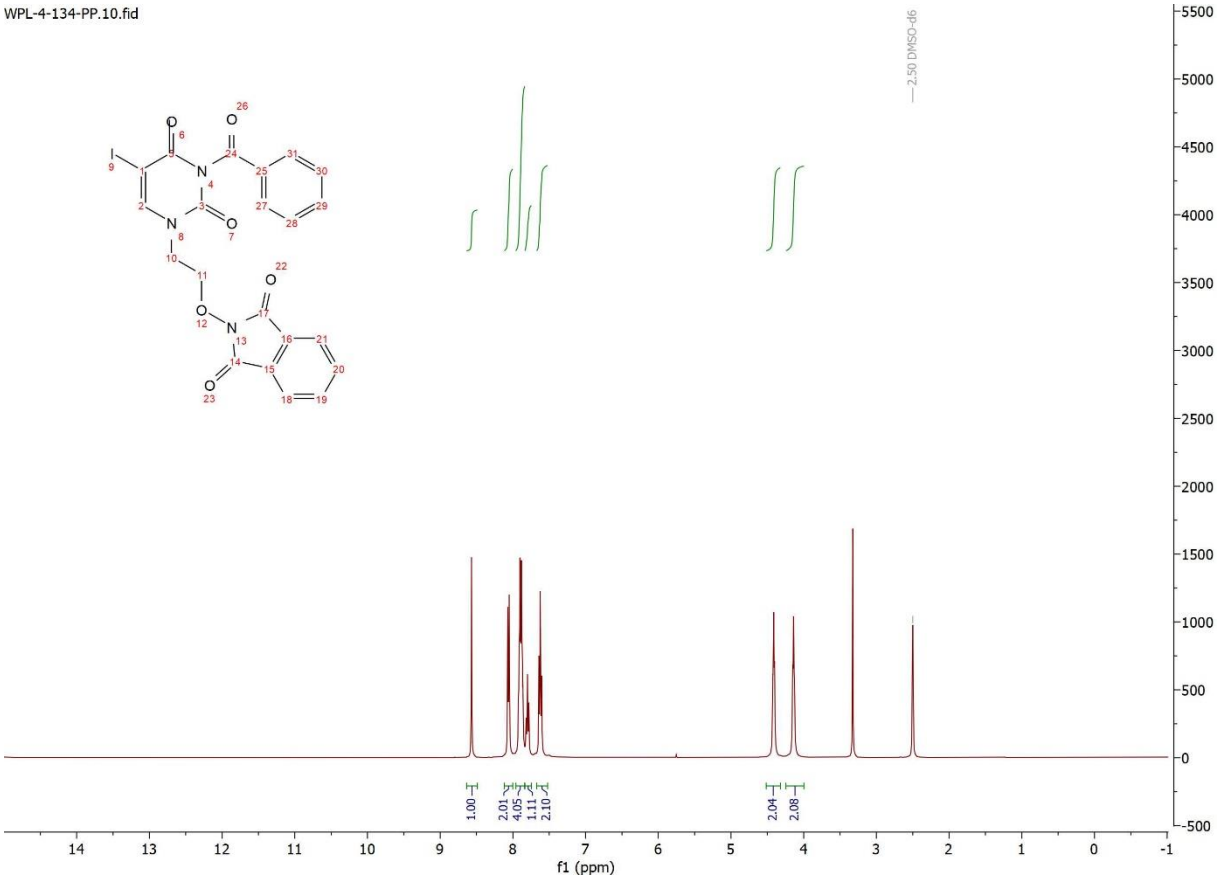


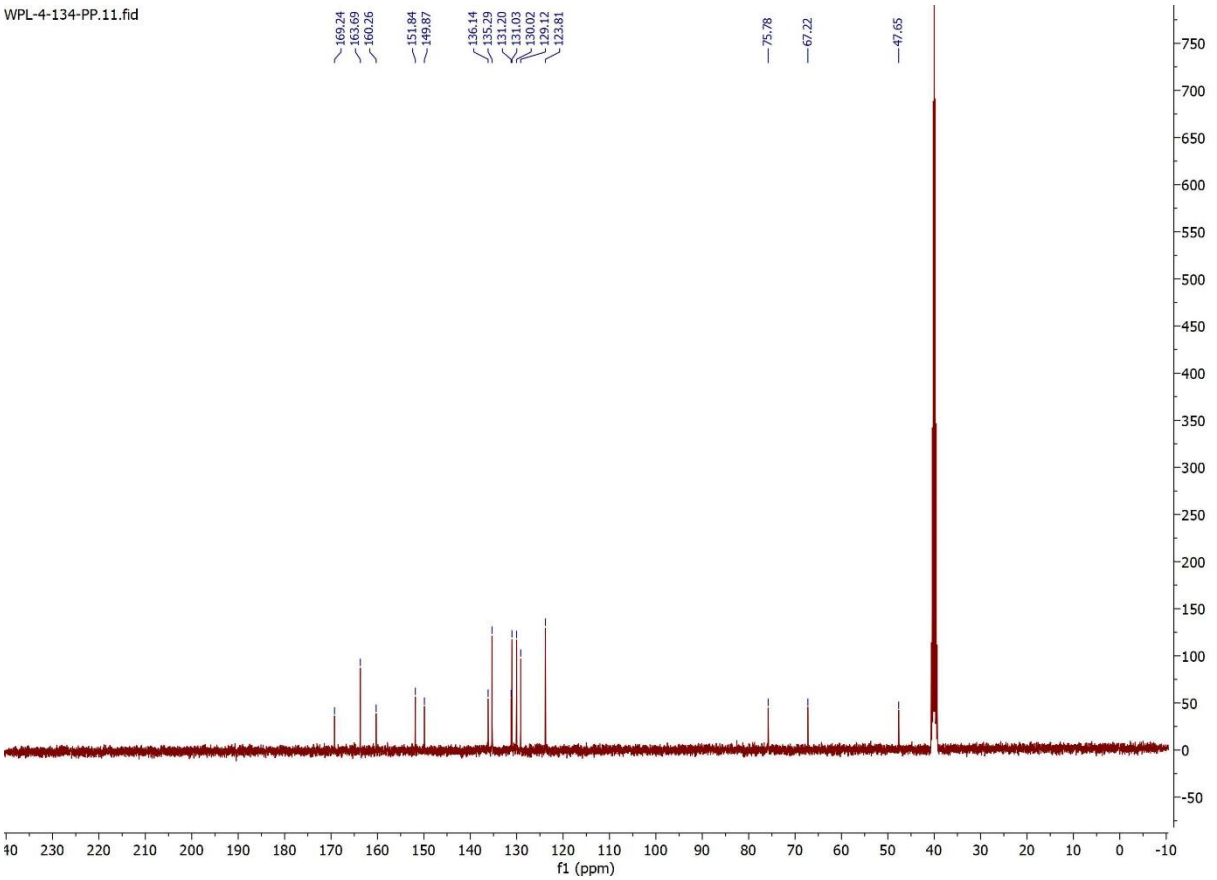

WPL-4-223-P-C.10.fid
169.75
160.61
150.31
148.36
136.03
131.40
130.91
129.98
67.15
f1 (ppm)

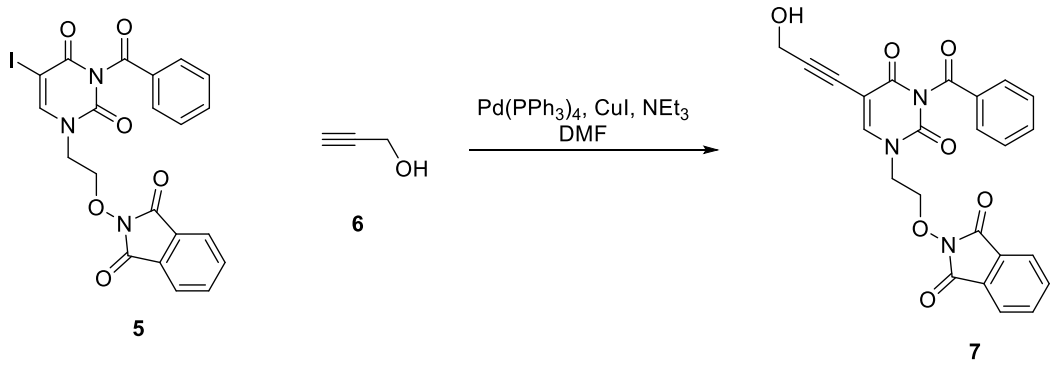


Synthesis of compound **5**: To a stirred mixture solution of compound **3** (1.368 g, 4.0 mmol) and *N*-(2-bromoethoxy) phthalimide (**4**, 1.080 g, 4.0 mmol) in DMSO (15 mL) was added potassium carbonate (552 mg, 4 mmol). The resulting mixture was stirred for 3 h at room temperature before being diluted with water. The mixture was extracted by ethyl acetate and the combined organic layers were washed with brine three times, dried over anhydrous sodium sulfate. Filtered and concentrated. The crude product was purified by flash column chromatography (eluting with 1:1 hexanes/acetone) to afford compound **5** (1.531 g, 90%) as a white foam.

^{1}H NMR (400 MHz, DMSO) δ 8.57 (s, 1H), 8.06 (d, *J* = 7.8 Hz, 2H), 7.90 (dtt, *J* = 8.8, 5.8, 3.7 Hz, 4H), 7.80 (t, *J* = 7.4 Hz, 1H), 7.62 (t, *J* = 7.7 Hz, 2H), 4.41 (t, *J* = 5.0 Hz, 2H), 4.14 (t, *J* = 5.1 Hz, 2H). ^{13}C NMR (101 MHz, DMSO) δ 169.24, 163.69, 160.26, 151.84, 149.87, 136.14, 135.29, 131.20, 131.03, 130.02, 129.12, 123.81, 75.78, 67.22, 47.65. HRMS $C_{21}H_{15}IN_3O_6^+$ $[M+H]^+$ calculated 532.0000, found 532.0005.

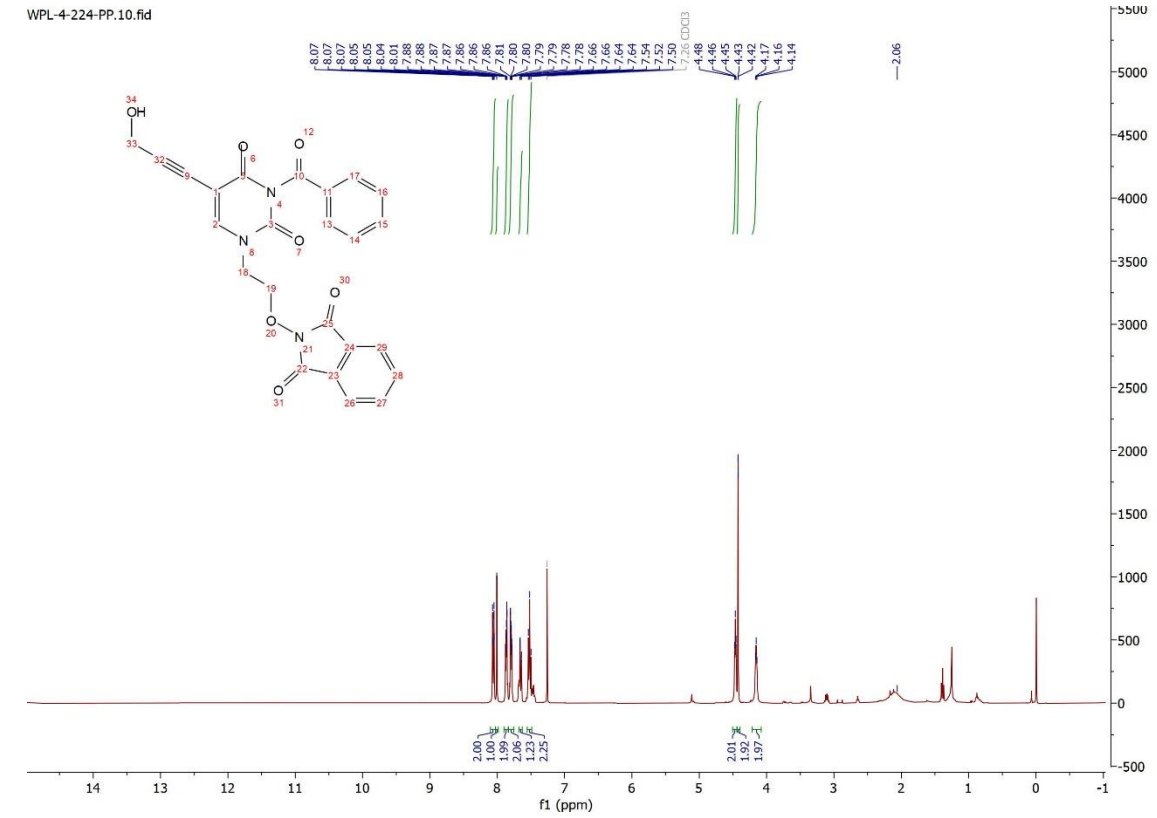

WPL-4-134-PP.10.fid
2.50 DMSO-d6
1.00
2.01
4.05
1.11
2.10
2.04
2.08
f1 (ppm)

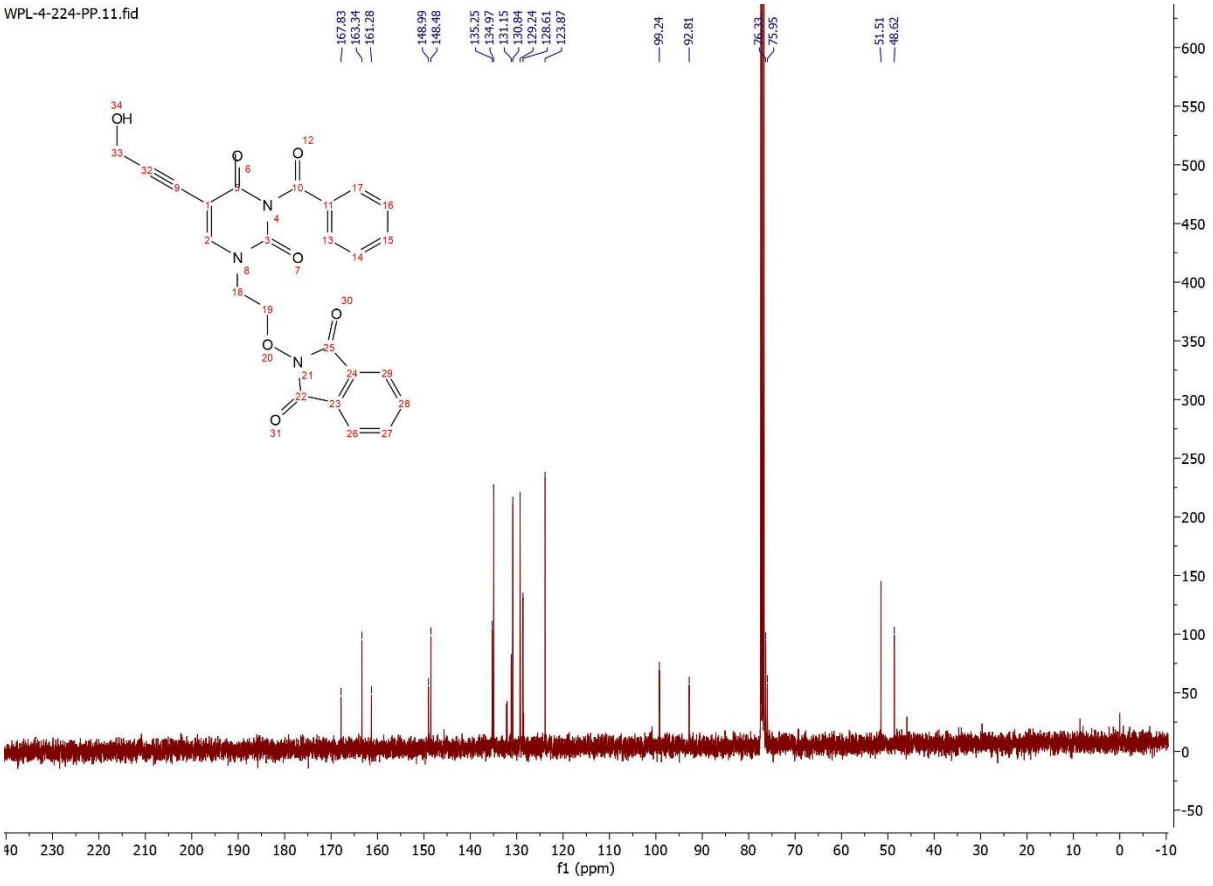

WPL-4-134-PP.11.fid
169.24
163.69
160.26
151.84
149.87
136.14
135.29
131.20
131.03
130.02
129.12
123.81
75.78
67.22
47.65
f1 (ppm)

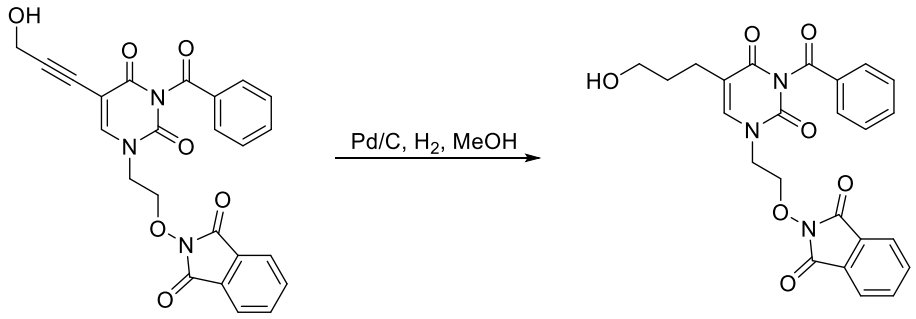


Synthesis of compound **7**: To a 25 mL sealed tube was added compound **5** (642 mg, 1.208 mmol), CuI (23 mg, 0.121 mmol), Pd $(PPh_3)_4$ (70 mg, 0.061 mmol) and degassed DMF (5 mL). The resulting solution was sequentially added 2-Propyn-1-ol (**6**, 140 μL, 2.417 mmol) and NEt_3 (353 μL, 2.537 mmol). The reaction mixture was avoided light and stirred for 24 h at room temperature. After that, the reaction mixture was quenched with water and extracted by ethyl acetate. The combined organic layers were washed with brine three times, dried over anhydrous sodium sulfate, and concentrated in vacuo. The crude product was purified by flash column chromatography (eluting with 1:1 hexanes/acetone) to afford product **7** (474 mg, 86%) as a white foam.

^{1}H NMR (400 MHz, $CDCl_3$) δ 8.06 (d, *J* = 7.1 Hz, 2H), 8.01 (s, 1H), 7.87 (dd, *J* = 5.5, 3.1 Hz, 2H), 7.80 (dd, *J* = 5.5, 3.1 Hz, 2H), 7.66 (dd, *J* = 7.5, 2.0 Hz, 2H), 7.52 (t, *J* = 7.7 Hz, 2H), 4.46 (t, *J* = 4.5 Hz, 2H), 4.42 (d, *J* = 0.9 Hz, 2H), 4.15 (t, *J* = 4.5 Hz, 2H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 167.83, 163.34, 161.28, 148.99, 148.48, 135.25, 134.97, 131.15, 130.84, 129.24, 128.61, 123.87, 99.24, 92.81, 76.33, 51.51, 48.62. HRMS $C_{24}H_{18}N_3O_7^+$ $[M+H]^+$ calculated 460.1139, found 460.1138.


WPL-4-224-PP.10.fid
f1 (ppm)

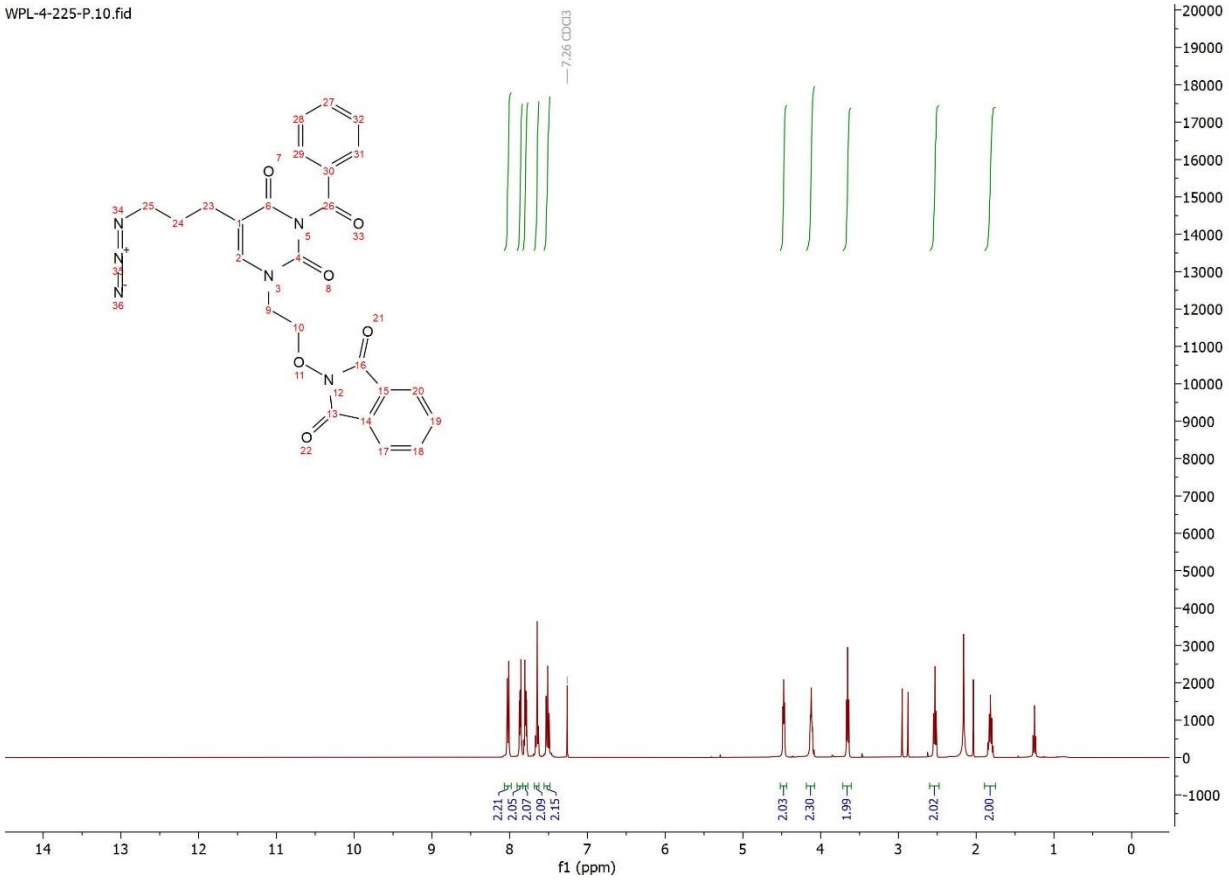

WPL-4-224-PP.11.fid
f1 (ppm)

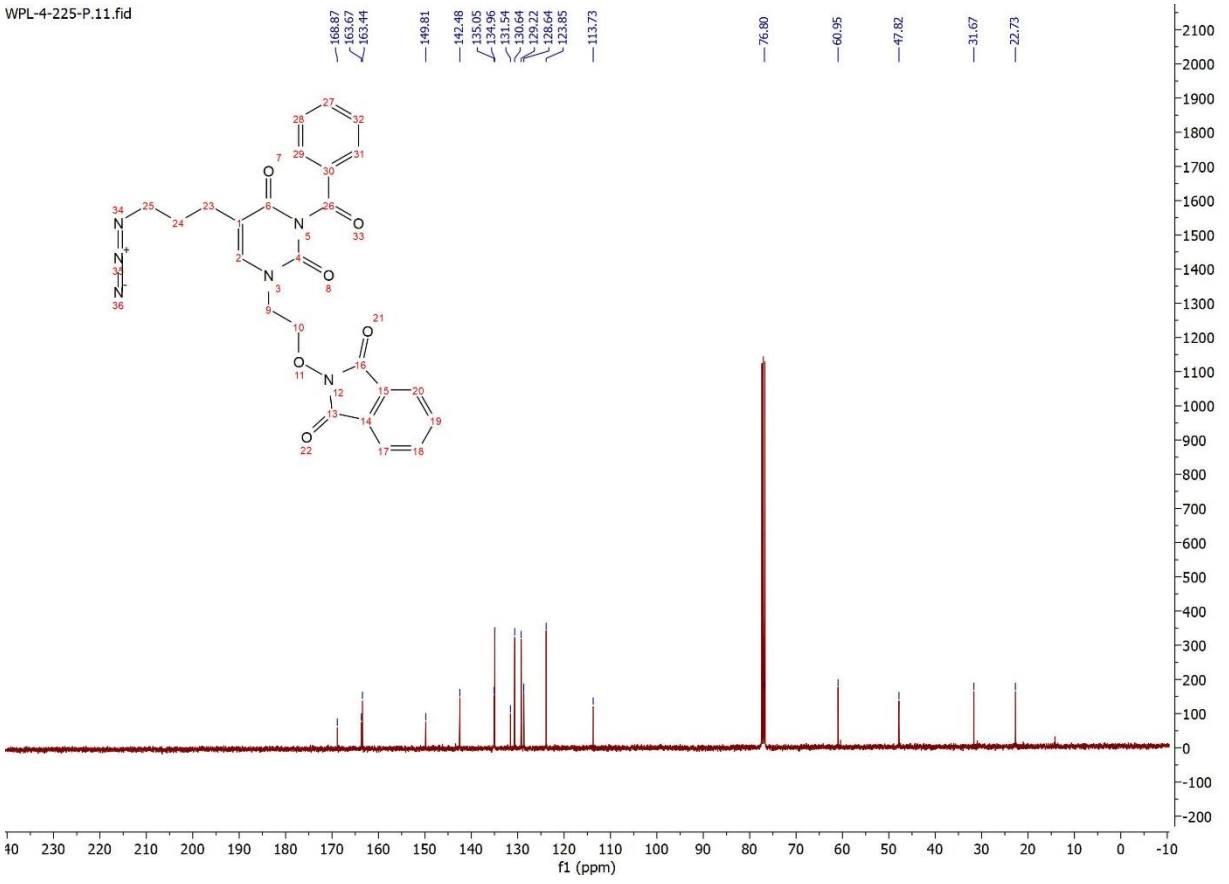


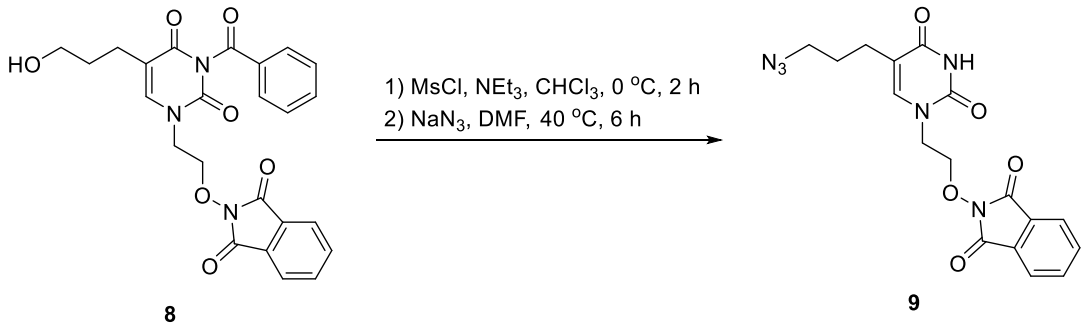
7 **8**

Synthesis of compound **8**: To a stirred solution of compound **7** (280 mg, 0.61 mmol) in a mixture of MeOH and acetone (20/1, v/v, 21 mL) was added Pd/C (28 mg, 10% wt). The resulting mixture was added a H_2 balloon and stirred for 30 min at room temperature. The mixture was filtered and evaporated all the volatiles. The crude product was purified by flash column chromatography (eluting with 1:1 hexanes/acetone) to afford compound **8** (122 mg, 44%) as a white foam.

^{1}H NMR (400 MHz, $CDCl_3$) δ 8.06 – 7.99 (m, 1H), 7.86 (dd, *J* = 5.5, 3.1 Hz, 1H), 7.79 (dd, *J* = 5.5, 3.1 Hz, 1H), 7.64 (d, *J* = 6.7 Hz, 1H), 7.51 (t, *J* = 7.8 Hz, 1H), 4.48 (t, *J* = 4.4 Hz, 1H), 4.16 – 4.06 (m, 1H), 3.65 (t, *J* = 6.0 Hz, 1H), 2.53 (t, *J* = 7.1 Hz, 1H), 1.82 (p, *J* = 6.6 Hz, 1H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 168.87, 163.67, 163.44, 149.81, 142.48, 135.05, 134.96, 131.54, 130.64, 129.22, 128.64, 123.85, 113.73, 60.95, 47.82, 31.67, 22.73. HRMS $C_{24}H_{22}N_3O_7^+$ $[M+H]^+$ calculated 464.1452, found 464.1460.


WPL-4-225-P.10.fid
7.26 CDCl3
2.21
2.05
2.09
2.15
2.03
2.30
1.99
2.02
2.00
f1 (ppm)


WPL-4-225-P.11.fid
168.87
163.67
163.44
149.81
142.48
135.05
134.96
131.54
130.64
129.22
128.64
123.85
113.73
76.80
60.95
47.82
31.67
22.73
f1 (ppm)

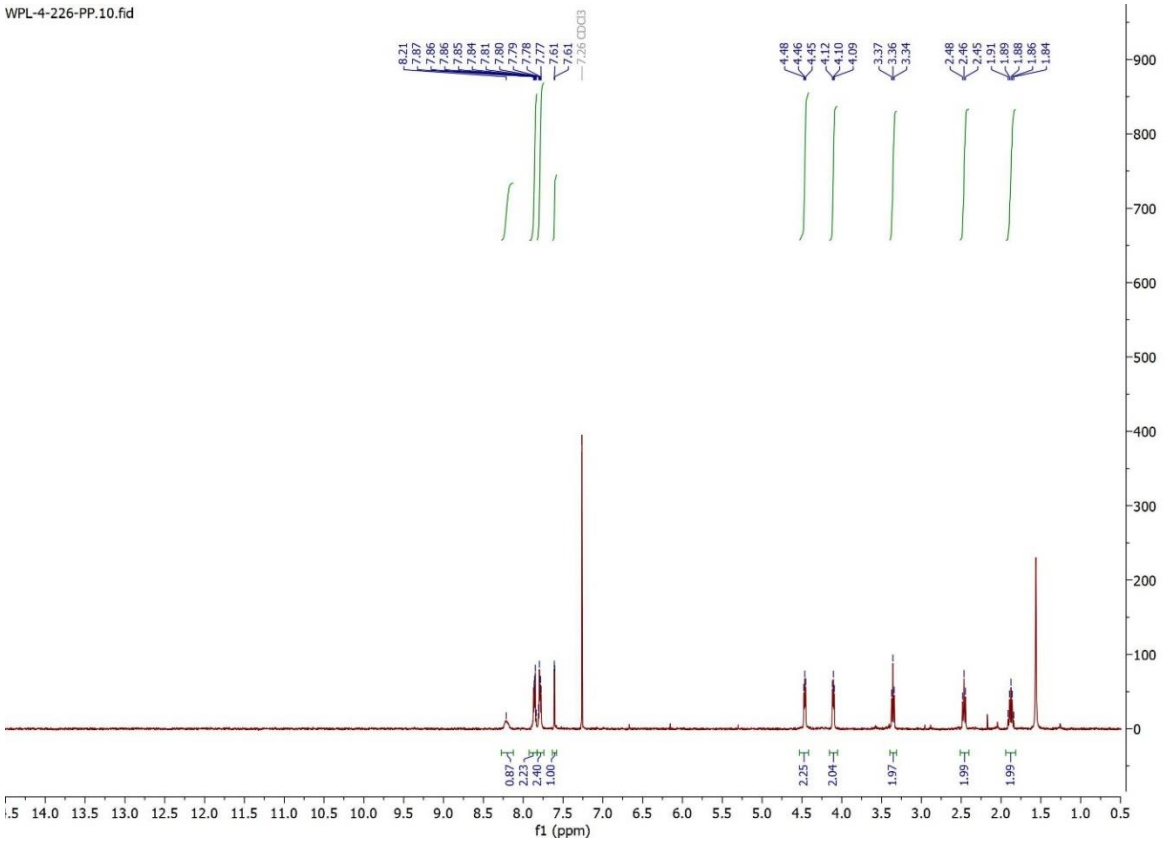


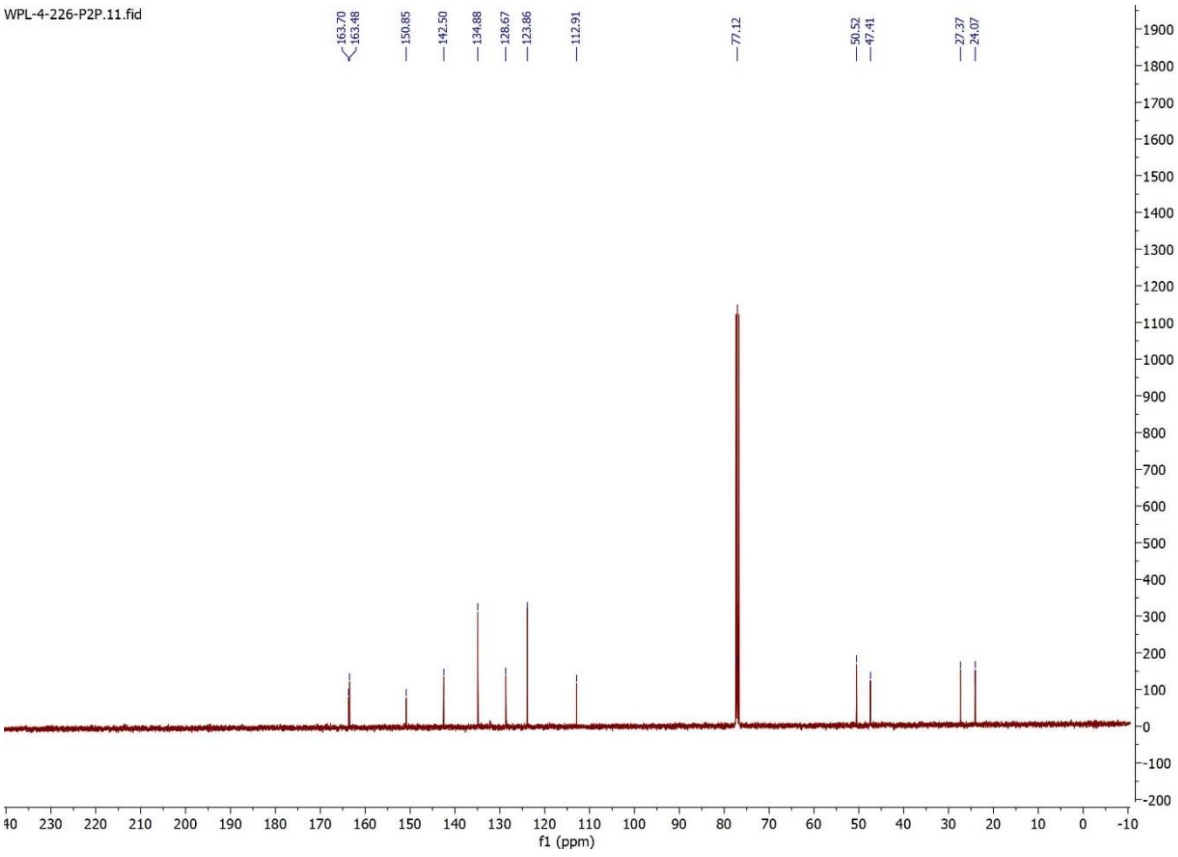
8 **9**

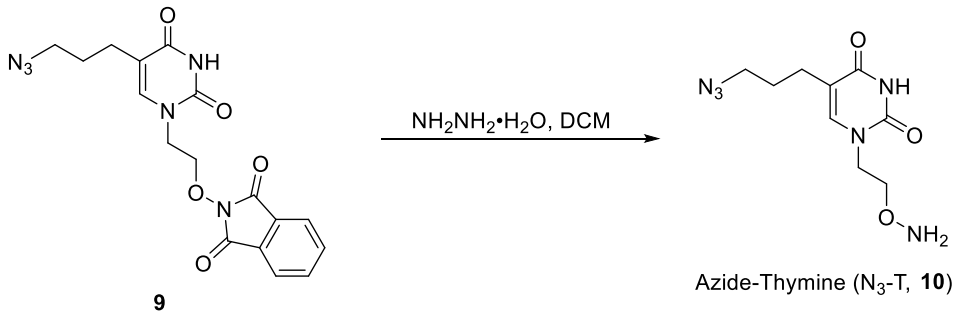
Synthesis of compound **9**: To a stirred solution of compound **8** (60 mg, 0.129 mmol) in $CHCl_3$ (5.0 mL) was cooled down to 0 °C in ice-water bath. Then the reaction solution was sequentially added NEt_3 (54 μL, 0.388 mmol) and MsCl (30 μL, 0.388 mmol). After that, the mixture warm to room temperature slowly. Keep stirring at room temperature for 1h and TLC starting materials 8was disappeared. Evaporation of all the volatiles, the residue was crude product and used directly without purification.

The crude methanesulfonyl product was dissolved in DMF (3.0 mL), directly. After that, the reaction solution was added NaN_3 (64 mg, 0.980 mmol). Then the resulting mixture was warmed to 40 °C and stirred for 6 h at this temperature. After being cooled to room temperature, the reaction mixture was quenched with water and extracted by ethyl acetate. The combined organic layers were washed with brine, dried over anthydrou sodium sulfide. Filtered and concentrated in vacuo. The crude product was purified by flash column chromatography (eluting with 1:1 hexanes/ethyl acetate) to afford compound **9** (30 mg, 61%) as a white foam.

^{1}H NMR (400 MHz, $CDCl_3$) δ 8.21 (s, 1H), 7.86 (dd, *J* = 5.5, 3.1 Hz, 2H), 7.79 (dd, *J* = 5.4, 3.2 Hz, 3H), 7.61 (d, *J* = 1.0 Hz, 1H), 4.50 – 4.43 (m, 2H), 4.11 (t, *J* = 4.6 Hz, 3H), 3.36 (t, *J* = 6.7 Hz, 2H), 2.46 (t, *J* = 7.3 Hz, 3H), 1.88 (p, *J* = 7.0 Hz, 3H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 163.70, 163.48, 150.85, 142.50, 134.88, 128.67, 123.86, 112.91, 50.52, 47.41, 27.37, 24.07. HRMS $C_{17}H_{17}N_6O_5^+$ $[M+H]^+$ calculated 385.1255, found 385.1251.

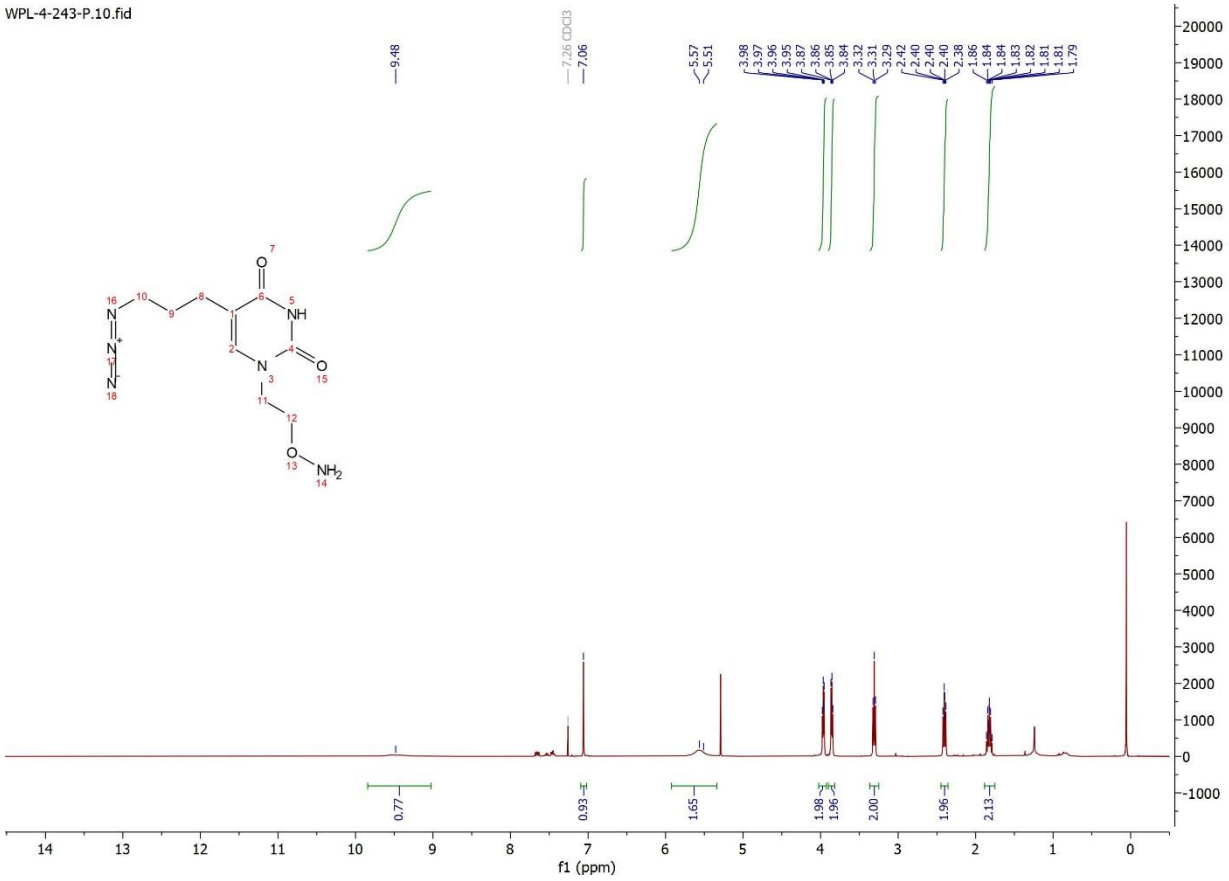

WPL-4-226-PP.10.fid
8.21
7.87
7.86
7.86
7.85
7.84
7.81
7.80
7.79
7.78
7.77
7.61
7.61
7.26 CDCl3
4.48
4.46
4.45
4.12
4.10
4.09
3.37
3.36
3.34
2.48
2.46
2.45
1.91
1.89
1.88
1.86
1.84
0.87
2.23
2.40
1.00
2.25
2.04
1.97
1.99
1.99
f1 (ppm)

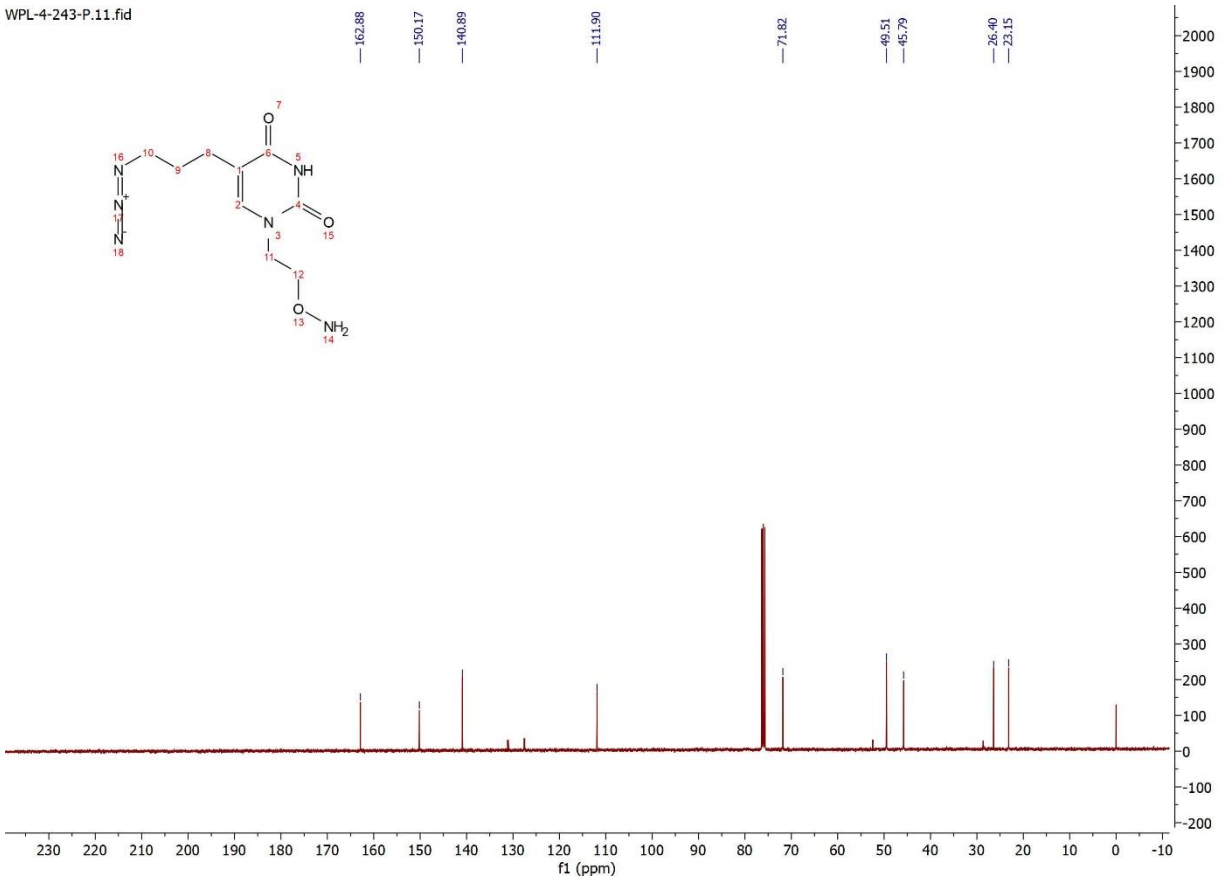

WPL-4-226-P2P.11.fid
163.70
163.48
150.85
142.50
134.88
128.67
123.86
112.91
77.12
50.52
47.41
27.37
24.07
f1 (ppm)



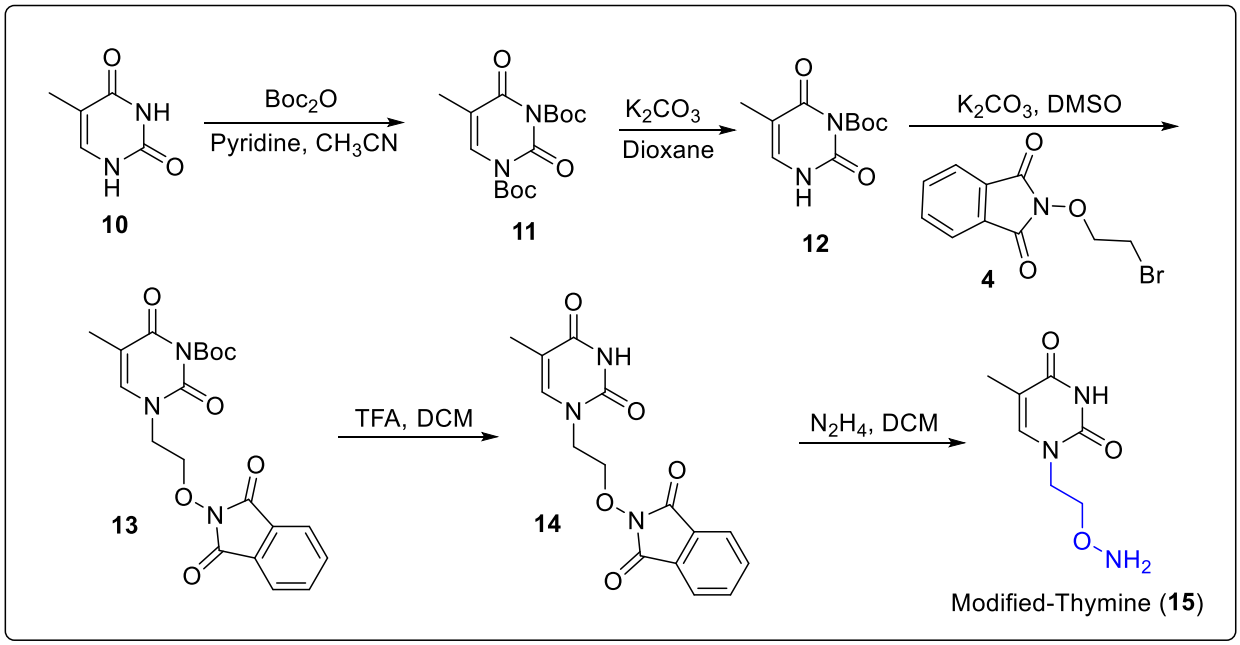
Synthesis of Azide-Thymine (N_3-T, **10**): To a stirred solution of compound **9** (48 mg, 0.125 mmol) in DCM (5.0 mL) was added hydrazinium hydroxide solution (8 μL, 0.25 mmol). The reaction mixture was stirred for 1 h at room temperature. After that, the mixture was filtered with 0.2 μm filter unit. Evaporation of all the volatiles, the residue was dissolved in 0.5 mL DCM and filtered with 0.2 μm filter unit. After that, evaporation of the DCM to afford the purified product azide-thymine (N3-T, **10**, 26 mg, 82%).

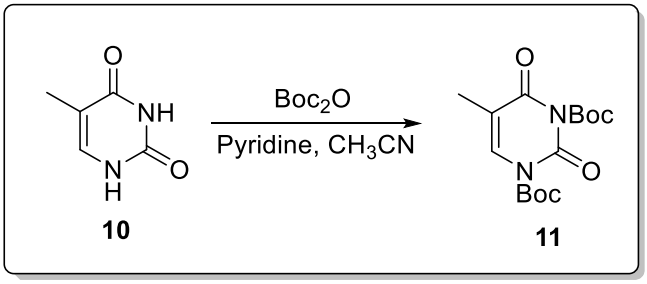
^{1}H NMR (400 MHz, $CDCl_3$) δ 9.48 (s, 1H), 7.06 (s, 1H), 5.57 (s, 2H), 3.96 (dd, *J* = 5.5, 3.9 Hz, 2H), 3.86 (dd, *J* = 5.4, 3.9 Hz, 2H), 3.31 (t, *J* = 6.5 Hz, 2H), 2.40 (dd, *J* = 8.0, 6.8 Hz, 2H), 1.82 (dq, *J* = 8.4, 6.6 Hz, 2H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 162.88, 150.17, 140.89, 111.90, 71.82, 49.51, 45.79, 26.40, 23.15. HRMS $C_9H_{15}N_6O_3^+$ $[M+H]^+$ calculated 255.1200, found 255.1201.


WPL-4-243-P.10.fid
9.48
7.26 CDCl3
7.06
5.57
5.51
3.98
3.97
3.96
3.95
3.87
3.86
3.85
3.84
3.32
3.31
3.29
2.42
2.40
2.40
2.40
2.38
1.86
1.84
1.84
1.83
1.82
1.81
1.81
1.79
0.77
0.93
1.65
1.98
1.96
2.00
1.96
2.13
f1 (ppm)

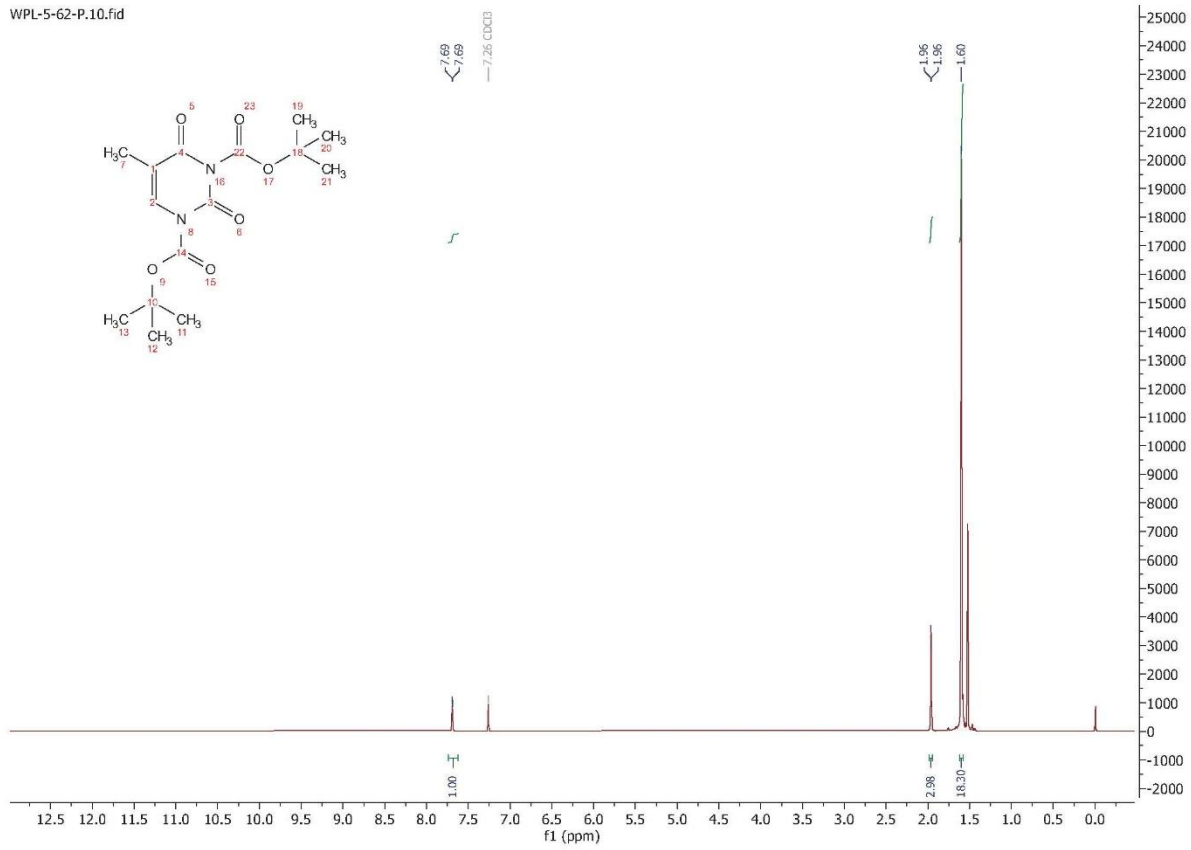

WPL-4-243-P.11.fid
162.88
150.17
140.89
111.90
71.82
49.51
45.79
26.40
23.15
f1 (ppm)

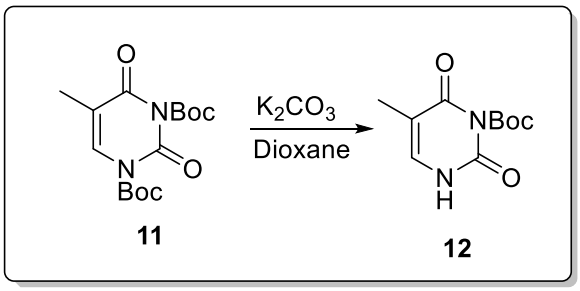
2.4.3.2 Synthesize of modified thymine





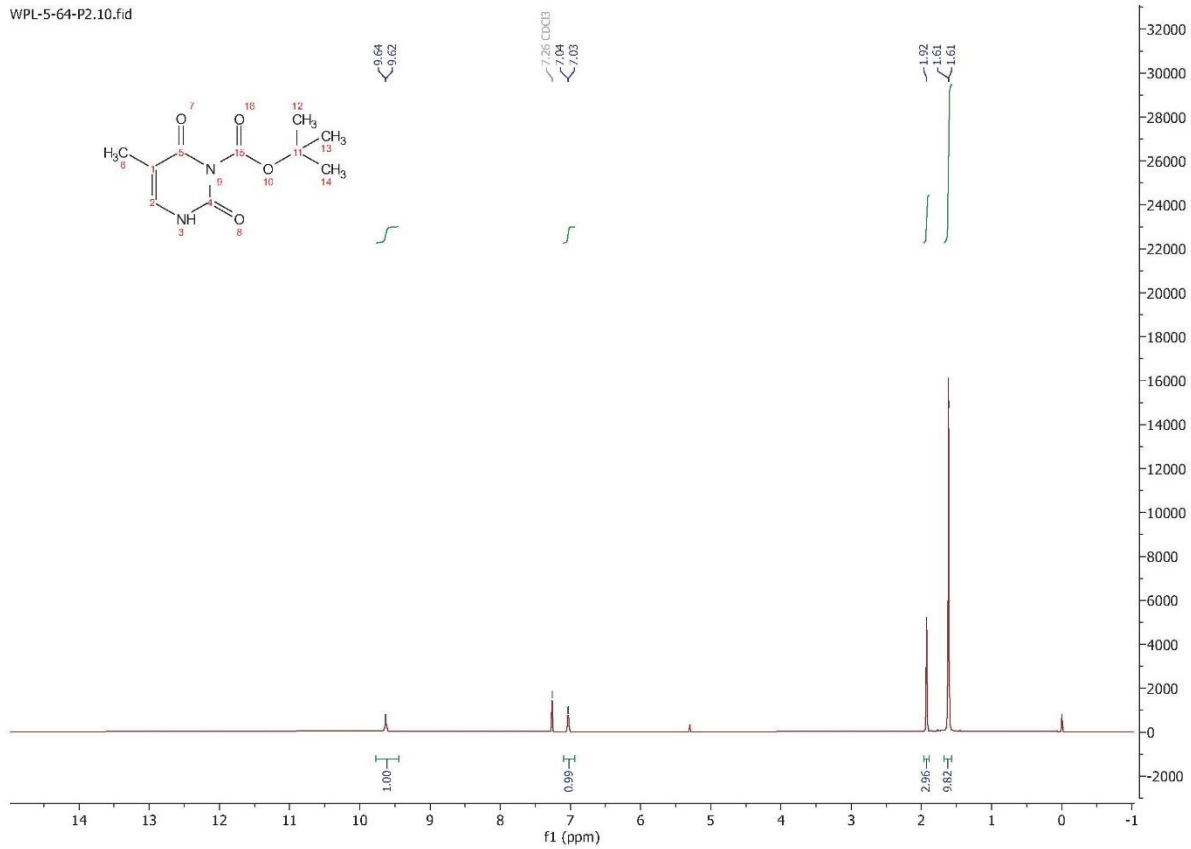
Synthesis of compound **11**: To a solution of thymine (**1**, 5.0 g, 39.64 mmol) in acetonitrile (75 mL) was added Boc_2O (25.92 g, 118.8 mmol). After the solid dissolved in the mixture solvents. The pyridine (15 mL) was added to the reaction and then the reaction warm to 50 °C for overnight. After evaporation of all the volatiles, the residue was purified by silica gel column chromatography (eluting with 5:1 hexanes/ethyl acetate) to give compound **11** (12.4 g, 95.8%) as a white foam. ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.69 (d, J = 1.3 Hz, 1H), 1.96 (d, J = 1.3 Hz, 4H), 1.60 (s, 18H).

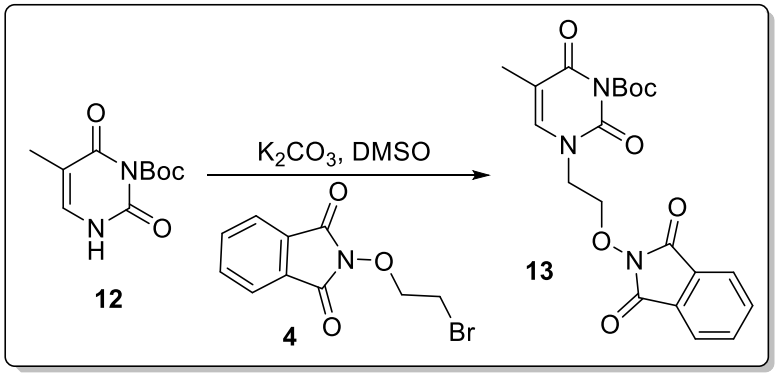




Synthesis of compound **12**: Di-tert-butyl 5-methyl-2,4-dioxopyrimidine-1,3(2H,4H)-dicarboxylate (**11**, 7.0 g, 21.46 mmol) in the 25 mL 1,4-dioxane at room temperature was added K_2CO_3 (4.45 g, 32.19 mmol) and the mixture was allowed to warm to 50 °C for 1 hour. TLC shows that the starting materials **11** was disappeared and cool the mixture to room

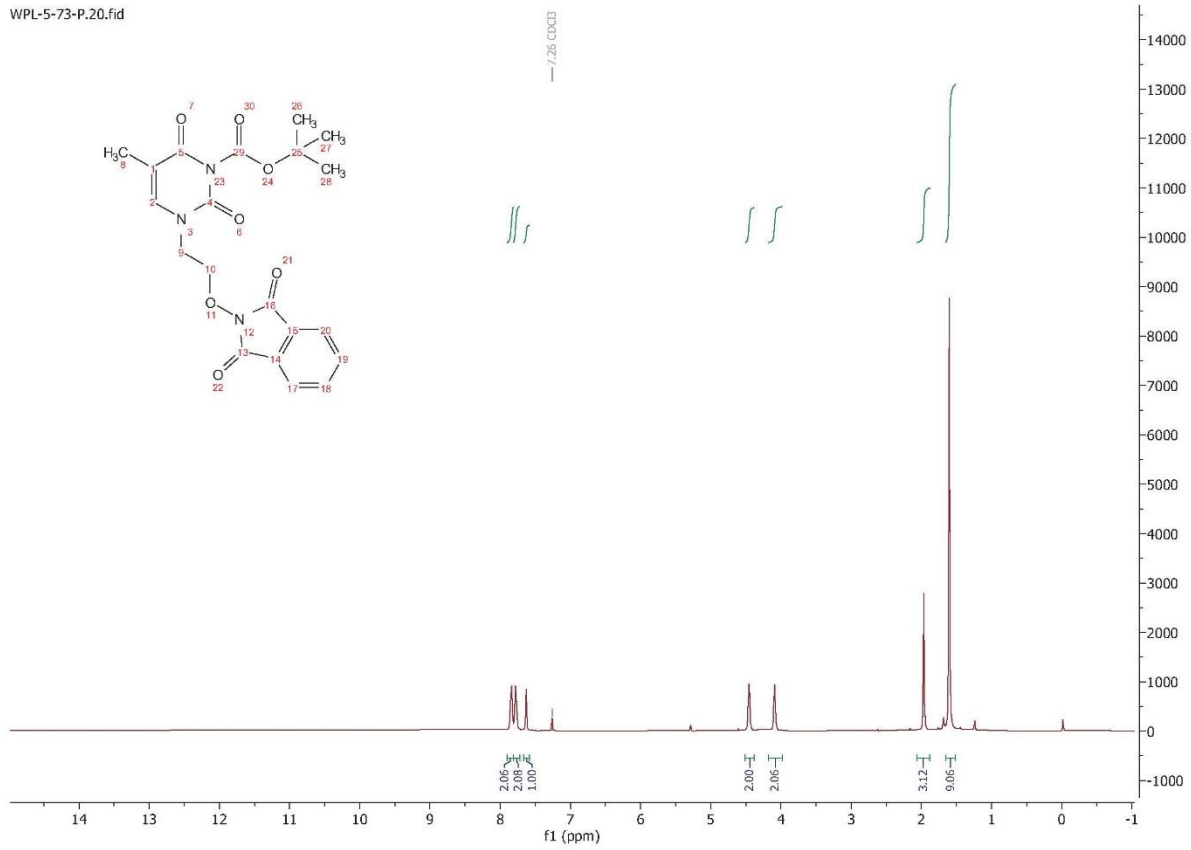
temperature. After that, the reaction mixture was quenched with water and extracted by dichloromethane. The combined organic layers were dried over anhydrous sodium sulfate and concentrated in vacuo. The crude product was purified by flash column chromatography (eluting with hexanes/acetone from 10:1 to 1:1) to afford product **12** (2.13 g, 44%) as white solid. ^{1}H NMR (400 MHz, $CDCl_3$) δ 9.63 (d, *J* = 5.5 Hz, 1H), 7.03 (d, *J* = 5.3 Hz, 1H), 1.92 (s, 3H), 1.64 – 1.59 (m, 9H).

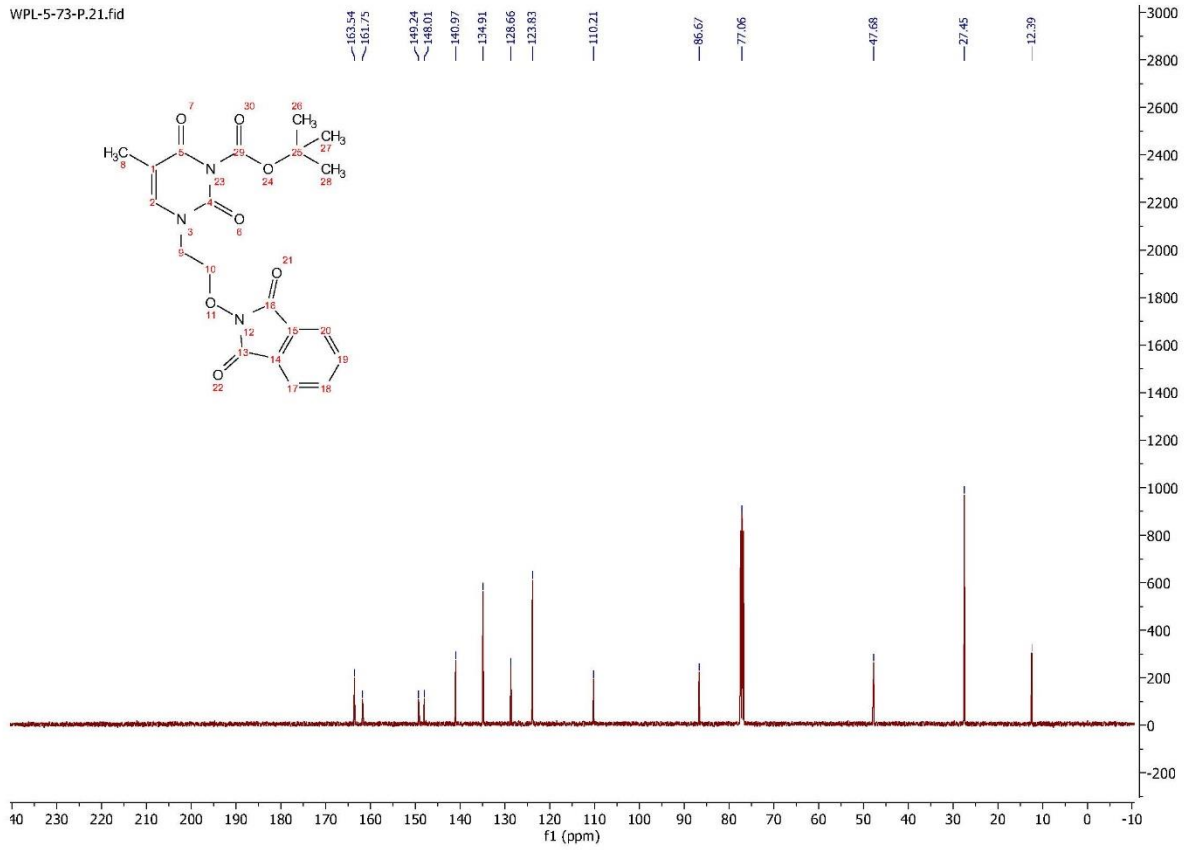


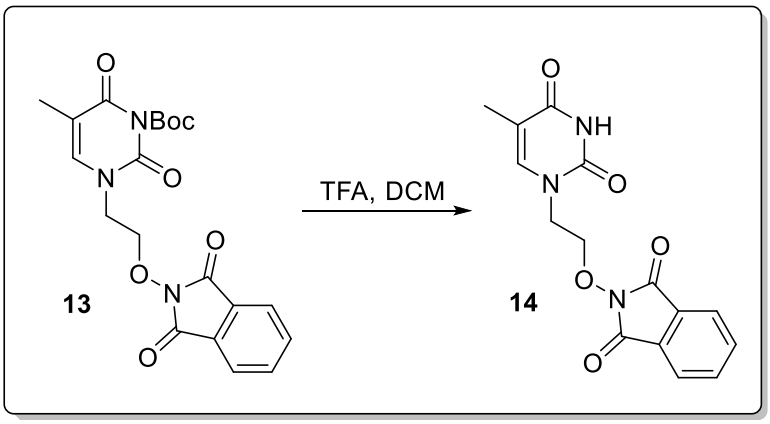


Synthesis of compound **13**: To a stirred mixture solution of compound **12** (226.09 mg, 1.0 mmol) and *N*-(2-bromoethoxy)phthalimide (**4**, 270.08 mg, 1.0 mmol) in DMSO (10 mL) was added potassium carbonate (138.21 mg, 1.0 mmol). The resulting mixture was stirred for 3 h at room temperature before being diluted with water. The mixture was extracted by ethyl acetate and the combined organic layers were washed with brine three times, dried over anhydrous sodium sulfate. Filtered and concentrated. The crude product was purified by flash column chromatography (eluting with 50:1 dichloromethane/methanol) to afford compound **13** (312 mg, 76%) as a white foam.

^{1}H NMR (400 MHz, $CDCl_3$) δ 7.84 (dd, *J* = 5.5, 3.2 Hz, 2H), 7.78 (dd, *J* = 5.5, 3.2 Hz, 2H), 7.63 (s, 1H), 4.46 (t, *J* = 4.5 Hz, 2H), 4.09 (t, *J* = 4.5 Hz, 2H), 1.97 (s, 3H), 1.60 (s, 9H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 163.54, 161.75, 149.24, 148.01, 140.97, 134.91, 128.66, 123.83, 110.21, 86.67, 77.06, 47.68, 27.45, 12.39. HRMS $C_{20}H_{22}N_3O_7^+$ $[M+H]^+$ calculated 416.1452, found 416.1446.

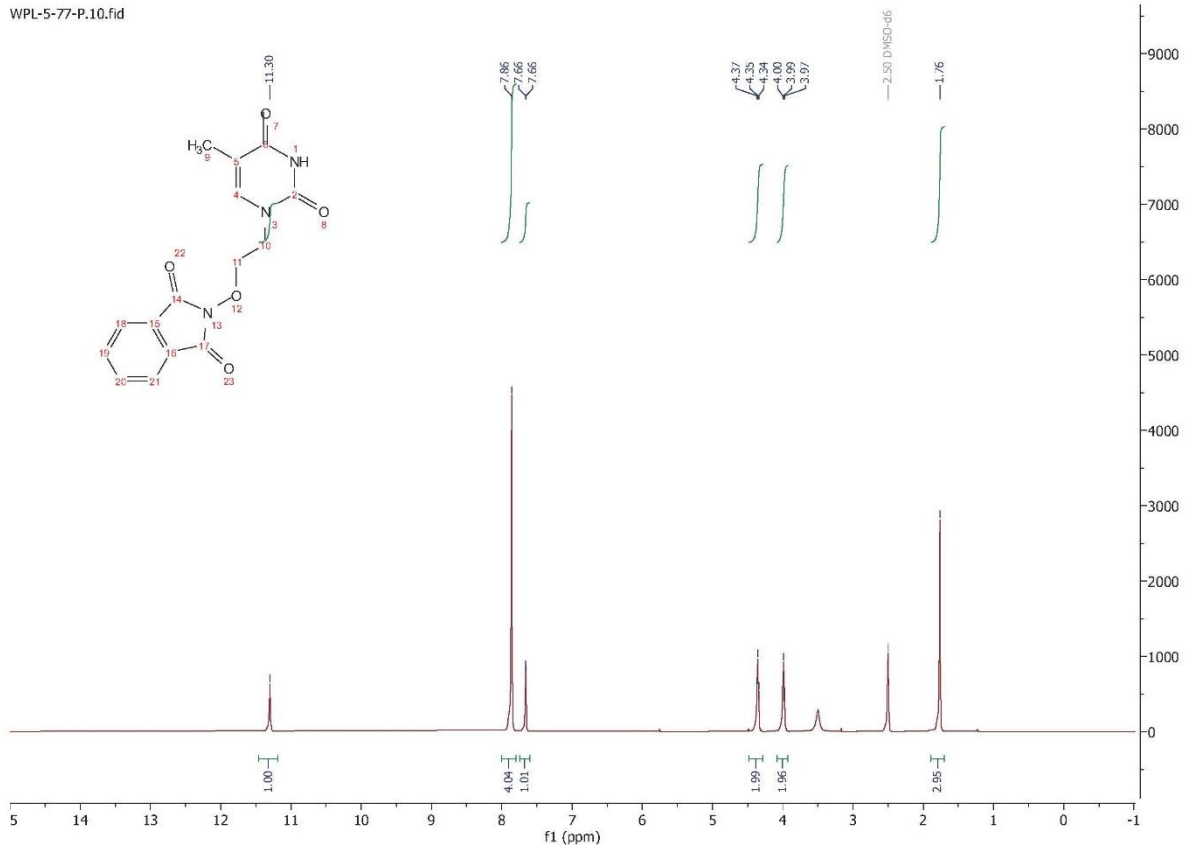


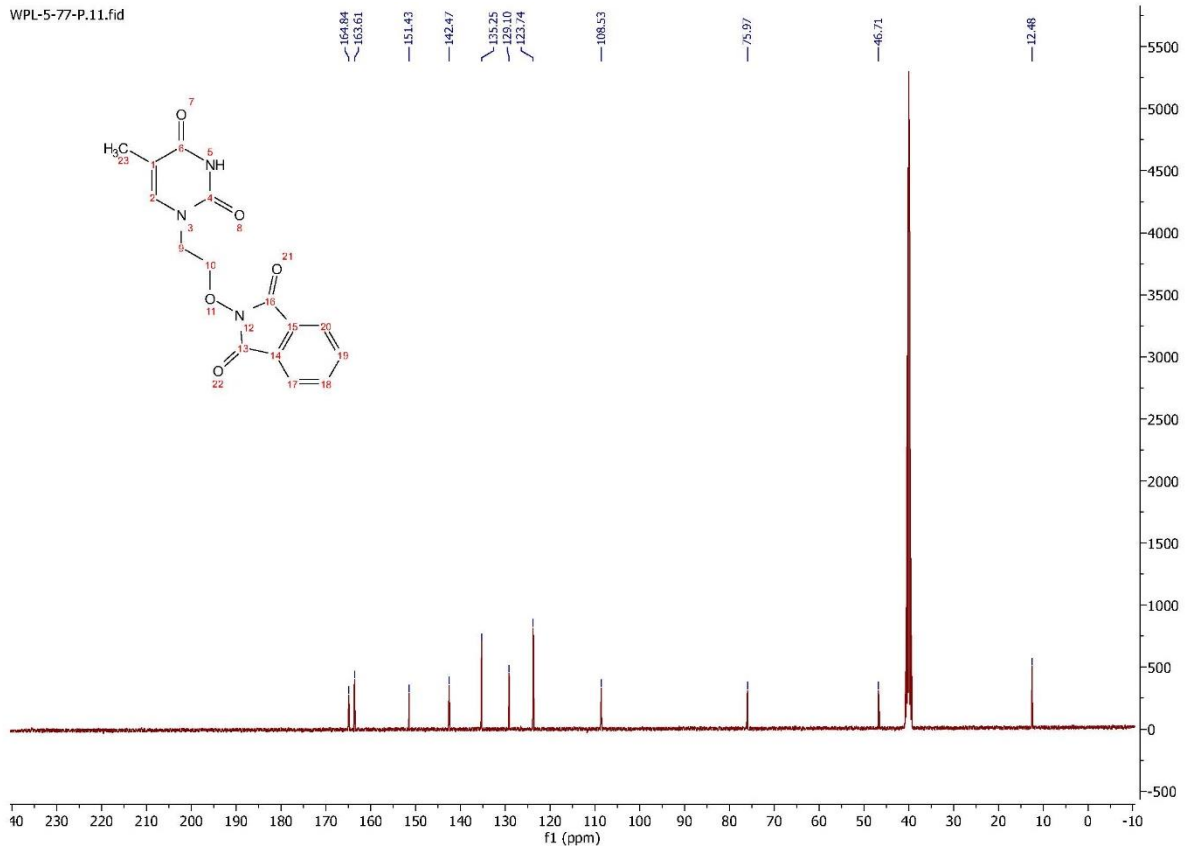


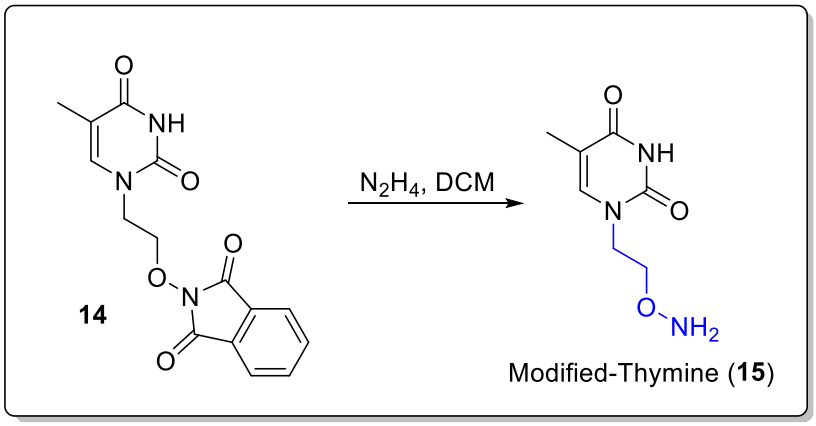


Synthesis of compound **14**: To a solution of compound **13** (120 mg, 0.24 mmol) in dichloromethane (6 mL), trifluoroacetic acid was added (1 mL, 17.16 mmol) at 0 °C and the mixture was allowed to reach room temperature and stirred for 2 hours. The mixture was concentrated to obtain a residue which was purified by flash column chromatography (eluting with 2:1 acetone/hexane) to afford compound **14** (75 mg, quant.).

^{1}H NMR (400 MHz, DMSO) δ 11.30 (s, 1H), 7.86 (s, 4H), 7.68 – 7.63 (m, 1H), 4.35 (t, *J* = 5.1 Hz, 2H), 3.99 (t, *J* = 5.1 Hz, 2H), 1.76 (s, 3H). ^{13}C NMR (101 MHz, DMSO) δ 164.84, 163.61, 151.43, 142.47, 135.25, 129.10, 123.74, 108.53, 75.97, 46.71, 12.48. HRMS $C_{15}H_{14}N_3O_5^+$ $[M+H]^+$ calculated 316.0928, found 316.0932.

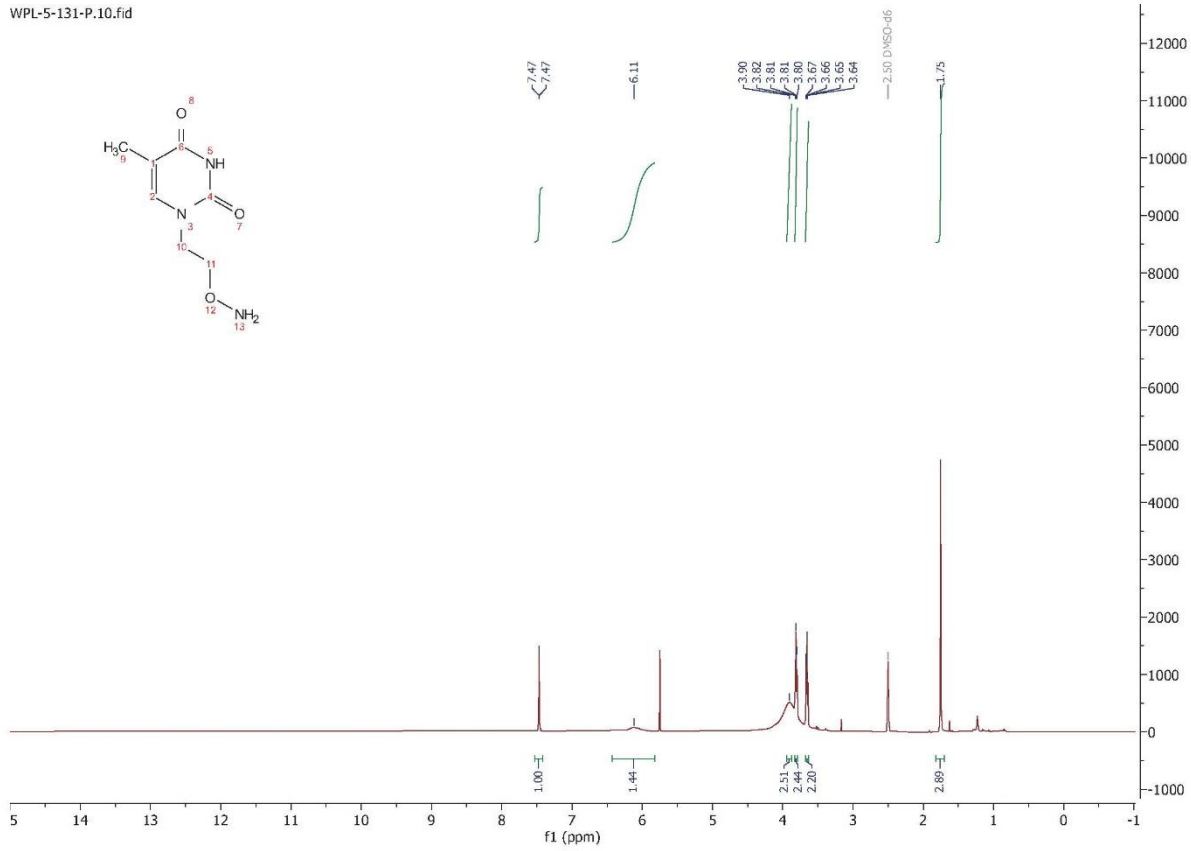


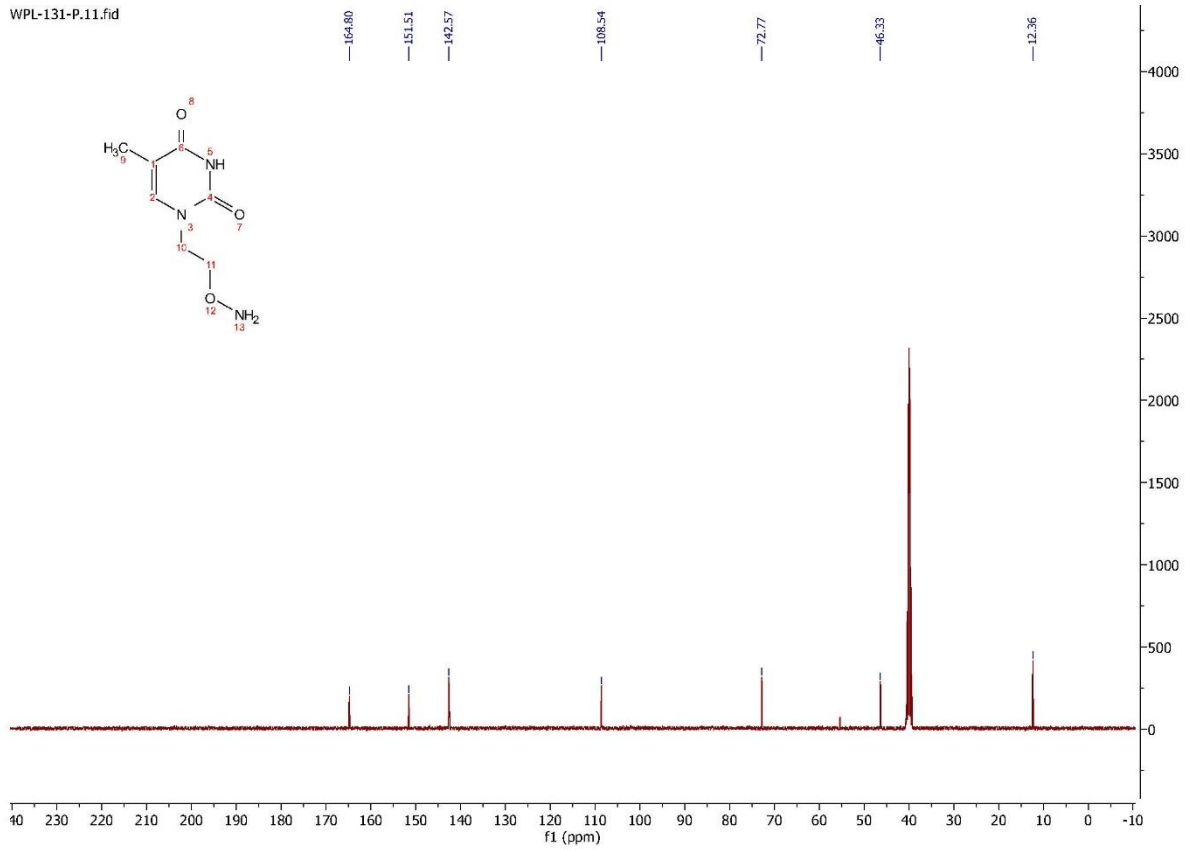




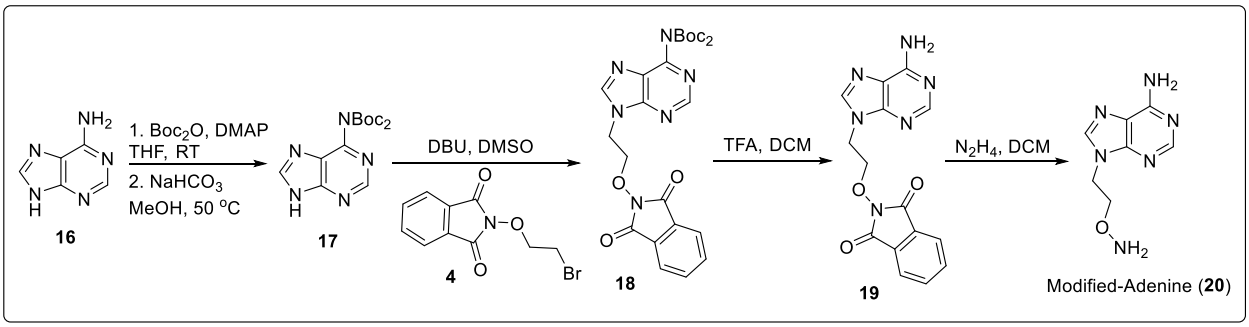
Synthesis of modified-thymine (**15**): To a stirred solution of compound **14** (80 mg, 0.254 mmol) in dichloromethane (5.0 mL) was added hydrazinium hydroxide solution (8.2 µL, 0.254 mmol). The reaction mixture was stirred for 3 hours at room temperature. After that, the mixture was filtered with 0.2 µm filter unit. Evaporation of all the volatiles, the residue was dissolved in 0.5 mL DCM and filtered with 0.2 µm filter unit, again. After that, evaporation of the DCM to afford the purified product modified-thymine (**15**, 28 mg, 60%).

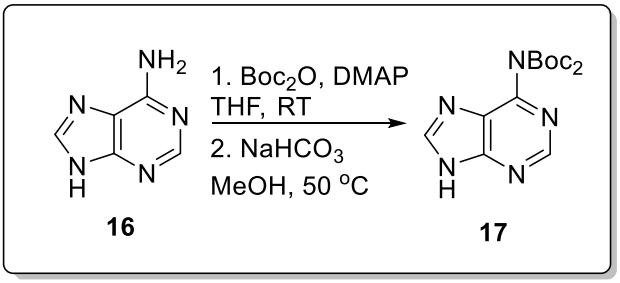
^{1}H NMR (400 MHz, DMSO) δ 7.47 (d, *J* = 1.3 Hz, 1H), 6.11 (s, 1H), 3.90 (s, 5H), 3.84 – 3.77 (m, 2H), 3.69 – 3.62 (m, 2H), 1.75 (s, 2H). ^{13}C NMR (101 MHz, DMSO) δ 164.80, 151.51, 142.57, 108.54, 72.77, 46.33, 12.36. HRMS $C_7H_{12}N_3O_3^+$ $[M+H]^+$ calculated 186.0873, found 186.0876.





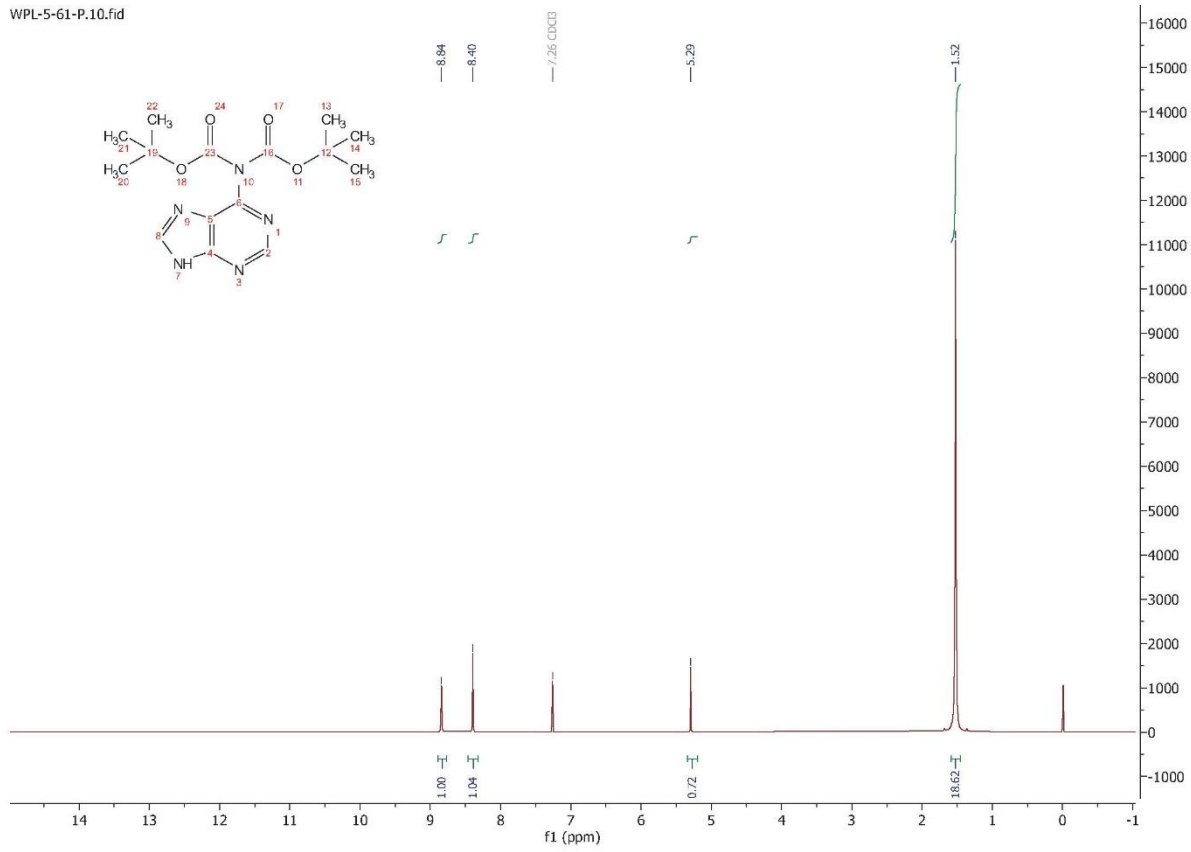
2.4.3.3 Synthesize of modified adenine

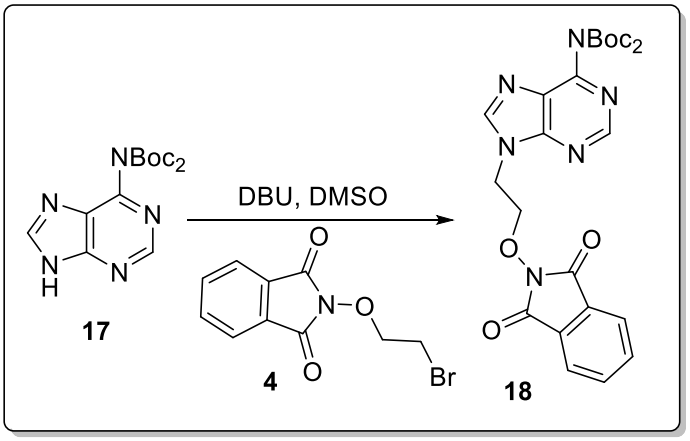




Synthesis of compound **17**: To a solution of adenine (**16**, 2.70 g, 20.0 mmol) in THF (100 mL) was added Boc_2O (17.77 g, 86 mmol) and 4‑(dimethylamino)pyridine (244 mg, 2.00 mmol). After 16 hours stirring at room temperature the solvent was removed under reduced pressure and the mixture taken up in ethyl acetate (500 mL). The organic phase was washed with HCl (1 M, 100 mL) and NaCl (sat., 100 mL) and dried over anhydrous sodium sulfate. Then filtered and concentrated under reduced pressure. The resulting slurry was redissolved in MeOH (200 mL), treated with $NaHCO_3$ (sat., 90 mL) and stirred for 1 h at 50 °C. The MeOH was then removed under reduced pressure, the residue solution diluted with H_2O and extracted with dichloromethane (3 x 100 mL). The organic phase was washed with HCl (1 M, 100 mL) and NaCl (sat., 100 mL) and then dried under reduced pressure affording the crude product which was purified by flash column chromatography (eluting with 1:1 hexanes/acetone) to afford compound **17** (4.12 g, 62%).

^{1}H NMR (400 MHz, $CDCl_3$) δ 8.84 (s, 1H), 8.40 (s, 1H), 5.29 (s, 1H), 1.52 (s, 18H).

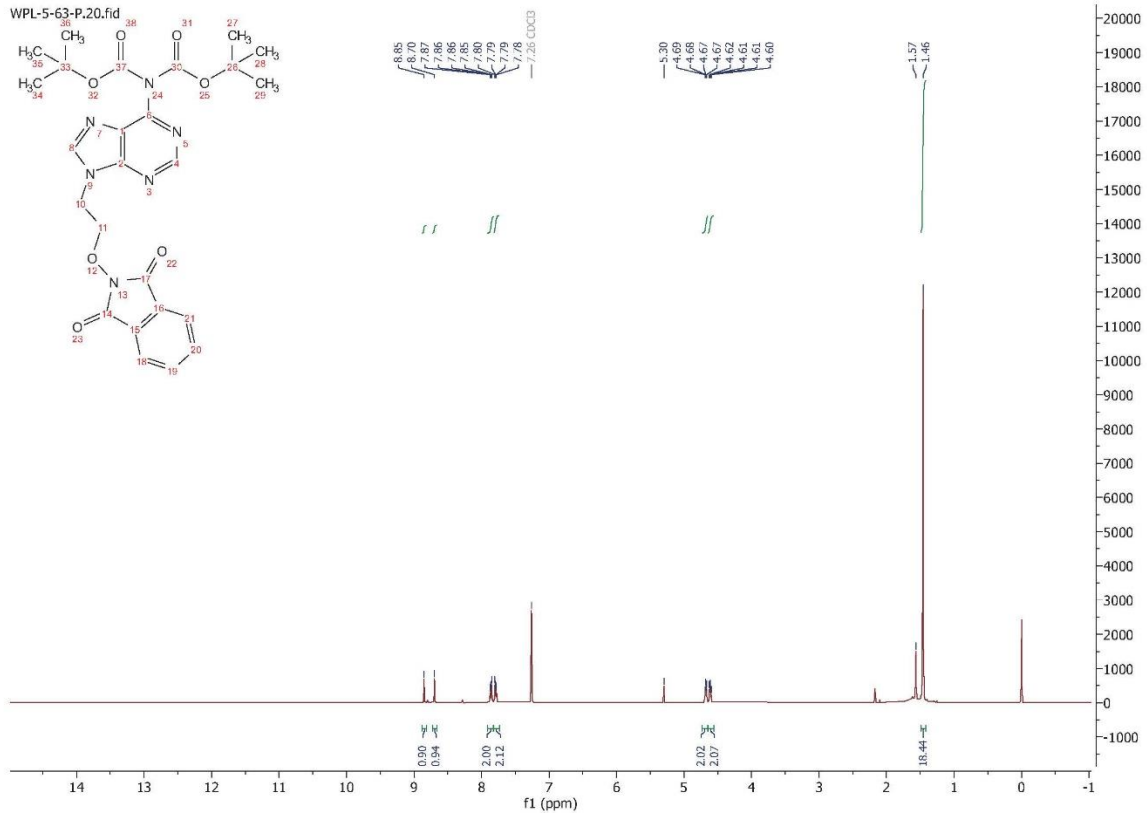


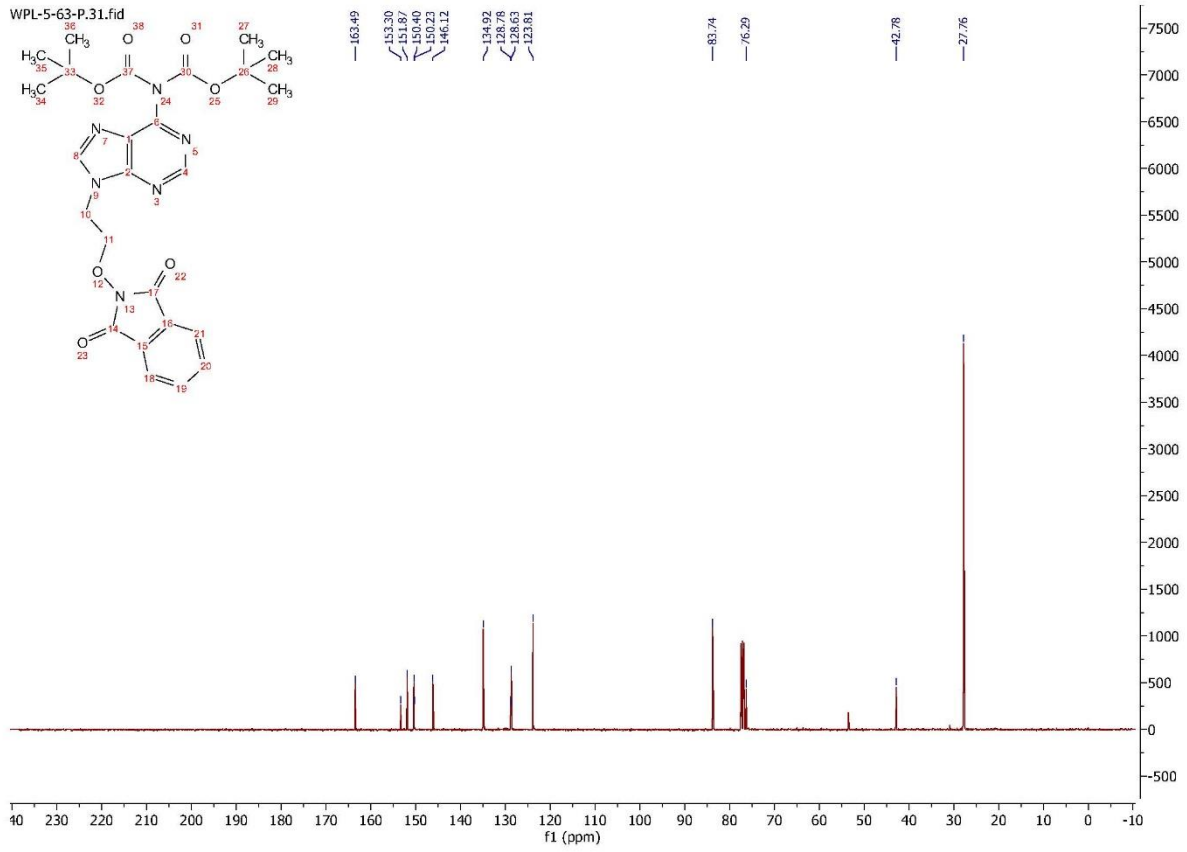


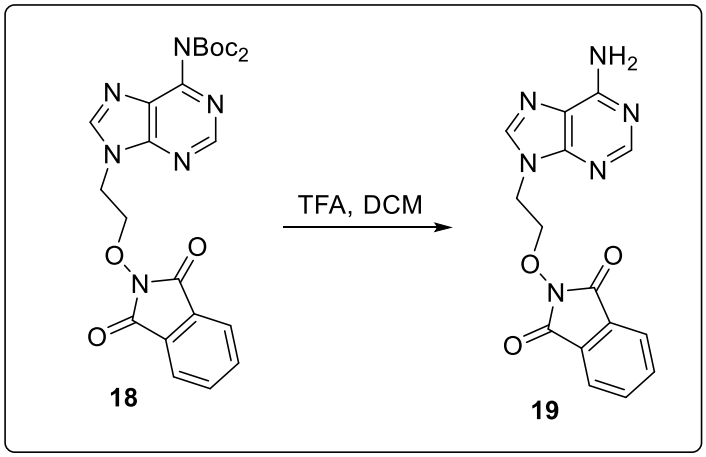
Synthesis of compound **18**: To a stirred mixture solution of 6 - N(Boc)$_2$ - adenine (**17**, 1.8 g, 5.37 mmol) and *N*-(2-bromoethoxy)phthalimide (**4**, 1.45 g, 5.37 mmol) in DMSO (60 mL) was added K_2CO_3 (742 mg, 5.37 mmol) and Cs_2CO_3 (1.75g, 5.37 mmol) and tetrabutylammonium iodide (198 mg, 0.54 mmol). The resulting mixture was stirred for 3 h at room temperature before being diluted with water. The mixture was extracted by ethyl acetate and the combined organic layers were washed with brine three times, dried over anhydrous

sodium sulfate. Filtered and concentrated. The crude product was purified by flash column chromatography (eluting with 1:1 hexanes/acetone) to afford compound **18** (2.05 g, 73%) as a white foam.

^{1}H NMR (400 MHz, $CDCl_3$) δ 8.85 (s, 1H), 8.70 (s, 1H), 7.86 (dd, *J* = 5.4, 3.2 Hz, 2H), 7.79 (dd, *J* = 5.6, 3.1 Hz, 2H), 4.68 (dd, *J* = 5.7, 3.9 Hz, 2H), 4.61 (dd, *J* = 5.6, 4.0 Hz, 2H), 1.46 (s, 18H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 163.49, 153.30, 151.87, 150.40, 150.23, 146.12, 134.92, 128.78, 128.63, 123.81, 83.74, 76.29, 42.78, 27.76 HRMS $C_{25}H_{29}N_6O_7^+$ $[M+H]^+$ calculated 525.2092, found 525.2101.

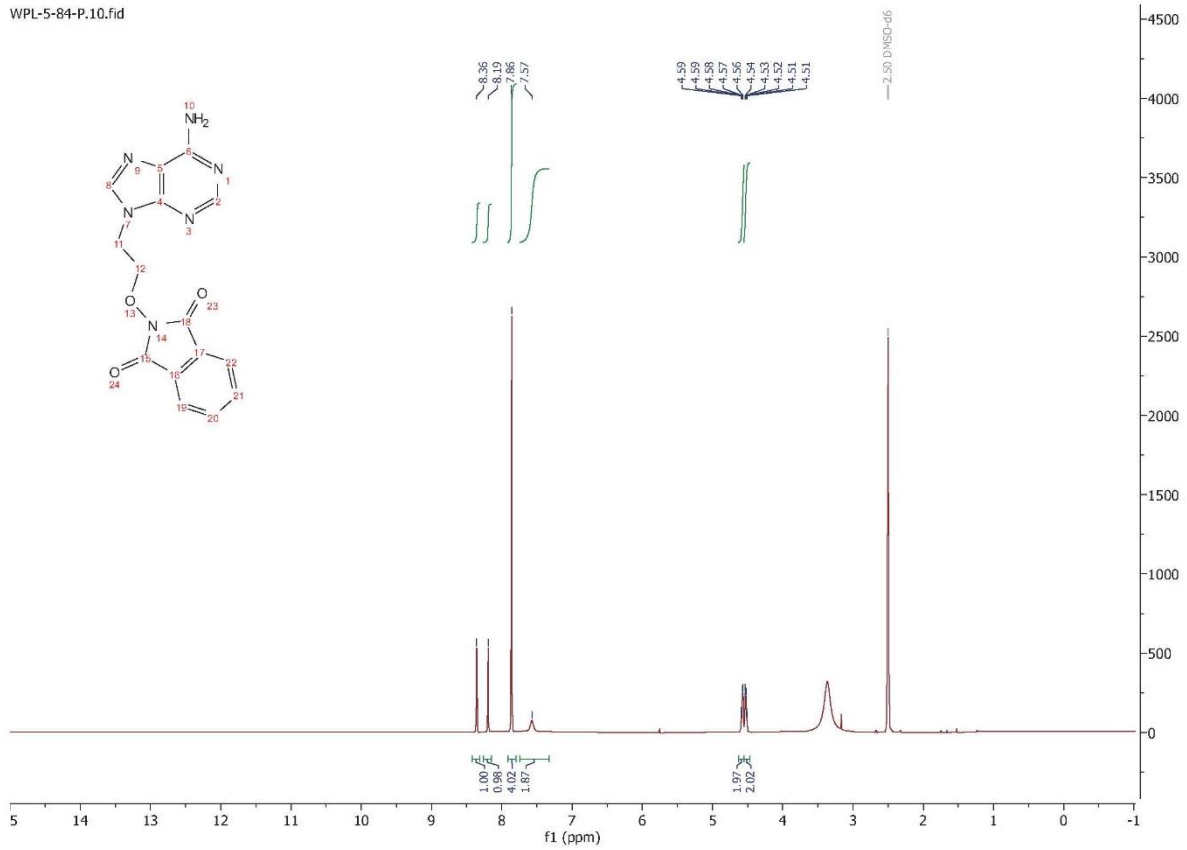


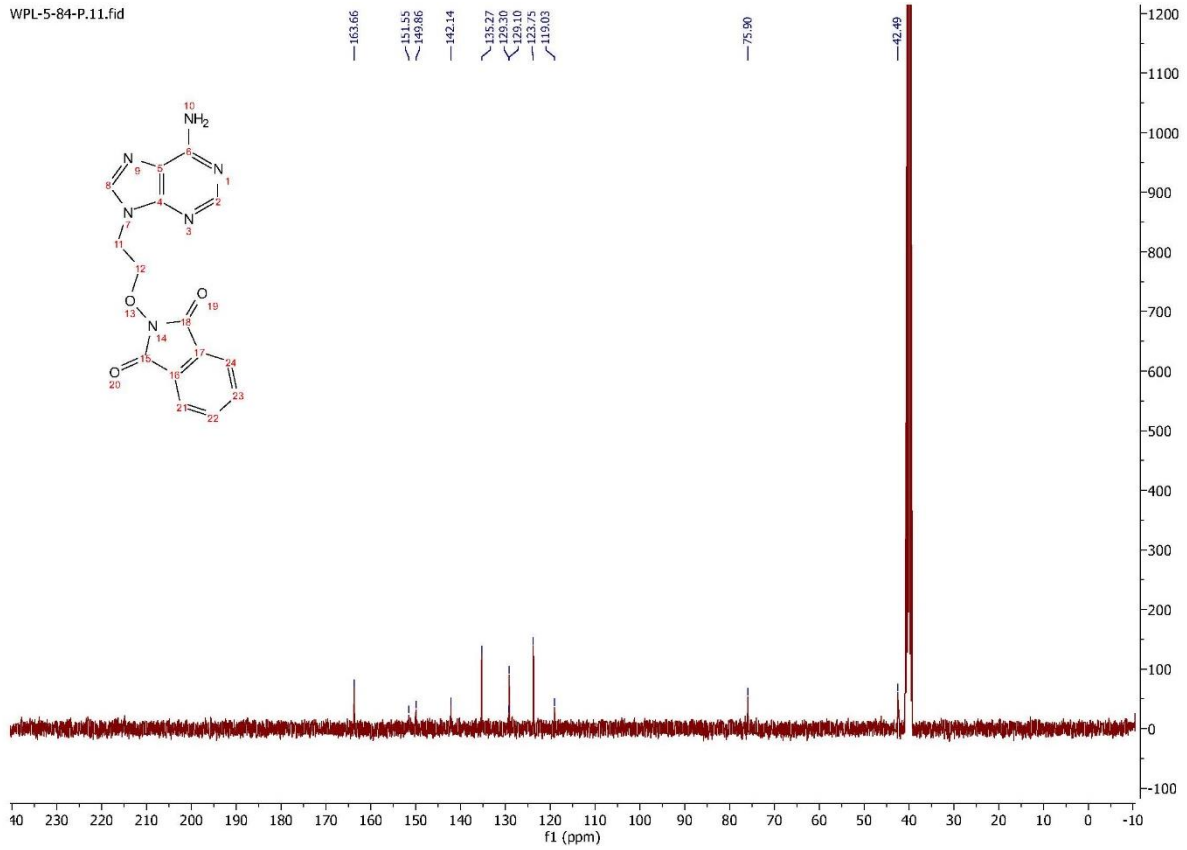


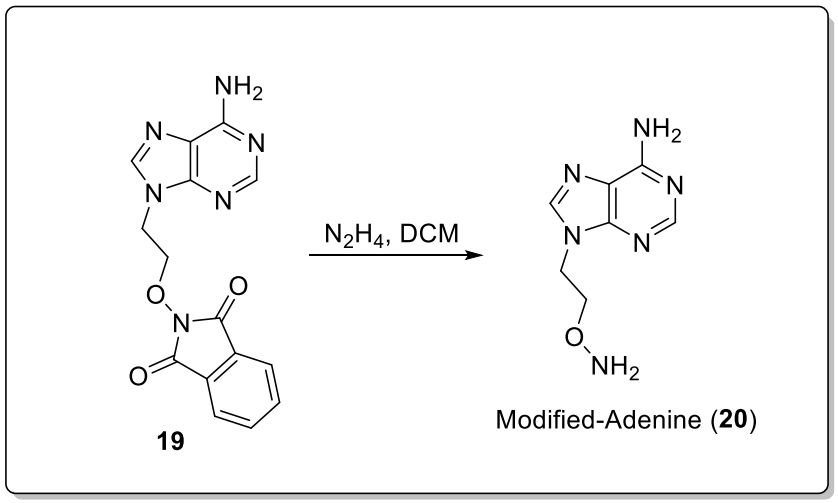


Synthesis of compound **19**: To a solution of compound **18** (300 mg, 0.57 mmol) in dichloromethane (4 mL), trifluoroacetic acid was added (1 mL, 17.16 mmol) at 0 °C and keep at 0 °C for 30min. Then the mixture was allowed to reach room temperature and stirred for 2 hours. The mixture was concentrated to obtain a residue which was purified by flash column chromatography (eluting with 2:1 acetone/hexane) to afford compound **19** (164 mg, 89%).

^{1}H NMR (400 MHz, DMSO) δ 8.36 (s, 1H), 8.19 (s, 1H), 7.86 (s, 4H), 7.57 (s, 2H), 4.57 (dd, *J* = 5.2, 3.9 Hz, 2H), 4.53 (d, *J* = 4.9 Hz, 2H). ^{13}C NMR (101 MHz, DMSO) δ 163.66, 151.55, 149.86, 142.14, 135.27, 129.30, 129.10, 123.75, 119.03, 75.90, 42.49. HRMS $C_{15}H_{13}N_6O_3^+$ $[M+H]^+$ calculated 325.1044, found 325.1057.

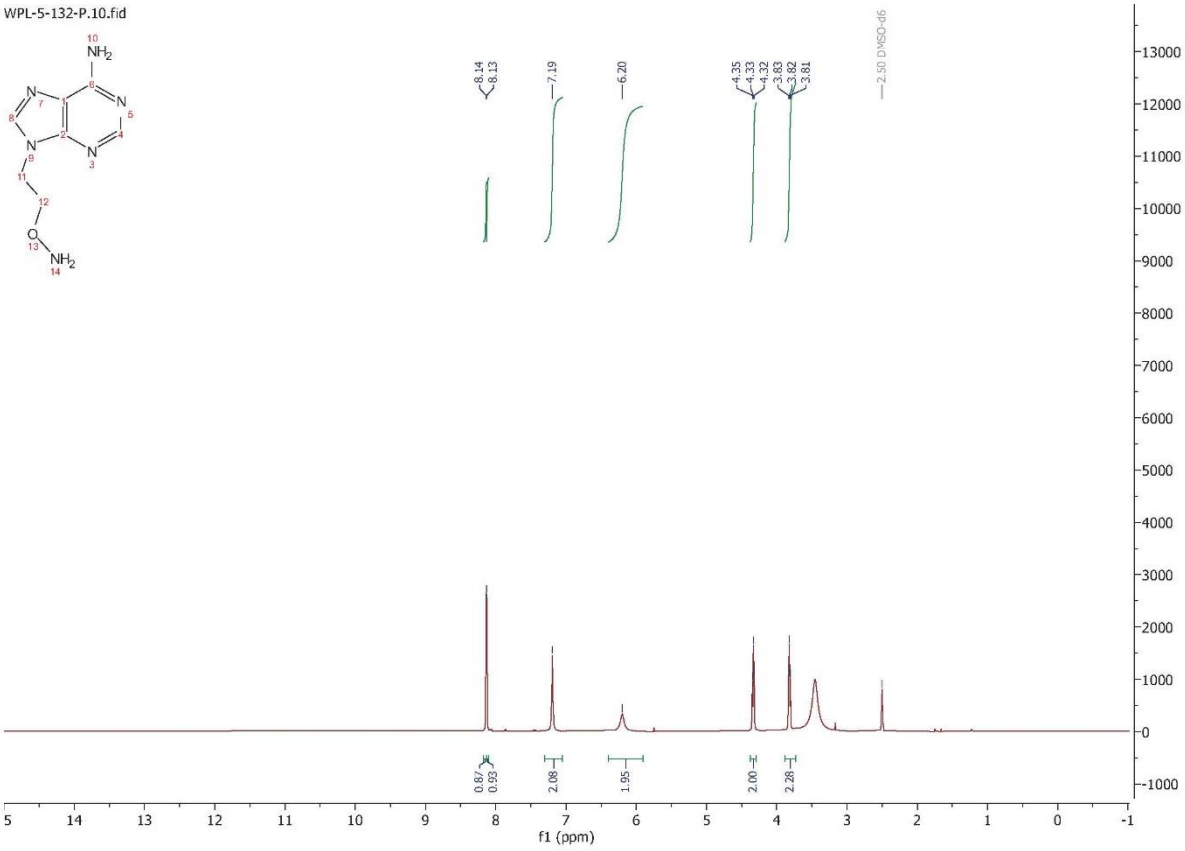


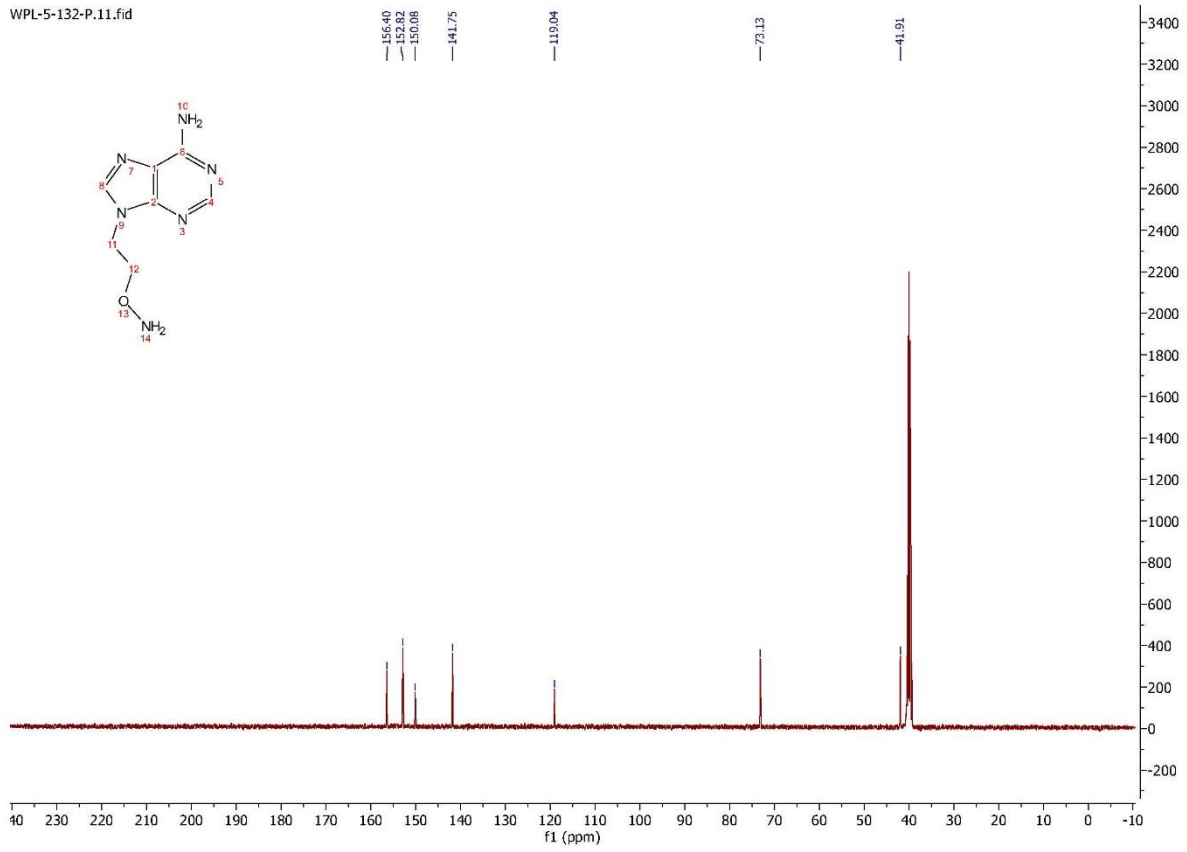




Synthesis of compound modified-adenine (**20**): To a stirred solution of compound **19** (10 mg, 0.031 mmol) in dichloromethane (3.0 mL) was added hydrazinium hydroxide solution (10 μL, 0.310 mmol). The reaction mixture was stirred for 3 hours at room temperature. After that, the mixture was filtered with 0.2 μm filter unit. After that, evaporation of the DCM to afford the purified product modified-thymine (**15**, 28 mg, 60%).

^{1}H NMR (400 MHz, DMSO) δ 8.14 (s, 1H), 8.13 (s, 1H), 7.19 (s, 2H), 6.20 (s, 2H), 4.33 (t, *J* = 5.1 Hz, 2H), 3.82 (t, *J* = 5.1 Hz, 2H). ^{13}C NMR (101 MHz, DMSO) δ 156.40, 152.82, 150.08, 141.75, 119.04, 73.13, 41.91. HRMS $C_7H_{11}N_6O^+$ $[M+H]^+$ calculated 195.0989, found 195.0997.





2.4.4 Preparation of synthetic DNA templates

Oligonucleotides containing 5fC, and 5caC were prepared using Applied Biosystems 392 DNA synthesizer. 5-Formyl-dC-CE phosphoramidite and 5-Carboxyl-dC-CE phosphoramidite (Glen Research) were used to incorporate 5fC and 5CaC at the desired position during solid-phase synthesis, followed by post synthetic deprotection by treatment with 30% ammonium hydroxide first and then 25–30% wt./wt. solution of sodium methoxide in methanol (Alfa Aesar) overnight at 25 °C. The DNA was purified by reversed-phase HPLC and confirmed by MALDI-TOF. Other oligonucleotides containing 5mC and oligonucleotides without modification were purchased from IDT.

2.4.5 Mouse embryonic stem cells (mESCs) cell culture

Mouse embryonic stem cells (mESCs) were cultured in feeder-free gelatin-coated plates in Dulbecco's Modified Eagle Medium (DMEM) (Invitrogen Cat. No. 11995) supplemented with 15% FBS (GIBCO), 2 mM L-glutamine (GIBCO), 0.1 mM 2-mercaptoethanol (Sigma), 1 × nonessential amino acids (GIBCO), 1,000 units/ml LIF (Millipore Cat. No. ESG1107), 1 × pen/strep (GIBCO), 3 mM CHIR99021 (Stemgent), and 1 mM PD0325901 (Stemgent).

2.4.6 TT-5mC-seq

2.4.6.1 Mouse ten-eleven translocation (TET) dioxygenases oxidation

Up to 500 ng gDNA was prepared for mTET treatment. The reaction mixture, with a final volume of 50 µL, included the following components: 5 µL of Tet1 protein (5.0 mg/mL), 25 µL of 2× reaction mix, and 2.5 µL of 2 mM freshly prepared ammonium iron(II) sulfate. The 2× reaction mix contains 100 mM HEPES buffer (pH 8.0), 2 mM α-ketoglutarate, 4 mM ascorbic acid, 5 mM dithiothreitol, 200 mM NaCl, and 2.4 mM ATP. The reaction was incubated at 37°C for 80 minutes, followed by an indefinite hold at 4°C.

After the initial incubation, 1 µL of Proteinase K at 20 mg/mL was introduced to the mixture, and the mixture was incubated at 50°C for 60 minutes to remove the excess mTET. The purification of the reaction product was carried out using a choice of methods: Zymo DNA Clean & Concentrator (DCC) Kit, 1.8× AMPure XP beads, or ethanol precipitation based on type of DNA input.

2.4.6.2 Thymine DNA glycosylase (TDG) excision

The thymine DNA glycosylase (TDG) excision was performed as previously described in Zhang et al.[73] , with some modifications.

For the TDG excision, the reaction was performed in a buffer containing 25 mM HEPES (pH 7.4), 0.5 mM EDTA, 0.5 mg/mL BSA, and 0.5 mM DTT to stabilize enzyme activity. To this mixture, TDG enzyme was added to a final concentration of 2 μM. After the initial incubation, 1 μL of Proteinase K at 20 mg/mL was introduced to the mixture, and the mixture was incubated at 50°C for 60 minutes to remove the excess TDG. The purification of the reaction product was carried out using a choice of methods: Zymo DNA Clean & Concentrator (DCC) Kit, 1.8× AMPure XP beads, or ethanol precipitation based on type of DNA input.

2.4.6.3 Addition of nucleotide alternatives

Nucleotide alternatives were dissolved in DMSO to generate stocks with a final concentration of 50 mM. For the reaction, 3 μL of the stock was combined with the purified DNA with abasic site and 25 μL of 2× MES buffer (pH 6.0, 100 mM) to reach a total volume of 50 μL.The reaction was incubated at 37°C for 2 hours to enable the incorporation of the modified base into the DNA. The purification of the reaction product was carried out using a choice of methods: Zymo DNA Clean & Concentrator (DCC) Kit, 1.8× AMPure XP beads, or ethanol precipitation based on type of DNA input.

2.4.6.4 DBCO-PEG4-Biotin conjugation

337 mg of dibenzocyclooctyne-PEG4-biotin (DBCO-PEG4-Biotin) was dissolved in 1 mL of ddH_2O to give around a concentration of 0.45 mM. 1 μL of DBCO-PEG4-Biotin solution was added to the purified DNA, and the mixture was incubated at 37°C for 2 hours. The purification of the reaction product was carried out using a choice of methods: Zymo DNA Clean & Concentrator (DCC) Kit, 1.8× AMPure XP beads, or ethanol precipitation based on type of DNA input.

2.4.6.5 5hmC blocking

DNA was combined with βGT and UDP-Glc in a glucosylation buffer consisting of 50 mM HEPES buffer (pH = 8.0) and 25 mM $MgCl_2$. This mixture was incubated at 37°C for 2 hours. The purification of the reaction product was carried out using a choice of methods: Zymo DNA Clean & Concentrator (DCC) Kit, 1.8× AMPure XP beads, or ethanol precipitation based on type of DNA input.

2.4.7 MALDI-TOF characterization

For matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry, the analysis was conducted by combining 1 μL of the reaction product solution with 1 μL of a matrix solution. The matrix solution was composed of 2'4'6'-trihydroxyacetophenone (THAP) at a concentration of 10 mg/mL in a 50% acetonitrile/water mixture, and ammonium citrate at a concentration of 50 mg/mL in water, mixed at a ratio of 8:1 (v/v). The resulting mixture was then applied to a MALDI sample plate and allowed to dry in the air. Subsequent analysis was performed using a Bruker Ultraflextreme MALDI-TOF/TOF mass spectrometer.

2.4.8 Quantitative polymerase chain reaction (qPCR) assay

Quantitative PCR was performed to optimize the number of amplification cycles required for final library construction. The amplification reactions were set up using SYBR Green PCR Master Mix in a final volume of 20 μL. Each reaction contained 9 μL of template DNA, 1 μL of primer mix, and 10 μL of 2x SYBR Green PCR Master Mix. Reactions were run on a 7300 Real-Time PCR System (Applied Biosystems).

The thermal cycling conditions were initiated with a 10-minute enzyme activation step at 95°C, followed by 40 cycles of 15 seconds at 95°C for denaturation, and 1 minute at

60°C for annealing and extension. Melting curve analysis was conducted post-amplification to confirm the specificity of the PCR products.

The cycle threshold (Ct) values were determined using the 7300 System SDS software, and the optimal cycle number for library amplification was established by identifying the Ct value that corresponded to the mid-exponential phase of the PCR. This ensured minimal bias and the maintenance of library complexity.

2.4.9 Genome DNA extraction and purification

Genomic DNA was extracted from mouse embryonic stem cells (mESCs) using the DNeasy Blood & Tissue Kit (Qiagen) according to the manufacturer's instructions. Briefly, cells were lysed in a solution containing proteinase K and Buffer AL provided by the kit. The lysate was then incubated at 56°C until complete lysis was achieved. The lysed samples were subsequently ethanol-precipitated and loaded onto DNeasy Mini spin columns for purification.

The bound DNA was washed with the Wash Buffers AW1 and AW2 to remove contaminants. High-quality genomic DNA was eluted in Buffer AE. The concentration of the eluted DNA was quantified using a Qubit fluorometer and DNA solution were stored at -80 °C.

2.4.10 DNA polymerase selection

2.4.10.1 Platinum Taq DNA Polymerase

Starting with 5 ng of DNA template in 5 µL of water, the following components were added: 1x PCR buffer, 2 µL of 10 mM dNTP Solution Mix, 1.5 mM $MgCl_2$, and 0.2 µM of each primer. 1 µL of The Platinum Taq DNA Polymerase was added last. Amplification conditions were as follows: initial denaturation at 95°C for 5 minutes, followed by 5 cycles of

95°C for 30 seconds, 55°C for 30 seconds, and 72°C for 1 minute, with a final extension at 72°C for 10 minutes. The PCR product was then analyzed via Sanger sequencing to assess the conversion rate of 5mC.

2.4.10.2 EpiMark Hot Start Taq DNA Polymerase

Starting with 5 ng of DNA template in 5 µL of water, the following components were added: 1x EpiMark Hot Start Taq Reaction Buffer, 2 µL of 10 mM dNTP Solution Mix, 1.5 mM MgCl2, and 0.2 µM of each primer. 1 µL of EpiMark Hot Start Taq DNA Polymerase was added to the reaction mixture last. The thermal cycling conditions were set as follows: initial denaturation at 94°C for 30 seconds, followed by 30 cycles of 94°C for 30 seconds, 55°C for 30 seconds, and 68°C for 1 minute, with a final extension step at 68°C for 5 minutes. The PCR product was then analyzed via Sanger sequencing to assess the conversion rate of 5mC.

2.4.10.3 Q5U Hot Start High-Fidelity DNA Polymerase

Starting with 5 ng of DNA template in 5 µL of water, we included 1x Q5U Reaction Buffer, 2 µL of 10 mM dNTP Solution Mix, 0.2 µM of each primer, and 1 µL of Q5U Hot Start High-Fidelity DNA Polymerase in the reaction. The amplification profile was initiated with an activation step at 98°C for 2 minutes, followed by 30 cycles of 98°C for 10 seconds, 55°C for 30 seconds, and 72°C for 30 seconds, concluding with a final extension at 72°C for 2 minutes. The PCR product was then analyzed via Sanger sequencing to assess the conversion rate of 5mC.

2.4.10.4 Bst DNA Polymerase

Starting with 5 ng of DNA template in 5 µL of water, the reaction mixture was prepared to include 1x Isothermal Amplification Buffer, 2 µL of 10 mM dNTP Solution Mix, 0.2 µM of each primer, and 1 µL of Bst DNA Polymerase. The amplification was carried out at 65°C for 60 minutes. The PCR product was then amplified by other DNA polymerases.

2.4.10.5 Taq DNA Polymerase

Starting with 5 ng of DNA template in 5 µL of water, the mixture was prepared with 1x Taq Reaction Buffer, 2 µL of 10 mM dNTP Solution Mix, 1.5 mM $MgCl_2$, and 0.2 µM of each primer, adding 1 µL of Taq DNA Polymerase last. The PCR conditions were as follows: an initial denaturation at 94°C for 3 minutes; 35 cycles of 94°C for 30 seconds, 55°C for 30 seconds, and 72°C for 1 minute; and a final extension at 72°C for 10 minutes. The PCR product was then analyzed via Sanger sequencing to assess the conversion rate of 5mC.

2.4.10.6 Phusion High-Fidelity DNA Polymerase

For the reaction, 5 ng of DNA template 5 µL of water, was mixed with 1x Phusion HF Buffer, 2 µL of 10 mM dNTP Solution Mix, 0.2 µM of each primer, and 1 µL of Phusion High-Fidelity DNA Polymerase. The PCR program included an initial denaturation at 98°C for 30 seconds, followed by 30 cycles of 98°C for 10 seconds, 55°C for 30 seconds, and 72°C for 30 seconds, and a final extension at 72°C for 10 minutes.

2.4.10.7 HIV Reverse Transcriptase (HIVRT)

1.0 µL of primer mix was added to ~10 ng purified DNA, followed by 95°C 2 min denature. The DNA/primer mixture was further mixed with 2 µL 10X AMV RT Buffer, 2 µL of 10 mM dNTP Solution Mix, and 1 µL of HIV RT. The reaction was diluted to 20 µL and

incubated at 37 °C for 60 mins. The rt product was then amplified by other DNA polymerases.

2.4.11 Primer extension assay

To assess the read-through capacity of DNA polymerases on 5-methylcytosine (5mC) modified sites, a primer extension assay was conducted. An 82-mer oligonucleotide template containing 5mC residues was synthesized (Integrated DNA Technologies). The oligonucleotide sequence was designed to generate a distinct band pattern upon successful extension by the polymerase.

Primer annealing was performed by mixing the oligonucleotide template (10 nM) with a 5' FAM-labeled primer (10 nM) in 1× NEBuffer 2 (New England Biolabs, product #B7002S) at 95°C for 5 minutes, followed by a ramp-down to 25°C at a rate of 1°C per minute. For the extension reaction, each annealed primer-template set was divided into separate aliquots for each enzyme. The reaction mixes for Taq Platinum DNA Polymerase, EpiMark Hot Start Taq DNA Polymerase, Q5U Hot Start High-Fidelity DNA Polymerase, Bst DNA Polymerase, Phusion High-Fidelity DNA Polymerase, and HIV Reverse Transcriptase (HIV RT) were prepared according to the manufacturers' specifications with the inclusion of dNTPs and an appropriate $MgCl_2$ concentration (Chapter **2.4.10**).

The extension reactions were incubated at the manufacturers' recommended condition for optimal polymerase activity. After the extension, reactions were terminated by the addition of loading dye and denatured at 95°C for 5 minutes. The samples were immediately chilled on ice and then loaded onto 6% TBE-UREA GEL 1.0mM (Life Technologies) Electrophoresis was conducted at a constant voltage until the dye front reached the bottom of the gel. The gels were then scanned on a fluorescence imager to visualize the FAM signal.

2.4.12 Dot blot assay

DNA samples with a concentration of 100 ng/μL were prepared for blotting. For each sample, 1 μL was carefully loaded onto the Amersham Hybond-N+ membrane (GE Healthcare, RPN119B). Membranes were air-dried and then crosslinked by UV stratalinker 2400 at 150 mJ/cm^2, and the process was repeated twice to ensure effective binding of the DNA to the membrane. The membranes were then blocked overnight in 5% fatty-acid free BSA in PBST (0.1% Tween-20) to prevent non-specific binding during the detection process. On the following day, the membrane was washed and incubated in streptavidin-HRP (Thermo, S-911) in PBST supplemented with 3% fatty-acid free BSA. The membrane was washed in PBST for 5 times before developed by SuperSignal™ West Pico PLUS Chemiluminescent Substrate (Thermo, 34577).

3 Human 5hmC tumor tissue map profiling

3.1 Introduction

The epigenetic modification of DNA via 5-hydroxymethylcytosine (5hmC) has become increasingly recognized for its role in regulating gene expression and its involvement in a range of biological processes, including mammalian development and the pathogenesis of diseases, particularly cancer[70, 74-76].The significance of 5hmC as a sensitive and reliable epigenetic biomarker in cancer has been underscored by several studies demonstrating its utility in reflecting the types and stages of cancers[77, 78]. Moreover, the tissue-specific nature of 5hmC and its dynamic changes in cancer highlight its potential as a marker for cancer initiation and progression[79].

Despite its importance, the comprehensive mapping and understanding of 5hmC's involvement in cancer, particularly in metastasis, have been limited. This shortfall stems, in part, from a historical focus on 5mC. The past decade's shift towards acknowledging 5hmC's distinct regulatory roles has begun to reshape our understanding of the cancer epigenome, marking it as a frontier for epigenetic research[80].

Central to the exploration of this frontier is the 5hmC-seal and nano-5hmC-seal technology, which facilitates sensitive, valid, and reliable genome-wide profiling of 5hmC[81, 82],(**Figure 3.1**). This technique leverages a viral enzyme, β-glucosyltransferase (β-GT), which catalyzes the transfer of a glucose part from uridine diphosphoglucose (UDP-Glu) to the hydroxyl group of 5hmC. This enzymatic action results in the formation of β-glucosyl-5-hydroxymethylcytosine (5-gmC) within the DNA structure (**Figure 3.1a**). Expanding on this, we harnessed β-GT to attach a modified glucose, specifically 6-N_3-glucose, to 5hmC, enabling selective and bio-orthogonal labeling of 5hmC in genomic DNA (**Figure 3.1b**). The presence of an azide group facilitates the attachment of a biotin tag or

other labels through click reaction, broadening the scope for various applications in enrichment, detection, and sequencing of 5hmC.

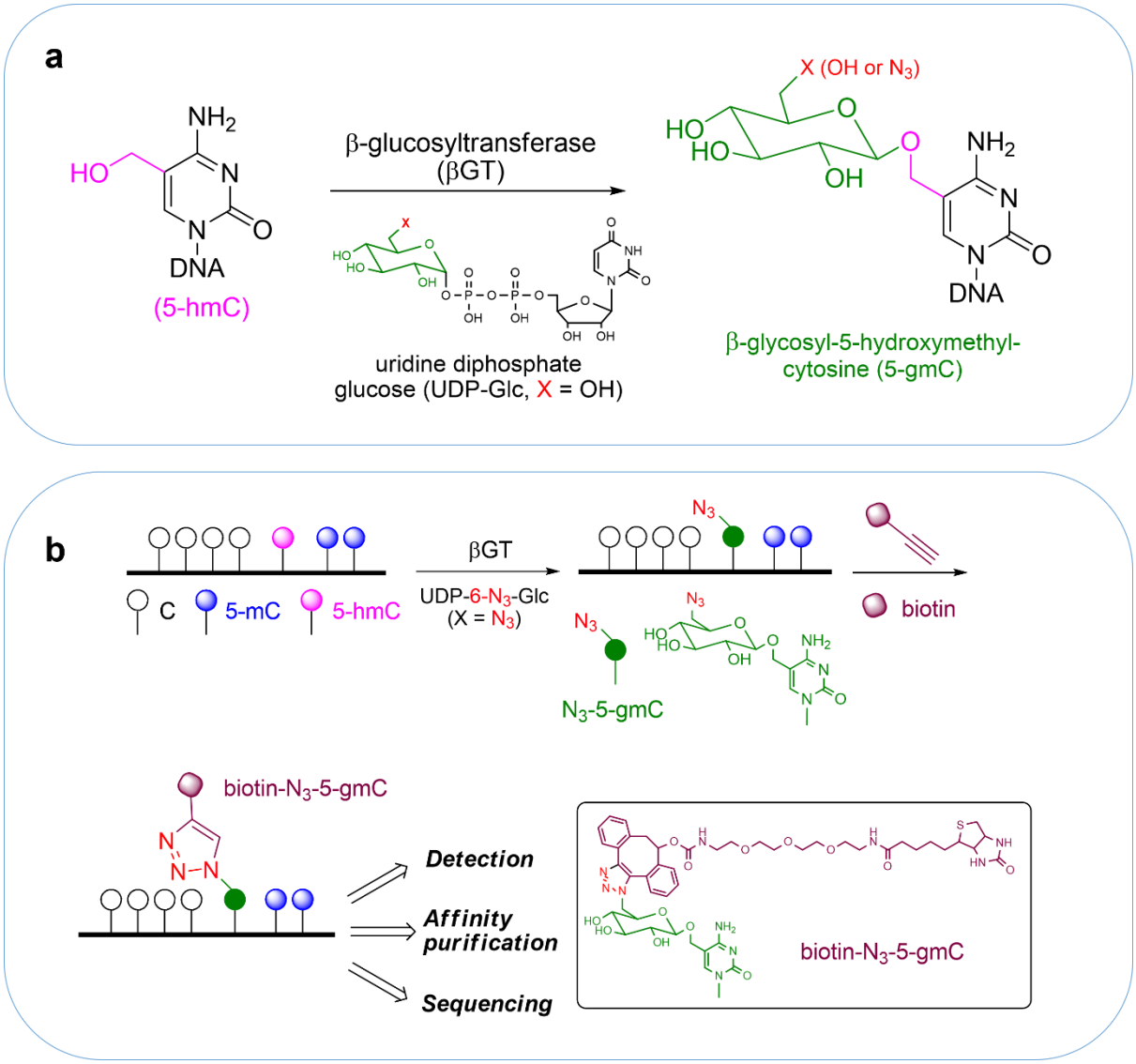


Figure 3.1 Selective labeling of 5-hmC in genomic DNA
(a) In duplex DNA, the hydroxyl group on 5-hydroxymethylcytosine (5-hmC) can undergo glucosylation by the enzyme β-glucosyltransferase (β-GT), using UDP-glucose as a cofactor to produce β-glucosyl-5-hydroxymethylcytosine (5-gmC). (b) By employing a chemically modified version of UDP-glucose, specifically UDP-6-azido-glucose (UDP-6-N3-Glu), an azide functional group can be introduced to 5-hmC. This modified nucleotide can subsequently be tagged with biotin through click chemistry reactions, enabling its detection, enrichment, and sequencing.

The 5hmC-seal approach, thus, acts as a crucial technique in our study, providing a sensitive, valid, and reliable method for genome-wide 5hmC profiling. This technology not only aids in mapping the 5hmC distribution across diverse cancer types but also paves the

way for future research endeavors aimed at understanding the epigenetic intricacies of cancer and exploring new avenues in cancer diagnostics and therapeutics.

3.2 Result and discussion

3.2.1 Project design and workflow

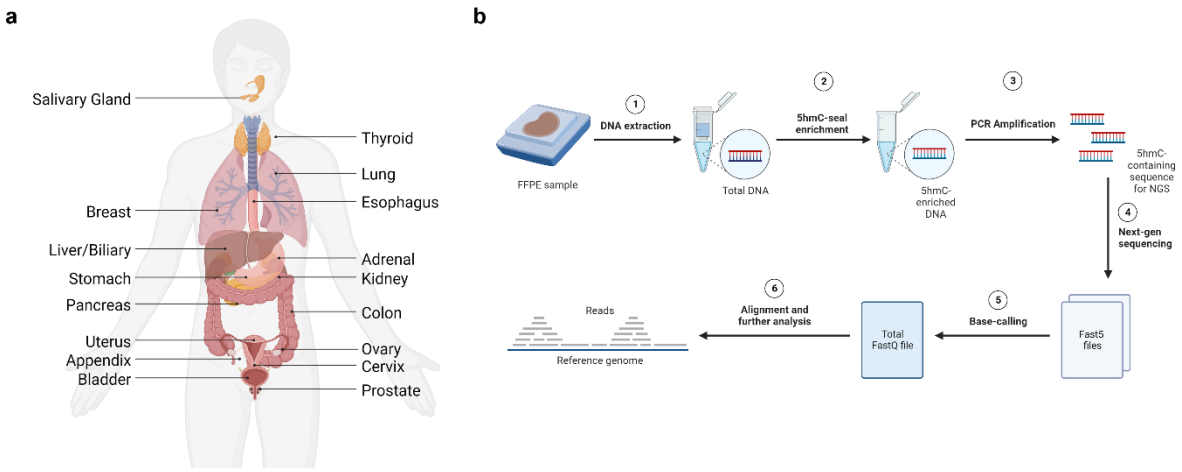


Figure 3.2 Human 5hmC tumor tissue map profiling based on 5hmC-seal.
(a) Schematic plot showing all the organ tissues analyzed in this study. (b) Workflow of FFPE sample treatment

Our study undertakes a comprehensive survey of the epigenomic terrain across a variety of human tissues, with a special focus on the distribution of 5-hydroxymethylcytosine (5hmC) within both tumorous and non-tumorous samples. he workflow adopted for this analysis is rigorously designed, beginning with the extraction and purification of DNA from formalin-fixed paraffin-embedded (FFPE) tumor samples, kindly provided by Dr. Feng Yue of Northwestern University (**Figure 3.2b**). These samples constitute a crucial foundation for our analysis.

The DNA was first extracted and purified for subsequent use, then subjected to the 5hmC-seal enrichment process. This technique selectively tags 5hmC residues, allowing us to isolate and amplify regions of interest through following PCR amplification. The amplified products are then sequenced using next-generation sequencing technologies, which generate

comprehensive reads that are subsequently base-called for identification of individual nucleotides.

The sequencing data are aligned to reference genomes, enabling us to discern the distribution patterns of 5hmC across different samples. For this study, we have analyzed genomic DNA from a total of 138 human tissue samples. This extensive collection includes 92 tumor samples, showcasing 20 distinct cancer types spanning 16 organ systems. Additionally, we have included 46 benign samples to serve as a comparative baseline, also derived from 16 organ systems (**Figure 3.2a**). Through this dataset, our aim is to uncover distinct 5hmC patterns that may be associated with cancer metastasis.

3.2.2 Distribution of 5-hydroxymethylcytosine in human tissues

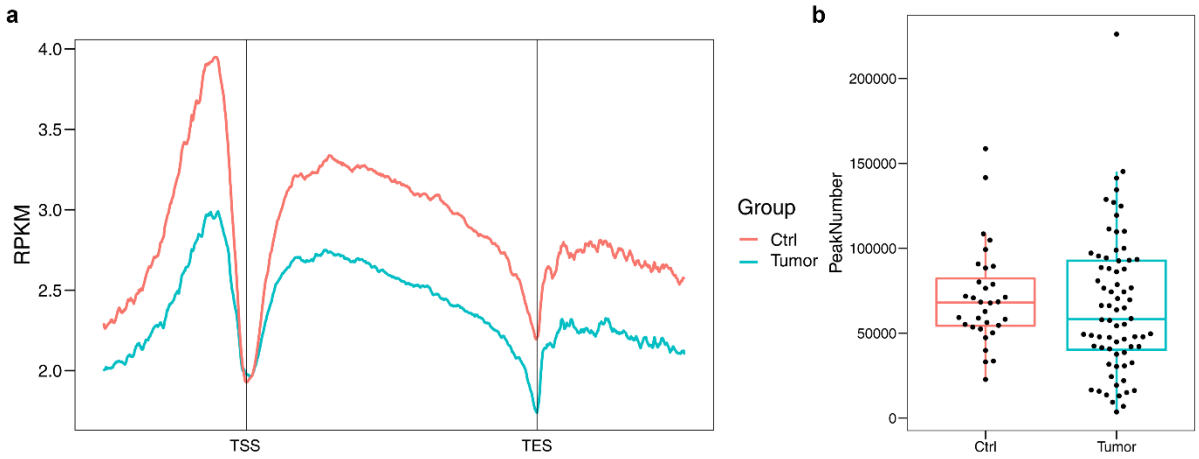


Figure 3.3 Distribution of 5-hydroxymethylcytosine in human tissues
(a) Metagene profile showing the distribution of 5hmC modifications around transcription start sites (TSS) and transcription end sites (TES) (b) The distribution of the number of 5hmC peaks detected in the control and tumor samples.

Our genome-wide analysis of 5hmC has provided a detailed map of its distribution across various human tissues. Metagene analysis of 5hmC at transcription start sites (TSS) and transcription end sites (TES) was shown (**Figure 3.3a**). We observed the expected pattern of 5hmC deficiency at TSSs and enrichment across gene bodies, consistent with

established profiles in mammalian genomes. The reduced 5hmC signals at TSSs suggest potential regulatory roles of this modification in gene expression.

In the tumor samples, the 5hmC distribution exhibits a slight deviation from control samples, particularly in regions flanking the TSS, which may reflect the altered epigenetic regulation within tumor tissues. We also noticed the global decrease of 5hmC levels on gene bodies for tumor sample, which indicates the disruption of normal epigenetic signatures during tumorigenesis.

Furthermore, an examination of the 5hmC peak counts between control and tumor samples (**Figure 3.3b**) shows a lower median peak count in tumors. This decrease could point to a dysregulation of 5hmC marking in response to cancer development. The greater spread of peak numbers within tumor tissues diverse epigenetic landscape across different cancer types.

3.2.3 Reduction of 5hmC peaks in coding and regulatory genomic regions were observed in tumor tissues

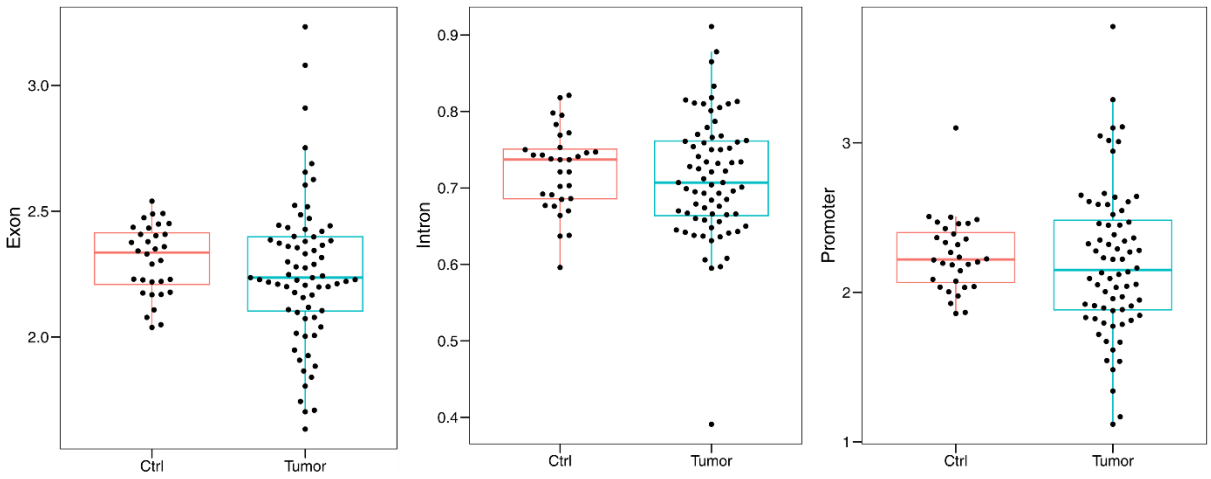


Figure 3.4 Comparative analysis of 5hmC peak distributions in genomic regions. Boxplots showing genomic enrichment of 5hmC peaks distribution at exons, introns, and promoters in $\log_2$(obs/exp).

Our analysis extends to the quantitative assessment of 5hmC modifications across exons, introns, and promoters, revealing a universal downtrend in tumor samples (**Figure**

3.4). The log_2(observed/expected) ratios of 5hmC peaks were systematically lower in tumor samples compared to controls across all examined genomic regions. The decreased levels of 5hmC in exons could have implications for the altered gene expression profiles characteristic of tumor cells, as these regions are integral to mRNA synthesis and processing. A similar reduction in promoters and introns points to a potential disruption of the regulatory networks governing gene transcription and post-transcriptional modifications.

3.2.4 Repetitive elements and intergenic regions in tumor tissues show contrasting 5hmC peak dynamics

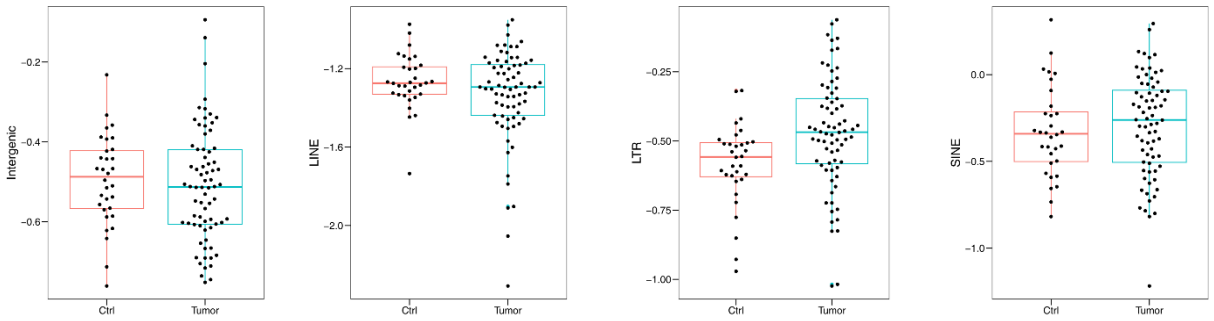


Figure 3.5 Differential distribution of 5hmC peaks in repetitive genomic elements and intergenic region
Boxplots showing genomic enrichment of 5hmC peaks distribution at intergenic region, LINEs, LTRs, and SINEs in log_2(obs/exp).

We then examined the 5mC distribution across intergenic regions, long interspersed nuclear elements (LINEs), short tandem repeats (STRs), and short interspersed nuclear elements (SINEs) (**Figure 3.5**). Tumor tissues exhibit a decrease in 5hmC levels within LINEs and intergenic regions. This reduction could signify a repression of these genomic areas, which might be linked to alterations in chromatin structure and transcriptional regulation associated with cancer. Conversely, an increase in 5hmC was detected in SINEs and LTRs in tumor tissues, potentially signaling their activation during oncogenesis. This epigenetic activation could play a role in disrupting normal gene regulation. The contrasting

patterns of 5hmC modification between intergenic and repetitive elements underscore the complexity of tumor epigenetics.

3.2.5 Consistency of 5hmC distribution in gene bodies

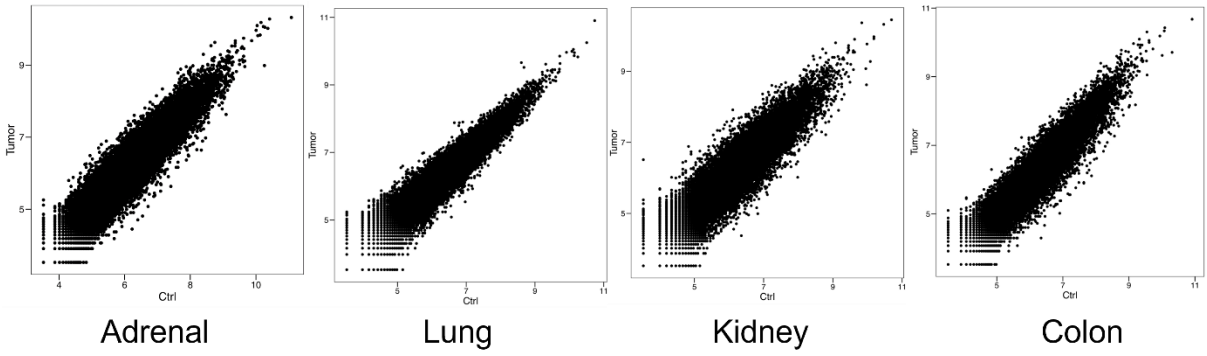


Figure 3.6 Normalized 5hmC levels on gene bodies in different tissues
This scatterplot illustrates the normalized levels of 5hmC in gene bodies, comparing control (X-axis) and adrenal tumor (Y-axis) samples. Each point represents a gene, with its normalized 5hmC level measured in the control and corresponding tumor tissue.

We then investigated the patterns of 5hmC distribution across gene bodies and conducted a comprehensive scatterplot analysis to compare the normalized levels of 5hmC between control and tumor samples in several tissues (**Figure 3.6**).

Although there is an observable decrease in overall 5hmC levels on gene bodies in tumor samples, the scatterplot reveals that the relative distribution of 5hmC among the genes remains largely unchanged. This observation indicates that, despite the transformation into tumor tissue, most genes retain the relative modification levels characteristic of their normal tissue states. This persistence in the relative ordering of 5hmC modifications suggests the possibility of using 5hmC as a tool for tracing tissue origins and hints at a potential conservation of epigenetic regulatory mechanisms at the gene level, even as the tissue undergoes malignant transformation.

3.2.6 Multidimensional scaling of 5hmC signatures across tissue types

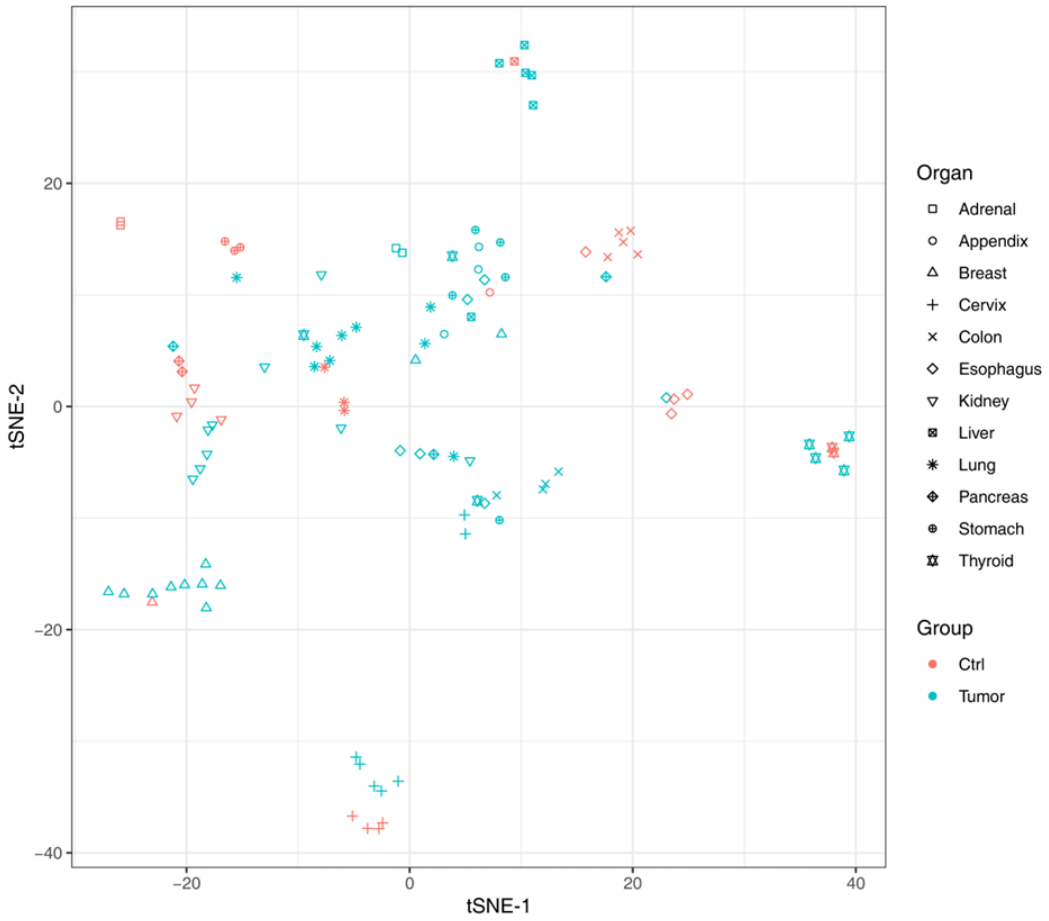


Figure 3.7 t-SNE clustering of genomic 5hmC distributions using all gene bodies. The t-SNE plot highlight the distinct clustering of tissue types and relationship between control and tumor samples. Each symbol on the plot corresponds to a different organ type, while the color denotes the group—red for control and blue for tumor samples.

In the pursuit of elucidating the epigenetic distinctions between various human tissues and their tumor counterparts, we employed t-distributed stochastic neighbor embedding (t-SNE), presents a t-SNE plot that synthesizes the high-dimensional epigenetic data into a two-dimensional space, facilitating an intuitive understanding of the relationships and variances within the dataset.

The scatterplot generated through t-SNE analysis reveals noticeable clusters according to tissue type, suggesting that despite the heterogeneity within the samples, there are inherent 5hmC signatures unique to each tissue. Interestingly, when contrasting control and tumor

samples, we observe different patterns, reflecting the profound epigenetic remodeling that accompanies tumorigenesis, some group has distinct separations while the others do not, reflecting potential epigenetic remodeling that accompanies tumorigenesis.

3.2.7 Distinct epigenetic profiles of tissue-specific genes in normal and tumor samples

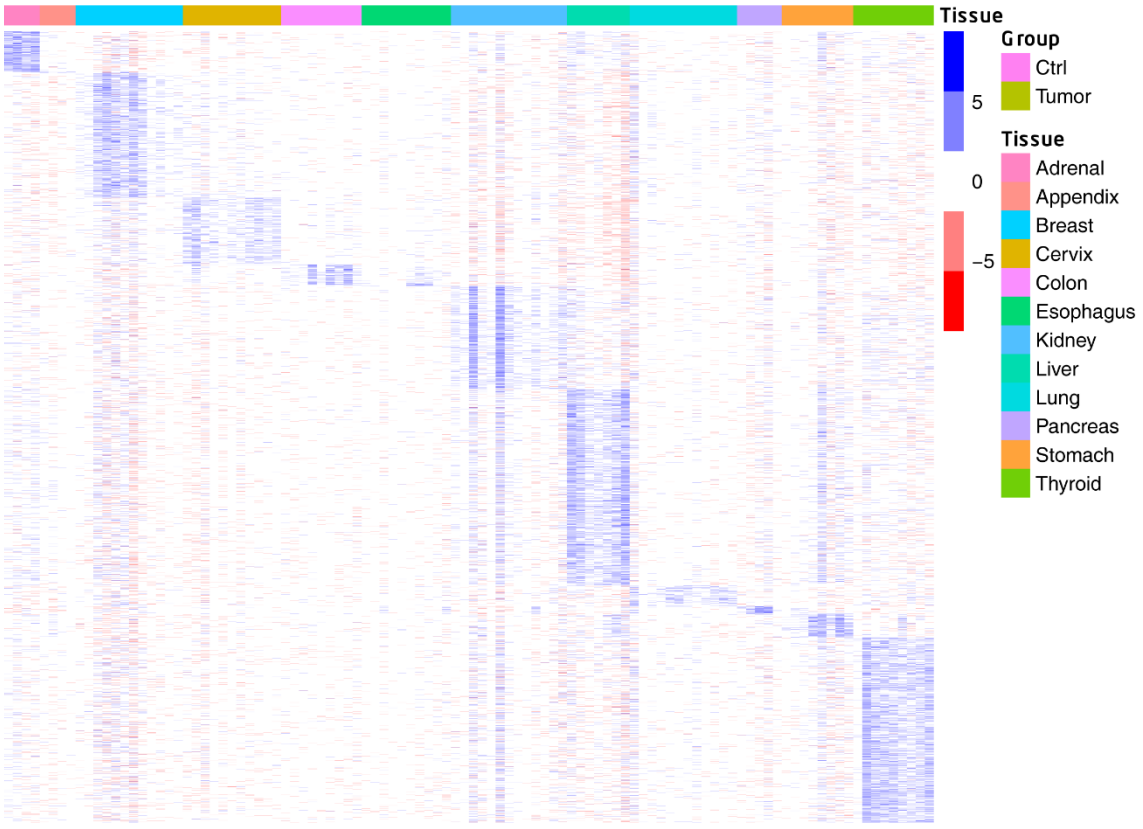


Figure 3.8 Relative 5hmC-modification levels on tissue-specific 5hmC-enriched genes in each tissue type.
Each row corresponding to a gene and each column to a tissue type. The color scale indicates the degree of 5hmC modification, with blue denoting lower and red indicating higher levels of modification.

We next asked whether the 5hmC modifications marked tissue-specific genes. To elucidate the complexity of tissue-specific epigenetic regulation, we explored the distribution of 5hmC within gene bodies across different tissues and their respective tumor types. The heatmap visualizes the 5hmC modification levels in gene bodies, showcasing distinctive patterns among different tissues.

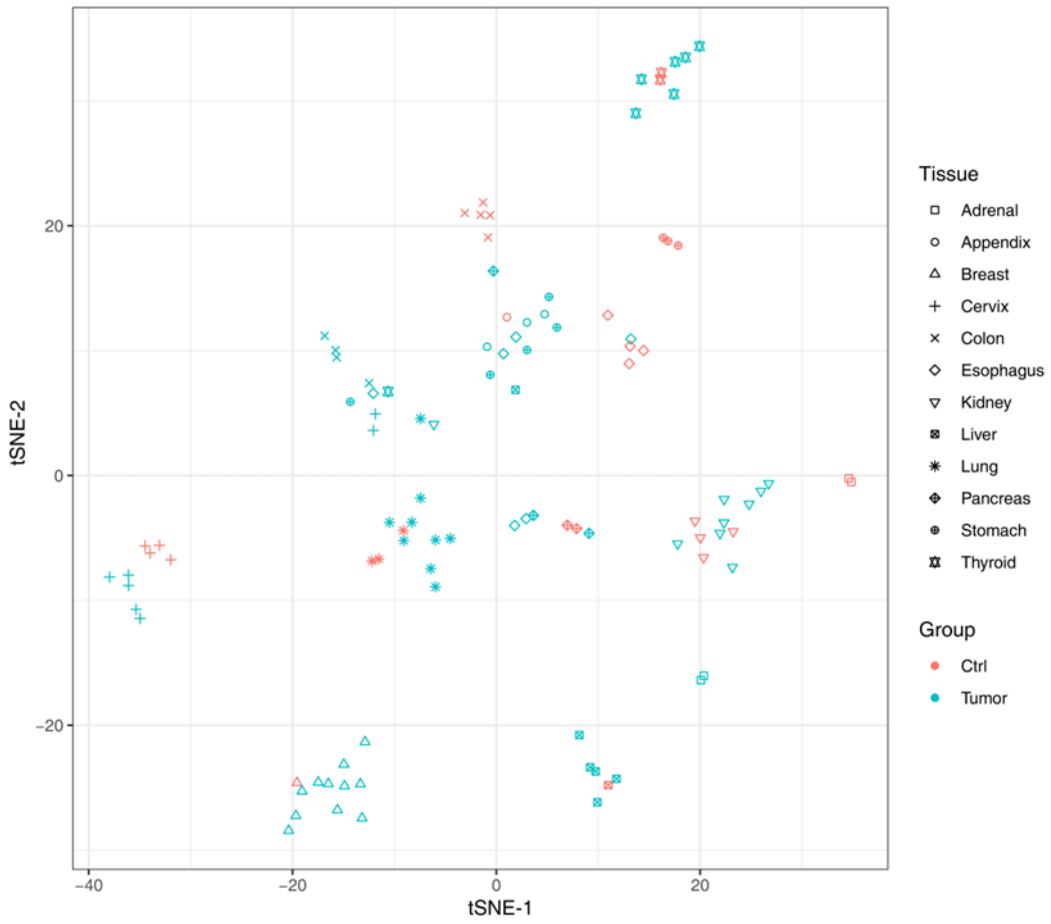


Figure 3.9 t-SNE clustering of genomic 5hmC distributions using tissue-specific gene bodies.
The t-SNE plot highlight the distinct clustering of tissue types and relationship between control and tumor samples. Each symbol on the plot corresponds to a different organ type, while the color denotes the group—red for control and blue for tumor samples.

In our continued exploration of the epigenetic differences between normal and cancerous tissues, we employed t-SNE to analyze the 5hmC modifications of tissue-specific gene bodies (**Figure 3.9**) and leverages the distinctive epigenetic signatures to achieve a more refined separation of tissue types.

When focusing on tissue-specific genes, the t-SNE plot reveals a clear distinction between various tissues, with normal and tumor samples from the same tissue type forming discrete clusters. This pattern indicates that despite the epigenetic alterations that occur in tumorigenesis, the defining characteristics of tissue-specific gene expression regulation are largely retained in tumors.

The clustering pattern observed underscores the robustness of tissue-specific epigenetic landscapes against the backdrop of cancer's genomic instability. Notably, the close clustering of normal and tumor samples within the same tissue type suggests that the primary tissue-specific gene expression programs remain a dominant feature of the epigenetic profile, even in the transformed state.

This analysis provides compelling evidence for the persistence of tissue identity at the epigenetic level in tumors and offers a potential avenue for the development of diagnostic markers that reflect the tissue of origin. Moreover, the distinct separation of tissue types based on tissue-specific gene bodies may guide the identification of epigenetic targets for tailored therapeutic interventions.

3.3 Discussion and future perspective

Our comprehensive investigation into the epigenetic modifications across a range of human tissues has unveiled the nuanced landscape of 5-hydroxymethylcytosine (5hmC) in the context of health and disease.

The consistent 5hmC patterns observed in gene bodies and repetitive genomic elements highlight the complexity of epigenetic regulation in tumorigenesis. While the overall levels of 5hmC are reduced in tumor tissues, the relative distribution among various genes remains stable, suggesting that the epigenetic hierarchy is preserved despite the disease state. This conservation of 5hmC patterning offers a glimpse into the resilience of epigenetic landscapes, even as they undergo cancer-associated reprogramming.

Intriguingly, our findings also point to the relation between the increase of the 5hmC level and the activation of certain repetitive elements, such as SINEs and LTRs, in tumors. This contrasts with the general decrease in 5hmC within LINEs and intergenic regions, implying a selective and differential epigenetic remodeling process. The functional

implications of these alterations remain to be fully elucidated; however, they may be linked to the activation of oncogenes or the suppression of tumor suppressor genes.

Furthermore, through the application of t-SNE analysis and analysis on relative 5hmC levels on tissue-specific 5hmC-enriched genes in each tissue type, we have delineated the distinct 5hmC signatures that differentiate among different tissue types.

Looking ahead, there are several perspectives for future research. A deeper exploration into the biological consequences of 5hmC changes will be crucial for understanding their role in gene expression regulation and cancer development. The epigenetic consistency observed in tissue-specific gene bodies opens up potential for diagnostic advancements, allowing for the detection and classification of tumors based on their epigenetic profiles.

Additionally, the role of 5hmC in the activation of repetitive elements warrants further investigation. Understanding the mechanisms behind this phenomenon could reveal new targets for epigenetic therapy, offering strategies to modulate these modifications to inhibit tumor growth or progression.

In conclusion, the study of 5hmC modifications presents a rich landscape for discovery in the field of epigenetics. Our work lays the foundation for future investigations on the role of DNA modification in cancer metastasis and the epigenetic regulation of cancer that will ultimately contributing to the advancement of precision medicine in oncology.

3.4 Experimental section

3.4.1 Genome DNA extraction and purification from FFPE samples

Formalin-fixed paraffin-embedded tumor samples were provided by Feng Yue (Northwestern University). Genomic DNA extraction from pre-sliced FFPE samples was performed using a modified protocol based on QIAamp DNA FFPE Tissue Kit handbook, optimized to yield high-quality DNA suitable for downstream sequencing and PCR analysis.

Each pre-sliced FFPE sample was treated with 1 mL of xylene, followed by vortexing for 10 seconds to ensure complete dissolution of paraffin. The samples were then centrifuged at 21,000× g for 2 minutes at room temperature. Post centrifugation, the supernatant was carefully discarded, leaving behind a transparent pellet. The pellets were washed with 1 mL of 100% ethanol to remove any residual xylene. This step involved vortex mixing followed by centrifugation at 21,000× g for another 2 minutes. The supernatant was then carefully removed, ensuring the pellet remained undisturbed. To evaporate the residual ethanol, the pellet was air-dried on a heating block at 56°C for 10 minutes. The dried pellet was then resuspended in 180 μL of Buffer ATL, to which 20 μL of proteinase K (New England Biolabs) was added. The mixture was vortexed thoroughly and then incubated at 56°C overnight. Following incubation, the samples were heated at 90°C for 1 hour to reverse formaldehyde modifications on the nucleic acids. After a brief centrifugation to remove any condensate, the samples were allowed to cool to room temperature.

To further purify the DNA, 2 μL of 100 mg/mL RNase A was added to each sample and incubated for 2 minutes at room temperature. Subsequently, 200 μL of Buffer AL was added to each sample, followed by thorough vortex mixing. An equal volume of 96-100% ethanol was then added, and the mixture was immediately and thoroughly mixed. The lysate was then transferred to a QIAamp MinElute column, centrifuged, and washed as per the

standard protocol. The final step involved the elution of DNA using either pre-warmed 56 °C nuclease-free water for immediate use samples The concentration of the eluted DNA was quantified using a Qubit fluorometer and DNA solution were stored at -80 °C

3.4.2 5hmC-Seal of FFPE sample

The 5hmC-Seal was performed as previously described in Han et al. [83] , with some modifications.

3.4.2.1 Buffer preparation

Prepare buffers as following:

Buffer 1: 5 mM Tris-HCl (pH = 7.5), 0.5 mM EDTA, 1 M NaCl, and 0.2% Tween 20

Buffer 2: Contains 5 mM Tris-HCl (pH = 7.5), 0.5 mM EDTA, and 0.2% Tween 20

Buffer 3: 5 mM Tris-HCl (pH = 9.0), 0.5 mM EDTA, 1 M NaCl, and 0.2% Tween 20

Buffer 4: 5 mM Tris-HCl (pH = 9.0), 0.5 mM EDTA, and 0.2% Tween 20

Binding Buffer:

10 mM Tris-HCl (pH = 7.5), 1 mM EDTA, 2 M NaCl, and 0.4% Tween 20

3.4.2.2 gDNA fragmentation

The DNA fragmentation process was initiated using the KAPA HyperPrep Kit. Based on different concentrations of each sample, library construction start with 20-100 ng of gDNA in a total volume of 12 µL DNase-free water, mixed with 2 µL of diluted condition buffer and 2 µL of KAPA fragmentation buffer and 4 µL of KAPA fragmentation enzyme. The mixture was proceeded to a PCR cycle with the following conditions: initial cooling at 4°C for 2 minutes, incubation at 37°C for 30 minutes, followed by indefinite holding at 4°C.

3.4.2.3 End repair and A-tailing

To the fragmented DNA solution (20 μL from the previous step), 2.8 μL of end repair and A-tailing buffer mix and 1.2 μL of end repair and A-tailing enzyme mix were added. The reaction was incubated at 20°C for 30 minutes, followed by a 30-minute incubation at 65°C to facilitate end repair and the addition of an adenosine (A) to the 3' ends of the DNA fragments.

3.4.2.4 Adapter ligation and DNA purification

For adapter ligation, the 24 μL solution from the E&A reaction was combined with 2.8 μL of nuclease-free water. Subsequently, 12 μL ligation buffer mix, 4 μL adapter ligase, and 1.2 μL Illumina indexed adapter (NextFlex) were added to the mixture. The reaction was conducted at 20°C for a duration ranging from 1 to 4 hours based on the amount of input, with the lid temperature set to 40°C.

The DNA purification was performed using the Zymo DNA Clean & Concentrator (DCC) Kit post-adapter ligation. The process started by adding 305 μL of DNA binding buffer to the ligation mixture (44 μL) along with 1 μL of salmon sperm DNA (1 mg/mL) as a carrier. This was applied to the Zymo-Spin Column for selective DNA binding. The column-bound DNA was then washed by DNA wash buffer to eliminate unbound components. Finally, the DNA was eluted in 20 μL of pre-warmed (56°C) nuclease-free water, concentrating the DNA for subsequent applications.

3.4.2.5 Selective 5hmC chemical labeling

The fragmented DNA was combined with βGT and N_3-UDP-Glc in a glucosylation buffer consisting of 50 mM HEPES buffer (pH = 8.0) and 25 mM $MgCl_2$. This mixture was incubated at 37°C for 2 hours. The DNA was purified using Zymo DNA Clean &

Concentrator (DCC) Kit following the manufacturer's instructions and elute in 30 μL of pre-warmed nuclease-free water.

337 mg of Dibenzocyclooctyne-PEG4-biotin (DBCO-PEG4-Biotin) was dissolved in 1 mL of ddH_2O to give around a concentration of 0.45 mM. 1 μL of DBCO-PEG4-Biotin solution was added to the purified DNA, and the mixture was incubated at 37°C for 2 hours. The DNA was then purified using Zymo DNA Clean & Concentrator (DCC) Kit following the manufacturer's instructions and elute in 30 μL of pre-warmed nuclease-free water.

3.4.2.6 Streptavidin-bead-based 5hmC enrichment

Bead preparation: For the streptavidin-bead-based 5hmC enrichment, Dynabeads M-270 Streptavidin (Invitrogen) were prepared by first washing them three times with Buffer 1. Each wash involved adding 2.5 μL of M270 beads per reaction to Buffer 1. After the washes, the beads were incubated in 100 μL of Buffer 1 containing 1 μL of 1 mg/mL salmon sperm DNA for 30 minutes, was aimed at blocking non-specific binding sites on the beads. Following the incubation, the beads were washed three additional times with Buffer 1 to remove any unbound salmon sperm DNA. Finally, the beads were resuspended in 30 μL of Binding Buffer for each reaction, preparing them for the subsequent pulldown process.

5hmC pulldown and wash: Adding 30 μL of the prepared DNA solution to the prepared beads in Binding Buffer. This mixture was incubated for 30 minutes to allow the 5hmC-enriched DNA to bind to the beads. The bead-DNA complex was then washed twice with 100 μL of each Buffer 1-4 sequentially, ensuring the removal of non-specifically bound DNA. After each wash, the beads were carefully transferred to a new tube to avoid carryover of washing buffers. Following the final wash with Buffer 4, the DNA-beads complex was resuspended in 23.8 μL of nuclease-free water for following PCR process. All binding and washing were done at room temperature with gentle rotation.

The captured DNA fragments (23.8 μL) was combined with 1.2 μL primer mix and 25 μL 2x enzyme mix supplied in the KAPA HyperPrep Kit. then amplified with 10–15 cycles of PCR amplification based on the amount of input gDNA. The PCR products were purified using 1.0× AMPure XP beads according to the manufacture's instruction. The libraries were quantified by a Qubit fluorometer (Life Technologies) and sequenced on NovaSeq platform.

3.4.2.7 5hmC-Seal data analysis

Paired-end sequencing reads were first trimmed by Trim_Galore to remove adaptor sequences and low-quality nucleotides. Clean reads were then aligned to hg19 reference genome by Bowtie2, with only uniquely mapped reads retained for the following analysis. PCR duplicates were removed by Samtools. MACS2 was used to call 5hmC peaks across the genome, and Bedtools was used to filter out 5hmC peaks in hg19 blacklisted regions. Fragment counts on gene bodies were calculated by featureCounts. Tissue-specific genes were generated by comparison of certain tissue samples with samples from all other tissues using DESeq2.

References

1. Waddington, C.H., *The epigenotype. 1942.* Int J Epidemiol, 2012. **41**(1): p. 10-3.
2. Berger, S.L., et al., *An operational definition of epigenetics.* Genes Dev, 2009. **23**(7): p. 781-3.
3. Jeltsch, A., J. Broche, and P. Bashtrykov, *Molecular Processes Connecting DNA Methylation Patterns with DNA Methyltransferases and Histone Modifications in Mammalian Genomes.* Genes, 2018. **9**(11): p. 566.
4. Andrea Fuso, T.R., Michela Orticello, Marco Lucarelli, *The complex interplay between DNA methylation and miRNAs in gene expression regulation.* Biochimie, 2020.
5. Luger, K., et al., *Crystal structure of the nucleosome core particle at 2.8 A resolution.* Nature, 1997. **389**(6648): p. 251-60.
6. Suzuki, M.M. and A. Bird, *DNA methylation landscapes: provocative insights from epigenomics.* Nat Rev Genet, 2008. **9**(6): p. 465-76.
7. Hsu, P.J., H. Shi, and C. He, *Epitranscriptomic influences on development and disease.* Genome Biology, 2017. **18**(1).
8. Srinageshwar, B., et al., *Role of Epigenetics in Stem Cell Proliferation and Differentiation: Implications for Treating Neurodegenerative Diseases.* Int J Mol Sci, 2016. **17**(2).
9. Kyriakou, G. and M. Melachrinou, *Cancer stem cells, epigenetics, tumor microenvironment and future therapeutics in cutaneous malignant melanoma: a review.* Future Oncol, 2020. **16**(21): p. 1549-1567.
10. Hardy, T.M. and T.O. Tollefsbol, *Epigenetic diet: impact on the epigenome and cancer.* Epigenomics, 2011. **3**(4): p. 503-18.
11. Ito, S., et al., *Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine.* Science, 2011. **333**(6047): p. 1300-1303.

12. Wyatt, G.R., *Recognition and estimation of 5-methylcytosine in nucleic acids.* Biochem J, 1951. **48**(5): p. 581-4.

13. Holliday, R. and J.E. Pugh, *DNA modification mechanisms and gene activity during development.* Science, 1975. **187**(4173): p. 226-32.

14. Riggs, A.D., *X inactivation, differentiation, and DNA methylation.* Cytogenet Cell Genet, 1975. **14**(1): p. 9-25.

15. Kalousek, F. and N.R. Morris, *Deoxyribonucleic acid methylase activity in rat spleen.* J Biol Chem, 1968. **243**(9): p. 2440-2.

16. Roy, P.H. and A. Weissbach, *DNA methylase from HeLa cell nuclei.* Nucleic Acids Res, 1975. **2**(10): p. 1669-84.

17. Bestor, T.H. and V.M. Ingram, *Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA.* Proc Natl Acad Sci U S A, 1983. **80**(18): p. 5559-63.

18. Chen, T., et al., *Complete inactivation of DNMT1 leads to mitotic catastrophe in human cancer cells.* Nat Genet, 2007. **39**(3): p. 391-6.

19. Ratel, D., et al., *N6-methyladenine: the other methylated base of DNA.* Bioessays, 2006. **28**(3): p. 309-15.

20. Campbell, J.L. and N. Kleckner, *E. coli oriC and the dnaA gene promoter are sequestered from dam methyltransferase following the passage of the chromosomal replication fork.* Cell, 1990. **62**(5): p. 967-79.

21. Collier, J., H.H. McAdams, and L. Shapiro, *A DNA methylation ratchet governs progression through a bacterial cell cycle.* Proc Natl Acad Sci U S A, 2007. **104**(43): p. 17111-6.

22. Low, D.A., N.J. Weyand, and M.J. Mahan, *Roles of DNA adenine methylation in regulating bacterial gene expression and virulence.* Infect Immun, 2001. **69**(12): p. 7197-204.

23. Luo, G.Z., et al., *DNA N(6)-methyladenine: a new epigenetic mark in eukaryotes?* Nat Rev Mol Cell Biol, 2015. **16**(12): p. 705-10.

24. Boulias, K. and E.L. Greer, *Means, mechanisms and consequences of adenine methylation in DNA.* Nat Rev Genet, 2022. **23**(7): p. 411-428.

25. Li, E. and Y. Zhang, *DNA Methylation in Mammals.* Cold Spring Harbor Perspectives in Biology, 2014. **6**(5): p. a019133-a019133.

26. Stewart, C.L., et al., *De novo methylation, expression, and infectivity of retroviral genomes introduced into embryonal carcinoma cells.* Proc Natl Acad Sci U S A, 1982. **79**(13): p. 4098-102.

27. Okano, M., et al., *DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development.* Cell, 1999. **99**(3): p. 247-57.

28. Okano, M., S. Xie, and E. Li, *Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases.* Nat Genet, 1998. **19**(3): p. 219-20.

29. Jabbari, K. and G. Bernardi, *Cytosine methylation and CpG, TpG (CpA) and TpA frequencies.* Gene, 2004. **333**: p. 143-9.

30. Clough, D.W., L.M. Kunkel, and R.L. Davidson, *5-Azacytidine-induced reactivation of a herpes simplex thymidine kinase gene.* Science, 1982. **216**(4541): p. 70-3.

31. Jones, P.A. and S.M. Taylor, *Cellular differentiation, cytidine analogs and DNA methylation.* Cell, 1980. **20**(1): p. 85-93.

32. Mohandas, T., R.S. Sparkes, and L.J. Shapiro, *Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation.* Science, 1981. **211**(4480): p. 393-6.

33. Li, E., C. Beard, and R. Jaenisch, *Role for DNA methylation in genomic imprinting.* Nature, 1993. **366**(6453): p. 362-5.

34. Meehan, R.R., et al., *Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs.* Cell, 1989. **58**(3): p. 499-507.

35. Hendrich, B. and A. Bird, *Identification and characterization of a family of mammalian methyl-CpG binding proteins.* Mol Cell Biol, 1998. **18**(11): p. 6538-47.

36. Bird, A.P. and A.P. Wolffe, *Methylation-induced repression--belts, braces, and chromatin.* Cell, 1999. **99**(5): p. 451-4.

37. Meissner, A., et al., *Genome-scale DNA methylation maps of pluripotent and differentiated cells.* Nature, 2008. **454**(7205): p. 766-70.

38. Zhang, Y., et al., *DNA methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution.* PLoS genetics, 2009. **5**(3): p. e1000438.

39. Iguchi-Ariga, S.M. and W. Schaffner, *CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation.* Genes Dev, 1989. **3**(5): p. 612-9.

40. Kovesdi, I., R. Reichel, and J.R. Nevins, *Role of an adenovirus E2 promoter binding factor in E1A-mediated coordinate gene control.* Proc Natl Acad Sci U S A, 1987. **84**(8): p. 2180-4.

41. Jones, P.A., *Functions of DNA methylation: islands, start sites, gene bodies and beyond.* Nat Rev Genet, 2012. **13**(7): p. 484-92.

42. Chodavarapu, R.K., et al., *Relationship between nucleosome positioning and DNA methylation.* Nature, 2010. **466**(7304): p. 388-392.

43. Laurent, L., et al., *Dynamic changes in the human methylome during differentiation.* Genome Res, 2010. **20**(3): p. 320-31.

44. Hellman, A. and A. Chess, *Gene body-specific methylation on the active X chromosome.* Science, 2007. **315**(5815): p. 1141-3.

45. Ito, S., et al., *Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification.* Nature, 2010. **466**(7310): p. 1129-33.

46. Kriaucionis, S. and N. Heintz, *The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain.* Science, 2009. **324**(5929): p. 929-30.

47. Tahiliani, M., et al., *Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1.* Science, 2009. **324**(5929): p. 930-5.

48. Ito, S., et al., *Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine.* Science, 2011. **333**(6047): p. 1300-3.

49. He, Y.F., et al., *Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA.* Science, 2011. **333**(6047): p. 1303-7.

50. Maiti, A. and A.C. Drohat, *Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites.* J Biol Chem, 2011. **286**(41): p. 35334-35338.

51. Krokan, H.E. and M. Bjørås, *Base excision repair.* Cold Spring Harb Perspect Biol, 2013. **5**(4): p. a012583.

52. Cortázar, D., et al., *Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability.* Nature, 2011. **470**(7334): p. 419-423.

53. Kumar, R., et al., *AID stabilizes stem-cell phenotype by removing epigenetic memory of pluripotency genes.* Nature, 2013. **500**(7460): p. 89-92.

54. Rai, K., et al., *DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45.* Cell, 2008. **135**(7): p. 1201-1212.

55. Yan, X.X., et al., *The deficiency of 5-methylcytosine (5mC) and its ramification in the occurrence and prognosis of colon cancer.* Medicine (Baltimore), 2023. **102**(34): p. e34860.

56. Walker, N.J., et al., *Hydroxymethylation profile of cell-free DNA is a biomarker for early colorectal cancer.* Sci Rep, 2022. **12**(1): p. 16566.

57. Hu, J., et al., *5mC regulator-mediated molecular subtypes depict the hallmarks of the tumor microenvironment and guide precision medicine in bladder cancer.* BMC Medicine, 2021. **19**(1): p. 289.

58. Easwaran, H., et al., *A DNA hypermethylation module for the stem/progenitor cell signature of cancer.* Genome research, 2012. **22**(5): p. 837-849.

59. Skvortsova, K., et al., *DNA hypermethylation encroachment at CpG island borders in cancer is predisposed by H3K4 monomethylation patterns.* Cancer cell, 2019. **35**(2): p. 297-314. e8.

60. Scourzic, L., E. Mouly, and O.A. Bernard, *TET proteins and the control of cytosine demethylation in cancer.* Genome medicine, 2015. **7**(1): p. 1-16.

61. Russler-Germain, D.A., et al., *The R882H DNMT3A mutation associated with AML dominantly inhibits wild-type DNMT3A by blocking its ability to form active tetramers.* Cancer Cell, 2014. **25**(4): p. 442-54.

62. Bera, R., et al., *Genetic and Epigenetic Perturbations by DNMT3A-R882 Mutants Impaired Apoptosis through Augmentation of PRDX2 in Myeloid Leukemia Cells.* Neoplasia, 2018. **20**(11): p. 1106-1120.

63. Takeshima, H. and T. Ushijima, *Accumulation of genetic and epigenetic alterations in normal cells and cancer risk.* NPJ precision oncology, 2019. **3**(1): p. 7.

64. Claus, R., et al., *Quantitative DNA methylation analysis identifies a single CpG dinucleotide important for ZAP-70 expression and predictive of prognosis in chronic lymphocytic leukemia.* J Clin Oncol, 2012. **30**(20): p. 2483-91.

65. van den Bent, M.J., *Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective.* Acta Neuropathol, 2010. **120**(3): p. 297-304.

66. 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, 1992. **89**(5): p. 1827-31.

67. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-322.

68. Clark, S.J., et al., *Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq).* Nat Protoc, 2017. **12**(3): p. 534-547.

69. Farlik, M., et al., *Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics.* Cell Rep, 2015. **10**(8): p. 1386-97.

70. Miura, F., et al., *Amplification-free whole-genome bisulfite sequencing by post-bisulfite adaptor tagging.* Nucleic Acids Res, 2012. **40**(17): p. e136.

71. Liu, Y., et al., *Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution.* Nat Biotechnol, 2019. **37**(4): p. 424-429.

72. Vaisvila, R., et al., *Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA.* Genome Res, 2021. **31**(7): p. 1280-1289.

73. Zhang, L., et al., *Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA.* Nature Chemical Biology, 2012. **8**(4): p. 328-330.

74. Stroud, H., et al., *5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells.* Genome biology, 2011. **12**: p. 1-8.

75. Cimmino, L., et al., *TET family proteins and their role in stem cell differentiation and transformation.* Cell Stem Cell, 2011. **9**(3): p. 193-204.

76. Ficz, G. and J.G. Gribben, *Loss of 5-hydroxymethylcytosine in cancer: Cause or consequence?* Genomics, 2014. **104**(5): p. 352-357.

77. Ficz, G. and J.G. Gribben, *Loss of 5-hydroxymethylcytosine in cancer: cause or consequence?* Genomics, 2014. **104**(5): p. 352-7.

78. Pfeifer, G.P., et al., *The role of 5-hydroxymethylcytosine in human cancer.* Cell Tissue Res, 2014. **356**(3): p. 631-41.

79. Cui, X.L., et al., *A human tissue map of 5-hydroxymethylcytosines exhibits tissue specificity through gene and enhancer modulation.* Nat Commun, 2020. **11**(1): p. 6161.

80. Zeng, C., et al., *Towards precision medicine: advances in 5-hydroxymethylcytosine cancer biomarker discovery in liquid biopsy.* Cancer Communications, 2019. **39**(1): p. 12.

81. Song, C.-X., et al., *Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine.* Nature Biotechnology, 2011. **29**(1): p. 68-72.

82. Han, D., et al., *A highly sensitive and robust method for genome-wide 5hmC profiling of rare cell populations.* Molecular cell, 2016. **63**(4): p. 711-719.

83. Han, D., et al., *A Highly Sensitive and Robust Method for Genome-wide 5hmC Profiling of Rare Cell Populations.* Molecular Cell, 2016. **63**(4): p. 711-719.