The recognition, generation and detection of 5-hydroxymethylcytosine

Ying Bi

This thesis is submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Clinical Medicine

St. Cross College

Ludwig Institute for Cancer Research

Nuffield Department of Clinical Medicine

University of Oxford

Trinity 2019

This thesis is submitted to the Department of Medical Science, University of Oxford, in fulfilment of the requirements for the degree of Doctor of Philosophy. This thesis is the result of my own work unless otherwise stated and does not constitute part of any other thesis.

The recognition, generation and detection of 5-hydroxymethylcytosine

Ying Bi

Trinity 2019

Doctor of Philosophy in Clinical Medicine

St. Cross College

Abstract

5-hydroxymethylcytosine (5hmC) is a DNA modification generated from the oxidation of 5-methylcytosine (5mC) by ten-eleven translocation (TET) enzymes. 5hmC is enriched at promoters, gene bodies and enhancers. The enrichment of 5hmC at the gene body is positively correlated with gene expression. Low levels of 5hmC are observed in various solid tumours and haematological malignancies. The distribution of 5hmC in the genome is strand asymmetric at CG sequences. There are high levels of adjacent and opposing hydroxymethylated and methylated CpG (H/M) sites in the genome. 5hmC in H/M sites account for around 60% of all 5hmC. In this thesis, proteins that specifically recognise H/M sites were identified using DNA pull-down followed by LC-MS/MS. However, the binding specificity requires more biochemical validations. The enzymatic pathway that generates H/M site was studied and preliminarily results showed that TET2 have strand selective activity on double-strand DNA, which might be a possible way to generate H/M sites. To better understand the distribution of 5hmC in the genome, a 5hmC selective oxidation reaction was applied to map 5hmC on DNA and RNA. The method can be further expended to 5mC mapping in RNA in combination with TET oxidation of 5mC to 5hmC.

Acknowledgements

I would like to thank my supervisors Chunxiao Song and Skirmantas Kriaucionis for giving me the opportunity to do this project, having insightful ideas and fruitful discussions. I thank all the former and current lab members of both labs: Anandhakumar Chandran, Ashvina Segaran, Fang Yuan, Gergana Velikova, Hiromi Tagoh, Jingfei Cheng, Lei Chen, Lin Chen, Mankgopo Kgatle, Marketa Tomkova, Martin Cusack, Michael McClellan, Olena Yavorska, Paolo Spingardi
Paulina Siejka-Zielinska, Sophie Kirschner, Wenjun Huang and Yibin Liu for scientific suggestions and discussion as well as friendship and accompany. Specifically, I would like to thank Fang Yuan for working together on the RNA modification mapping project, Paulina Siejka-Zielinska for suggestion and discussion on protein purifications, Paolo Spingardi for help with the LC-MS/MS, Hiromi Tagoh for teaching me nuclear protein extraction from brain and Michael McClellan for providing me with mouse brain for testing. I would like to thank Georgina Berridge from Benedikt Kessler group for processing proteomics samples of DNA pull-down.

I would like to acknowledge the China Scholarship Council and the University of Oxford for funding of my study and research.

List of abbreviations

2HG	2-hydroxyglutarate
5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
A	adenine
ACE-seq	APOBEC-coupled epigenetic sequencing
a-KG	a-ketoglutarate
AML	acute myeloid leukaemia
BER	base excision repair
BS-Seq	bisulfite sequencing
βGT	β-glucosyltransferase
C	cytosine
C/C sites	unmodified CpG sites
CAPS	chemical-assisted pyridine borane sequencing
CD	catalytic domain
CDS	coding sequence
cf-DNA	cell-free DNA
CGI	CpG island
CIMP	CpG island methylator phenotype
DHU	dihydrouracil
DMEM	Dulbecco's Modified Eagle Medium
DNA	deoxyribonucleic acid
DNMTs	DNA methyltransferases
DSBH	double-stranded β helix
EMSA	Electrophoretic mobility shift assay
G	guanine
GO	Gene ontology
H/C sites	hemi-hydroxymethylated CpG sites
H/H sites	symmetrically hydroxymethylated CpG sites
H/M sites	adjacent and opposing methylated and hydroxymethylated CpG sites
hm^5C	5-hydroxymethylcytidine
IDH	Isocitrate dehydrogenase

LC-MS/MS	liquid chromatography-tandem mass spectrometry
LIC	ligation independent cloning
M/C sites	hemimethylated CpG sites
M/M sites	symmetrically methylated CpG sites
m^5C	5-methylcytidine
m^6A	N6-methyladenosine
MBD	methyl-CpG-binding domain
mESCs	mouse embryonic stem cells
NgTET	*Naegleria gruberi Tet-like oxygenase*
NuRD	Nucleosome Remodelling Deacetylase
oxBS-Seq	oxidative bisulfite sequencing
PDEI	snake venom phosphodiesterase I
PGCs	primordial germ cells
SAM	S-adenosyl-methionine
SEM	the standard error of the mean
SD	standard deviation
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SRA	SET- and RING finger-associated
T	thymine
TAB-Seq	TET-assisted bisulfite sequencing
TAPS	TET-assisted pyridine borane sequencing
TAWO-Seq	TET-Assisted WO-Seq
TDG	thymine DNA glycosylase
TET	Ten-eleven translocation enzyme
TKO	triple-knockout
TSS	transcription start sites
WO-Seq	peroxotungstate oxidation sequencing

Table of Contents

1 Chapter 1 Introduction 1

1.1 Mammalian DNA modifications 1

1.2 5mC 2

1.2.1 DNA methyltransferases 5

1.2.2 5mC binding proteins 7

1.3 5hmC 12

1.3.1 TET proteins 13

1.3.2 5hmC and DNA demethylation 16

1.3.3 Methods for genome-wide 5hmC mapping 18

1.3.4 Genomic distribution of 5hmC 19

1.3.5 5hmC binding proteins 22

1.3.6 5hmC and cancer 25

1.4 Aim of the study 26

2 Chapter 2 Materials and Methods 28

2.1 Preparation of DNA and RNA 28

2.1.1 BDNF probes for DNA pull-down 28

2.1.2 Probe1 for DNA pull-down and model DNA for TET reactions 28

2.1.3 Probe2 for DNA pull-down 28

2.1.4 Probes for electrophoretic mobility shift assay 29

2.1.5 Non-specific DNA competitor 29

2.1.6 Model DNA for peroxotungstate reactions 29

2.1.7 DNA Annealing 30

2.1.8 Model RNA in vitro transcription 30

2.1.9 RNA purification with TRIzol Reagents 31

2.2 Plasmids 31

2.2.1 Ligation independent cloning 32

2.3 Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and Western blot 33

2.3.1 SDS-PAGE 33

2.3.2 Western blot 34

2.4 293T cell culture and transfection 34

2.5 Mouse embryonic stem cell culture 35

2.6 Nuclear protein extraction from mESCs 35

2.7 Nuclear protein extraction from mouse brain 36

2.8 DNA pull-down 37

2.9 Protein expression and purification 38

2.9.1 Protein expression in E. coli 38

2.9.2 Purification of ASH2L, UHRF1, YTHDF3 from E coli. 38

2.9.3 Purification of YBX1 from E. coli 39

2.9.4 Protein expression in Expi293F™ Cells 40

2.9.5 Protein purification from Expi293F™ Cells 41

2.10 Electrophoretic mobility shift assay (EMSA) 41

2.11 TET reaction 42

2.11.1 Reaction 42

2.11.2 Separating strand 42

2.11.3 DNA hydrolysis for LC-MS/MS 43

2.11.4 LC-MS/MS 44

2.12 Peroxotungstate oxidation reaction 45

2.12.1 Preparation of peroxotungstate 45

2.12.2 Peroxotungstate oxidation reaction on DNA 45

2.12.3 Peroxotungstate oxidation reaction on RNA 46

2.13 TET oxidation of RNA 46

2.14 Reverse transcription 47

2.15 Quantification of conversation rate 47

2.15.1 Restriction enzyme digestion method 47

2.15.2 TA cloning method 48

2.16 Mass spectrometry analysis of DNA fragments 48

2.17 Enzymatic DNA hydrolysis 48

2.17.1 PDE1 and Benzonase + PDE1 methods 48

2.17.2 DNase I + PDE1 methods 49

2.17.3 DNA Degradase Plus methods 49

2.17.4 Nucleoside digestion mix methods 49

2.17.5 Nuclease P1 methods 49

2.18 Notes of statistical analysis 50

3 Chapter 3 H/M sites binding proteins 51

3.1 Introduction 51

3.2 Identifications of potential H/M sites binding proteins using DNA pull-down followed by mass spectrometry 53

3.2.2 Optimisation of pull-down conditions 55

3.2.3 Samples for mass spectrometry analysis 62

3.2.4 Mass spectrometry analysis results 65

3.3 Validation of H/M site binding proteins 72

3.3.1 YBX1 72

3.3.2 RBBP5 74

3.3.3 UHRF1 78

3.3.4 TET1 83

3.3.5 YTHDF3 89

3.4 Discussion 91

3.4.1 DNA pull-down and MS 91

3.4.2 Validation of candidate H/M site binding proteins 94

4 Chapter 4 Generation of H/M sites 98

4.1 Introduction 98

4.2 Results 100

4.3 Discussion 105

5 Chapter 5 5hmC detection methods 107

5.1 Cell-free DNA 5hmC signature and cancer detection 107

5.2 Base-resolution 5mC and 5hmC mapping methods 108

5.3 Peroxotungstate selective 5hmC oxidation 110

5.4 Reaction with single-strand DNA 111

5.5 Reaction optimisation on double-strand DNA 113

5.6 TET-assisted pyridine borane sequencing 117

5.7 Comparison of commonly used methods for DNA hydrolysis 120

5.8 5hmC and 5mC mapping in RNA 123

5.9 Discussion 125

6 Chapter 6 Conclusion and Discussion 127

6.1 The recognition of H/M sites 127

6.2 The generation of H/M sites 132

6.3 Detection of 5hmC 134

7 References 137

1 Chapter 1 Introduction

1.1 Mammalian DNA modifications

The genetic information in cells is stored in deoxyribonucleic acid (DNA), which consists of four bases: cytosine (C), guanine (G), adenine (A) and thymine (T). DNA modifications are covalent chemical modifications on the DNA bases. The modifications discussed in this thesis are the ones generated actively by enzymes rather than the result of unwanted DNA damage. These modifications do not alter base-pairing properties of the DNA bases but affect the protein-DNA interaction of DNA binding proteins, thereby regulating chromatin structure and gene expression.

The most abundant DNA modification in mammals is 5-methylcytosine (5mC). This modification is deposited and maintained by DNA methyltransferases (DNMTs) (Li et al., 1992; Okano et al., 1999). The level of 5mC is around 3% to 5% of all cytosines in mammalian genomes (Ehrlich et al., 1982; Fraga et al., 2002; Ito et al., 2011; Kuo et al., 1980; Tomkova et al., 2016; Wilson et al., 1986). Ten-eleven translocation (TET) family proteins oxidise 5mC into 5-hydroxymethylcytosine (5hmC) (Ito et al., 2010; Tahiliani et al., 2009), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (He et al., 2011; Ito et al., 2011) in a step-wise manner. 5fC and 5caC but not 5hmC can be recognised and excised by thymine DNA glycosylase (TDG), and further repaired to unmodified cytosines by base excision repair (BER) pathway (Figure 1-1) (He et al., 2011; Maiti and Drohat, 2011; Weber et al., 2016). 5hmC is more abundant

than 5fC and 5caC in the genome (about 1000, 20 and 3 every million cytosines respectively in mouse embryonic stem cells) (Figure 1-1) (Ito et al., 2011).

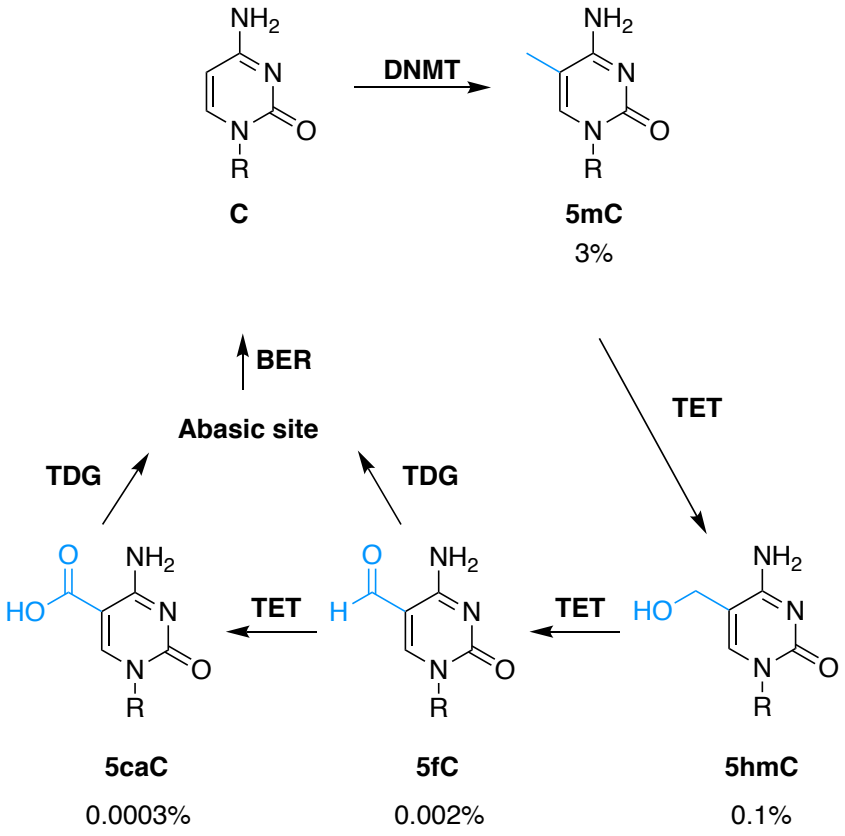


Figure 1-1 Enzymatic conversion of cytosine modifications. 5mC is deposited and maintained by DNA methyltransferase (DNMT). Ten-eleven translocation (TET) enzyme oxidises 5mC to 5hmC, 5fC, 5caC in a stepwise manner. 5fC and 5caC can be recognised and removed by the thymine-DNA glycosylase (TDG), the abasic site left is filled in with unmodified cytosine via DNA base-excision repair (BER). The abundance of each modification in mouse embryonic stem cells is labelled.

1.2 5mC

The existence of 5mC in mammalian DNA was first reported in 1948 from calf thymus DNA (Hotchkiss, 1948). Follow-up studies confirmed the finding (Wyatt,

1951; Wyatt and Markham, 1950). 5mC exists primarily in the sequence context of CpG dinucleotides, with the exception of embryonic stem cells and neurons where about 25% methylated cytosines are not in CpG context (Guo et al., 2013; Lister et al., 2009). Mammalian genome is globally methylated with a small proportion unmethylated (Bird et al., 1985). Whole-genome base-resolution mapping of DNA methylation in mouse embryonic stem cells (mESCs) showed that around 90% of CpGs are highly methylated (methylation level higher than 50%) and only about 6.5% of CpGs are lowly methylated (methylation level lower than 10%) (Stadler et al., 2011).

CpG sites are not evenly distributed in the genome. There are regions of high CpG frequencies called CpG island (CGI) which are frequently unmethylated (Bird, 1986). CGIs are mainly located at gene promoters (Antequera and Bird, 1993). The CpG sites are depleted at the rest of the genome, with around 20% of the excepted frequency (Lander et al., 2001). The depletion is not observed in species that lack cytosine methylations including fruit flies and bees (Bird, 1980). This is attributed to the fact that methylated cytosine is more prone to mutation than unmethylated ones. The spontaneous deamination rate of 5mC is higher than that of unmodified cytosine (Lindahl and Nyberg, 1974; Shen et al., 1994). Moreover, the deamination of 5mC results in T:G mismatch which is less efficiently repaired than the U:G mismatch generated from unmodified cytosine, due to the risk of excising the wrong base (Lindahl, 1993).

Early studies showed a transcriptional repression role of 5mC. *In vivo,* artificial manipulation of the methylation status of CpG sites upstream of a reporter gene showed that methylations repress gene expression in *cis* (Siegfried et al., 1999). *In vitro* methylated DNA fragments, after inserted into cells and integrated into the genome, formed a less accessible chromatin structure than the unmethylated counterparts (Keshet et al., 1986). However, the genome-wide analysis showed the promoter methylation is negatively correlated with CpG content regardless of the gene expression level, suggesting that 5mC is unlikely to have a global role in gene expression (Weber et al., 2007). Despite the lack of global correlation with gene expression, 5mC represses the expression of a subset of genes such as imprinted genes and transposable elements. For example, parental-specific hypermethylation upstream of H19 gene silences the parental allele and ensures maternal-specific expression (Thorvaldsen et al., 1998; Tremblay et al., 1995). Intracisternal A-particle (IAP) retrotransposons are heavily methylated in somatic tissues and demethylation activates the transcription of IAP (Walsh et al., 1998). 5mC is also associated with X chromosome inactivation. Demethylation results in re-expression of X chromosome genes (Mohandas et al., 1981; Venolia et al., 1982). The methylation of inactive X occurs after the inactivation, suggesting methylation does not initiate the inactivation but plays a role in maintaining the inactivated state (Lock et al., 1987). In summary, 5mC do not affect gene expression globally but have a repressive role on a subset of genes.

In cancer cells, the methylation profile is altered. Globally, there is a loss of 5mC (Vidal et al., 2017). Hypomethylation promotes chromosomal instability (Eden et al.,

2003). DNMT1/DNMT3B double knockout in human cancer cell line HCT116 results in genomic instability and chromosomal translocations (Karpf and Matsui, 2005). In mice, genomic hypomethylation caused by hypomorphic DNMT1 leads to aneuploidy and T cell lymphomas (Gaudet et al., 2003). Locally, promoter hypomethylation activates oncogenes, for example, hypomethylation of cyclin D2 gene in gastric carcinoma (Oshimo et al., 2003), whereas in other cancer types, promoter CpG island hypermethylation causes the silence of tumour suppressor genes, such as p16/CDKN2 in bladder cancer (Gonzalez-Zulueta et al., 1995), VHL in renal carcinoma (Herman et al., 1994), APC gene in colorectal carcinoma (Hiltunen et al., 1997). A subset of colorectal cancer is characterised by promoter hypermethylation, termed CpG island methylator phenotype (CIMP) (Toyota et al., 1999). CIMP is mediated by BRAF mutations and associated with poor survival in colorectal cancers (Advani et al., 2018; Fang et al., 2014). CIMP was also studied in other cancer types. In clear cell renal cell carcinomas, CIMP is associated with more aggressive tumour phenotypes (Arai et al., 2012) while in breast cancer and gliomas, it is associated with better survival (Fang et al., 2011; Noushmehr et al., 2010). In summary, global loss of 5mC is associated with genome instability and promotes cancer whereas locally promoter methylation is associated with gene silencing but its effect on cancer is related to the functions of the specific gene being silenced.

1.2.1 DNA methyltransferases

In mammals, there are six DNA methyltransferases (DNMTs): DNMT1, DNMT2, DNMT3A, DNMT3B, DNMT3L and DNMT3C (Figure 1-2). DNMT1, DNMT3A, DNMT3B and DNMT3C catalyse the methylation of cytosine using S-adenosyl-methionine (SAM) as a methyl group donor with a base-flipping mechanism that flips

the substrate cytosine out from the DNA double-helix into the catalytic pocket (Cheng and Roberts, 2001; Klimasauskas et al., 1994; Song et al., 2012). DNMT2 (Okano et al., 1998b) and DNMT3L (Chedin et al., 2002) do not have the canonical DNA methylation activity. DNMT2 catalyses the methylation of tRNA (Goll et al., 2006). DNMT3L stimulates the activity of DNMT3A and DNMT3B (Chedin et al., 2002; Gowher et al., 2005; Suetake et al., 2004). DNMT1 was initially identified as a *de novo* methyltransferase to methylate unmethylated CpGs (Bestor et al., 1988). Later it was found to have higher methylation activity on hemimethylated CpGs than unmethylated ones (Yoder et al., 1997). The *de novo* methylation activity of DNMT1 is not worse than that of DNMT3A and DNMT3B but DNMT1 has much higher activity on hemimethylated CpGs in an *in vitro* methylation assay. In the assay, DNMT1 catalyse the incorporation of ^{3}H labelled methyl group onto unmethylated DNA at a similar rate as DNMT3A and DNMT3B, whereas on hemimethylated DNA, DNMT1 is more than twenty times more efficient than DNMT3A and DNMT3B (Okano et al., 1998a). Therefore, DNMT1 is mainly a maintenance methyltransferase that methylates hemimethylated CpGs. In contrast, DNMT3A and DNMT3B have similar activities on unmethylated and hemimethylated CpGs and are classified as *de novo* methyltransferases (Okano et al., 1998a). In addition to CpG sites, DNMT3A was also found to generate methylated CpA and CpT sites in embryonic stem cells (Ramsahoye et al., 2000). DNMT3C is found in rodents but not humans. DNMT3C is expressed specifically in male germ cells and silences the young transposons, preserving male fertility (Barau et al., 2016).

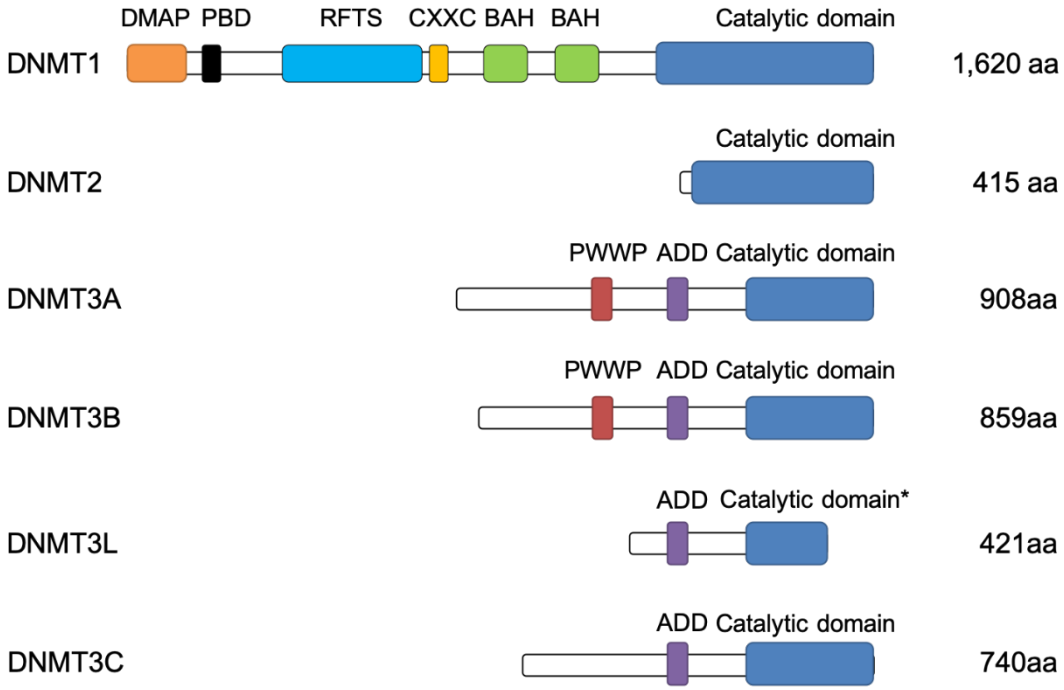


Figure 1-2 A schematic illustration of the DNMTs. DNMT proteins share a similar C terminal catalytic domain. * The catalytic domain of DNMT3L is truncated and do not have methyltransferase activity. DMAP, Dnmt1-associated protein binding domain; PBD, proliferating cell nuclear antigen (PCNA)-binding domain; RFTS, replication foci targeting sequence; CXXC, CXXC zinc finger domain; BAH, bromo-adjacent homology domain; PWWP, proline-tryptophan-tryptophan-proline motif; ADD, ATRX-DNMT3-DNMT3L domain.

1.2.2 5mC binding proteins

DNA modifications including 5mC act on gene expression in two ways: affecting the binding of transcription factors or attracting specific binding proteins. A systematic study of the effect of 5mC on DNA binding affinity of transcription factors showed that of the 350 transcription factors studied, binding of 117 transcription factors was inhibited by 5mC while binding of 175 was enhanced (Yin et al., 2017).

5mC is recognised and interpreted by specific binding proteins, which can be classified into three families based on structural similarities of methylation recognition domains: the methyl-CpG-binding domain (MBD), the Kaiso-like zinc finger domain, and the SET- and RING finger-associated (SRA) domain (Figure 1-3). Studies of 5mC binding proteins provide insights into the biological functions of 5mC and also serve as examples of how specific DNA modification binding proteins can be identified. The identification and characterisation of the three 5mC binding proteins families are discussed below.

1.2.2.1 MBD family proteins

MeCP2 is the first protein identified in the MBD family. It was detected as an 84 kDa band on SDS-PAGE to specifically bind DNA containing a single symmetrically methylated CpG and was purified by DNA affinity purification (Lewis et al., 1992). MeCP2 binds to methylated CpG and functions as a transcriptional repressor (Meehan et al., 1992; Nan et al., 1997). MeCP2 is part of the NCoR/SMRT co-repressor complexes (Lyst et al., 2013). MeCP2 is associated with a neurological disorder, Rett syndrome. MeCP2 was found to be mutated in 65% of Rett syndrome patients (Bienvenu et al., 2000). The phenotype of mice with *Mecp2* mutations resembles the syndrome of Rett patients with normal early development followed by cognitive regression (Chen et al., 2001; Guy et al., 2001). Restoration of *Mecp2* in the mouse model reverses the neurological defects (Guy et al., 2007; Robinson et al., 2012).

The DNA binding domain of MeCP2 was subsequently identified and named MBD (methyl-CpG-binding domain) (Nan et al., 1993). More MBD family proteins were identified based on sequence homology, including MBD1, MBD2, MBD3, MBD4,

(Hendrich and Bird, 1998), MDB5 and MBD6 (Roloff et al., 2003). MBD1 binds to methylated DNA at heterochromatin and represses transcription. MBD1 also has CXXC domains that bind unmethylated CpGs and can still be targeted to heterochromatic foci in the cells that are deprived of DNA methylation (Fujita et al., 1999; Jørgensen et al., 2004). However, later studies with genome-wide profiling of MBD1 binding sites showed that the targeting of MBD1 by its CXXC domain to the unmethylated region only happened in the absence of MBD domain (Baubec et al., 2013), indicating the genomic localisation of wild-type MBD1 in wild-type cells are predominantly directed by the MBD domain while the function of CXXC domain requires further exploration. Both MBD2 and MDB3 are components of the NuRD (Nucleosome Remodelling Deacetylase) complex, which functions as a transcriptional repressor by deacetylating histone (Feng and Zhang, 2001). MBD2 direct the NuRD complex to methylated DNA (Zhang et al., 1999) while MBD3 does not bind to methylated DNA due to alteration of two highly conserved amino acids in its MBD domain (Saito and Ishikawa, 2002). Recent studies also propose MDB3 as a 5hmC binding protein and will be discussed in more detail in later sections. MDB4 has a glycosylase domain in addition to the MBD domain and is proposed to repair the mismatch mutations at methylated CpG sites (Hendrich et al., 1999; Millar et al., 2002; Petronzelli et al., 2000). MBD5 and MBD6 do not bind methylated DNA *in vitro* and their genomic localisation in cells is independent of DNA methylation (Laget et al., 2010). Thus, MBD5 and MDB6 are unlikely to be 5mC binding proteins. In summary, the studies of MDB family proteins provide mechanistic understandings of the repression role of 5mC.

1.2.2.2 Kaiso-like zinc finger family proteins

The zinc finger domain-containing 5mC binding proteins include Kaiso, ZBTB4 and ZBTB38. Kaiso was detected by its ability to bind specifically to DNA probes with at least two methylated CpGs and was further purified with DNA affinity chromatography. Kaiso was showed to have methylation-dependent repressive activity (Prokhortchouk et al., 2001). In addition to methylated DNA, Kaiso was found to recognise unmethylated consensus sequence TCCTGCNA as well (Daniel et al., 2002). An analysis of Kaiso binding sites in the genome of various cell lines showed that it binds unmethylated regions and associated with active gene transcription, putting debate on the role of Kaiso as a 5mC binding protein (Blattler et al., 2013). ZBTB4 and ZBTB38 were identified based on sequence similarity to Kaiso and also bind to methylated DNA and repress expression (Filion et al., 2006). However, the functions of ZBTB4 and ZBTB38 remain poorly characterised. Whether Kaiso-like zinc finger family proteins are 5mC binding proteins or not requires further study.

1.2.2.3 SRA domain-containing proteins

UHRF1 was identified as a 5mC binding protein from methylated DNA pull-down followed by liquid chromatography-tandem mass spectrometry (LC-MS/MS) (Unoki et al., 2004). It was later found to have higher affinities for hemimethylated DNA than full methylated DNA and play a role in DNA methylation maintenance by recruiting DNMT1 to hemimethylated sites (Bostick et al., 2007; Sharif et al., 2007). The SRA domain of UHRF1 recognises 5mC using base-flipping mechanism, a mechanism that is also used by DNMTs (Arita et al., 2008). UHRF2 is the other protein that has the SRA domain, but it recognises 5hmC rather than 5mC (Zhou et al., 2014).

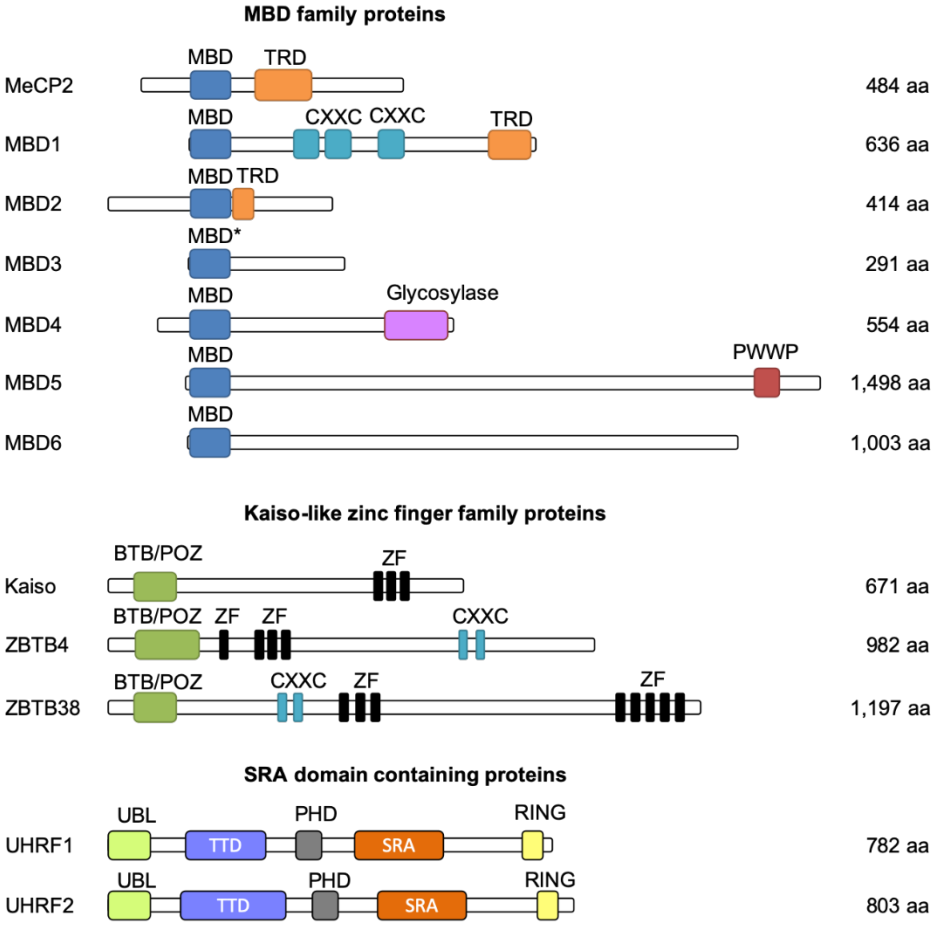


Figure 1-3 A schematic illustration of 5mC binding proteins. MBD family proteins share a similar N terminal MBD domain which recognises methylated CpG. *, there are alterations of two amino acids in the MBD domain of MBD3 so MBD3 does not bind to methylated DNA. Kaiso-like zinc finger family proteins have conserved BTB/POZ domain that is involved in protein-protein interactions and zinc finger domain that may bind 5mC. UHRF1 and UHRF2 have similar SRA domains. However, UHRF1 recognises hemimethylated CpG while UHRF2 binds 5hmC. MBD, methyl-CpG-binding domain; TRD, transcriptional repression domain; CXXC, CXXC zinc finger domain; PWWP, proline-tryptophan-tryptophan-proline motif; BTB/POZ, broad complex, tramtrack and bric a brac/poxvirus and zinc finger domain; ZF, zinc finger; UBL, ubiquitin-like domain; TTD, tandem tudor domain; PHD, plant homeodomain; SRA, SET, and RING finger associated domain; RING, really interesting new gene.

1.3 5hmC

5hmC was first detected in the mammalian genome in 1972 (Penn et al., 1972) but the results had not been reproduced by others (Kothari and Shankar, 1976) until 2009 when 5hmC was detected and confirmed in Purkinje cells, granule cells and mouse embryonic stem cells (Kriaucionis and Heintz, 2009; Tahiliani et al., 2009). Oxidation of 5mC by TET enzymes is the major pathway to generate 5hmC as lacking all three DNMTs or all three TETs results in undetectable amounts of 5hmC (Dawlaty et al., 2014; Ficz et al., 2011; Szwagierczak et al., 2010). However, there are reports of other possible ways to generate 5hmC. *In vitro*, DNMT enzyme was observed to catalyse the formation of 5hmC from unmodified cytosine when using aldehydes as cofactors and the reverse reaction when no cofactor was provided (Liutkevičiūtė et al., 2009). AlkB family enzymes were reported to oxidise 5mC to 5hmC, 5fC and 5caC *in vitro* (Bian et al., 2019). 5hmC can be generated from 5mC and also converted back to unmodified cytosine upon exposure to UV light. However, the condition used in the experiment was not physiological: free nucleotide instead of DNA was used and the UV exposure was for two hours (Privat and Sowers, 1996). Unlike 5mC, the level of which is similar among tissue types (Ehrlich et al., 1982; Fraga et al., 2002; Ito et al., 2011; Kuo et al., 1980; Tomkova et al., 2016; Wilson et al., 1986), the levels of 5hmC varies between different tissues, from more than 1% of all cytosines in human brain to about 0.1% in thymus, spleen and blood (Bachman et al., 2014; Chen et al., 2019a; Globisch et al., 2010; Liu et al., 2013b; Nestor et al., 2012; Szwagierczak et al., 2010; Tomkova et al., 2016).

1.3.1 TET proteins

TET proteins were discovered by a computational search for mammalian homology protein of known protozoa Fe(II) and α-ketoglutarate (α-KG) dependent dioxygenases (Iyer et al., 2009). Three human proteins TET1, TET2, TET3 were identified from the search and experimentally proved to be able to oxidise 5mC to 5hmC (Tahiliani et al., 2009). Subsequent studies found that TET enzymes can further oxidise 5hmC to 5fC and 5caC (He et al., 2011; Ito et al., 2010; Ito et al., 2011). The mammalian TET family proteins share a similar C-terminal catalytic domain including double-stranded β helix (DSBH) and a cysteine-rich region (Hu et al., 2013). TET1 and TET3 but not TET2 have N-terminal CXXC domains (Figure 1-4). The TET1 CXXC domain preferentially binds to unmethylated CpG, which is in line with its role in preventing aberrant methylation at CGIs rather than oxidising 5mC permissively (Jin et al., 2014; Xu et al., 2011; Zhang et al., 2010). The TET3 CXXC domain binds 5caC and targets TET3 to the transcription start sites (TSS) of lysosome-related genes and may regulate active removal of 5mC (Jin et al., 2016; Stroynowska-Czerwinska et al., 2018). TET2 has lost the CXXC domain during the evolution through chromosome inversion. The CXXC domain is now encoded by a separate gene, IDAX. IDAX binds to DNA containing unmethylated CpGs, interact with TET2 directly and potentially recruit TET2 to DNA. Besides, IDAX activates caspase, which leads to the degradation of TET2 and IDAX (Ko et al., 2013). However, the biological signification of IDAX requires further study.

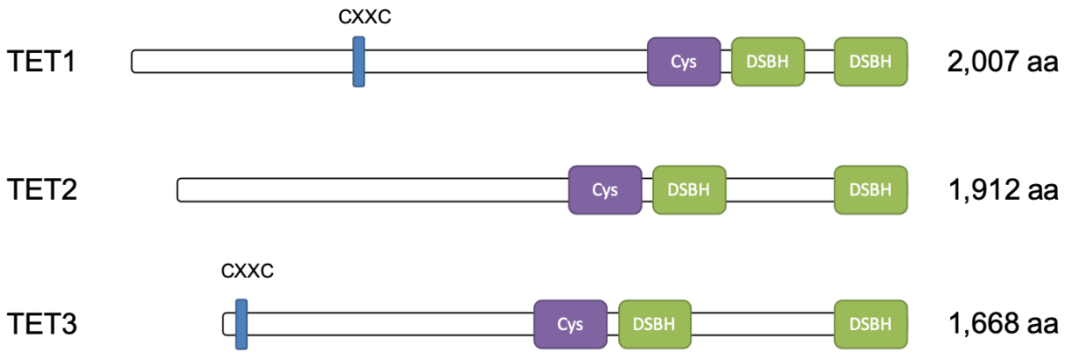


Figure 1-4 A schematic illustration of the TET proteins. TET proteins share conserved C terminal catalytic domain including a cysteine-rich region and DSBH. CXXC, CXXC zinc finger domain; Cys, cysteine-rich region; DSBH, double-stranded β-helix.

The crystal structure of TET2 catalytic domain (CD) in complex with DNA showed that TET2 specifically recognises CpG dinucleotide with the 5mC flipped out from the DNA double-helix and inserted into the catalytic pocket of TET2 (Hu et al., 2013). In line with this, *in vitro* biochemical studies showed that TET2CD has higher activity on 5mC in CpG than CpC or CpA context (Hu et al., 2013). Outside the CpG sites, TET2CD is only in contact with the phosphate groups of the DNA, indicating that TET2CD do not have sequence selectivity outside the CpG site (Hu et al., 2013). TET2CD has higher enzymatic activity on 5mC than 5hmC or 5fC but has similar affinities towards 5mC, 5hmC and 5fC. The crystal structure shows that the difference in catalytic activity is a result of different organisations of 5mC from 5hmC and 5fC within the catalytic pocket of TET2CD (Hu et al., 2015). The activity of TET2CD is not affected by the modification status of the cytosine on the opposite strand. It has similar activity to 5mC opposite to unmodified cytosine, 5mC, 5hmC, 5fC and 5caC (Crawford et al., 2016). There are two studies on the chemical processivity of TET2CD. One showed that TET2CD catalyses the multiple-step oxidation of 5mC all

the way from 5hmC, 5fC to 5caC within one enzyme-substrate encounter (Crawford et al., 2016) while the other study had contradictory observations that TET does not retain on the substrate after each oxidation step (Tamanaha et al., 2016). It is possible that TET2CD can work in both processive and non-processive ways depending on the reaction condition. After one oxidation, TET2CD is not more likely to oxidase nearby sites on the same piece of DNA. The same property was also observed for TET1CD and full-length *Naegleria gruberi* Tet-like oxygenase (NgTET) (Tamanaha et al., 2016). Biochemical studies on TET family proteins were mostly done with TET2CD. As TET family proteins have homologous oxygenase domains (Iyer et al., 2009), it is possible TET1 and TET3 have similar biochemical properties as TET2, but further experiments on other TET family proteins, both catalytic domain and full-length, are required to confirm this.

The TET family proteins have distinct functions. In mESCs, TET1 mainly regulates 5hmC levels at promoters whereas TET2 is primarily responsible for 5hmC levels at gene bodies and enhancers (Hon et al., 2014; Huang et al., 2014). Functionally, TET1 is involved in regulating the maintenance and lineage differentiation potential of mESCs (Ito et al., 2010; Koh et al., 2011) while the loss of TET2 results in delayed gene activation during differentiation (Hon et al., 2014). TET1 and TET2 are involved in the methylation reprogramming in primordial germ cells (PGCs) (Hackett et al., 2013; Yamaguchi et al., 2013). In brains, TET1 is related to memory and hippocampal neurogenesis (Rudenko et al., 2013; Zhang et al., 2013) while TET3 is related to neuron plasticity (Li et al., 2014; Yu et al., 2015). TET2 regulates self-renewal and differentiation of hematopoietic stem cells. *Tet2* deficiency in mice leads to increased

hematopoietic progenitor cells and myeloid malignancies (Ko et al., 2011; Li et al., 2011a; Quivoron et al., 2011). TET3 is involved in demethylation of the paternal pronucleus after fertilization (Gu et al., 2011; Iqbal et al., 2011; Wossidlo et al., 2011).

1.3.2 5hmC and DNA demethylation

The observation of genome-wide demethylation after fertilisation (Kafri et al., 1992; Monk et al., 1987; Santos et al., 2002) led people to explore the mechanism of demethylation. One mechanism is DNA replication-dependent passive demethylation through cell replication. DNMT1 maintains DNA methylation by methylating the hemimethylated CpGs generated after DNA replication (Bestor and Ingram, 1983; Gruenbaum et al., 1982; Stein et al., 1982; Wigler et al., 1981). Failing to methylate the newly synthesised DNA strand leads to the dilution of DNA methylation (Wu and Zhang, 2010). However, some observations cannot be explained by passive demethylation, for example, the demethylation of the paternal genome after fertilisation occurs before DNA replication (Mayer et al., 2000; Oswald et al., 2000), indicating the existence of active demethylation pathways. There are several active demethylation mechanisms proposed. MDB2 was the first cytosine demethylase reported (Bhattacharya et al., 1999). However, this observation is debatable as other studies reported MBD2 as a 5mC binding protein (Hendrich and Bird, 1998; Zhang et al., 1999) and MBD2 knockout mice have normal DNA methylation patterns (Hendrich et al., 2001). The excision of 5mC by TDG was also proposed to contribute to demethylation, but the activity of TDG on 5mC/G pair is about 30 times lower than that on T/G pair (Zhu et al., 2000). The deamination of 5mC by AID and APOBEC1 to thymine followed by the repair of the T/G mismatch is another proposed pathway for demethylation (Morgan et al., 2004). AID null mice have increased DNA

methylation level in PGCs compared to wild-type mice, supporting the role of AID in demethylation (Popp et al., 2010). Despite the increase, the methylation level in AID deficient PGCs is still lower than ESCs or somatic cells, indicating the involvement of other demethylation pathways (Wu and Zhang, 2010). The discovery of 5hmC and TET enzyme provides mechanistic insight into both passive and active DNA demethylation. 5hmC contributes to passive demethylation by affecting DNMT1 activity. DNMT1 has lower activity on hemi-hydroxymethylated substrates compared to hemimethylated ones, resulting in impaired DNA methylation maintenance and loss of methylation through DNA replication (Hashimoto et al., 2012; Ji et al., 2014; Otani et al., 2013; Seiler et al., 2018; Valinluck and Sowers, 2007). In the active demethylation pathway, 5mC is oxidized to 5hmC, 5fC and 5caC by TET (He et al., 2011; Ito et al., 2010; Ito et al., 2011; Tahiliani et al., 2009). 5fC and 5caC can be excised by TDG (He et al., 2011; Shen et al., 2013; Song et al., 2013). The abasic site generated by TDG is further repaired to unmodified cytosines by BER pathway (He et al., 2011; Maiti and Drohat, 2011; Weber et al., 2016). Apart from removing by TGD, 5fC was shown to be directly converted to unmodified cytosine through C-C bond cleavage in stem cells (Iwan et al., 2017). 5caC can go through direct decarboxylation catalysed by DNMT3A/DNMT3L complex *in vitro* (Liutkevičiūtė et al., 2014). The study of TET enzyme provides a mechanism for post-fertilisation demethylation, which is through TET3 mediated oxidation of 5mC followed by replication-dependent dilution (Gu et al., 2011; Inoue et al., 2011; Inoue and Zhang, 2011; Iqbal et al., 2011; Wossidlo et al., 2011). However, a recent study showed that the early loss of 5mC in mouse zygotes is independent of TET3 and 5hmC generation, indicating the involvement of other demethylation pathways (Amouroux et al., 2016).

In conclusion, 5hmC and TET mediated oxidation of 5mC play a major role in demethylation while AID/APOBCE1 and other unknown mechanisms may contribute as well.

1.3.3 Methods for genome-wide 5hmC mapping

Methods for genome-wide 5hmC mapping can be classified into two categories: enrichment methods and base-resolution methods. Enrichment methods use antibody (hmeDIP) (Nestor and Meehan, 2014) or selective chemical labelling (hMe-Seal) (Song et al., 2011) to specifically enrich 5hmC containing DNA fragments, but do not give absolute quantities or the positions of 5hmC within the fragments. The basic principle of base-resolution methods is to convert either modified or unmodified cytosine to thymine, making it distinguishable by sequencing. In conventional bisulfite sequencing, unmodified cytosine is converted into uracil which is replicated as T during PCR amplification while both 5mC and 5hmC are resistant to bisulfite conversion and read as C (Huang et al., 2010; Jin et al., 2010; Nestor et al., 2010). Modified bisulfite methods that distinguish 5mC and 5hmC include TAB-seq (Yu et al., 2012a; Yu et al., 2012b) and oxBS-seq (Booth et al., 2012; Booth et al., 2013). TAB-seq converts 5mC to T while 5hmC remains as C, allowing for the direct mapping of 5hmC. oxBS-seq maps 5mC directly. 5hmC information can be deduced by comparing mapping from bisulfite sequencing and oxBS-seq. Recently, a bisulfite-free method, APOBEC-coupled epigenetic sequencing (ACE-seq) has been developed. In ACE-seq 5hmC is unconverted and mapped as C while 5mC and unmodified cytosine are converted to T by enzymatic deamination, which is less destructive than the chemical conversion of bisulfite sequencing (Schutsky et al., 2018). Both bisulfite sequencing-based methods and ACE-seq convert all unmodified

cytosines, which account for around 95% of all cytosines. As a result, the sequence complexity of the genome is reduced, resulting in lower sequencing quality and mapping rate. Recently, several bisulfite free methods have been developed: hmC-CATCH (Zeng et al., 2018) and CAPS (Liu et al., 2019). These methods convert 5hmC to T, leaving 5mC and unmodified cytosine unchanged, thus solving the sequence complexity problem.

1.3.4 Genomic distribution of 5hmC

Base resolution mapping of 5hmC showed that at per base level, the abundance of 5hmC is lower than that of 5mC. In ESCs, the median abundance of 5hmC and 5mC is 20.9% and 59.0%, respectively (Yu et al., 2012b). Similarly, in brains, 5hmC occurs at a lower frequency than 5mC (30% and 60% respectively) (Wen et al., 2014). In embryonic stem cells, the majority of 5hmC is in CpG context (99.89% in hESCs and 98.7% in mESCs). Despite the symmetry of the CpG sequence, only 21% of 5hmC are symmetrical. There is a modest enrichment of local guanine content on the 5hmC containing strand around 50 base pairs upstream and downstream of 5hmC sites (Yu et al., 2012b). The non-symmetrical modification of 5hmC is observed in another study using single-molecular imaging, where 5hmC was observed to be in the form of adjacent and opposing hydroxymethylated and methylated CpG (H/M) sites, roughly accounting for 60% of 5hmC in various mouse tissues (cerebellum, kidney, heart and liver) and mESCs (Song et al., 2016). The function and formation of H/M sites require further study.

Genome-wide profiling of 5hmC has mainly been done in brain and embryonic stem cells. Genomic distribution of 5hmC shows correlations with genes and regulatory

elements which will be discussed in the below sections. Although the molecular mechanism awaits further study, the correlations provide insights into the biological functions of 5hmC.

1.3.4.1 Gene bodies

In both brains and ESCs, 5hmC is enriched at gene bodies, especially exons (Ficz et al., 2011; Jin et al., 2011; Pastor et al., 2011; Song et al., 2011; Stroud et al., 2011; Szulwach et al., 2011b; Wang et al., 2012; Williams et al., 2011; Wu et al., 2011; Xu et al., 2011). The enrichment of 5hmC at gene body is positively correlated with gene expression (Jin et al., 2011; Lister et al., 2013; Mellen et al., 2012; Song et al., 2011; Wen et al., 2014; Wu et al., 2011; Xu et al., 2011), whereas the gene body 5mC is negatively correlated with gene expression (Mellen et al., 2012). Additionally, in brains, a transcription related strand bias towards sense strand of 5hmC was observed while 5mC is biased towards antisense strand (Wen et al., 2014). It requires further study to understand if gene body 5hmC is a cause or result of active transcription. In human and mouse brain, 5hmC is enriched at the exon-intron boundary, indicating its potential function related to splicing (Khare et al., 2012; Wen et al., 2014). The observation that 5hmC level is lower at alternatively spliced exons than constitutive exons further supports the role of 5hmC in splicing (Khare et al., 2012).

1.3.4.2 Promoters

5hmC was found to be enriched at promoters of lowly expressed genes in mESCs (Wu et al., 2011; Xu et al., 2011) and hESCs (Stroud et al., 2011). In another study in mESCs, 5hmC high promoters were further classified into 5mC high and 5mC low. 5hmC high 5mC low genes have higher expression than average of all genes while the expression of 5hmC high 5mC high genes is lower than average (Ficz et al., 2011).

Taken these studies together, it is likely that 5hmC at promoters has a gene activating effect, but as 5mC and 5hmC often co-exist (Williams et al., 2011; Yu et al., 2012b), the effect of 5hmC is cancelled out by 5mC. More analysis that integrates 5hmC and 5mC information is needed to further elucidate this.

In mESCs, 5hmC is enriched at promoters with low CpG densities but depleted at high CpG density promoters (Williams et al., 2011; Wu et al., 2011; Xu et al., 2011; Yu et al., 2012b). This is in line with the observation that high CpG density promoter is depleted of 5mC, which is the precursor of 5hmC. In mESCs 5hmC is enriched at bivalent promoters (marked by both H3K27me3 and H3K4me3) of genes that are silenced in ESCs but activated upon differentiation (Ficz et al., 2011; Pastor et al., 2011). A comparison of 5hmC in brains from P7 and adult mice showed that 5hmC was high at the TSS of developmental activated genes (Szulwach et al., 2011b). The change of 5mC to 5hmC at these promoters may act as a pre-activating mark.

Unlike in ESCs, in brain 5hmC is depleted around TSS (Mellen et al., 2012; Song et al., 2011; Szulwach et al., 2011b; Wen et al., 2014), but enriched at around 1 kb upstream of TSS (Mellen et al., 2012; Song et al., 2011; Szulwach et al., 2011b). Antisense RNA is also enriched at upstream of TSS (Bornelöv et al., 2015). There may be an association between 5hmC and antisense RNA transcription. However, further studies are required to validate this hypothesis.

1.3.4.3 Distal regulatory regions

5hmC is enriched at enhancers marked by H3K4me1 in mESCs(Pastor et al., 2011; Wu et al., 2011; Yu et al., 2012b), human ES cells (Stroud et al., 2011; Szulwach et

al., 2011a), human brain (Wen et al., 2014) and mouse brain (Lister et al., 2013), indicating a connection of 5hmC and distal regulation elements. In mESCs, 5hmC is enriched at CTCF binding sites (Wu et al., 2011; Yu et al., 2012b). Base-resolution mapping of 5hmC showed that 5hmC is absent at the transcription factor binding sites but enriched at flanking regions; the enrichment of 5hmC is correlated with the depletion of 5mC (Stadler et al., 2011; Yu et al., 2012b). In summary, there is an association of transcription factor binding and local conversion of 5mC to 5hmC.

1.3.5 5hmC binding proteins

Proteins act as bridges between the modifications on DNA and the changes in gene expressions. 5hmC modification functions through either reducing the binding of certain protein factors or recruiting specific binding proteins. 5hmC reduces the binding of methyl-CpG binding proteins MBD1, MBD2 and MBD4 (Hashimoto et al., 2012; Jin et al., 2010). Studies on characterisation of 5hmC binding proteins can be classified into two categories: biased approaches where people study the 5hmC binding potential of known 5mC binding proteins including UHRF1 and MBD, unbiased approaches where 5hmC containing DNA probes were used for pull-down from nuclear protein extracts and the pull-down proteins were identified by LC-MS/MS.

1.3.5.1 5hmC binding potential of known 5mC binding proteins

UHRF1 is known to bind hemimethylated (M/C) DNA better than unmodified (C/C) or symmetrically methylated (M/M) DNA (Bostick et al., 2007; Sharif et al., 2007). There are several studies comparing UHRF1 binding affinity of hemimethylated (M/C) DNA to that of hemi-hydroxymethylated (H/C) or fully hydroxymethylated (H/H) DNA. Three studies found UHRF1 has higher affinity to M/C than H/C or H/H

(Hashimoto et al., 2012; Otani et al., 2013; Zhou et al., 2014) while one study found similar affinity of UHRF1 to M/C, H/C and H/H (Frauer et al., 2011). The study that showed UHRF1 had similar affinities to M/C and H/C also showed that it has similar affinity to M/M and M/C, which is controversial to previous reports. Taken together, UHRF1 is not likely to be an H/H or H/C specific binding protein.

The studies on MBD3 showed controversial results. Biochemically, one study showed that MBD3 is a 5hmC binding protein (Yildirim et al., 2011), while others did not find binding preference of MBD3 towards 5hmC (Cramer et al., 2014; Hashimoto et al., 2012; Iurlaro et al., 2013; Spruijt et al., 2013). In HeLa cells, MBD3 co-localised with 5fC and 5caC but not 5hmC, detected by an antibody-based assay (Cui and Irudayaraj, 2015). In mESCs, one study showed that genomic localisation of MBD3 was dependent on 5hmC and loss of MBD3 resulted in reduction of 5hmC level (Yildirim et al., 2011), but another study found that the localisation of MDB3 was not affected in *Dnmt1*, *Dnmt3a*, and *Dnmt3b* triple-knockout (TKO) mESCs (Baubec et al., 2013) but the comparison was only restricted to MDB3 binding in low-methylated regions. A later study re-analysing the data from Baubec et al. using all MDB3 binding regions found a reduction of MBD3 binding in the TKO mESCs. Besides, TET1 activity is required for MBD3 binding (Hainer et al., 2016). Taken together, MBD3 can be a 5hmC binding protein but its function requires further study.

1.3.5.2 Identification of 5hmC bind proteins through unbiased DNA pull-down

There are several studies looking for 5hmC binding proteins with DNA pull-down followed by LC-MS/MS (Iurlaro et al., 2013; Mellen et al., 2012; Spruijt et al., 2013; Takai et al., 2014; Xiong et al., 2016). In these studies, 5hmC, 5mC or unmodified

cytosine containing DNA probes were used for pull-down from nuclear protein extracts. The proteins enriched by the pull-down was identified by LC-MS/MS.

Most of the proteins identified in these studies have not been followed up with two exceptions, MeCP2 and UHRF2. MeCP2 was identified as a 5mC and 5hmC binding protein in brains. The binding of MeCP2 to 5hmC in the genome increases chromatin accessibility (Mellen et al., 2012). MeCP2 recognize 5hmC in CpA context (Gabel et al., 2015). A further structural and biochemical study confirmed its binding to 5hmCpA (Sperlazza et al., 2017). 5hmC in none CpG context including CpA context is rare in ESCs with about 0.1% of all 5hmC (Yu et al., 2012b). In line with the involvement of MeCP2 in brain function, the abundance of none CpG 5hmC is around 2.5%, which is higher than that of ESCs (Wen et al., 2014). UHRF2 was identified as a 5hmC binding protein in neuronal progenitor cells (Spruijt et al., 2013). The crystal structure of UHRF2 SRA domain with 5hmC containing DNA provided structural evidence for the binding (Booth et al., 2012). Another biochemical study showed that UHRF2 did not bind to 5hmC containing DNA specifically but had increased ubiquitination activity towards histone H3 peptides in the presence of 5hmC containing oligos (Vaughan et al., 2018). When co-expressed with TET1 in HEK293T cells, UHRF2 stimulated TET1 activity (Spruijt et al., 2013), but the effect of UHRF2 on TET1 activity was not observed in mouse embryonic fibroblast cells and HeLa cells (Chen et al., 2017). In *Uhrf2* null mice, there is a reduction of 5hmC and change in gene expression in brain and impaired brain function (Chen et al., 2017), supporting the role of UHRF2 as a 5hmC binding protein. In summary, both MeCP2 and UHRF2 are 5hmC binding proteins and are involved in brain functions.

1.3.6 5hmC and cancer

TET1 was initially discovered as part of a fused gene with MLL in 11q23 translocations in acute myeloid leukaemia (AML) patients (Lorsbach et al., 2003; Ono et al., 2002). Later it was found that the MLL-TET1 fuse protein directly upregulates the expression of the original TET1 and overexpression of TET1 plays an oncogenic role in the 11q23 translocations associated AML (Huang et al., 2013). In proneural glioblastoma, high levels of 5hmC and TET1 were observed and TET1-mediated production of 5hmC is required for glioblastoma genesis (Takai et al., 2014). TET1 is also found to be a tumour suppressor in other cancer types. Downregulation of TET1 was observed in hepatocellular carcinoma (Liu et al., 2013a), prostate cancer and breast cancer (Hsu et al., 2012) and are correlated with poor survival. In summary, TET1 can be either an oncogene or a tumour suppressor depending on the tissue type and cellular context.

Tet2 deficiency in mouse leads to increase of hematopoietic progenitor cells and stem cell self-renewal (Ko et al., 2011; Moran-Crusio et al., 2011; Quivoron et al., 2011). About one-third of *Tet2* null mouse developed myeloid malignancies within one year of age (Li et al., 2011b). The fact that not all *Tet2* knockout mice develop cancer indicates that additional mutations might be required for tumorigenesis. Chromosomal deletion involving *TET2* locus was found in AML (Viguié et al., 2005). TET2 mutation is frequently found in T-cell lymphomas (around 40%) (Lemonnier et al., 2012) and myeloid malignancies (from 6.05 to 27.36% in AML patients and from 6.05 to 36.23% in AML with normal cytogenetics) and associated with poor survival (Wang et al., 2019). 5hmC level is lower in genomic DNA from the bone marrow of

myeloid cancer patients with *TET2* mutation than that of healthy controls (Ko et al., 2010). Isocitrate dehydrogenase (IDH) mutation was also found in AML (Figueroa et al., 2010; Weissmann et al., 2011). The IDH mutation leads to a neomorphic enzymatic activity that generates an aberrant metabolite 2-hydroxyglutarate (2HG). 2HG inhibits TET activities by competing with α-KG (Figueroa et al., 2010). TET2 and IDH mutation rarely co-exists in AML (Weissmann et al., 2011). When TET2-mediated oxidation of 5mC to 5hmC is inhibited by 2HG, TET2 mutation is not required in cancer, indicating a role of low 5hmC levels in myeloid cancer tumorigenesis.

Cancer-associated *TET3* gene expression changes are often observed together with changes of other TET family genes. *TET1*, *TET2* and *TET3* are down-regulated in breast carcinoma, hepatocellular carcinoma (Yang et al., 2013) and melanoma (Lian et al., 2012). *TET1* and *TET3* are down-regulated in ovarian cancer (Ye et al., 2016) and head and neck cancer (Misawa et al., 2018). It is not clear if the downregulation of *TET3* is a cause of cancer or a result of global gene expression dysregulation in cancers.

1.4 Aim of the study

This thesis aims to understand the recognition, generation and detection of 5-hydroxymethylcytosine. In the mouse genome, around 60% of 5hmC is in H/M sites (Song et al., 2016). However, the function and the generation of H/M sites remain largely unknown. Previous studies looking for 5hmC binding proteins used symmetrically hydroxymethylated CpG sites as probes. The specific binding proteins

of H/M sites have not been identified so far. To understand the function of H/M sites, an unbiased search for H/M site binding proteins was performed with DNA pull-down followed by LC-MS/MS (**Chapter 3**). In **Chapter 4**, the potential of TET protein in generating H/M sites was explored. At the time when I started my project, there is no bisulfite-free method available for base-resolution mapping of 5hmC. Therefore, in **Chapter 5**, a 5hmC selective oxidation reaction (Okamoto et al., 2011) was applied to the base-resolution mapping of 5hmC. In collaboration with my colleague Fang Yuan, the use of this reaction was extended to the mapping of 5hmC and 5mC in RNA.

2 Chapter 2 Materials and Methods

2.1 Preparation of DNA and RNA

2.1.1 BDNF probes for DNA pull-down

PCR primer to generate BDNF probes:

Forward: Biotin-GCGTGAATTTGCTAGGACTGG

Reverse: GAATTACCAGAATCAGAATTCCG

BDNF probe was generated by PCR with Taq DNA polymerases using unmodified dCTP or 5-methyl-dCTP together with dATP, dTTP and dGTP from a plasmid containing mouse *Bdnf* gene promoter sequence. The probe is 120 bp long and modified at 5' end with biotin.

2.1.2 Probe1 for DNA pull-down and model DNA for TET reactions

Forward:

TCGATGTAGTGCGTCACXGGATGATAGCTGTACTCA

Reverse:

Biotin-TGAGTACAGCTATCATCXGGTGACGCACTACATCGA

X = C, 5mC or 5hmC

2.1.3 Probe2 for DNA pull-down

Forward:

CCCTCATCGCAGCTTTCCCAXGATGGCTGCXGATTAGCXGAGGTGCGCG

TTGGATGGAAGCCACTATCTCTCAAGGAAGTGCTTGCTCCT

Reverse:

Biotin-

AGGAGCAAGCACTTCCTTGAGAGATAGTGGCTTCCATCCAACGCGCACC

TXGGCTAATXGGCAGCCATXGTGGGAAAGCTGCGATGAGGG

X = C, 5mC or 5hmC

2.1.4 Probes for electrophoretic mobility shift assay

Forward:

TCGATGTAGTGCGTCACXGGATGATAGCTGTACTCA

Reverse:

FAM-TGAGTACAGCTATCATCXGGTGACGCACTACATCGA

X = C, 5mC or 5hmC

2.1.5 Non-specific DNA competitor

200 μL 300 ng/μL lambda DNA (dam–, dcm–) (Thermo Scientific™) was fragmented using an ultrasonic processor (SONICS® Vibra-Cell Processors VCX 130) with following parameters: 20% amplitude, 5 s on, 15 s off, 120 s on in total. The fragment length was around 500 bp.

2.1.6 Model DNA for peroxotungstate reactions

DNA1:

GCGGCGTGATACTGGTCCCGAG(hmC)CTGAAGTTAGGCC(hmC)GGGATG

ACTGA(hmC)AGTCTTCCGAGACCGACGACACAGGTCTCCCTATAGTGAG

TCGTATTA

DNA2:

CCCTTACCTACCACTTCGTGTATGTAGATAGGTAGTATA(hmC)AATTGAT

AT(hmC)GAAATGAGTA(hmC)GTAGATAGTAGAAAGTAAGATGGAGGTGA
GAGTGAGAGT

DNA3:

CCCTTACCTACCACTTCGTGATGTAGATAGAGTG(hmC)GTATTGATAGTA
(hmC)GAATGAT(hmC)GAGTAGAGAAGTT(hmC)GTAAGTAGAGAGGAGGT
GAGAGTGAGA GT

2.1.7 DNA Annealing

The annealing was performed in a reaction containing 10 μM forward oligo, 10 μM reverse oligo, 5 mM $MgCl_2$, 25 mM NaCl and 5 mM Tris-HCl pH 7.5. The reaction was incubated in a thermal cycler at 90˚C for 5 min followed by 60 cycles of 1˚C decrease per minute.

2.1.8 Model RNA in vitro transcription

DNA template

TTCCCTTACCTACCACTTCCATCACGTACTCATTTCGATATCAATTGTATA
CATACTCTCACTCTCACCTCCCTATAGTGAGTCGTATTA

The sequence that is complementary to the T7 promoter sequence is underlined.

The DNA template was annealed with a short oligo of the T7 promoter sequence. In vitro transcription was performed overnight with HiScribe™ T7 High Yield RNA Synthesis Kit (NEB) using unmodified CTP, 5-methyl-CTP or 5-hydroxymethyl-CTP together with ATP, TTP and GTP. 40 pmol DNA template was used for a 20 μL reaction. After the transcription step, 4 units of DNase I, reaction buffer and water was added to the reaction, making a final volume of 50 μL. The DNase I treatment was performed at 37˚C for 30 min. The RNA was purified with TRIzol reagent.

RNA sequence

GGGAGGUGAGAGUGAGAGUAUGUAUA(hmC)AAUUGAUAU(hmC)GAAA

UGAGUA(hmC)GUGAUGGAAGUGGUAGGUAAGGGAA

2.1.9 RNA purification with TRIzol Reagents

To purify the in-vitro transcribed RNA or RNA from cells, 500 μL TRIzol Reagent and 100 μL chloroform were added to 50 μL reaction or 2×10^7 cells. Samples were mixed, incubated at room temperature for 5 min and centrifuged at 4˚C 13,500 × g for 20 min. The upper (water) phase was mixed with an equal volume of isopropanol and incubated at -20 ˚C for 1 h. The RNA precipitate was washed twice with 400 μL ice-cold 75% ethanol, air-dried and dissolved in 50 μL nuclear-free water.

2.2 Plasmids

Table 2-1 Summary of plasmids used in this thesis

Name	Plasmid	Insert	Tag	Expression	Source
pNIC28-YBX1	pNIC28	Mouse YBX1	6 × His	E. coli	Cloned
pNIC28-ASH2L	pNIC28	Mouse ASH2L	6 × His	E. coli	Cloned
pNIC28-UHRF1	pNIC28	Mouse UHRF1	6 × His	E. coli	Cloned
pGTVL2-YTHDF3	pGTVL2	Mouse YTHDF3	6 × His-GST	E. coli	Cloned
pCDNA3-FLAG-Tet1	pCDNA3	Mouse TET1	FLAG	Mammalian	addgene #60938
pEF1a-FLAG-HA-Tet2	pEF1a	Mouse TET2	FLAG-HA	Mammalian	addgene #41710
pCDNA3-FLAG-mTet1CD	pCDNA3	Mouse TET1CD	FLAG	Mammalian	Cloned

Table 2-2 PCR primers and templates used for cloning

Primers	Sequences	Templates
Ybx1-LIC-F	TACTTCCAATCCatgagcagcgaggccgag	MGC Ybx1 cDNA Accession: BC061634 Clone ID: 6530605
Ybx1-LIC-R	TATCCACCTTTACTGTCActcagccccgccctg	
Ash2l-LIC-F	TACTTCCAATCCatggcggcggctgga	MGC Ash2l cDNA Accession: BC012957 Clone ID: 4211433
Ash2l-LIC-R	TATCCACCTTTACTGTCAgggttcccagggtggact	
Uhrf1-LIC-F	TACTTCCAATCCatgtggatccaggttcgaactatgg	cDNA from mESCs
Uhrf1-LIC-R	TATCCACCTTTACTGTCAccggccgctgccata	
Ythdf3-LIC-F	TACTTCCAATCCatgtcagccactagcgtg	cDNA from mESCs
Ythdf3-LIC-R	TATCCACCTTTACTGTCAttgcttgtttctatttctctcc	
KpnI-Flag-mTet1CD-F	GGGGTACCatggactacaaagacgatgacgacaaggaagctgcaccctgtga	cDNA from mESCs
BamHI-mTet1CD-R	CGGGATCCttagacccaacgattgtagg	

2.2.1 Ligation independent cloning

Ligation independent cloning (LIC) utilise the 3'- 5' exonuclease activity of T4 DNA polymerase to create overhangs. When supplemented with dGTP, the incorporation dGTP restrict of the exonuclease activity after the first complementary C residue. In LIC cloning, overlapping sequences were designed between vectors and inserts and joint together by the overhangs generated by the T4 DNA polymerase. The assembled fragment has 4 nicks which will be repaired by E. coli after transformation (Haun et al., 1992)

The CDS of target genes were prepared by PCR using Phusion High-Fidelity PCR Master Mix with HF Buffer (Thermo Scientific™, F531S) as detailed in Table 2-2. The PCR product was treated with DpnI to remove the plasmid used as the PCR

template. The vectors (pNIC28 or pGTVL2) were digested with BsaI. The PCR products and digested vectors were purified with Zymo-Spin IC Columns with Buffer PB (QIAGEN) and eluted in nuclease-free water. Both PCR products and vectors were treated with T4 polymerase to generate 5'-overhang. The reaction was performed in NEB buffer 2.1, 100 μg/ml BSA, 5 mM DTT, 2.5 mM dGTP (for vector) or 2.5 mM dCTP (for insert), 0.75 units T4 DNA polymerase (NEB, M0203S) in 10 μL reaction and incubated for 30 minutes at 22˚C, then 20 minutes at 75˚C in a thermocycler. The T4 polymerase treated insert and plasmid was mixed in a 1:3 molar ratio and incubated at room temperature for 30 min, transferred to ice and transformed to One Shot™ Mach1™ T1 Phage-Resistant Chemically Competent E. coli (Invitrogen™) following manufacture's instructions. The transformed E. coli were plated on LB agar plate containing 50 μg/mL kanamycin and 5% sucrose and incubated at 37˚C overnight. Single colonies were picked and cultured in 5 mL LB with 50 μg/mL kanamycin at 220 rpm 37˚C overnight. Miniprep was performed with Zyppy Plasmid Miniprep Kit (Zymo, D4019). The sequence of the insert was checked with Sanger sequencing.

2.3 Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and Western blot

2.3.1 SDS-PAGE

NuPAGE™ 4-12% Bis-Tris Protein Gels (1.0 mm) were run in Mini Gel Tank (Invitrogen) with NuPAGE™ MES SDS Running Buffer for 35 min at 200 V. For each gel 4 μL Precision Plus Protein™ Dual Color Standards was used. For

Coomassie stain, InstantBlue™ Protein Stain was used. SilverQuest™ Silver Staining Kit was used for silver stain following manufacture's instruction.

2.3.2 Western blot

After SDS-PAGE, proteins were transferred to PVDF membrane (Immobilon-P, IPVH00010, Millipore) in Mini Trans-Blot® Cell (Bio-Rad) filled with transfer buffer (10% methanol, 10 mM Tris, 100 mM Glycine, 0.005% SDS) at 180 mA for 60 min. The membrane was blocked in 5% milk in PBST (137 mM NaCl, 2.7 mM KCl, 4.3 mM Na_2HPO_4, 1.47 mM KH_2PO_4, 0.1% TWEEN® 20). After blocking the membrane was placed in a 50 mL Falcon tube with 5 mL primary antibody (MeCP2 (D4F3) XP® Rabbit mAb #3456, Cell Signaling Technology, 1: 000 dilution) in 5% milk PBST at 4˚C overnight with rotation. The membrane was washed three times with PBST and incubated with HRP-conjugated secondary antibody (Goat Anti-Rabbit IgG (H + L)-HRP Conjugate, Bio-Rad, 170-6515) in 5% milk PBST at room temperature for 2 h and washed three times with PBST. Clarity Western ECL Substrate (Bio-Rad) was used to detect the HPR activity. Gels were imaged with ChemiDoc gel imaging system (Bio-Rad).

2.4 293T cell culture and transfection

HEK 293T cells were maintained in DMEM (Gibco), supplemented with 10 % fetal serum (Sigma), 100 U/mL penicillin, and 100 mg/mL streptomycin (Gibco) at 37 ˚C under 5 % CO2 atmosphere. The coding sequence of mouse Tet1 catalytic domain (amino acid 1367-2039) was cloned into pcDNA3 vector and transfected into 293T cells using polyethylenimine. Cells were harvested 40 hours after transfection.

Genomic DNA was extracted from the cultured cells using TRIzol Reagent (Invitrogen).

2.5 Mouse embryonic stem cell culture

Mouse embryonic stem cells (mESCs) E14 were cultured on gelatine-coated plates in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 15 % FBS (Biosera FB-1001/500, Lot 014BS386), 2 mM L-glutamine, 1 × non-essential amino acids (Gibco 11140), 1 × penicillin-streptavidin (Gibco 15140), 0.1 mM β-mercaptoethanol, 10 ng/mL LIF, 1 μM PD0325901, and 3 μM CHIR99021 at 37 ˚C under 5% CO_2.

2.6 Nuclear protein extraction from mESCs

Two methods were compared for nuclear protein extraction. The buffers used are listed in Table 2-3. Method 2 (adapted from (Xiong et al., 2016)) was used for preparing samples for the DNA pull-down experiment. Cells were trypsinised and washed twice with PBS. 6×10^7 cells were incubated in 600 μL cytoplasmic protein extraction buffer for 10 minutes on ice and centrifuged at 6000 × g for 5 minutes. The pellet was washed twice, each time with 300 μL wash buffer, incubated in nuclear protein extraction buffer at 4 ˚C with gentle rotation (5 rpm) for 1 hour and centrifuged at 16,000 × g 4˚C for 20 min. The concentration of nuclear protein extract was determined with Qubit™ Protein Assay Kit.

Table 2-3 Buffers for nuclear protein extraction

	Method 1	Method 2
Cytoplasmic protein extraction buffer	5 mM PIPES (pH 8.0) 85 mM KCl 0.5 % NP-40 1 mM DTT	20 mM Tris-HCl (pH 8.0) 10 mM KCl 1.5 mM MgCl2 0.5 mM DTT 10% glycerol
Wash buffer	20 mM HEPES (pH 7.5) 10 mM KCl 1 mM MgCl2 0.1 % Triton X-100 1 mM DTT	20 mM Tris-HCl (pH 8.0) 10 mM KCl 1.5 mM MgCl2 0.5 mM DTT 10% glycerol
Nuclear protein extraction buffer	20 mM HEPES (pH 7.5) 10 mM KCl 300 mM NaCl 1 mM MgCl2 0.1 % Triton X-100 1 mM DTT	20 mM Tris-HCl (pH 8.0) 0.42 M NaCl 0.5 mM DTT 10% glycerol

2.7 Nuclear protein extraction from mouse brain

Each mouse brain (Seralab, MSE-BRAIN-C57) was homogenised in 5 mL homogenisation buffer (0.25 M sucrose, 150 mM KCl, 5 mM $MgCl_2$, 20 mM tricine-KOH pH 7.8, 0.15 mM spermine, 0.5 mM spermidine, cOmplete™ Protease Inhibitor Cocktail) and mixed with an equal volume of 50% iodixanol. The homogenate was

layered onto a 2-mL cushion of 28% iodixanol and centrifuged at 10,000 × g for 30 min to pellet the nuclei. The nuclei were resuspended in 5 volumes of buffer containing 20 mM HEPES pH 7.9, 150 mM KCl, 1 mM EDTA, 0.5 mM DTT, Halt™ Protease and Phosphatase Inhibitor Cocktail. Nuclear proteins were extracted by adding 3 M KCl to a final concentration of 400 mM incubating at 4˚C for 1h and centrifuging at 16,000 × g 4˚C for 10 min. The protein concentration was measured with Qubit™ Protein Assay Kit.

2.8 DNA pull-down

60 μmol of 5'-biotinylated C/C, M/M or H/M DNA probes were immobilised on 30 μL Dynabeads M-280 Streptavidin (Invitrogen) beads by incubating the probes with beads in Bind & Wash Buffer (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl, 0.2% TWEEN® 20) for 1 h at room temperature with gentle rotation (5 rpm) and washed twice with Bind & Wash Buffer, twice with H_2O. The probe-coupled beads were incubated with 300 μg of mESC nuclear extract or 120 μg mouse brain nuclear extract in pull-down buffer (20 mM HEPES pH 7.9, 150 mM KCl, 0.5% Triton X-100, 1 mM EDTA, 0.5 mM DTT, Halt™ Protease and Phosphatase Inhibitor Cocktail) in the presence of non-specific DNA binding competitors (21 μg sonicated lambda phage DNA fragments) at 4˚C for 2.5h. After that, the beads were washed six times with pull-down buffer, and the proteins were eluted by heating at 95℃ for 5 minutes in Laemmli sample buffer.

2.9 Protein expression and purification

2.9.1 Protein expression in E. coli

Stater culture was prepared by adding 1 μL glycerol stock of BL21-CodonPlus (DE3)-RIL E. coli containing pNIC28-YBX1, pNIC28-ASH2L, pNIC28-UHRF1 or pGTVL2-YTHDF3 plasmids to 10 mL LB with 50 μg/mL kanamycin and incubate at 37˚C overnight and 220 rpm. 10 mL of overnight starter culture was inoculated into 800 mL LB containing 50 μg/mL kanamycin, incubate at 37 ˚C and 200 rpm for 3 to 4 hours until OD600 was between 0.8 to 1.0. Add IPTG to a final concentration of 0.4 mM, incubate at 18˚C and 180 rpm overnight. The cells were pelleted by centrifuging 4000 g at 4˚C for 20 min.

2.9.2 Purification of ASH2L, UHRF1, YTHDF3 from E coli.

Cell pellets were re-suspended in lysis buffer (20 mM HEPES pH 7.5, 300 mM NaCl, 10 mM imidazole, 1 mM DTT, 1 mM PMSF) and homogenised with a high-pressure homogeniser and clarified by centrifuging 35,000 × g at 4˚C for 1h. The supernatant was incubated with Ni-NTA resin for 1 h at 4 ˚C with gentle shaking. The resin-lysate mixture was loaded onto a gravity chromatography column (Econo-Column® Chromatography Columns, #7372522) and drained. Then the resin was washed twice with 20 mL wash buffer (20 mM HEPES pH 8.0, 300 mM NaCl, 40 mM imidazole). The protein was eluted with 8 × 2 mL of elution buffer (20 mM HEPES pH 8.0, 300 mM NaCl, 500 mM imidazole). 10 μL of flow-though, each wash and each elution was separated on SDS-PAGE and stained with InstantBlue™ Coomassie Protein Stain. Fractions enriched with target proteins were pooled and concentrated with 30 KDa Amicon® Ultra-4 Centrifugal Filter Unit to less than 5 mL. HiLoad 16/60

Superdex 200 gel filtration column was equilibriumed with SEC buffer (20 mM HEPES pH 8.0, 500 mM NaCl, 1 mM DTT, 0.22 uM filtered) and the elusions from Ni-NTA resin were run on the column with a flow rate of 1 mL/min. Fractions are collected. The fractions enriched with target proteins were pooled and dialysed to pull-down buffer (20 mM HEPES pH 8.0, 150 mM KCl, 1 mM EDTA, 0.5 mM DTT) and concentrated with 30 KDa Amicon® Ultra-4 Centrifugal Filter Unit. Glycerol was added to a final concentration of 20%. Protein concentration was determined by nanodrop. The final protein was aliquoted, froze in liquid nitrogen and stored at -80˚C.

2.9.3 Purification of YBX1 from E. coli

The purification of YBX1 was modified from a previous paper (Selivanova et al., 2010). Cell pellets were re-suspended in lysis buffer (20 mM Tris-HCl pH 7.6, 2 M NaCl, 0.5 mM PMSF) and sonicated with an ultrasonic processor (SONICS® Vibra-Cell Processors VCX 750) using following parameters: 40% amplitude, 15 s on, 30 s off for 60 cycles and clarified by centrifuging 35,000 × g at 4˚C for 1h. The supernatant was diluted 4 times with 20 mM Tris-HCl pH 7.6 and incubated with Ni-NTA resin for 1 h at 4 ˚C with gentle shaking. The resin-lysate mixture was loaded onto a gravity chromatography column (Econo-Pac® Chromatography Column, Bio-rad, 7321010) and drained. Then the resin was washed wash buffer (20 mM Tris-HCl pH 7.6, 500 mM NaCl, 20 mM imidazole) and treated with Benzonase® Endonuclease (E1014-5KU, Millipore) in 1.5 mL $MgCl_2$, 20 mM Tris-HCl pH 8.0, 150 mM NaCl at 4˚C for 30 min. The protein was washed three times with wash buffer and eluted with 5 × 2 mL of elution buffer (20 mM Tris-HCl pH 7.6, 500 mM NaCl, 250 mM imidazole). 10 μL of flow-though, each wash and each elution was separated on SDS-

PAGE and stained with InstantBlue™ Coomassie Protein Stain. Fractions enriched with target proteins were diluted six times with Buffer A (20 mM Tris-HCl pH 7.6, 450 mM NaCl) and loaded onto HiTrap Heparin HP affinity columns (17040601, GE Healthcare). The target protein was eluted with an increasing concentration of NaCl from 450 mM to 2 M and fractions were collected. The fractions enriched with target proteins were dialysed in pull-down buffer (20 mM HEPES pH 8.0, 150 mM KCl, 1 mM EDTA, 0.5 mM DTT) and concentrated with 30 KDa Amicon® Ultra-4 Centrifugal Filter Unit. Glycerol was added to a final concentration of 20%. Protein concentration was determined by comparing with BSB control on Coomassie-stained SDS-PAGE gel. The final protein was aliquoted, froze in liquid nitrogen and stored at -80 ˚C.

2.9.4 Protein expression in Expi293F™ Cells

The plasmids were prepared with NucleoBond® Xtra Maxi Plus EF kit (MACHEREY-NAGEL). Expi293F™ Cells were seeded at 5×10^5/mL in 1 L FreeStyle™ 293 Expression Medium and cultured in a 2 L roller bottle at 37°C, 170 rpm, 5% CO_2 for around 24 h until cells density reached 1×10^6 cells/mL. The transfection reagent was prepared by diluting 1 mg plasmid in 100 mL PBS, mixing with 3 mL 1mg/mL polyethylenimine (Polysciences, 23966-1), vortexing for 15 s and incubating at room temperature for 20 min. The transfection reagent was added to the culture. After transfection, the cells were cultured at 37°C, 170 rpm 5% CO_2 for 48h. Cells were collected by centrifugation at 3000 × g, 4˚C for 20 min and stored at -80°C.

2.9.5 Protein purification from Expi293F™ Cells

Cells from 1 L culture were lysed by re-suspending in 100 mL lysis buffer (50 mM Tris-HCl pH 7.5, 500 mM NaCl, cOmplete™ Protease Inhibitor Cocktail, 1% Triton X-100) and incubating on ice for 20 min. The lysate was clarified by centrifuging at 30,000 × g, 4°C for 30 min. The supernatant was incubated with 1.2 mL ANTI-FLAG® M2 Affinity Gel (Sigma, A2220) at 4˚C for 30 min. The gel was load into an Econo-Column® chromatography column (Bio-Rad, 7372522) and drained. The gel was then washed with 12 mL TBS (50 mM Tris-HCl pH 7.5, 150 mM NaCl) for three times. The protein was eluted by adding 1.2 mL elution buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.1 mg/mL 3 × FLAG peptide, cOmplete™ Protease Inhibitor Cocktail) and incubating for 5 min. The elution was done five times. Eluted proteins were concentrated with 30 kDa Amicon® Ultra 0.5 mL Centrifugal Filters. The concentrated protein was pass through a Bio-Spin® P-30 Gel Column (Bio-Rad) with storage buffer (20 mM HEPES pH 7.4, 50 mM NaCl, 30% Glycerol) to exchange buffer and remove 3 × FLAG peptides. The proteins were aliquoted, froze in liquid nitrogen and stored at -80˚C.

2.10 Electrophoretic mobility shift assay (EMSA)

The polyacrylamide gel (5% acrylamide, bisacrylamide (37.5: 1), 90 mM Tris-borate, 2 mM EDTA, 0.1% ammonium persulfate, 0.08% tetramethylethylenediamine) was casted one day before the experiment and stored at 4˚C. The gel was pre-run in 0.5 × TBE (45 mM Tris-borate, 1 mM EDTA) at 4˚C 100 V for 30 min. Increasing amounts of proteins were incubated in pull-down buffer (20 mM HEPES pH 8.0, 150 mM KCl,

1 mM EDTA, 0.5 mM DTT) in 10 μL for 10 min at room temperature. 1 μL 50 nM DNA probe was adding. The reaction was mixed with a pipette and incubated at room temperature for 25 min. 2.2 μL of 6 × loading buffer (20 mM HEPES pH 8.0, 150 mM KCl, 1 mM EDTA, 0.5 mM DTT, 15% Ficoll® 400) was added. The samples were loaded on the gel and run at 4˚C 100 V for 1 h. The gel was imaged with a laser scanner. Quantification of bands was done with ImageJ.

2.11 TET reaction

2.11.1 Reaction

The DNA substrate was generated by annealing oligos. To remove the un-annealed single-strand DNA, 0.1 nmol DNA was treated with 20 U Exonuclease I (NEB, M0293S) at 30˚C for 45 min and purified with Oligo Clean & Concentrator (Zymo, D4060). 5 pmol DNA was incubated with TET1 or TET2 in a 50 μl reaction containing 50 mM HEPES pH 8.0, 100 μM ammonium iron(II) sulfate, 1 mM α-ketoglutarate, 2 mM ascorbic acid, 1 mM DTT, 100 mM NaCl, 1.2 mM ATP for the indicated amount of time. The reaction was terminated by adding 1 μL 0.5 M EDTA and incubating at 65˚C for 20 min.

2.11.2 Separating strand

After the reaction, 1.5 μL Dynabeads® MyOne™ Streptavidin C1 beads with 50 μL 2 × Bind & Wash Buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 2 M NaCl, 0.4% TWEEN® 20) was added to the reaction and incubated in at room temperature for 1 h with gentle rotation. The beads were washed with 100 μL Bind & Wash Buffer (5 mM Tris-HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl, 0.2% TWEEN® 20) for three times, each

time for 5 min. The beads were moved to a new tube and washed with 80 μL Bind & Wash Buffer for 5 min twice, followed by a wash with 20 μL TE buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA) for 5 min twice and with 20 μL 10 mM Tris-HCl pH 8.0 twice. After washing, the beads were treated with 10 U T7 Exonuclease in a 10 μL reaction with NEBuffer™ 4 at 25˚C for 1 h. The non-biotin strands were digested into nucleotides and remaining in the solution. For the biotin labelled strand, the beads were washed with 50 μL Bind & Wash Buffer for 5 min twice. The beads were moved to a new tube and washed with 50 μL Bind & Wash Buffer for 5 min twice, followed by a wash with 20 μL TE buffer (10 mM Tris-HCl, pH 8.0, 1 mM EDTA) for 5 min twice and with 20 μL 10 mM Tris-HCl pH 8.0 twice. The beads were re-suspended in 5 μL.

2.11.3 DNA hydrolysis for LC-MS/MS

To be quantified by LC-MS/MS, DNA samples were hydrolysed into nucleosides. The supernatant containing non-biotin strand was treated in a 50 μL reaction with 3 mU Phosphodiesterase I (Sigma, P3243-1VL), 1.2 U alkaline phosphatase (Sigma, P7923-10KU), 50 mM Tris-HCl pH 9.0, 10 mM $MgCl_2$, 10 μg/mL erythro-9-(2-hydroxy-3-nonyl)adenine (EHNA), 3.5 μM desferrioxamine mesilate (DFOM) at 37˚C for 4 h. The biotin-labelled strand on the beads was hydrolysed with 2.5 U DNA Degradase Plus (Zymo, E2020) supplemented with 10 μg/mL EHNA and 3.5 μM DFOM in a 20 μL at 37˚C for 4 h. Heavy labelling internal standards of deoxycytidine, 5-methyl-2'-deoxycytidine, 5-formyl-2'-deoxycytidine, 5-carboxy-2'-deoxycytidine were added into each sample to a final concentration of 12.5 nM each. Heavy labelled

5-formyl-2'-deoxycytidine and 5-carboxy-2'-deoxycytidine were gifts from Prof. Thomas Carell's group.

2.11.4 LC-MS/MS

The LC-MS/MS analysis was done with 1290 Infinity LC Systems (Agilent) coupled with a 6495B Triple Quadrupole Mass Spectrometer (Agilent). A ZORBAX Eclipse Plus C18 column (2.1 x 150 mm, 1.8-Micron, Agilent) was used. The column temperature was maintained at 40˚C, and the solvent system was 10 mM ammonium acetate pH 6.0 aqueous solution (Solvent A) and 100% methanol (Solvent B) with 0.4 ml/min flow rate. The gradient was: 0-5 min; 0 solvent B; 5-8 min; 0-5.63% solvent B; 8-9 min; 5.63% solvent B; 9-16 min; 5.63-13.66% solvent B; 16-17 min; 13.66-100% solvent B; 17-21 min; 100% solvent B; 21-24.3 min; 100-0% solvent B; 24.3-25 min; 0% solvent B. The dynamic multiple reaction monitoring mode (dMRM) of the MS was used for quantification. The source-dependent parameters were as follows: gas temperature 230˚C, gas flow 14 L/min, nebulizer 40 psi, sheath gas temperature 400 ˚C, sheath gas flow 11 L/min, capillary voltage 1500V in the positive ion mode, nozzle voltage 0V, high-pressure RF 110V and low-pressure RF 80V, both in the positive ion mode. The fragmentor voltage was 380 V for all compounds, while other compound-dependent parameters were summarised in Table 2-4. The quantification was performed with the calibration curves from nucleoside standards in the same run taken the heavy labelled internal standard into consideration.

Table 2-4 MS/MS parameters used for nucleosides quantification

RT: retention time, CE: collision energy; CAE: cell accelerator voltage. All the nucleosides were analysed in the positive mode.

Compound	Precursor Ion (m/z)	Product Ion (m/z)	RT (min)	Delta RT(min)	CE (V)	CAE (V)
mdC+H	242	126	9.05	1.5	10	4
mdC+H+3	245	129	9.05	1.5	10	4
hmdC+H	258	142	4.34	2	12	4
hmdC+H+3	261	145	4.34	2	12	4
fdC+Na	278	162	10.69	2	8	4
fdC+Na+2	280	164	10.69	2	8	4
cadC+Na	294	178	1.75	3	12	4
cadC+Na+2	296	180	1.75	3	12	4

2.12 Peroxotungstate oxidation reaction

2.12.1 Preparation of peroxotungstate

The preparation of dinuclear peroxotungstate ($K_2[[W(O)(O_2)_2(H_2O)]2(\mu\text{-}O)]\cdot 2H_2O$) was based on the reported procedure (Campbell et al., 1989). A solution of potassium tungstate oxide (Alfa Aesar, 14031) (K_2WO_4, 2.0 g) in water (5 mL) was treated with 30% H_2O_2 (10 mL). The yellow solution formed was treated with 1M HCl until just turning colourless (~500 μL). White crystals appeared on standing at 4℃ for 24h and were filtered off, washed with ethanol and air-dried.

2.12.2 Peroxotungstate oxidation reaction on DNA

The oxidation of single-strand DNA was performed as reported (Okamoto et al., 2011) by incubating 15ng DNA in 20μl of 50mM phosphate buffer (pH 7.0), 100mM NaCl and 5 mM peroxotungstate at 40℃ overnight. Double-stranded DNA was firstly denatured by incubation in water at 95℃ or incubation in 0.15M NaOH at room temperature for 10 minutes. After denaturing, the oxidation reaction was performed

in several modified single-strand reaction conditions. The condition that gave the highest conversion rate was to first denature 10 ng DNA in 1ml 0.15M NaOH at room temperature for 10 minutes, then acetic acid was added to adjust the pH to 7 and peroxotungstate was added to a final concentration of 5 mM. The oxidation reaction was incubated at 40℃ overnight. After the oxidation reaction, DNA fragments were PCR amplified and further analysed by restriction enzyme digestion or T/A cloning and Sanger sequencing.

2.12.3 Peroxotungstate oxidation reaction on RNA

The reaction was performed with 750 ng RNA, 5 mM peroxotungstate, 200 mM phosphate buffer pH 7.0, 40 U RiboLock RNase Inhibitor (Thermo Scientific™, EO0381) in a 50 μL reaction at 60˚C for 4 h. After the reaction, the RNA was purified with RNA Clean & Concentrator-5 (Zymo, R1015) following manufacturer's instructions. The purified RNA was oxidised purified again under the same condition.

2.13 TET oxidation of RNA

The TET oxidation was performed in a 20 μL reaction containing 50 mM MOPS buffer pH 6.9, 100 μM ammonium iron(II) sulfate, 1 mM α-ketoglutarate, 2 mM L-ascorbic acid, 1 mM DTT, 50 mM NaCl, 40 U RiboLock RNase Inhibitor (Thermo Scientific™, EO0381) and 10 μM NgTET1 at 37˚C for 1 h. The reaction was terminated by adding 0.8 U Proteinase K (NEB, P8107S) and incubating at 50˚C for 30 min. The product was purified with RNA Clean & Concentrator-5 (Zymo, R1015). To achieve a higher conversion rate, the reaction was performed twice. After TET oxidation, the RNA was treated with 250 mM sodium borohydride solution at room

temperature in the dark to reduce the over-oxidized 5fC to 5hmC. The reaction was quenched by sodium acetate pH 5.2 and purified with RNA Clean & Concentrator-5 (Zymo, R1015).

2.14 Reverse transcription

100 ng template RNA was mixed with 250 nM primer (TTC CCT TAC CTA CCA CTT CC), 10 mM DTT, 20 U RiboLock RNase Inhibitor (Thermo Scientific™, EO0381), 400 U TGIRT™-III Enzyme (InGex), 450 mM NaCl, 5 mM $MgCl_2$, 20 mM Tris-HCl pH 7.5 at room temperature for 30 min. 1.25 mM dNTPs were then added to the solution and incubated at 50˚C followed by 1˚C increase every minute until the temperature reached 60˚C and incubated at 60˚C for 20 min. The reaction was terminated by adding 5 M NaOH to a final concentration of 0.25 M and incubating at 95 ˚C for 3 min. The solution was neutralised with an equal amount of 5 M HCl. cDNA was purified with Micro Bio-Spin 6 column (Bio-Rad) followed by Oligo Clean & Concentrator (Zymo, D4060). The cDNA was further PCR amplified with Phusion High-Fidelity PCR Master Mix with HF Buffer (Thermo Scientific™, F531S).

2.15 Quantification of conversation rate

2.15.1 Restriction enzyme digestion method

The PCR product was treated with restriction enzyme for 30 min. After the digestion, DNA fragments were separated on E-Gel™ EX Agarose Gels, 2% (Invitrogen™) imaged by Chemidoc (Bio-rad). The quantification of band densities was done with Image Lab Software (Bio-rad).

2.15.2 TA cloning method

The PCR fragments were inserted into pCR™ 2.1-TOPO® TA vector using TOPO™ TA Cloning™ Kit (Invitrogen™, 450641) following manufactures instructions. The plasmids were transformed into to One Shot™ Mach1™ T1 Phage-Resistant Chemically Competent E. coli (Invitrogen™) following manufactures instruction. Single colonies were picked and cultured in 5 mL LB with 100 μg/mL ampicillin at 220 rpm 37˚C overnight. Plasmids were purified with Zyppy Plasmid Miniprep Kit (Zymo, D4019). Sanger sequencing was performed on the plasmid. The conversion rate was calculated by summarising the sequencing result of all plasmids.

2.16 Mass spectrometry analysis of DNA fragments

DNA was desalted with Micro Bio-Spin® 6 chromatography columns (BioRad) and serially diluted with matrix (9 mg/mL 2′,4′,6′-Trihydroxyacetophenone monohydrate and 5.5 mg/mL ammonium citrate in 50% acetonitrile 50% water). 1 μL DNA-matrix mix was put on the silver plate and air-dried. The samples were analysed with Voyager-DE MALDI-TOF (matrix-assisted laser desorption ionization time-of-flight) Biospectrometry Workstation in the negative mode.

2.17 Enzymatic DNA hydrolysis

2.17.1 PDE1 and Benzonase + PDE1 methods

1μg DNA was hydrolysed in a 50 μL reaction of 45 mM NaCl, 9 mM $MgCl_2$, 9 mM Tris-HCl pH 7.9, 5 mU Phosphodiesterase I (Sigma, P3243-1VL), 1.2 U alkaline phosphatase (Sigma, P7923-10KU), 10 μg/mL EHNA and 3.5 μM DFOM at 37 ^{o}C

for 2 hours. For the Benzonase + PDE1 method, DNA was incubated in the same reaction with the addition of 25 U Benzonase Nuclease (Sigma-Aldrich).

2.17.2 DNase I + PDE1 methods

1μg DNA was digested in a 20 μL reaction containing 5 U DNase I, DNase I Reaction Buffer (NEB), 2 mU Phosphodiesterase I, 1.2 U alkaline phosphatase (Sigma, P7923-10KU), 10 μg/mL EHNA and 3.5 μM DFOM at 37 ^{o}C for 2 hours.

2.17.3 DNA Degradase Plus methods

1μg DNA was treated in a 20 μL reaction with 5 U DNA Degradase Plus (Zymo), 10 μg/mL EHNA and 3.5 μM DFOM in DNA Degradase Reaction buffer (Zymo) at 37˚C for 2h.

2.17.4 Nucleoside digestion mix methods

1μg DNA was digested in a 20 μL reaction with 1μL Nucleoside digestion mix (NEB), 50 mM sodium acetate pH 5.4, 1 mM $ZnCl_2$, 10 μg/mL EHNA and 3.5 μM DFOM in at 37˚C for 2h.

2.17.5 Nuclease P1 methods

DNA was hydrolysed in two steps. In the first step, 1μg DNA was treated with 2U P1 nuclease, 10 μg/mL EHNA, 3.5 μM DFOM in 100 mM ammonium acetate buffer pH 5.0 at 37˚C overnight. After that, 15 μL of H2O, 4.3 μL of 10 × alkaline phosphatase buffer (50 mM Tris-HCl pH 9.0, 10 mM $MgCl_2$), 3 mU Phosphodiesterase I and 1.2 U alkaline phosphatase were added into reaction. The reaction was further incubated at 37˚C for 3h.

After the hydrolysis, the reaction was top-up to 100 μL of LC Solvent A(10 mM ammonium acetate pH 6.0 and the solution was filtered with Amicon Ultra-0.5 mL 10K centrifugal filters to remove the hydrolysis enzymes before LC-MS/MS analysis.

2.18 Notes of statistical analysis

Choice of Standard deviation or SEM

Standard deviation (SD) measures the variance of the sample and describes the variability between individuals in a sample. The standard error of the mean (SEM) describes how accurate the sample mean represents the population mean and is used for statistical inference of the population mean (Altman and Bland, 2005). Therefore, in the plots summarising replicates, SD was used to describe the dispersion of data.

When comparing TET activities on different DNA substrates, two-way ANOVA was used to determine whether time points and/or DNA have effects on TET activities. If ANOVA showed significant differences between DNA substrates, t-test was used to compare different DNA within each time point. Holm-Sidak method was used to correct the t-test *p*-value for multiple comparisons.

3 Chapter 3 H/M sites binding proteins

3.1 Introduction

Chemical modifications on DNA, like 5mC and 5hmC, can affect gene expression by either reducing the affinity of DNA binding proteins or by attracting specific binding proteins. 5mC can affect the binding of transcription factors (Yin et al., 2017). Similarly, 5hmC is known to reduce the binding of 5mC binding proteins (Hashimoto et al., 2012; Jin et al., 2010). 5mC can also be specifically recognised by proteins like methyl-CpG-binding (MBD) family proteins (Fournier et al., 2011). Likewise, there may be proteins that recognise 5hmC specifically. The aim of this chapter is to identify 5hmC binding proteins.

There are several ways to identify proteins that specifically recognise DNA modifications. In the case of 5mC binding proteins, MeCP2 and Kaiso were identified by DNA affinity purification (Lewis et al., 1992; Prokhortchouk et al., 2001). Later on, more 5mC binding proteins like MBD family proteins, ZBTB4 and ZBTB38 were identified based on sequence similarities to MeCP2 and Kaiso (Filion et al., 2006; Fournier et al., 2011). The advances in mass spectrometry (MS) have led to a more efficient way for identification of DNA binding proteins by DNA pull-down followed by MS (Carey et al., 2012). Several studies have identified 5hmC binding proteins using this method. UHRF2 was identified as 5hmC binding proteins in neuronal progenitor cells (Spruijt et al., 2013). The co-crystal structure of UHRF2 Set and RING finger-associated (SRA) domain with DNA further supported UHRF2 as a

5hmC binding protein (Zhou et al., 2014). MeCP2 was identified as a 5mC and 5hmC binding protein in mouse brain (Mellen et al., 2012) and was later proved to recognise 5hmC in a CA but not CG context (Gabel et al., 2015). Apart from UHRF2 and MeCP2, CHTOP was identified as 5hmC binding protein in glioblastoma (Takai et al., 2014); MSH6, SALL1 and SALL4A were identified in mESCs (Iurlaro et al., 2013; Xiong et al., 2016), but there was no follow-up study on these proteins.

5hmC has a distinct genomic distribution from 5mC. The base-resolution mapping of 5hmC in human embryonic stem cells showed that while most of (99.89%) 5hmCs are in CG context, unlike 5mC, of which 91.8% are symmetrically modified, only 21.0% of 5hmCs are symmetric (Yu et al., 2012b). With single-molecule imaging, high levels of adjacent and opposing hydroxymethylated and methylated CpG (H/M) sites were observed, roughly accounting for 60% of 5hmC in various mouse tissues (cerebellum, kidney, heart and liver) and mESCs (Song et al., 2016). Despite their high abundance, the biological functions of H/M sites are largely unknown. Identification of H/M site-specific binding proteins would provide insights on its biological role. Previous studies seeking to find 5hmC binding proteins used symmetrically hydroxymethylated probes with both cytosines in a CpG site hydroxymethylated or probes with all cytosine hydroxymethylated (Table 3-1), which may not represent the major state of 5hmC in the genome. Thus, this chapter will focus on identifying specific binding proteins of H/M sites.

Table 3-1 Summary of studies on 5hmC binding proteins

Identified proteins	DNA probe	Tissue type	Study
UHRF2	— hmC G — — G hmC —	Neuronal progenitor cells	(Spruijt et al., 2013)
CHTOP	— hmC G — — G hmC —	Glioblastoma	(Takai et al., 2014)
MeCP2	All Cs are 5hmC	Rat cerebella	(Mellen et al., 2012)
MSH6	All Cs are 5hmC	mESCs	(Iurlaro et al., 2013)
SALL1, SALL4	— hmC G — — G hmC —	mESCs	(Xiong et al., 2016)

3.2 Identifications of potential H/M sites binding proteins using DNA pull-down followed by mass spectrometry

3.2.1.1 Overview of DNA pull-down

The DNA pull-down was performed with nuclear protein extracts following similar principles as previous paper (Mellen et al., 2012; Xiong et al., 2016), with the modification that DNA probes were with H/M sites instead of fully hydroxymethylated (Figure 3-1 a, b). As a control, DNA probes containing fully methylated sites (M/M) or unmodified sites (C/C) were used. Pull-down without probes was also performed as a control for nonspecific protein binding to the beads (Figure 3-1 b). Mouse brains and embryonic stem cells were chosen as the materials for pull-down analysis because they have the highest levels of 5hmC in tissues and cell lines respectively (0.15% of total DNA in mouse brains and 0.027% in mESCs) (Ito et al., 2011; Li and Liu, 2011).

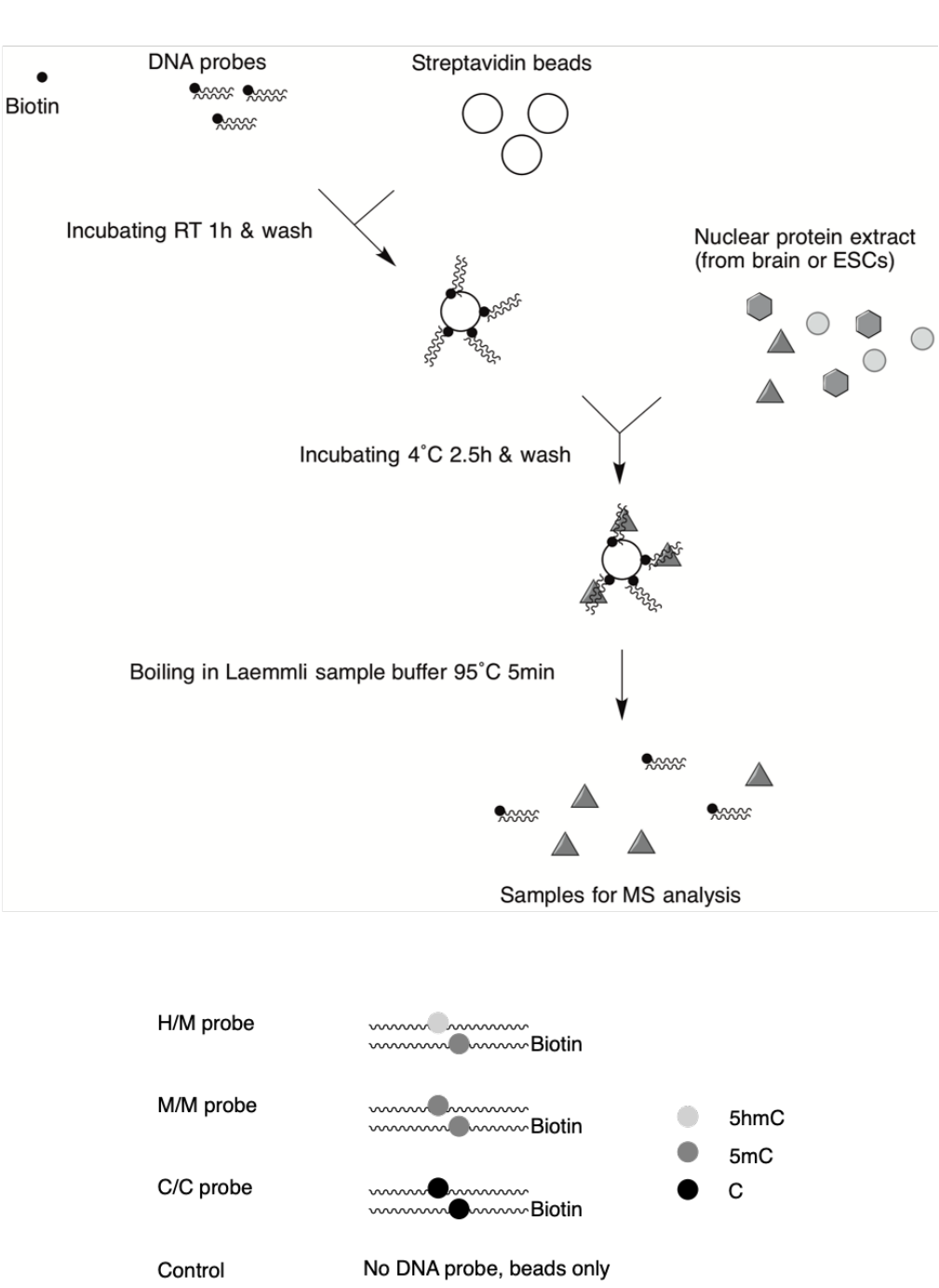


Figure 3-1 A schematic diagram of DNA affinity pull-down from nuclear extracts. a) The workflow of the pull-down. b) Probes and controls used in the pull-down.

3.2.2 Optimisation of pull-down conditions

To optimise the method for nuclear protein extraction from mESCs, two methods using different sets of buffers (Table 2-3) were compared. Cytoplasmic protein GAPDH, nuclear proteins TBP and MECP2 were blotted in cytoplasmic and nuclear extracts respectively (Figure 3-2). Method 1 gave a better removal of GAPDH in nuclear extract than Method 2 but was not efficient in extracting MECP2. With Method 2, MECP2 and TBP were enriched in nuclear extract. Although there were still GAPDH left in the nuclear extract, it showed depletion compared with the cytoplasmic extract. Therefore Method 2 was used for preparing samples for DNA pull-down.

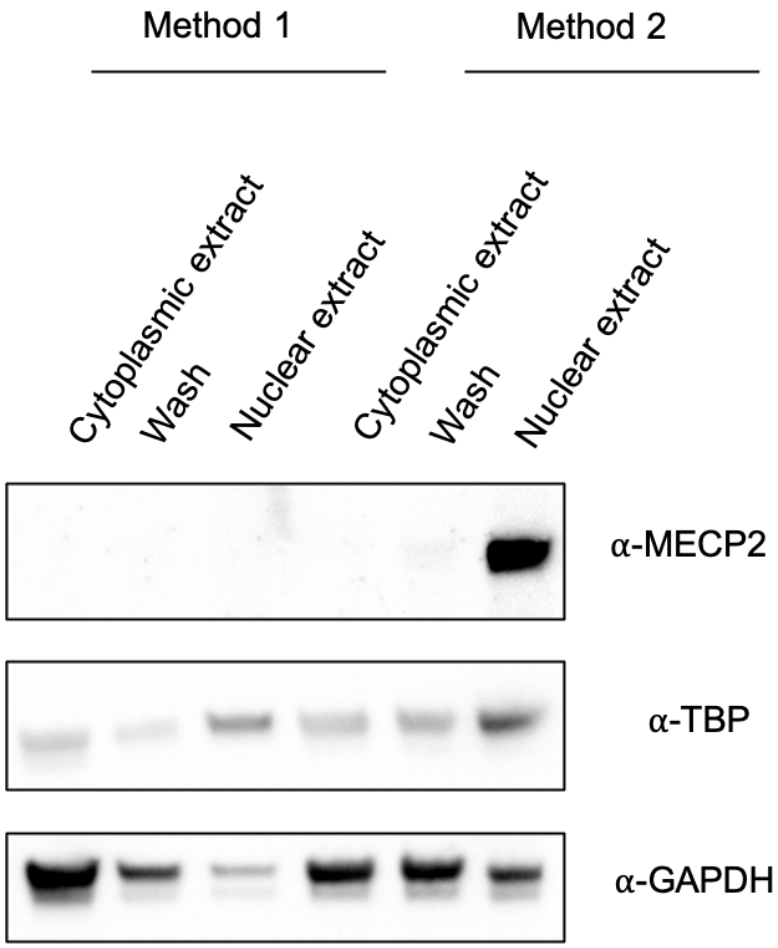
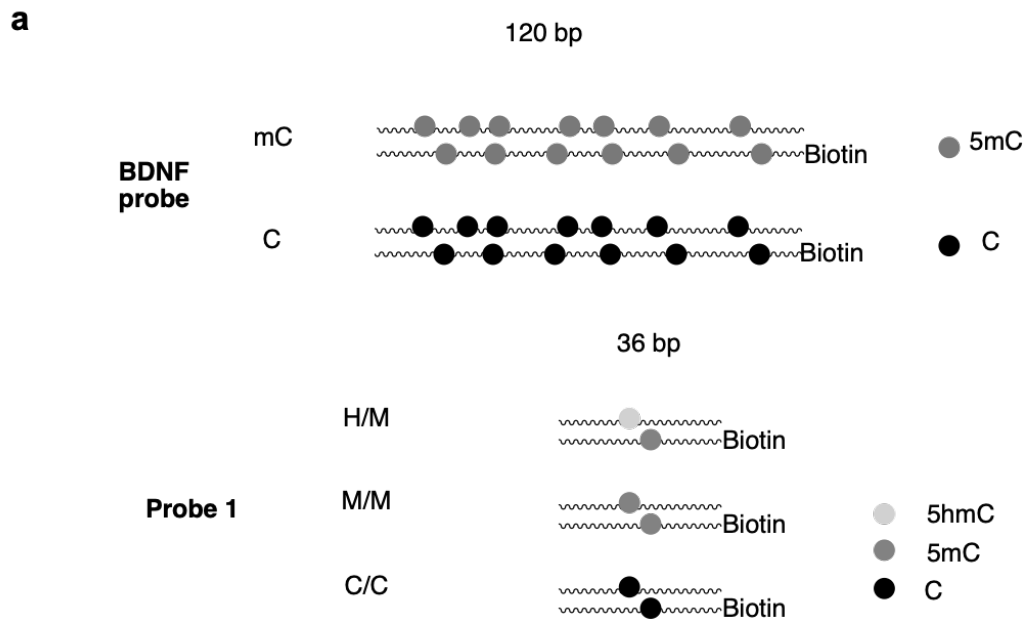
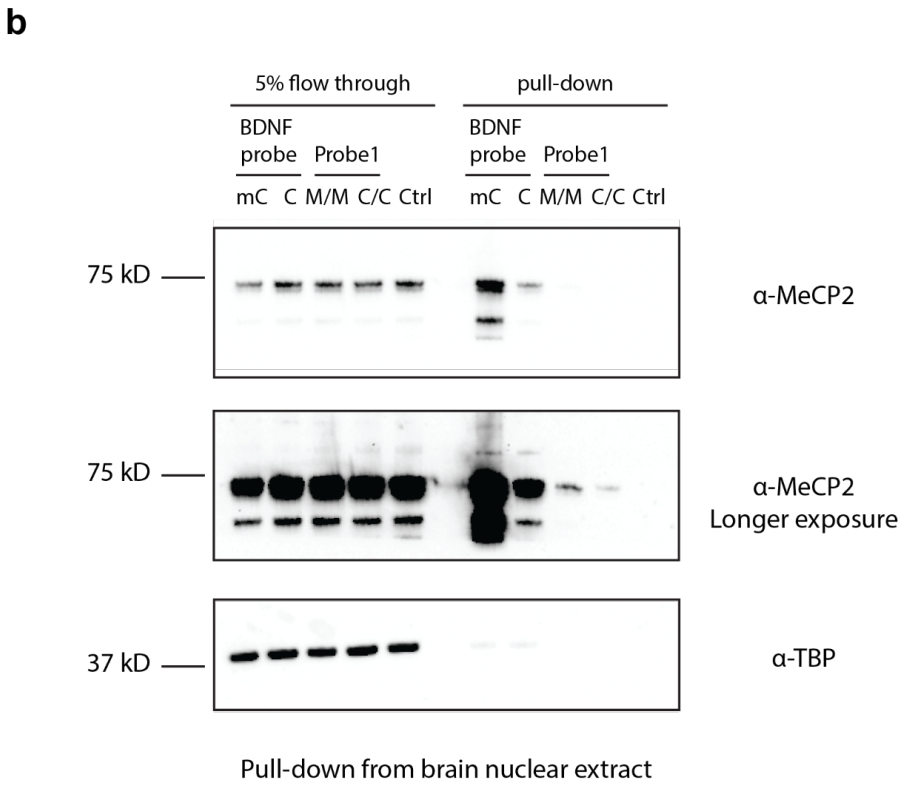


Figure 3-2 Comparison of nuclear protein extraction methods. Western blot of nuclear protein MECP2 and TBP, cytoplasmic protein GAPDH in cytoplasmic and nuclear extracts with Method 1 or Method 2.

To optimise the experimental conditions of pull-down, the known 5mC binding protein MECP2 was used as a positive control. It was reported that MECP2 can be enriched by a 120bp probe of the BDNF promoter sequence (BDNF probe) with all cytosines methylated but not the unmethylated equivalent (Mellen et al., 2012). The BDNF probe was used as a positive control to assess pull-down efficiencies (Figure 3-3 a). As shown in Figure 3-3, compared with the unmethylated probe, methylated BDNF probe was able to enrich MECP2 from both mouse brain and mESC nuclear

extracts. As the MECP2 protein is more abundant in mouse brains than in mESCs, the signal was stronger from the brain (Figure 3-3 b, c). A 36 bp probe (Probe 1) with one H/M site and its symmetric methylated (M/M) and unmodified (C/C) equivalents were designed for the pull-down of H/M site binding proteins (Figure 3-3 a). The methylated (M/M) and unmodified (C/C) Probe 1 were tested under the same pull-down condition as BDNF probe. However, the efficiency was much lower compared with the BDNF probe (Figure 3-3 b, c). The reason for this could be that Probe 1 has only one methylcytosine while the BDNF probe was methylated at all cytosines (26 cytosines in total) (Figure 3-3 a). Thus, a longer probe with more modified sites may improve pull-down efficiency.




b
5% flow through
pull-down
BDNF probe
Probe1
mC C M/M C/C Ctrl
mC C M/M C/C Ctrl
75 kD
α-MeCP2
75 kD
α-MeCP2
Longer exposure
37 kD
α-TBP
Pull-down from brain nuclear extract

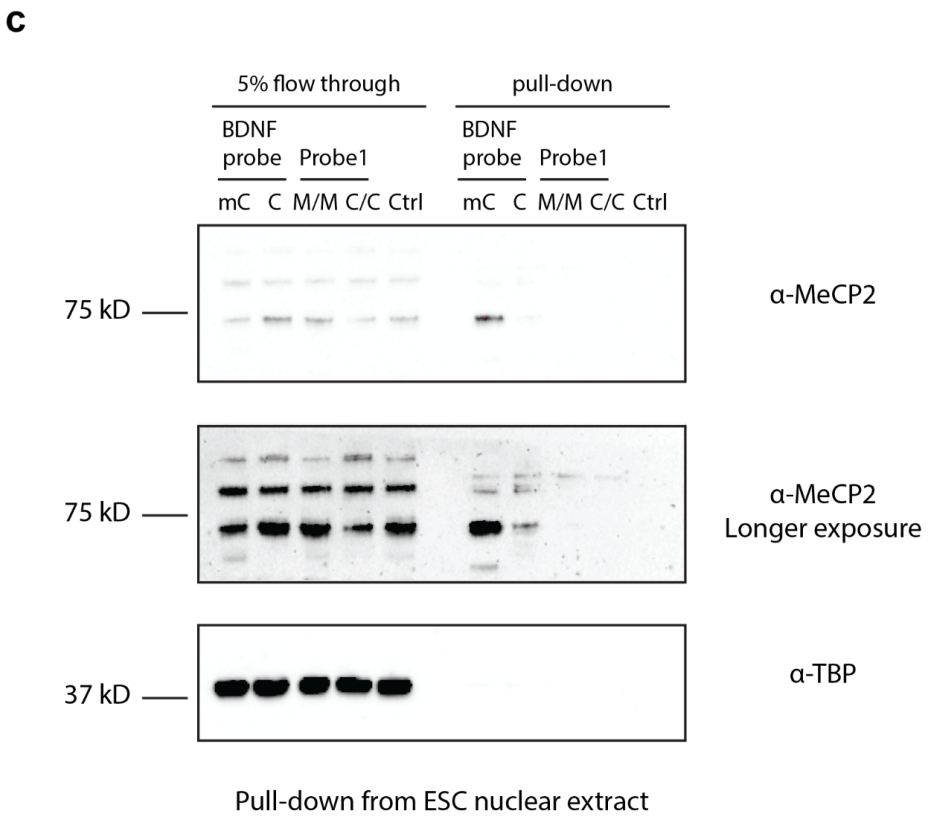

c
5% flow through
pull-down
BDNF probe
Probe1
mC C M/M C/C Ctrl
mC C M/M C/C Ctrl
75 kD
α-MeCP2
75 kD
α-MeCP2
Longer exposure
37 kD
α-TBP
Pull-down from ESC nuclear extract

Figure 3-3 Comparison of BDNF probe and Probe 1. a) A schematic diagram of BDNF probe and Probe 1. b, c) Pull-down was performed with methylated (mC), unmodified (C) BDNF probe, methylated (M/M), unmodified (C/C) Probe 1 and no DNA probe (Ctrl). The pull-downed proteins and flow through proteins from (b) mouse brain nuclear extract and (c) mESC nuclear extract were analysed using Western blot with MeCP2 antibody. There are two isoforms of MeCP2, which are 53.6 kDa and 52.3 KD respectively. It normally migrates higher than 50 kDa on an SDS-PAGE, around 70 kDa and 60 KD in this case.

Considering the technical limits of synthesising DNA oligos with 5hmC modifications, a 90 bp probe with three modified sites (Probe 2) was used. Probe 2 was designed based on H/M sites in the mESC genome. From currently available data, there is no base-resolution genomic mapping of 5mC and 5hmC simultaneously on the two strands of the same piece of DNA. Consequently, it was difficult to directly identify H/M sites in the genome. An approximate of this was to find 5hmC sites based on Tet-assisted bisulfite sequencing (TAB-Seq), 5mC and 5hmC sites based on bisulfite sequencing. Based on TAB-Seq (Yu et al., 2012b) and bisulfite sequencing (Lu et al., 2014) data in mESCs, CpG sites that had 5mC on one strand and 5hmC on the other strand were selected as potential H/M sites, even though it is uncertain if the 5mC and 5hmC co-exist on the same DNA molecule. A 90 bp sequence from intron 1 of Prdm16 gene was used as the Probe 2, which contains three potential H/M sites (Figure 3-4). The pull-down efficiency of Probe 2 was comparable to that of BDNF probe (Figure 3-5), so it was used in the pull-down to generate samples for MS analysis.

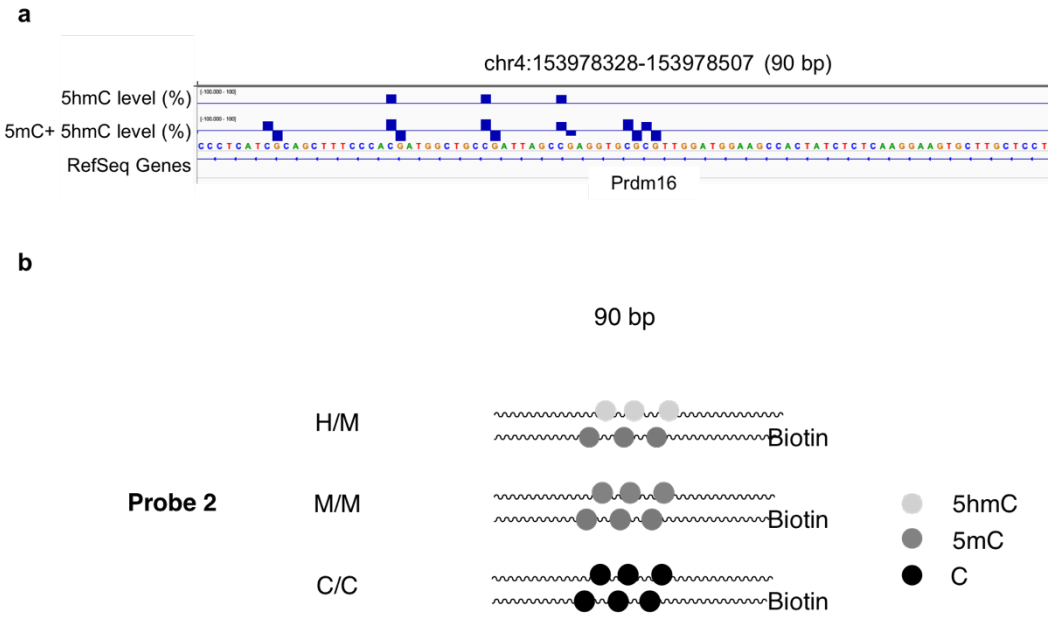


Figure 3-4 The design of Probe 2. a) Genome browser view of 5mC and 5hmC modifications of the probe region in mESCs. b) A schematic diagram of Probe 2

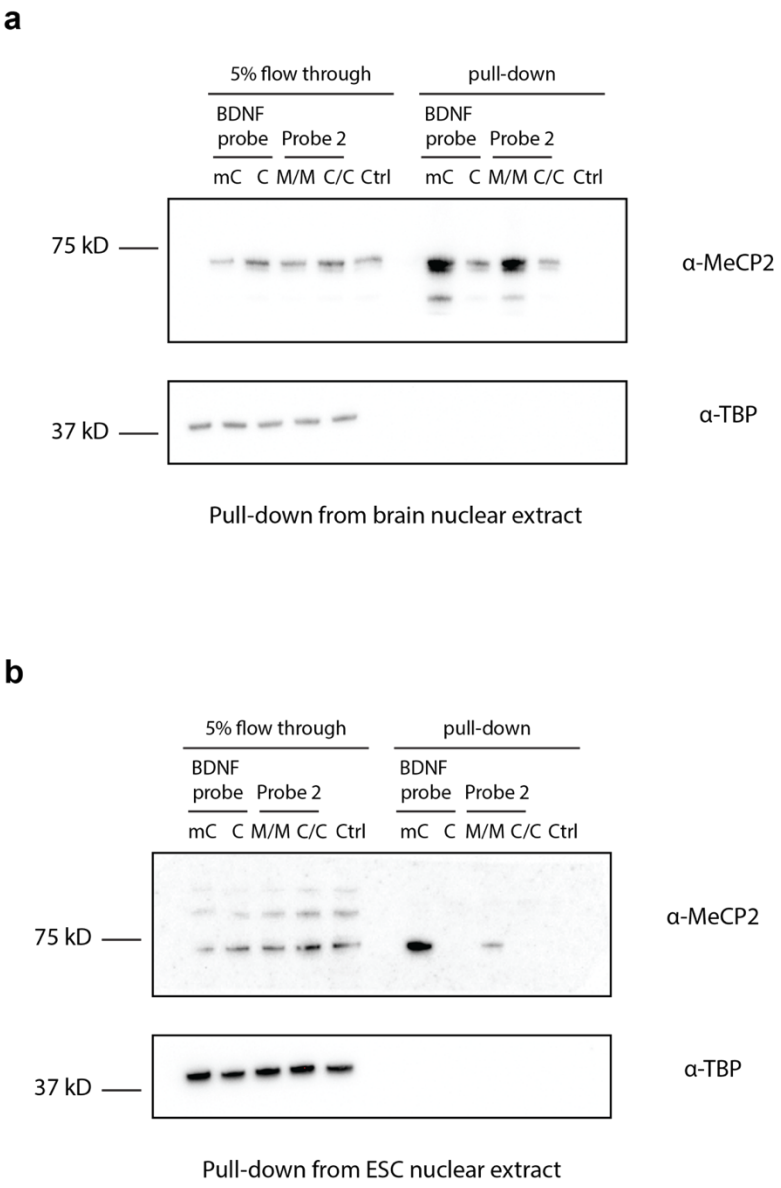
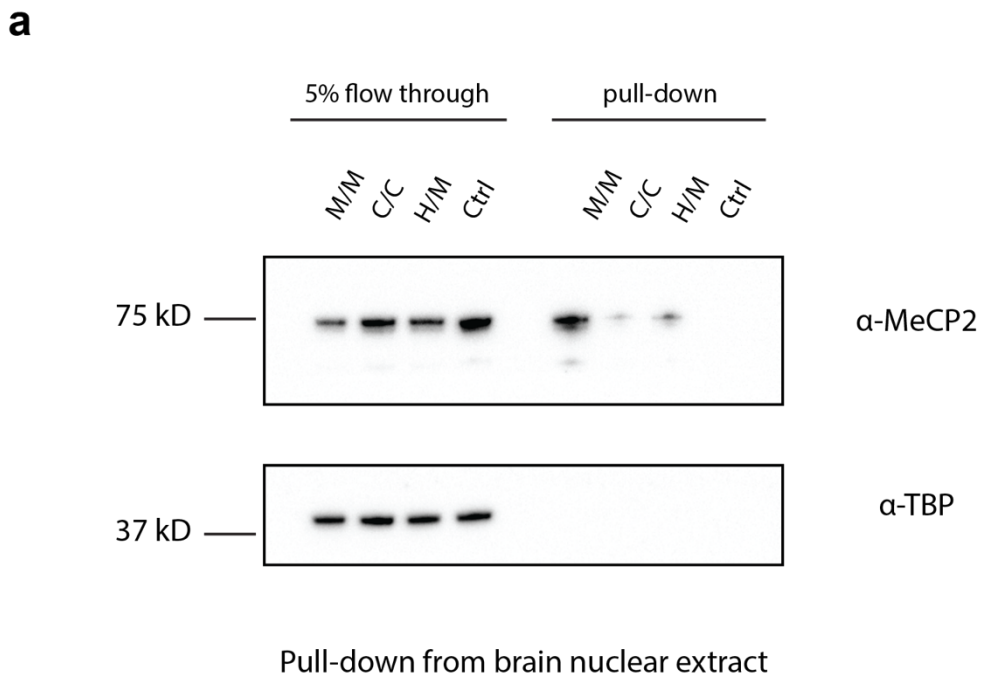
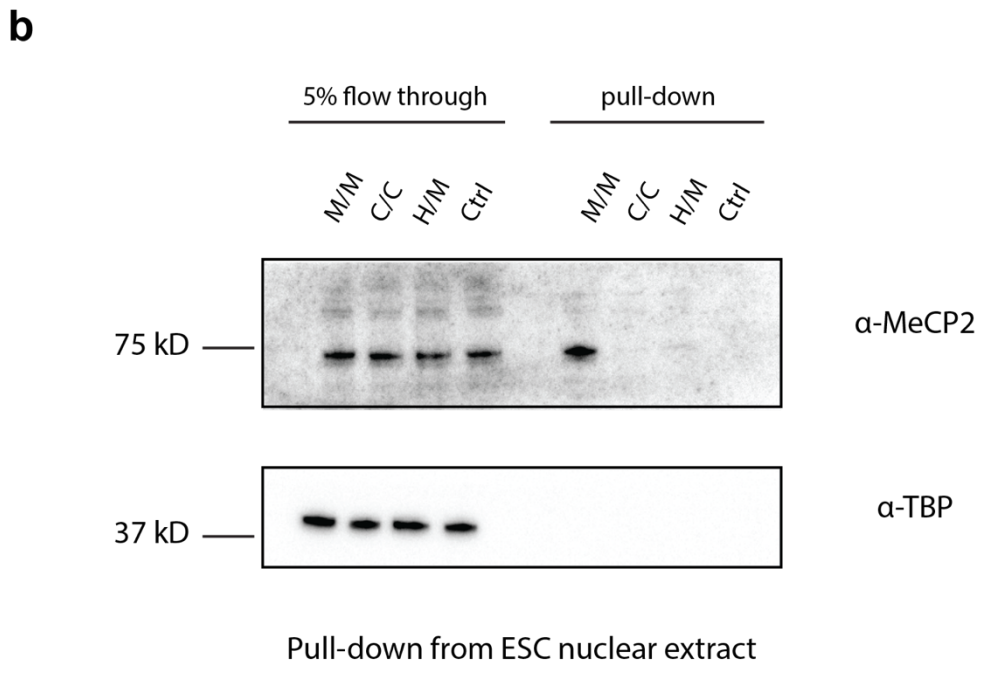


Figure 3-5 Comparison of BDNF probe and Probe 2. Pull-down was performed with methylated (mC), unmodified (C) BDNF probe, methylated (M/M), unmodified (C/C) Probe 2 and no DNA probe (Ctrl). The pull-downed proteins and flow through proteins from (b) mouse brain nuclear extract and (c) mESC nuclear extract were analysed using Western blot with MeCP2 antibody. There are two isoforms of MeCP2, which are 53.6 kDa and 52.3 KD respectively. It normally migrates higher than 50 kDa on an SDS-PAGE gel, around 70 kDa and 60 KD in this case.

3.2.3 Samples for mass spectrometry analysis

To generate samples for mass spectrometry analysis, the pull-down was scaled up and done in three replicates. One-sixth of pull-downed protein of each replicate was analysed with SDS-PAGE for quality control. To assess the pull-down specificity, samples were blotted for positive control protein MeCP2. As expected, MeCP2 was enriched by the methylated probe (Figure 3-6 a, b), validating the pull-down specificity. To check the protein abundance, the gel was stained with Coomassie Blue, followed by silver stain. There were several bands visible after Coomassie Blue stain (Figure 3-6 c), indicating sufficient amounts of protein for MS pre-processing. The overall amounts of proteins shown on the silver-stained gel were similar among different probes (Figure 3-6 d).


a
5% flow through
pull-down
M/M
C/C
H/M
Ctrl
M/M
C/C
H/M
Ctrl
75 kD
α-MeCP2
37 kD
α-TBP
Pull-down from brain nuclear extract


b
5% flow through
pull-down
M/M
C/C
H/M
Ctrl
M/M
C/C
H/M
Ctrl
75 kD
α-MeCP2
37 kD
α-TBP
Pull-down from ESC nuclear extract

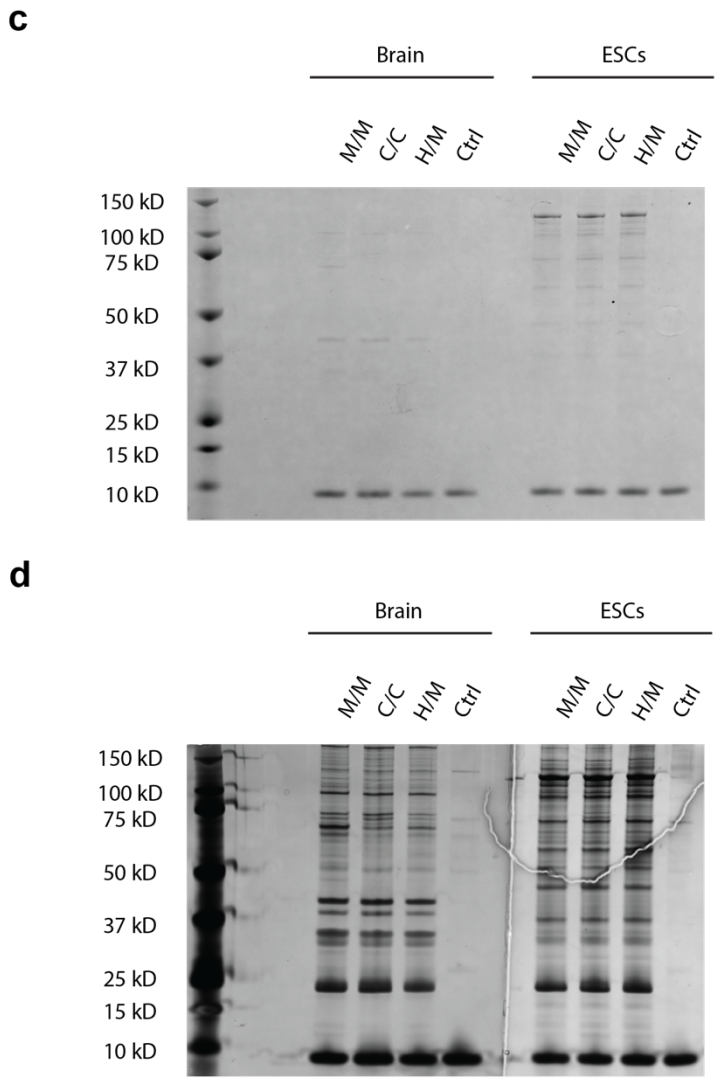


Figure 3-6 Samples for mass spectrometry analysis. Pull-down was performed with DNA probes containing H/M, M/M and C/C sites and no DNA probe (Ctrl). a, b) The pull-downed proteins and flow through proteins from (a) mouse brain nuclear extract and (b) mESC nuclear extract were analysed using Western blot using MeCP2 antibody. There are two isoforms of MeCP2, which are 53.6 kDa and 52.3 KD respectively. It normally migrates higher than 50 kDa on an SDS-PAGE gel, around 70 kDa and 60 KD in this case. c, d) the pull-downed proteins were separated by SDS-PAGE and stained with (c) Coomassie Blue and (d) silver stain.

3.2.4 Mass spectrometry analysis results

Proteins from DNA pull-down were subjected to trypsin digestion and analysed by LC-MS/MS. Label-free quantification of peptide ions was done with Progenesis QI. The normalisation of the raw abundance between samples was based on all identified peptides. Alignment scores between samples were checked to assess data quality. Two brains samples (M/M_3 and H/M_3) were found to have poor alignment scores (Table 3-2), which may indicate insufficient amounts of proteins in those two samples. The sample loss may be from sample preparation or sample injection. Thus, these two samples were not included in further analysis. As a quality control, pathway enrichment analysis was performed with all identified proteins before filtered by H/M sites binding preferences. The proteins were enriched in nucleus of gene ontology (GO) cellular component terms and DNA binding of GO molecular function terms as expect (Figure 3-7). The proteins were filtered by the number of unique peptides greater than the median of all identified proteins (> 2 for the brains, > 5 for the ESCs). ANOVA was performed to identify proteins that were significantly enriched or depleted by any of the probes. The proteins with ANOVA test *p*-value less than 0.05 were kept. As this step was an initial screen, loose criteria were used. The *p*-values used here were not corrected for multiple comparisons. The normalised abundance of proteins that passed the above filters was plotted (Figure 3-8). Proteins that were enriched by the H/M probe compared with both M/M and C/C probes were shown at the upper-right quadrant of the scatter plot (Figure 3-8 a, c). In the pull-down from brains, two proteins were selected for further validation: YBX1 and RBBP5. YBX1 showed the highest fold enrichment of all identified proteins. RBBP5 is part of a complex formed by four proteins: WDR5, RBBP5, ASH2L and DPY-30 (WRAD

complex) (Ernst and Vakoc, 2012). One of the proteins in the complex, ASH2L, although did not past the statistical test, also showed the trend of enrichment by H/M probe. From the mESC results, all five proteins with the $\log_2$ fold change of both H/M probe over M/M probe and H/M probe over C/C probe higher than 0.5 were selected for further validation.

Table 3-2 The percentage of matchable peaks of total ion chromatograms in each sample that are matched with the reference samples (C/C_2 for brain samples, C/C_1 for ESC samples)

Brain samples		ESC samples	
Sample Name	Alignment Score	Sample Name	Alignment Score
C/C_1	53.9%	C/C_1	NA (Reference)
C/C_2	NA (Reference)	C/C_2	88.7%
C/C_3	60.5%	C/C_3	68.7%
M/M_1	64.1%	M/M_1	69.4%
M/M_2	60.3%	M/M_2	61.1%
M/M_3	10.6%	M/M_3	77.3%
H/M_1	66.4%	H/M_1	80.7%
H/M_2	57.5%	H/M_2	80.0%
H/M_3	29.4%	H/M_3	77.4%

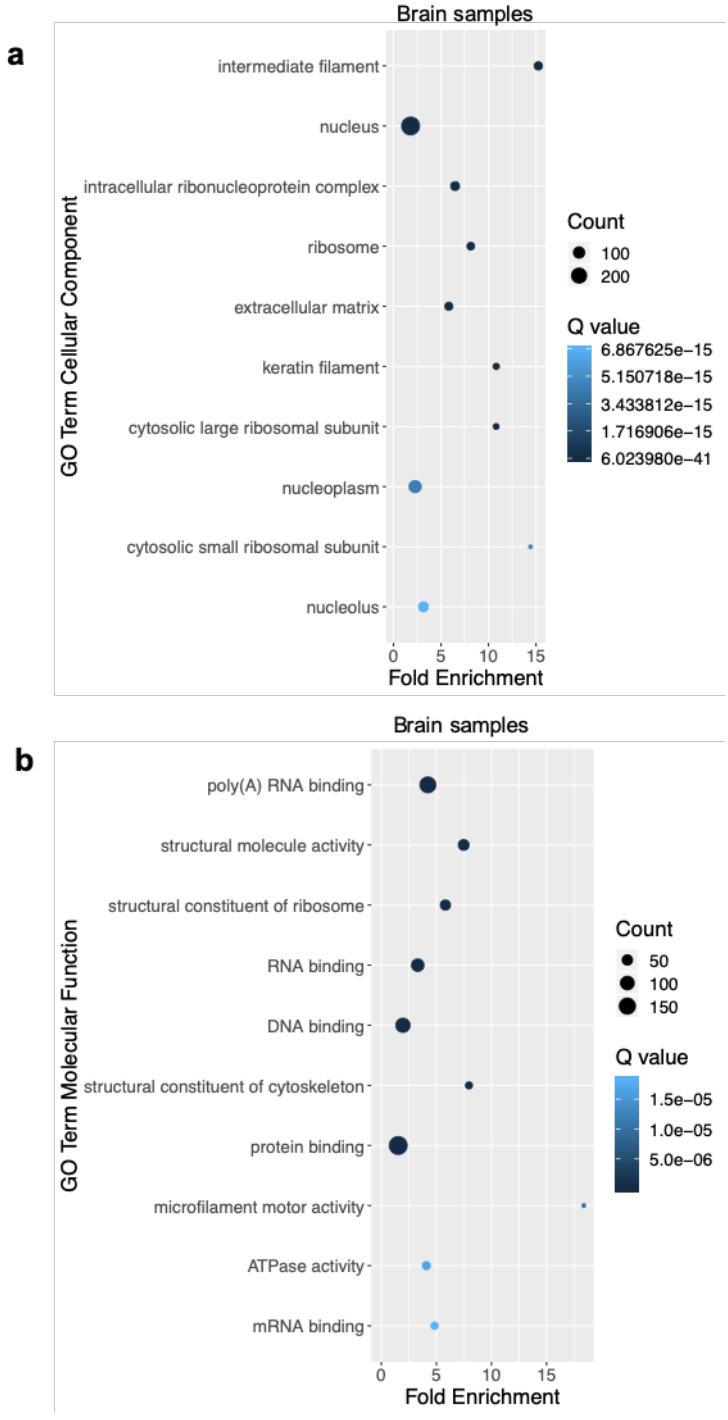
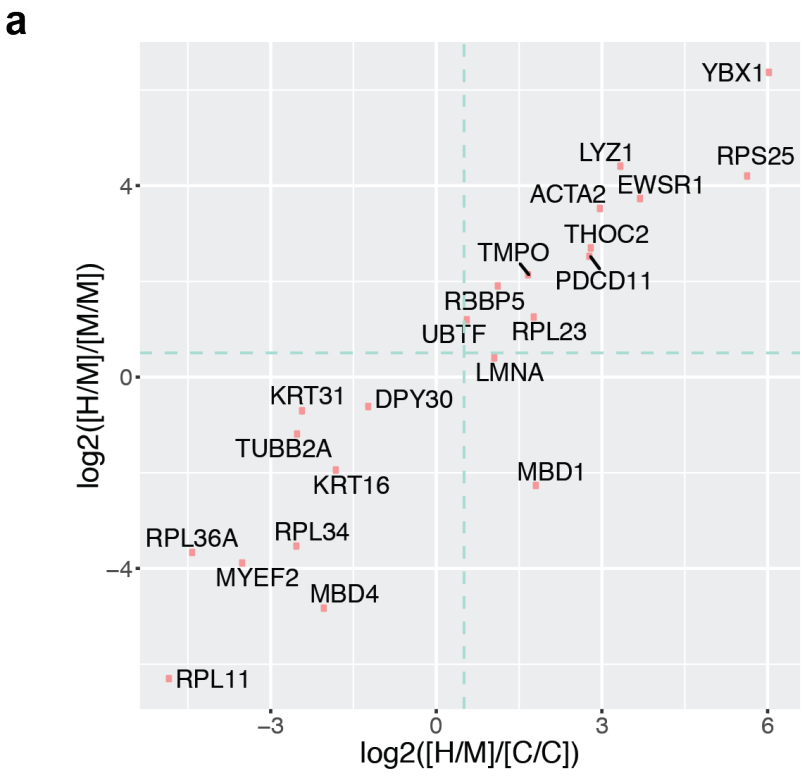
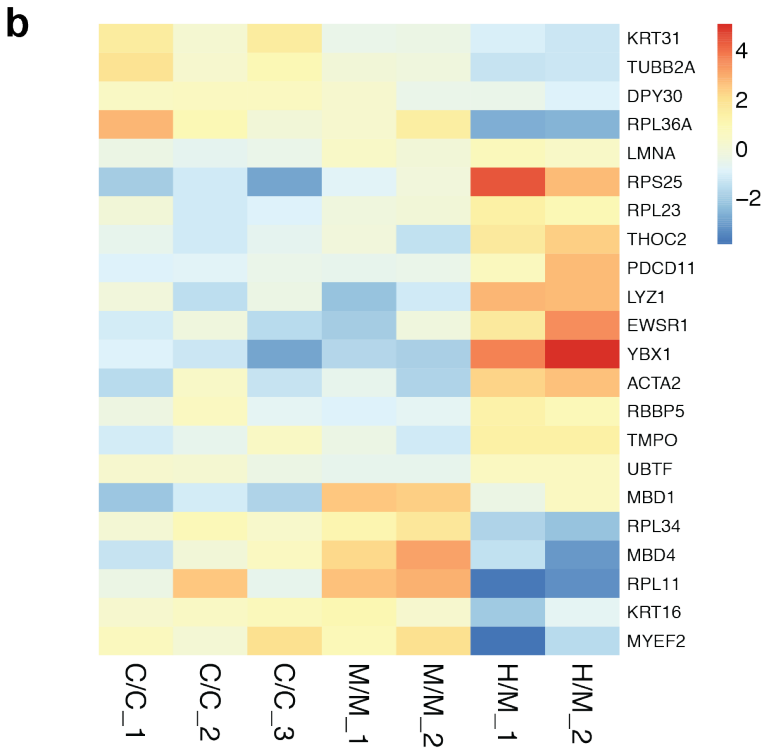

a
Brain samples
intermediate filament
nucleus
intracellular ribonucleoprotein complex
ribosome
extracellular matrix
keratin filament
cytosolic large ribosomal subunit
nucleoplasm
cytosolic small ribosomal subunit
nucleolus
GO Term Cellular Component
0
5
10
15
Fold Enrichment
Count
100
200
Q value
6.867625e−15
5.150718e−15
3.433812e−15
1.716906e−15
6.023980e−41
b
Brain samples
poly(A) RNA binding
structural molecule activity
structural constituent of ribosome
RNA binding
DNA binding
structural constituent of cytoskeleton
protein binding
microfilament motor activity
ATPase activity
mRNA binding
GO Term Molecular Function
0
5
10
15
Fold Enrichment
Count
50
100
150
Q value
1.5e−05
1.0e−05
5.0e−06

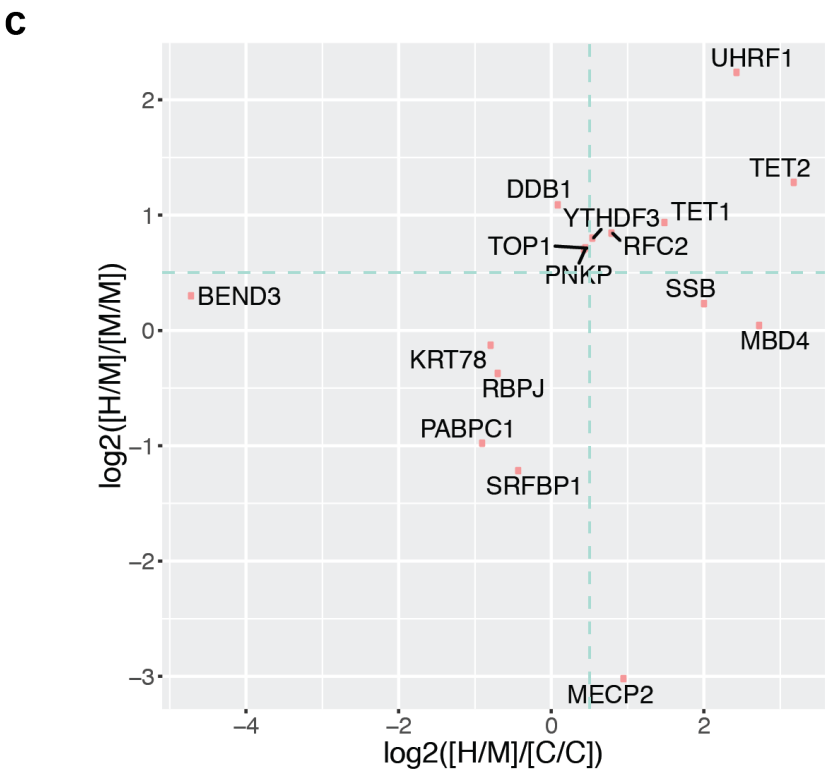
Figure 3-7 Pathway enrichment analysis of identified proteins from brains (a, b) and ESC (c, d) pull-down samples. The enrichment analysis was performed with two types of pathway annotations: cellular component (a, c) and molecular functions (b, d).



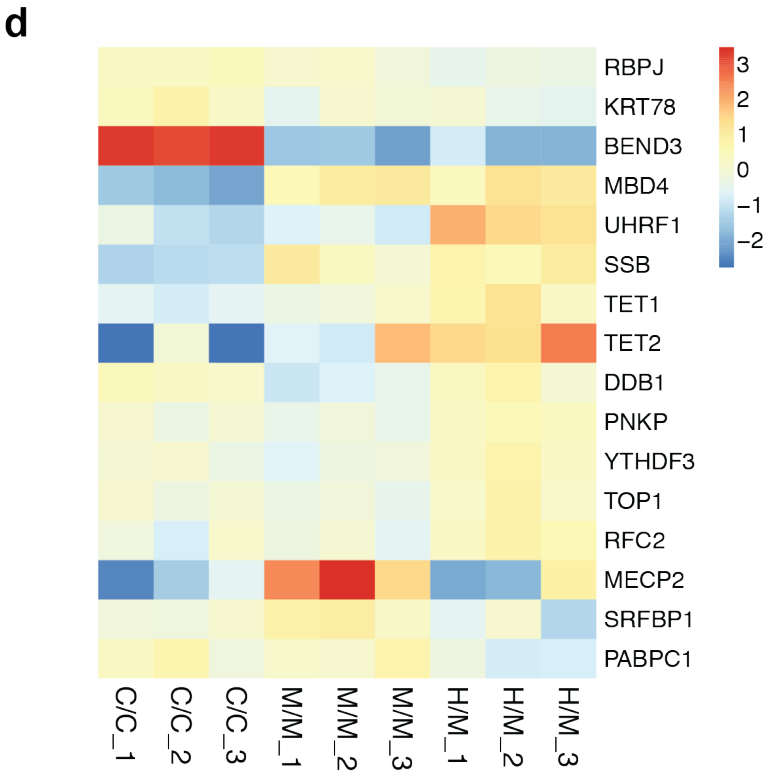
Filtered hits of pull-down from brain



Filtered hits of pull-down from brain



Filtered hits of pull-down from ESCs



Filtered hits of pull-down from ESCs

Figure 3-8 Plots of proteins that were enriched by at least one of the probes. Identified proteins were filtered by ANOVA tests to be significantly enriched by at least one of the probes, and the coverage of at least two unique peptides from the brain or six from the ESCs. a, c) Scatterplots of $\log_2$ average ratio of protein abundances pull-downed by DNA probes containing H/M, M/M or C/C sites. The dashed lines indicate the $\log_2$ ratio of 0.5. b, d) Heatmaps of normalised protein abundances after $\log_2$ transformation and normalised by row mean subtraction.

3.3 Validation of H/M site binding proteins

3.3.1 YBX1

YBX1 was identified as a Y-box (CCAAT) binding protein but later found to have no sequence specificity. As a DNA binding protein, it binds to AP site and mismatch, promoting strand separation (Eliseeva et al., 2011; Gaudreault et al., 2004; Hasegawa et al., 1991). Recently, YBX1 was found to be an RNA binding protein that recognises 5mC in RNA in human (Chen et al., 2019b) and zebrafish (Yang et al., 2019). There is no report so far about its affinity to 5hmC. In the DNA pull-down followed by MS, YBX1 was more than 60 times enriched by the H/M probe compared with the M/M or C/C probes (Figure 3-9). To validate the binding preference, recombinant YBX1 was purified for electrophoretic mobility shift assay (EMSA). The full-length mouse YBX1 coding sequence (CDS) was cloned into pNIC28 vector. His-tagged YBX1 was expressed in E. coli and purified with Ni-NTA resin followed by heparin column (Figure 3-10 a). The binding affinity of purified protein towards DNA probes with H/M, M/M or C/C sites was studied with EMSA (Figure 3-10 b). YBX1 showed similar affinity toward H/M, M/M and C/C sites (Figure 3-10 c).

a

Unique peptides	4
ANOVA *p* value	0.0030
T test *p* value	
H/M v.s. M/M	0.012
H/M v.s. C/C	0.0072
Fold change	
(H/M)/(M/M)	82.6
(H/M)/(C/C)	66.2

b

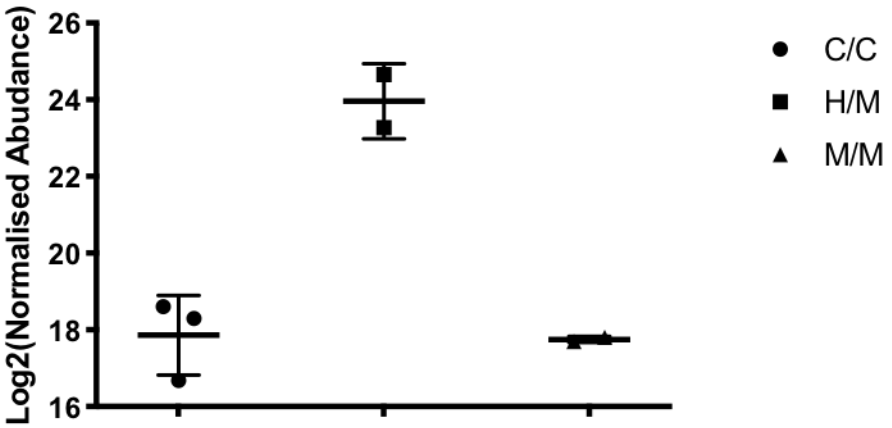


Figure 3-9 Summary of YBX1 MS results. a) Summary of MS results and statistics. b) Log_2 abundance of protein pull-downed by DNA probes containing H/M, M/M or C/C sites.

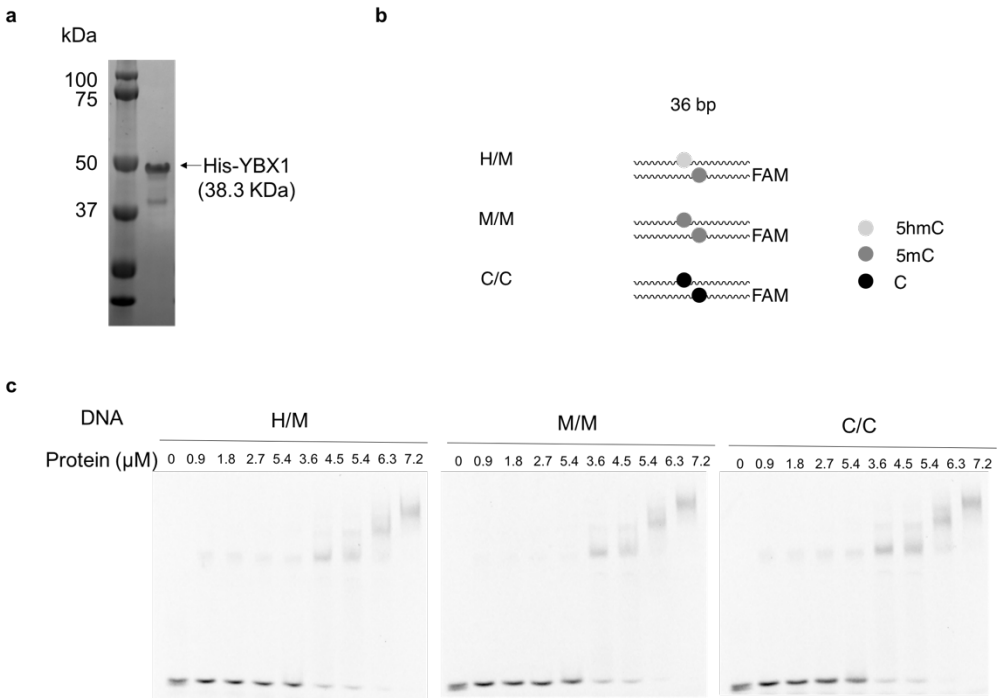


Figure 3-10 EMSA of His-YBX1. a) SDS-PAGE showing purified His-YBX1. Arrow pointing at the band of His-YBX1. The molecular weight of His-YBX1 is 38.3 kDa, but as YBX1 is a positively charged protein with an isoelectric point of around 9.7, on SDS-PAGE the protein normally migrates slower. b) A schematic diagram of DNA used in the EMSA. A 36-bp DNA fragment containing one H/M, M/M or C/C site was end-labelled with FAM. c) EMSA with increasing amounts of His-YBX1.

3.3.2 RBBP5

RBBP5 is part of a complex formed by four proteins: WDR5, RBBP5, ASH2L and DPY-30 (WRAD complex), required to activate the SET1 family histone H3 lysine 4 (H3K4) methyltransferases (Ernst and Vakoc, 2012). The MS result of proteins from the WRAD complex was summarised in Figure 3-11. Two of the proteins, RBBP5 and ASH2L, were enriched by the H/M probe while the other two proteins showed no preference or depletion by H/M probe. It is curious that the proteins from the same complex showed different binding preferences. The reason could be that the condition

used in the DNA pull-down was not intended to preserve complexes. There was 400 mM KCl in lysis buffer and 0.5% Triton X-100 in pull-down buffer. Therefore, the protein complex might have been disrupted to some extent. ASH2L but not RBBP5 has a known DNA binding domain (Ernst and Vakoc, 2012). Therefore, ASH2L was selected for validating the binding preference. The full-length mouse ASH2L CDS was cloned into pNIC28 vector. His-tag ASH2L was expressed in E. coli and purified with Ni-NTA resin followed by gel filtration. The protein showed two UV absorption peaks during gel filtration. When separated on SDS-PAGE, both peaks showed bands at the molecular weight of His-ASH2L so the first peak was likely to be aggregations of the proteins (Figure 3-12). Proteins from each peak were analysed separately with EMSA. Proteins from peak2 but not peak1 showed DNA binding. ASH2L showed similar affinity on to H/M, M/M and C/C site containing DNA (Figure 3-13).

a

	Unique peptides	ANOVA (*p*)
ASH2L	26	0.068
RBBP5	16	0.048
WDR5	11	0.29
DPY-30	7	0.029

b

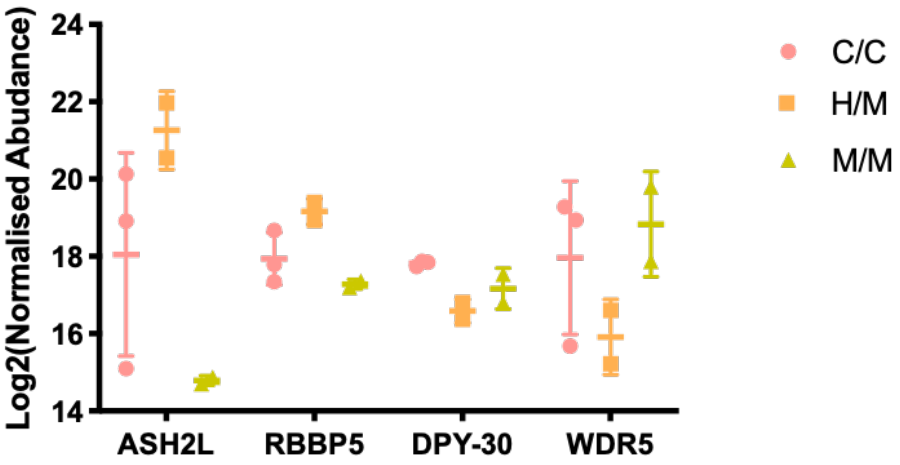


Figure 3-11 Summary of MS results of WRAD complex proteins. a) Summary of the number of unique peptides and *p*-values of the ANOVA test among three probes. b) Log_2 abundance of protein pull-downed by DNA probes containing H/M, M/M or C/C sites.

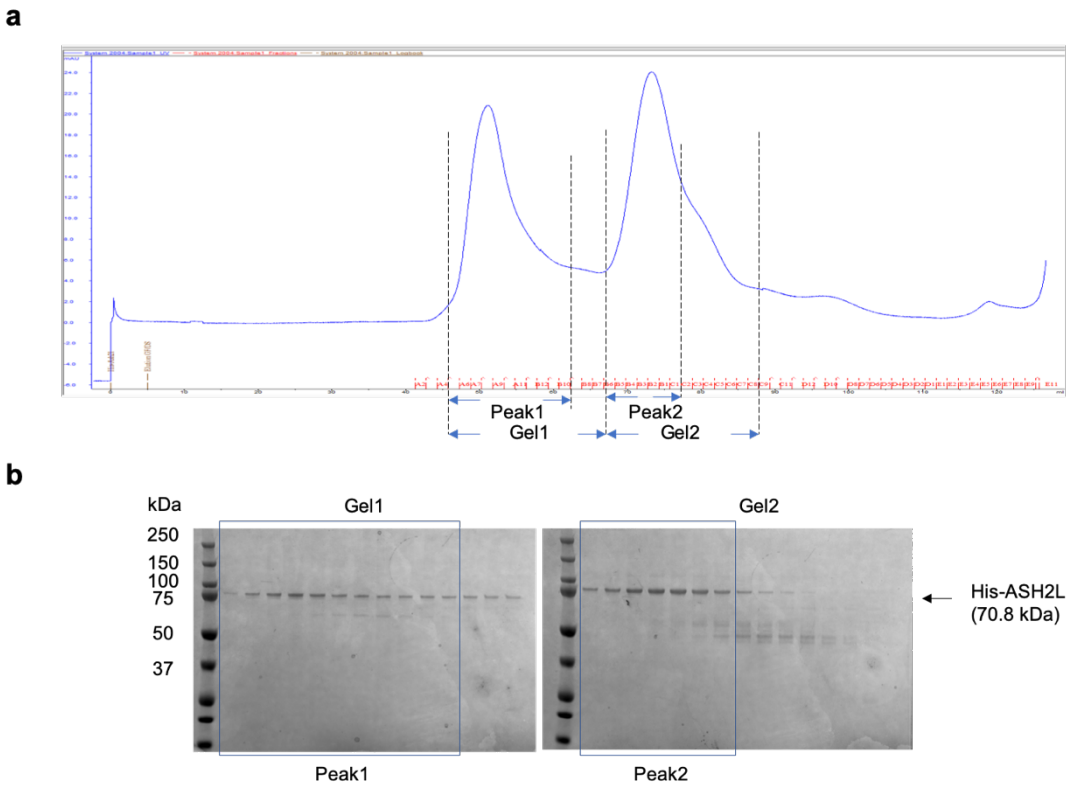


Figure 3-12 Gel filtration results of His-ASH2L. a) UV absorption profile of His-AS2L through a HiLoad™ 16/600 Superdex™ 200 pg column. b) SDS-PAGE showing proteins from peak1 and peak2. Arrow pointing at the size of His-ASH2L (70.8 kDa).

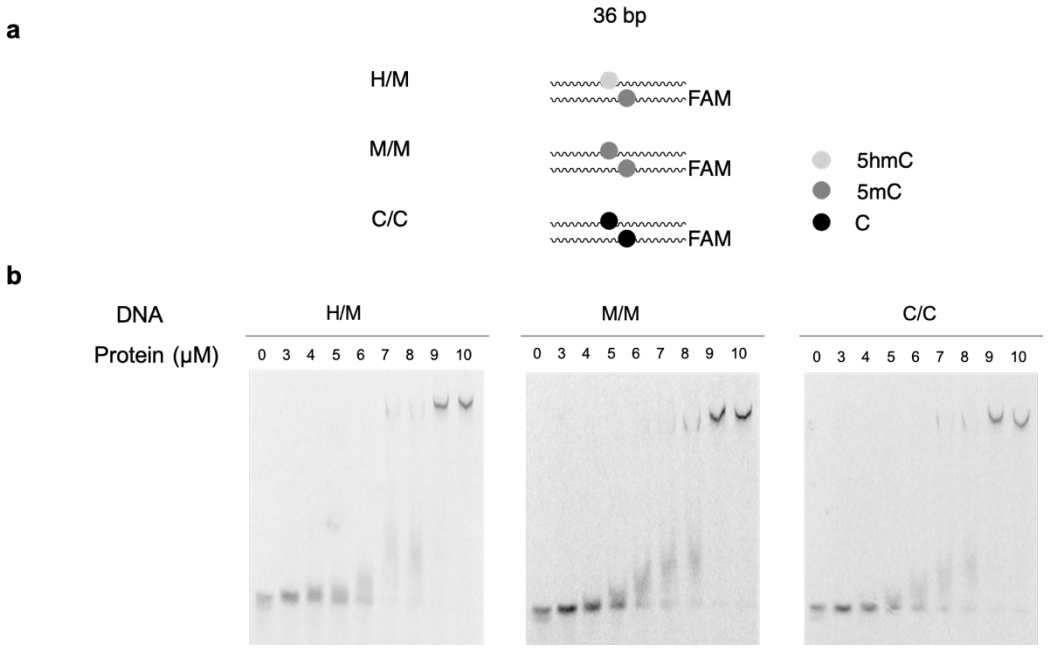


Figure 3-13 EMSA of His-ASH2L. a) A schematic diagram of DNA used in the EMSA. A 36-bp DNA fragment containing H/M, M/M or C/C sites was end-labelled with FAM. b) EMSA with increasing amounts of His-ASH2L.

3.3.3 UHRF1

UHRF1 aids DNMT1 in DNA methylation maintenance by recruiting DNMT1 to hemimethylated CpGs (Bostick et al., 2007; Sharif et al., 2007). UHRF1 is about 5-fold enriched by H/M probe compared with M/M or C/C probe in the MS analysis (Figure 3-14). The full-length mouse UHRF1 CDS was cloned into pNIC28 vector. UHRF1 with His-tag was expressed in E. coli and purified with Ni-NTA resin. The SDS PAGE of the purified protein showed several lower molecular weight bands, which can be either contaminants or degraded proteins (Figure 3-15 a). To improve the purity, elution with imidazole gradients was performed. However, the gradient elution did not help with removing the low molecular weight bands (Figure 3-15 b). The protein was further purified with gel filtration (Figure 3-16 a). To validate the

binding preference, EMSA was performed with the purified protein. It is known that UHRF1 binds to hemimethylated DNA (Bostick et al., 2007; Sharif et al., 2007). Therefore, DNA with hemimethylated CpG (C/M) sites was used as a positive control. To compare with the M/C site, DNA with hemi-hydroxymethylated sites (H/C) was also used (Figure 3-16 b). UHRF1 showed the highest affinity on M/C site, the lowest to unmodified CpG (C/C) sites. It showed similar affinity to H/M, M/M, H/C, which is in-between its affinity to C/M and C/C sites (Figure 3-16 c, d).

Unique peptides	41
ANOVA *p* value	0.00041
T test *p* value	
H/M v.s. M/M	0.0011
H/M v.s. C/C	0.0022
Fold change	
(H/M)/(M/M)	4.7
(H/M)/(C/C)	5.4

b

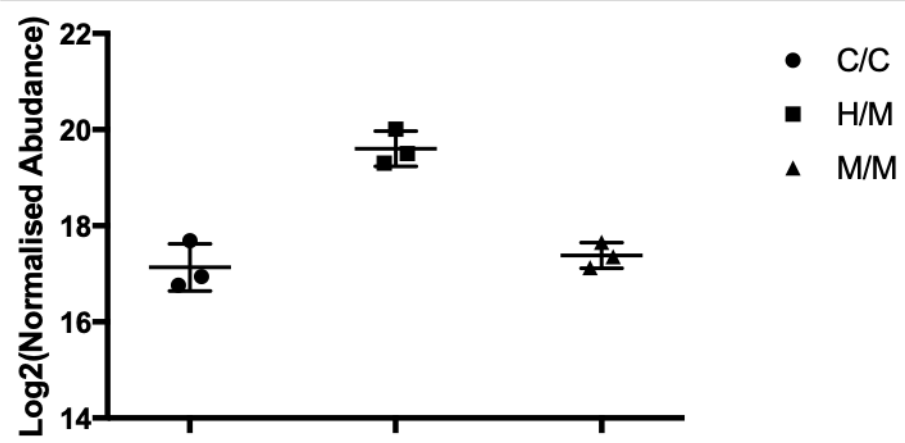


Figure 3-14 Summary of UHRF1 MS results. a) Summary of MS results and statistics. b) Log_2 abundance of protein pull-downed by DNA probes containing H/M, M/M or C/C sites.

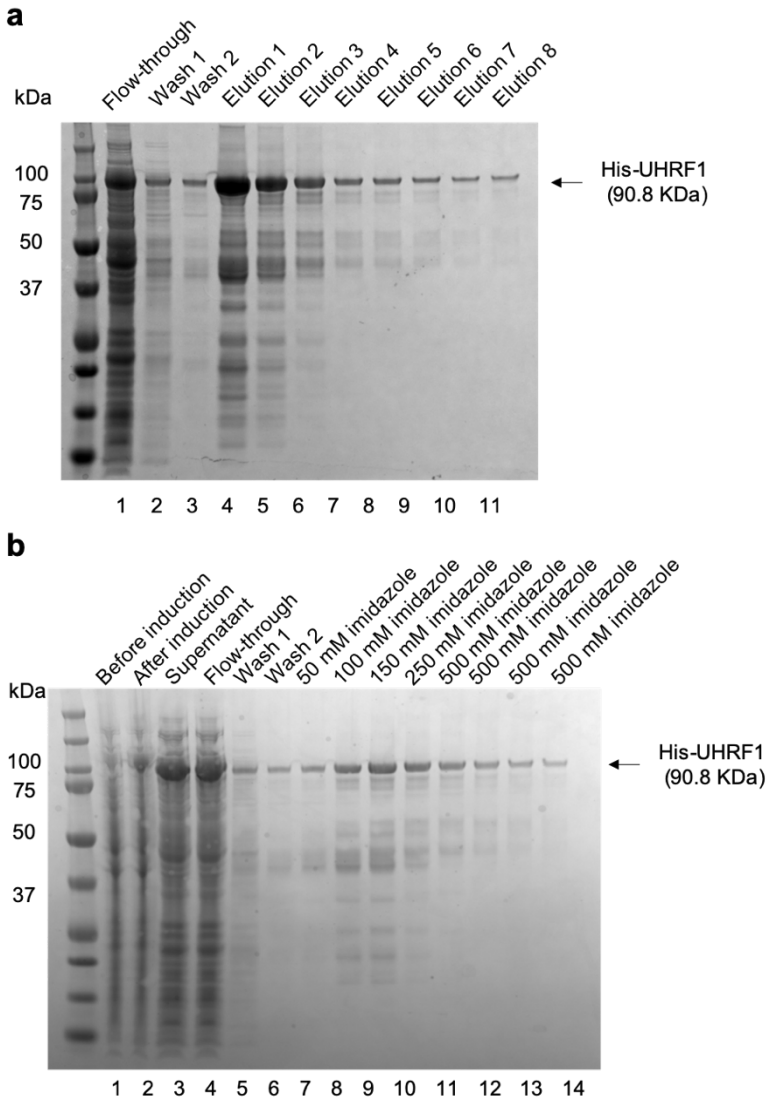


Figure 3-15 His-UHRF1 purification with Ni-NTA resin. a) Coomassie-stained SDS PAGE gel showing samples from His-UHRF1 purification steps. Lane 1: flow-through of cell lysates passing Ni-NTA resin; 2-3, washes of Ni-NTA resin; 4-11, elutions with 500 mM imidazole; Arrow pointing at the band of His-UHRF1 (90.8 kDa). b) Coomassie-stained SDS PAGE gel showing samples from His-UHRF1 purification steps with imidazole gradient elution. Lane 1: whole bacteria lysate before IPTG induction; 2, whole bacteria lysate after IPTG induction; 3, supernatant of cell lysates; 4, flow-through of cell lysates passing Ni-NTA resin; 5-6, washes of Ni-NTA resin; 7-10, elution with 50 mM, 100 mM, 150 mM, 250 mM imidazole; 11-14, elution with 500 M imidazole. Arrow pointing at the band of His-UHRF1 (90.8 kDa).

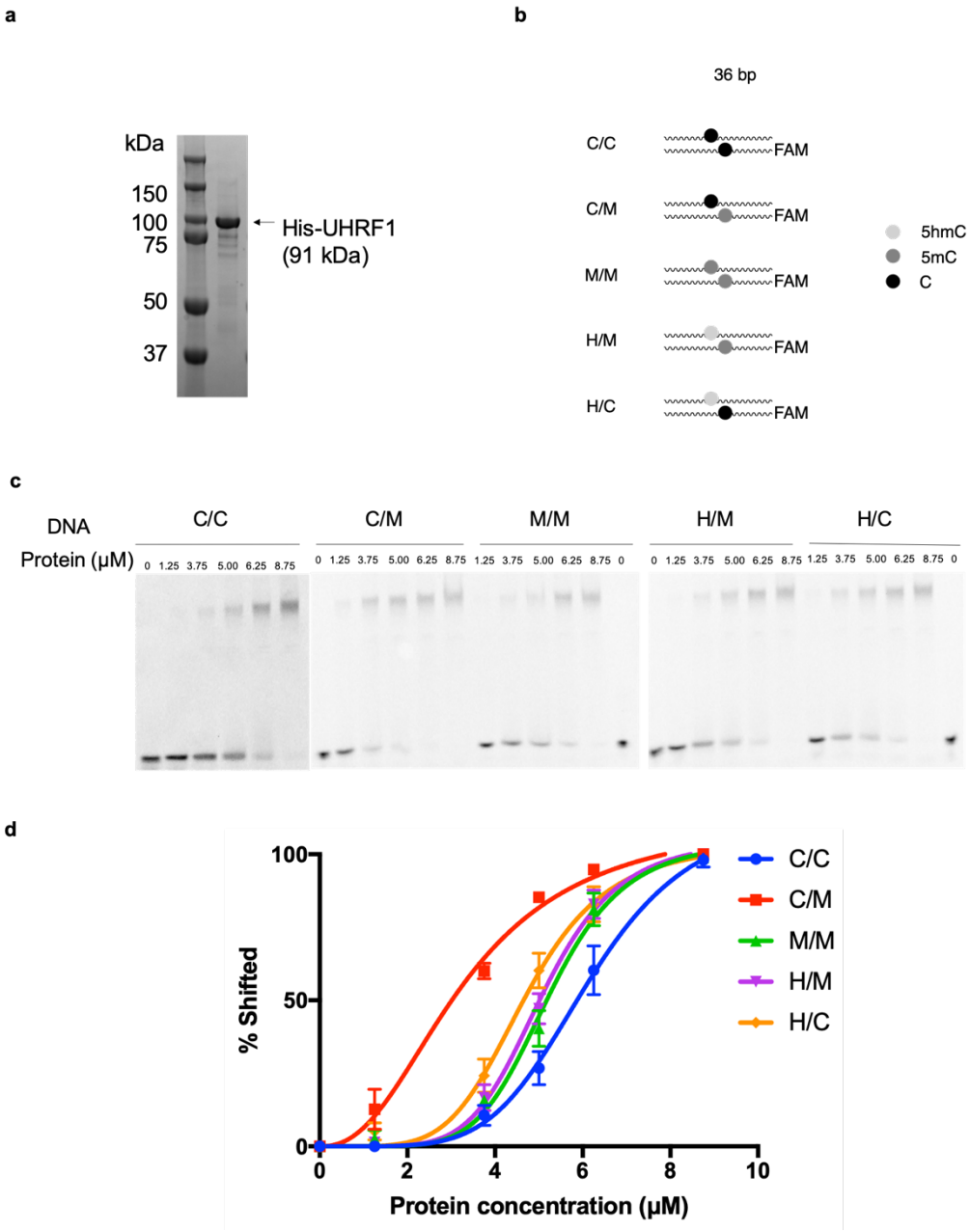


Figure 3-16 EMSA of His-UHRF1. a) SDS-PAGE showing purified His-UHRF1. Arrow pointing at the band of His-UHRF1 (91 kDa). b) A schematic diagram of DNA used in the EMSA. A 36-bp DNA fragment containing C/C, C/M, M/M, H/M or H/C sites was end-labelled with FAM. c) EMSA with increasing amounts of His-UHRF1. d) Quantification of the percentage of shifted bands from EMSA gels. Experiments were done in three replicates. Hill equation was used to fit the curve.

3.3.4 TET1

TET1 is a Fe(II) and α-ketoglutarate (α-KG) dependent dioxygenase that catalyses the oxidation of 5mC to 5hmC, 5fC and 5caC (Ito et al., 2011; Tahiliani et al., 2009). Apart from the catalytic domain, TET1 has a CXXC domain, which was reported to bind to unmethylated CpG sites (Xu et al., 2011; Zhang et al., 2010). In HEK293 cells TET1 was shown to be enriched at unmethylated CpG islands and the enrichment was abolished by mutated CXXC domain (Jin et al., 2014). TET1 was about 2-fold enriched by H/M probe compared to M/M or C/C probe in the MS analysis (Figure 3-17). The binding difference might be mediated by the CXXC domain and may further lead to differences in enzymatic activities. Recombinant full-length TET1 was used to study the enzymatic activities on H/M sites. FLAG-tagged full-length mouse TET1 was expressed in mammalian cells and purified with anti-FLAG affinity resin following a previous report (He et al., 2011) (Figure 3-18 a). The enzymatic activity of TET1 on DNA containing asymmetrically modified H/M or M/H sites was compared with DNA containing M/M or H/H sites. To compare TET activity on each strand separately, one strand of the synthetic DNA was biotin-labelled at the 5’ end. After the reaction, the strand with biotin can be separated from the strand without biotin (Figure 3-18 b). The amounts of 5mC or 5hmC and their oxidation products in each strand were quantified with LC-MS/MS. To compare methods for separating the two strands, a DNA fragment with 5hmC on the biotin strand and 5mC on the non-biotin strand was used. After the DNA fragments were immobilised on streptavidin beads, the non-biotin strand was peeled off by NaOH denaturation or 5' to 3' exonuclease (T7 exonuclease or exonuclease VIII) digestion. The amount of 5mC and 5hmC in the supernatant containing the non-biotin strand and on the beads containing

the biotin strand were quantified with LC-MS/MS to access the separation. If the biotin strand and the non-biotin strand were separated efficiently, all 5hmC will be on the beads and all 5mC will be in the supernatant. Among three methods, the method with T7 exonuclease gave the best separation, with the highest percentage of 5hmC on the beads as well as the highest percentage of 5mC in the supernatant (Table 3-3). After establishing the method, the TET1 reaction was performed on DNA containing asymmetrically modified or symmetrically modified sites, each strand from asymmetrically modified sites was compared with its equivalent from fully methylated or fully hydroxymethylated substrates. There was no significant difference observed between the reaction on H/M or M/H sites compared with symmetrically modified sites (Figure 3-18 c, d).

Table 3-3 Strand separation with different methods

Methods for strand separation	NaOH denaturation		T7 Exonuclease		Exonuclease VIII	
	supernatant	beads	supernatant	beads	supernatant	beads
5hmC percentage	29%	100%	1%	100%	0%	55%
5mC percentage	71%	0%	99%	0%	100%	45%

a

Unique peptides	19
ANOVA *p* value	0.005
T test *p* value	
H/M v.s. M/M	0.05
H/M v.s. C/C	0.0086
Fold change	
(H/M)/(M/M)	1.9
(H/M)/(C/C)	2.8

b

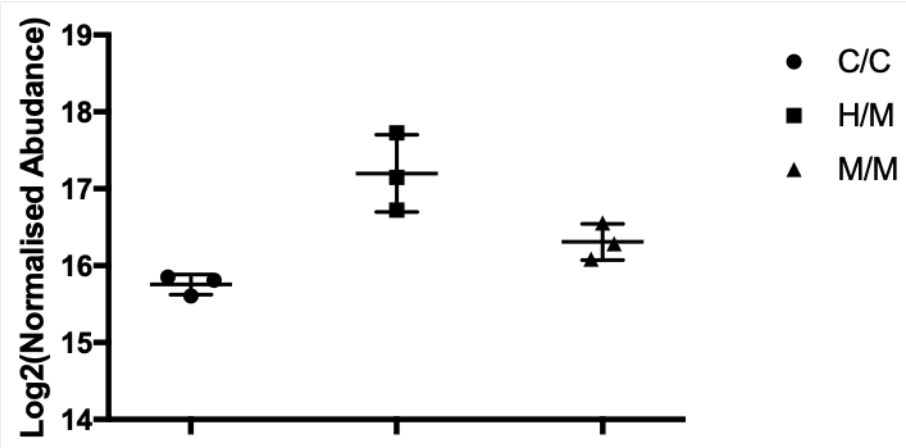
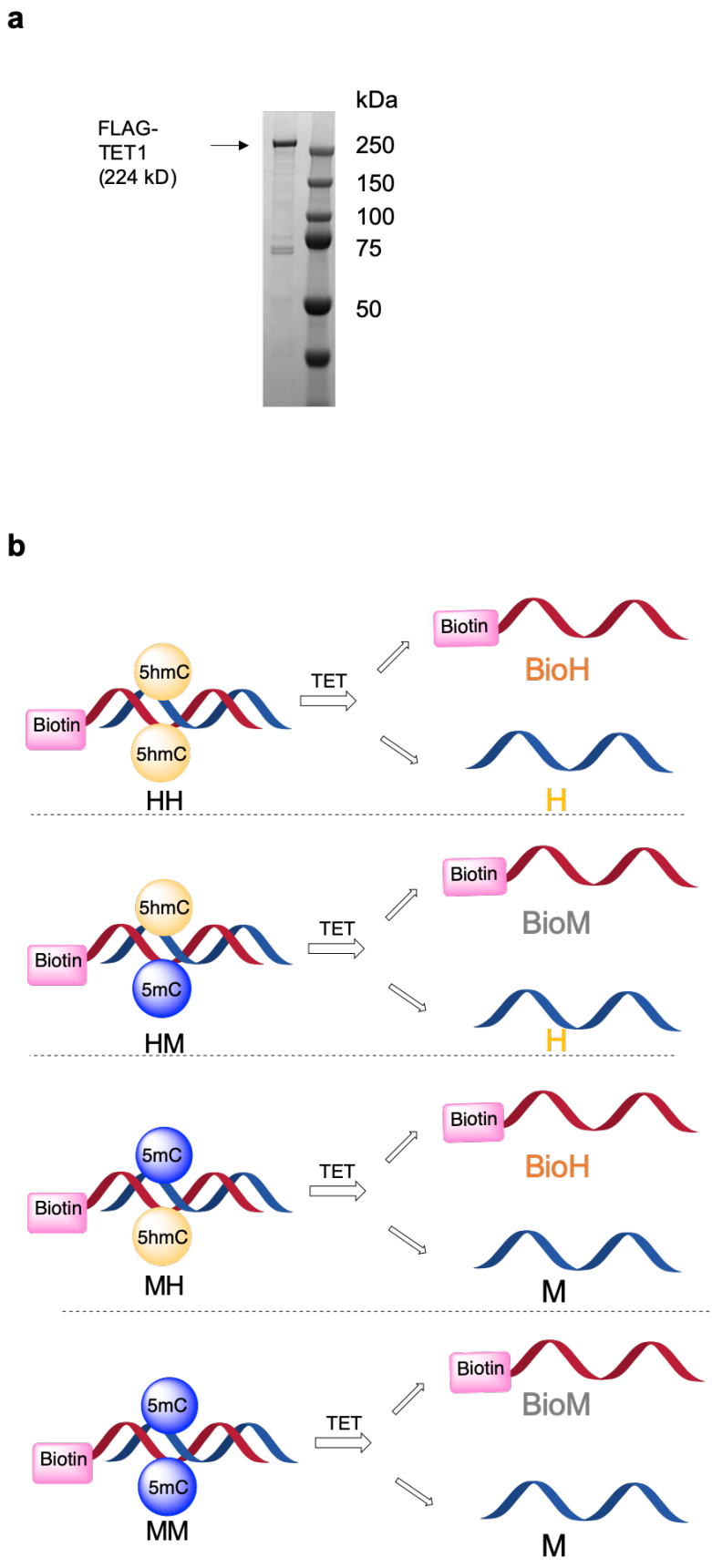
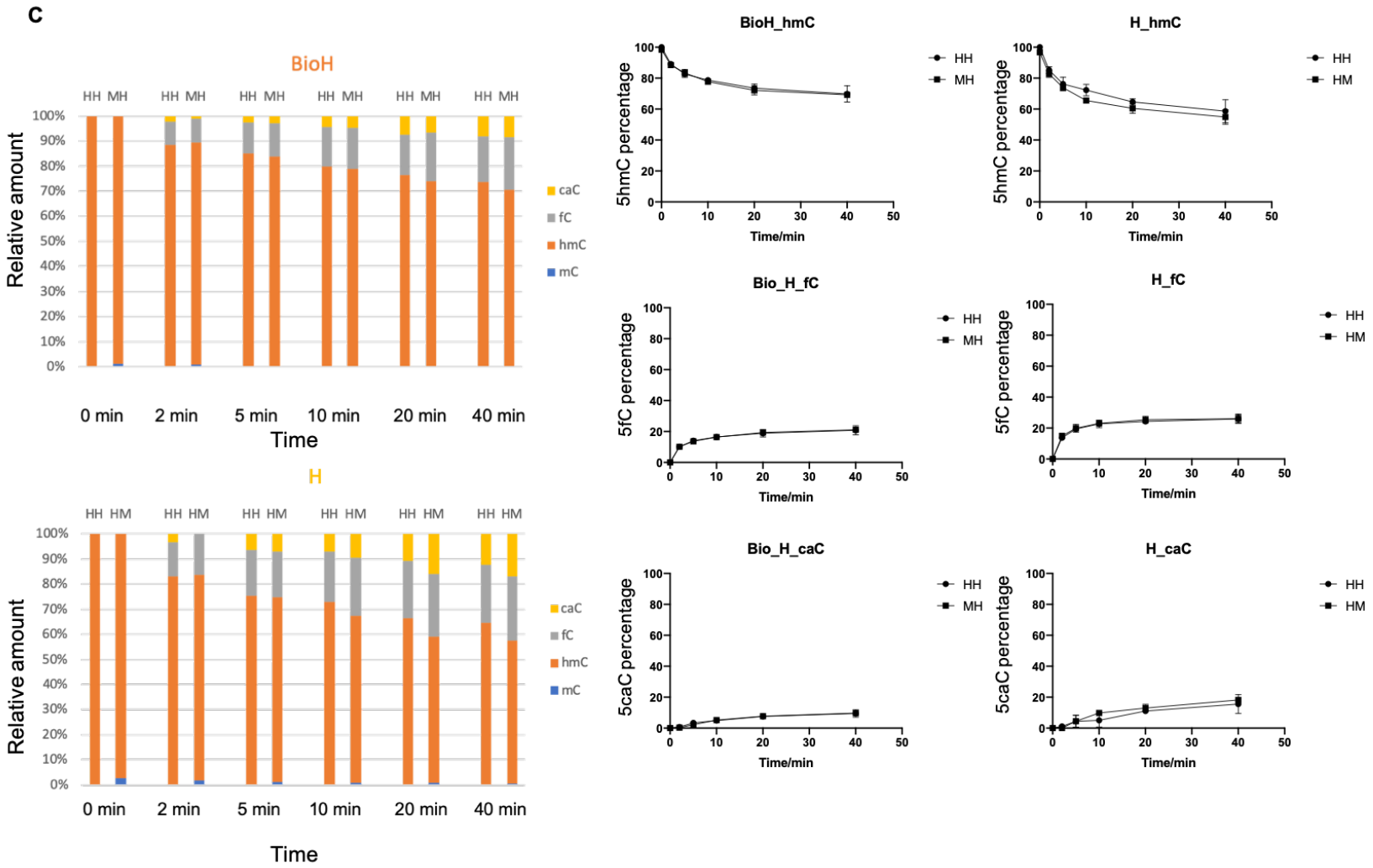


Figure 3-17 Summary of TET1 MS results. a) Summary of MS results and statistics. b) Log_2 abundance of protein pull-downed by DNA probes containing H/M, M/M or C/C sites.


a
FLAG-
TET1
(224 kD)
kDa
250
150
100
75
50
b
Biotin
5hmC
5hmC
HH
TET
BioH
H
Biotin
5hmC
5mC
HM
TET
BioM
H
Biotin
5mC
5hmC
MH
TET
BioH
M
Biotin
5mC
5mC
MM
TET
BioM
M


C
BioH
HH MH HH MH HH MH HH MH HH MH HH MH
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Relative amount
caC
fC
hmC
mC
0 min
2 min
5 min
10 min
20 min
40 min
Time
H
HH HM HH HM HH HM HH HM HH HM HH HM
Relative amount
caC
fC
hmC
mC
0 min
2 min
5 min
10 min
20 min
40 min
Time
BioH_hmC
5hmC percentage
Time/min
HH
MH
H_hmC
5hmC percentage
Time/min
HH
HM
Bio_H_fC
5fC percentage
Time/min
HH
MH
H_fC
5fC percentage
Time/min
HH
HM
Bio_H_caC
5caC percentage
Time/min
HH
MH
H_caC
5caC percentage
Time/min
HH
HM

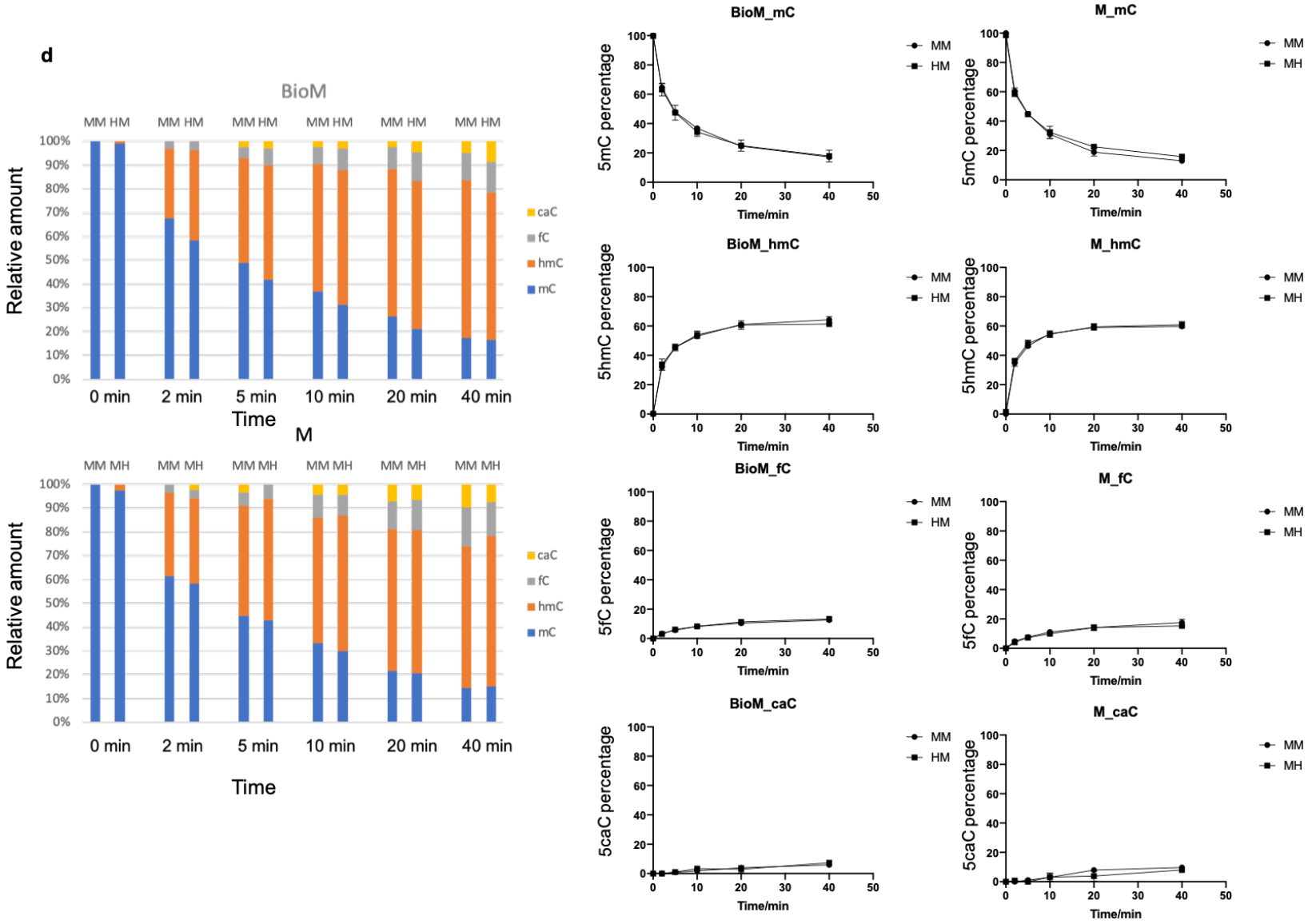

d
BioM
MM HM
MM HM
MM HM
MM HM
MM HM
MM HM
Relative amount
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0 min
2 min
5 min
10 min
20 min
40 min
Time
caC
fC
hmC
mC
M
MM MH
MM MH
MM MH
MM MH
MM MH
MM MH
Relative amount
0 min
2 min
5 min
10 min
20 min
40 min
Time
caC
fC
hmC
mC
BioM_mC
5mC percentage
Time/min
MM
HM
M_mC
5mC percentage
Time/min
MM
MH
BioM_hmC
5hmC percentage
Time/min
MM
HM
M_hmC
5hmC percentage
Time/min
MM
MH
BioM_fC
5fC percentage
Time/min
MM
HM
M_fC
5fC percentage
Time/min
MM
MH
BioM_caC
5caC percentage
Time/min
MM
HM
M_caC
5caC percentage
Time/min
MM
MH

Figure 3-18 TET1 enzymatic activity. a) SDS-PAGE showing purified FLAG-TET1. Arrow pointing at the band of FLAG-TET1 (224 kDa). b) A schematic diagram of the experimental design. DNA containing H/M, M/H, H/H or M/M sites was biotin-labelled on one strand. After the TET1 reaction, the biotin strand was separated from the non-biotin strand and analysed with LC-MS/MS separately. c, d) Comparison of TET1 oxidation of 5hmC (c) or 5mC (d) in H/M site to that of H/H (c) or M/M (d) sites. Left panels: representative results of LC-MS/MS quantification of TET1 oxidation products; right panels: summary of three replicates. Error bars indicate SD (n = 3). ANOVA test was performed for each plot and showed no significant difference.

3.3.5 YTHDF3

YTHDF3 is an N6-methyladenosine (m^6A) binding protein and promotes mRNA translation (Li et al., 2017a; Shi et al., 2017; Yang et al., 2017). There have not been reports about its interaction with DNA. In the DNA pull-down and MS from mESC nuclear extract, it showed about 1.5-fold enrichment by H/M sites compared with H/H and M/M sites (Figure 3-19). Full-length mouse YTHDF3 CDS was cloned into pGTVL2 vector. His-GST tagged YTHDF3 was purified with Ni-NTA resin followed by gel filtration (Figure 3-20 a). EMSA was performed with the purified protein. From the EMSA, the binding of YTHDF3 to H/M sites was not better than that to M/M or C/C sites (Figure 3-20 b, c).

a

Unique peptides	9
ANOVA *p* value	0.023
T test *p* value	
H/M v.s. M/M	0.025
H/M v.s. C/C	0.056
Fold change	
(H/M)/(M/M)	1.7
(H/M)/(C/C)	1.5

b

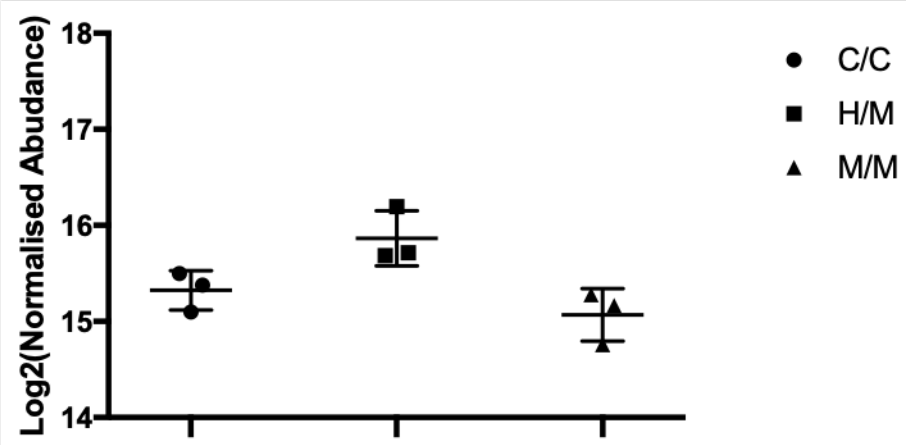


Figure 3-19 Summary of YTHDF3 MS results. a) Summary of MS results and statistics. b) Log_2 abundance of protein pull-downed by DNA probes containing H/M, M/M or C/C sites.

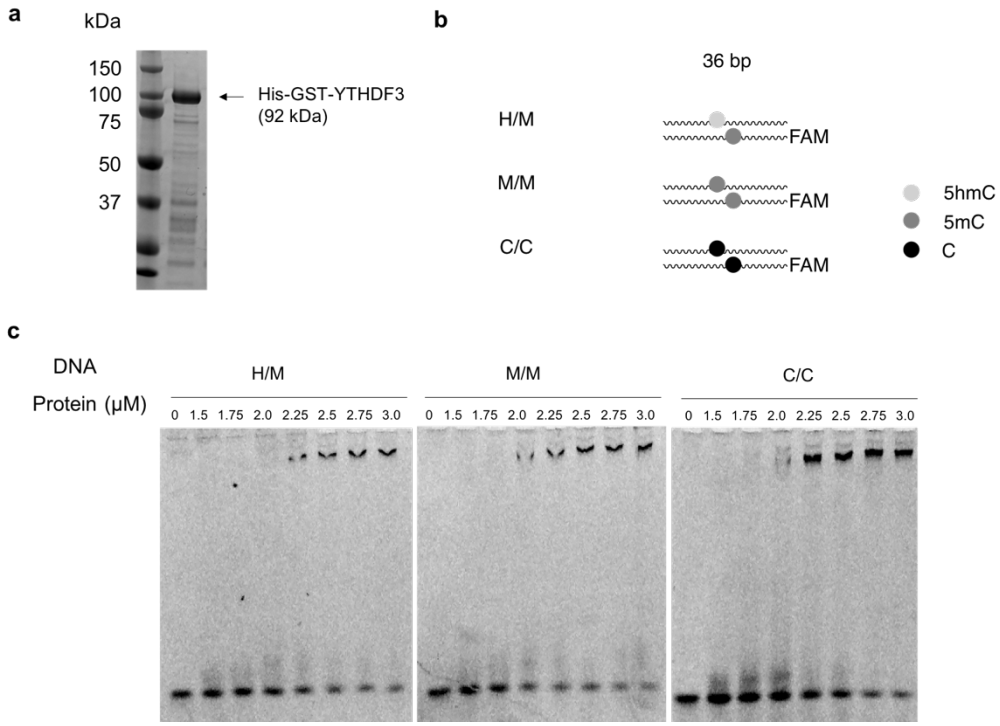


Figure 3-20 EMSA of His-GST-YTHDF3. a) SDS-PAGE showing purified His-GST-YTHDF3. Arrow pointing at the band of His-GST-YTHDF3 (92 kDa). b) A schematic diagram of DNA used in the EMSA. A 36-bp DNA fragment containing H/M, M/M or C/C sites was end-labelled with FAM. c) EMSA with increasing amounts of His-GST-YTHDF3.

3.4 Discussion

3.4.1 DNA pull-down and MS

A 5mC binding protein, MeCP2, was used as a positive control to access the pull-down specificity. Western blot of MeCP2 was performed for all replicates for MS, confirming the enrichment of MeCP2 by the M/M probe (Figure 3-6 a, b). In the MS results from mESCs, MeCP2 passed the ANOVA test and showed enrichment by the M/M probe (Figure 3-8 d). However, in the case of mouse brain samples, though MeCP2 was enriched by M/M probe, it did not pass the statistical test because of the

variation between two replicates of the H/M sample (Figure 3-21). With only two replicates, the variation between samples made statistical test challenging. Mathematically, ANOVA is valid with two replicates. However, samples with a small number of replicates are more prone to be affected by variance from random sampling and result in reduced statistical power. It is possible that there were other true hits missed because of the lack of statistical power. More replicates would favour the identification of more H/M site binding proteins.

a

Unique peptides	136
ANOVA *p* value	0.13
T test *p* value	
M/M v.s. C/C	0.0010
M/M v.s. H/M	0.5
Fold change	
(M/M)/(C/C)	12.6
(M/M)/(H/M)	10.3

b

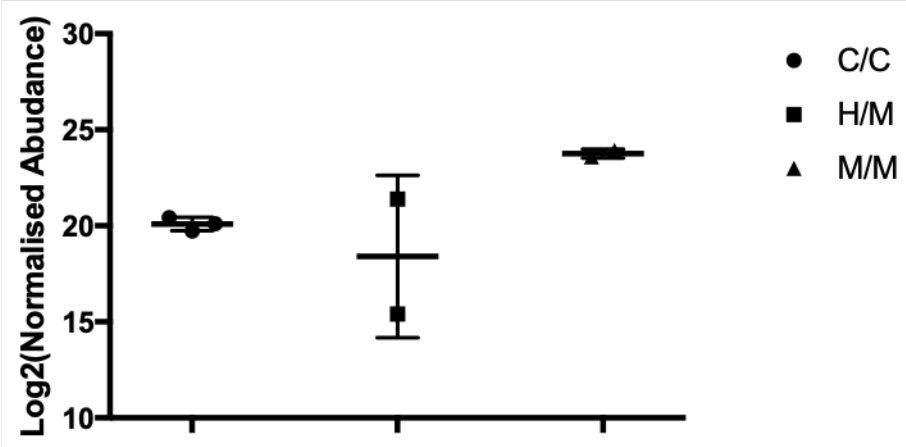


Figure 3-21 Summary of MECP2 MS results of pull-down with mouse brain nuclear extract. a) Summary of MS results and statistics. b) Log_2 abundance of protein pull-downed by DNA probes containing H/M, M/M or C/C sites.

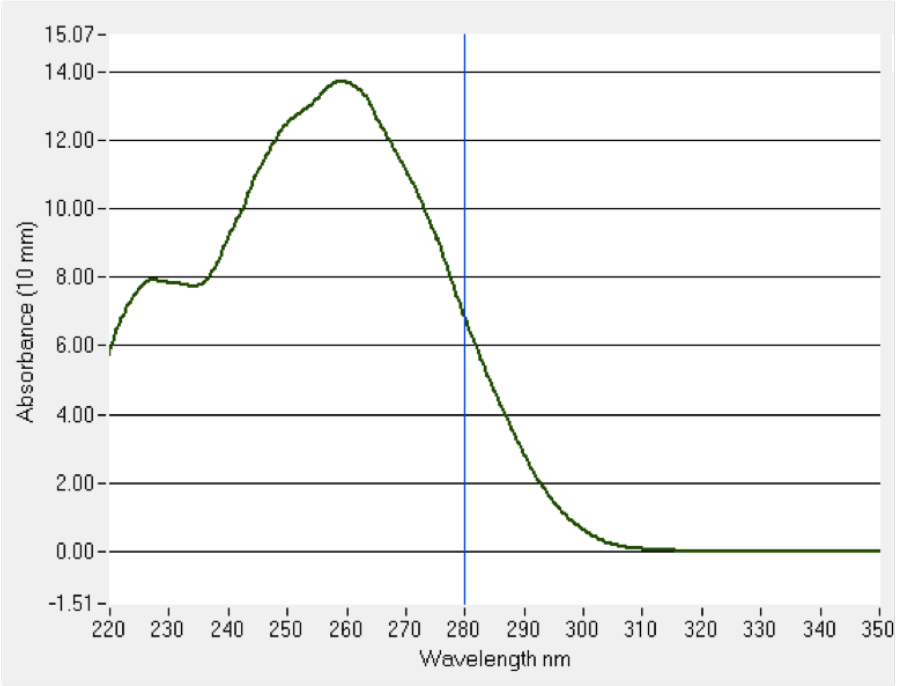
In a previous study, ASH2L and RBBP5 were shown to be enriched by unmodified DNA compared to 5mC or 5hmC containing DNA (Spruijt et al., 2013). In my pull-down, the enrichment of the two proteins by unmodified probe compared to 5mC probe was also observed, but they are more enriched by the H/M probe (Figure 3-11 b). The reason might be that the 5hmC containing DNA probes used in the previous

study was symmetrically modified at CpG sites whereas the 5hmC in the probe I used was in H/M sites.

3.4.2 Validation of candidate H/M site binding proteins

The recombinant His-YBX1 protein tends to be co-purified with nucleic acids when purified from bacteria. When purified with standard Ni-NTA followed by gel filtration procedure, the UV absorption of purified His-YBX1 was dominated by nucleic acids, peaking at 260 nm instead of 280 nm (Figure 3-22 a). Therefore, several purification steps have been modified to remove nucleic acids. The NaCl concentration was increased to 2 M in the lysis buffer to disrupt the interaction between protein and nucleic acids. Benzonase® Nuclease treatment was performed when protein is on the resin. Besides, heparin columns which have a similar structure and competes for binding with nucleic acids were used for purification. With the modified purification procedure, the nucleic acid contamination was reduced but still not completely removed (Figure 3-22 b). In the EMSA, the remaining nucleic acids in the protein might compete for binding with the probes. As the nucleic acid contamination was equally presented in the assay with different probes, the relative binding preference among different probes should not be affected.

a



b

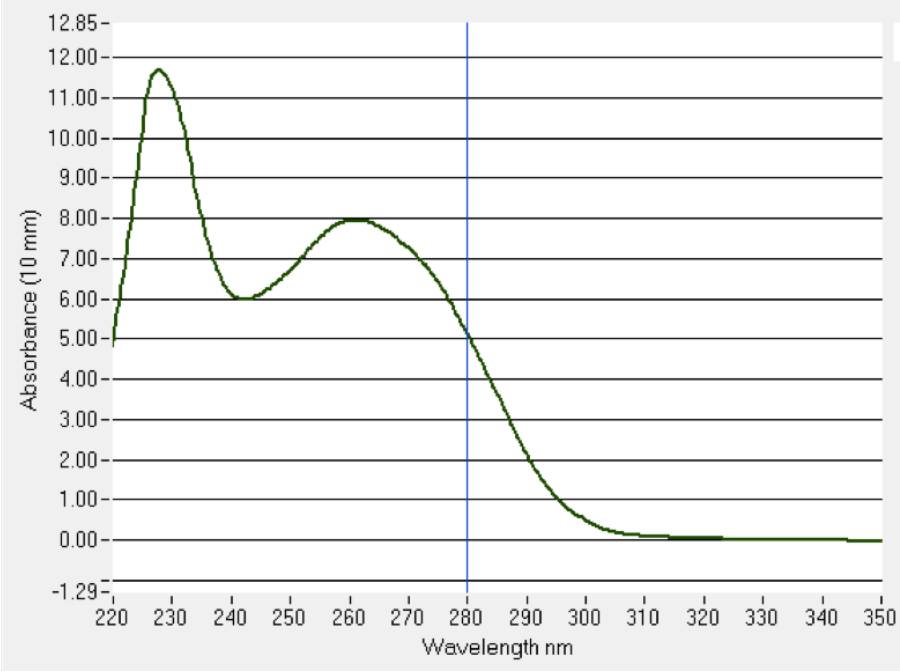


Figure 3-22 An absorbance spectrum of purified His-YBX1 protein with standard purification procedure (a) and with additional steps to remove nucleic acid (b).

The DNA used in the pull-down has a different sequence context from the DNA used in the EMSA. One possible reason why the H/M binding proteins identified from DNA-pull down and MS was not validated by EMSA is that the binding preference may be related to the sequence context. To test this hypothesis, EMSA with the oligos of the same sequence context can be used.

The binding preference of ASH2L identified from DNA pull-down and MS was not reproduced in the EMSA experiments. As this protein is part of a protein complex, it is possible that the binding preference is achieved together with other proteins from the complex, such as RBBP5. To further study this, RBBP5 could be used together with ASH2L in the EMSA.

The EMSA experiment of UHRF1 showed different binding preference from the MS. A previous study with fluorescence polarization assay showed that UHRF1 has similar affinity to H/C, M/M, H/M and H/H sites, which is in-between its affinity to M/C and C/C sites (M/C > H/C = M/M = H/M = H/H > C/C) (Hashimoto et al., 2012). This observation is consistent with my EMSA results (Figure 3-16). In the DNA pull-down experiment, the DNA was incubated with a variety of proteins from the nuclear extract, so the final pull-downed protein represented a combined effect of DNA affinity and competition with other proteins, whereas the EMSA experiment was performed in a more defined system where only the protein of interest was present, so the results were determined by binding affinity. More experiments need to be done to know if UHRF1 function is related to H/M sites in the context of living cells. UHRF1 functions in DNA methylation maintenance by recruiting methyltransferase DNMT1

to hemimethylated sites (Bostick et al., 2007; Sharif et al., 2007). The binding of UHRF1 to H/M sites might be related to its DNA methylation maintenance role. Thus, one way to study the relevance of UHRF1 to H/M sites could be knocking out UHRF1 and studying its effect on the abundance of H/M sites in the cells.

There was no difference observed in the enzymatic activity of TET1 on H/M sites compared with symmetrically modified sites. EMSA needs to be performed to validate the binding preference observed from DNA pull-down. After validating the binding difference, enzymatic activity test with the DNA substrate of the same sequence context as the ones used in DNA pull-down should be done to see if specific sequence context is needed. If there is still no difference in enzymatic activity, it means that the difference in binding affinity may not necessarily lead to different enzymatic activities. It may affect TET1 in other ways like directing the protein to specific regions of the genome.

The protein purification procedure of YTHDF3 requires further improvement to get more conclusive EMSA results. The protein was purified with GST-tag which is a relatively large tag (26 kDa) and may have an effect on the interaction of target protein and DNA. The tag was separated from the protein with a TEV protease cleavage site sequence, but the attempt of removing the tag with TEV was not successful. Alternatively, the protein can be expressed with a smaller tag like His-tag.

4 Chapter 4 Generation of H/M sites

4.1 Introduction

In mammals, DNA methylation patterns are established after fertilization and maintained after cell division. Genome-wide demethylation occurs after fertilization and the methylation pattern is re-established after the blastocyst stage in inner cells mass (Kafri et al., 1992; Monk et al., 1987; Santos et al., 2002). Knockout experiments showed that DNMT3A and DNMT3B are essential for *de novo* methylation and mouse development (Okano et al., 1999). DNA methylation pattern at CpG but not other sequence contexts can be maintained after cell division (Stein et al., 1982; Wigler et al., 1981). DNMT1 is the enzyme for DNA methylation maintenance and prefers hemimethylated CpG as a substrate (Bestor and Ingram, 1983; Gruenbaum et al., 1982). The mutation of DNMT1 cause loss of DNA methylation and failure in mouse embryonic development (Li et al., 1992). DNMT1 is targeted to replication site by its N-terminal targeting sequence (Leonhardt et al., 1992) and it is a component of DNA replication complex. PCNA interacts with DNMT1 and facilitates methylation (Chuang et al., 1997; Iida et al., 2002). UHRF1 binds to hemimethylated DNA and direct DNMT1-mediated DNA methylation maintenance during cell replication (Bostick et al., 2007; Sharif et al., 2007).

The maintenance of 5hmC is less well-studied compared with 5mC. The level of 5hmC is stable during cell cycles. 5hmC is formed not during DNA replication as with 5mC but slowly during the first 30 hours after DNA replication. The timing overlaps

with the S phase of the next cell cycle (Bachman et al., 2014). A study using fully hydroxymethylated episomal DNA found that after several rounds of cell division, 5hmC in CpG but not other sequence context became 5mC, indicating the unmodified cytosine at hemi-hydroxymethylated CpG sites generated by DNA replication was methylated (Kubosaki et al., 2012).

It is largely unknown whether and how H/M sites are maintained. After DNA replication, from an H/M site, one hemi-hydroxymethylated site and one hemimethylated site will be generated (Figure 4-1). DNMT3A and DNMT3B can methylate hemi-hydroxymethylated sites thus converting them back to H/M sites (Hashimoto et al., 2012; Otani et al., 2013). Biochemical studies showed that DNMT1 has much lower activity on hemi-hydroxymethylated sites than hemimethylated sites (Hashimoto et al., 2012; Otani et al., 2013; Seiler et al., 2018; Valinluck and Sowers, 2007). However, studies in cells showed that it may work on hemi-hydroxymethylated sites. Full hydroxymethylated episomal DNA was transfected to 293T cells, after several rounds of cell replication, the DNA became methylated at CpG sites but loss modifications in other sequence contexts (Kubosaki et al., 2012). In 293T cells, DNMT3A and DNMT3B are lowly expressed (Scherf et al., 2000; Thul et al., 2017) so it is likely in the episomal DNA, the hemi-hydroxymethylated sites generated after DNA replication are methylated by DNMT1, though further evidence in DNMT3A/DNMT3B double knockout cells is needed to support the hypothesis.

The hemimethylated site generated from H/M site through DNA replication can be methylated by DNMT1 and become a fully methylated site. To turn into H/M site,

one of the 5mC in the fully methylated site needs to be oxidised to 5hmC by TET enzymes. It is currently unknown if TET proteins have selective activity on a certain strand of double-strand DNA and which ones of TET1, TET2 and TET3 have such selective activity. This chapter aims to study the strand-selective activity of TET enzymes.

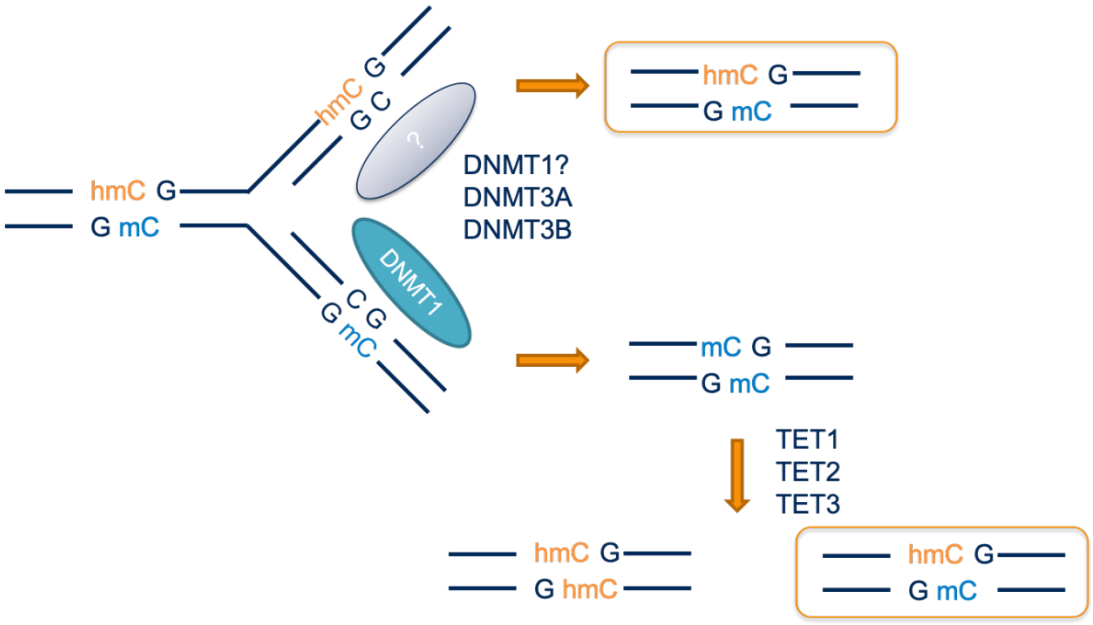
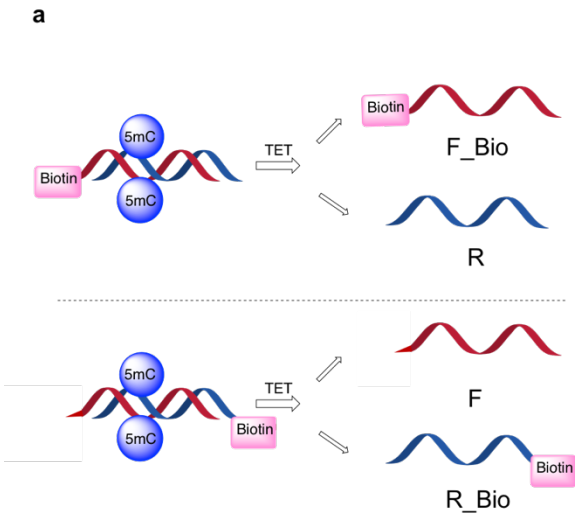


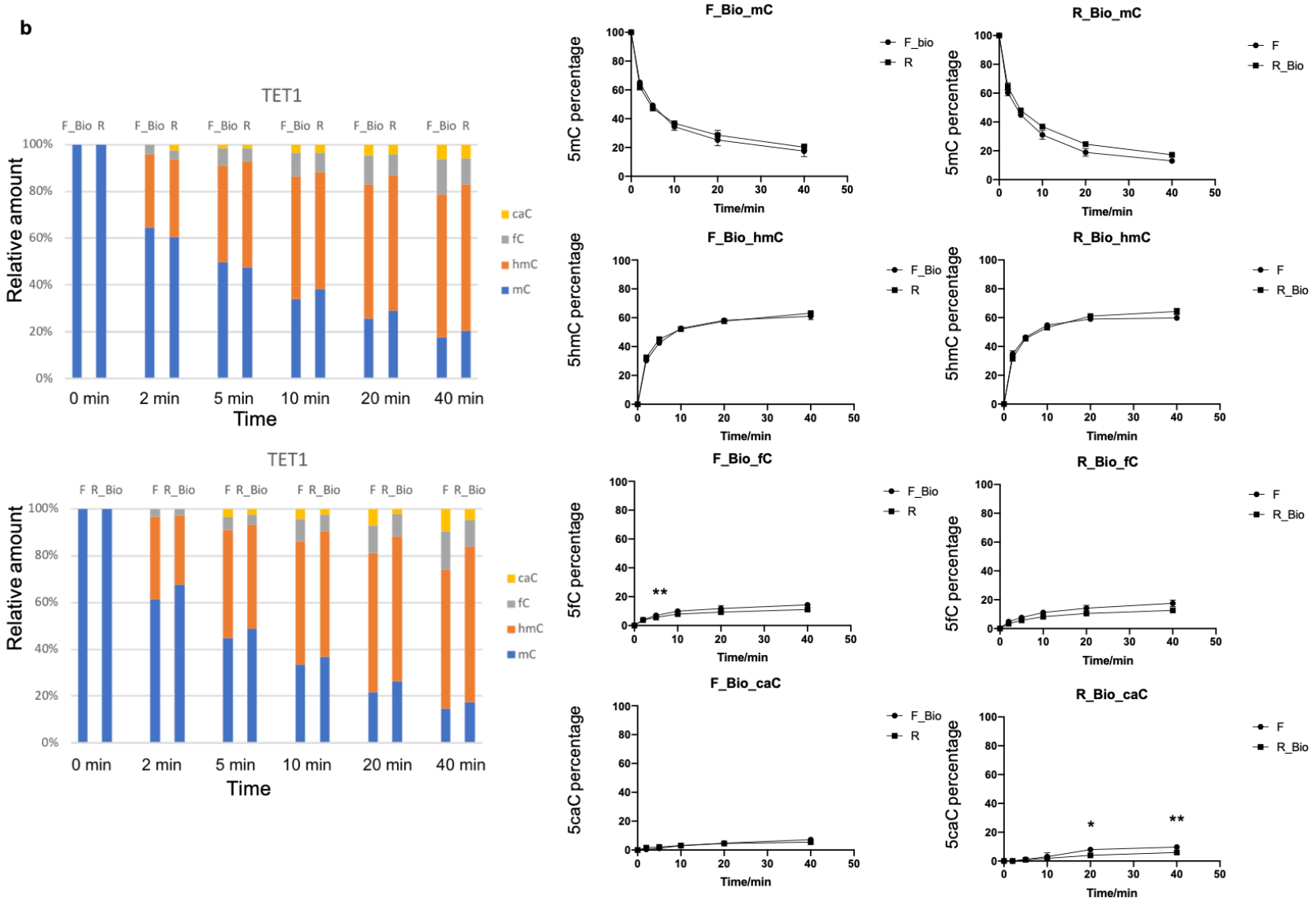
Figure 4-1 A schematic diagram of possible H/M sites maintenance pathways. After DNA replication, hemi-hydroxymethylated sites and hemimethylated sites are generated from H/M sites. Hemi-hydroxymethylated sites can be methylated by DNMT3A/3B and possibly DNMT1. Hemi-methylated sites will become fully methylated after DNMT1 methylation. From the fully methylated sites, it is unknown if any of the TET1, TET2 or TET3 will turn it into an H/M site.

4.2 Results

Local enrichment of guanine was observed on the 5hmC containing strand in base-resolution mapping of 5hmC in embryonic stem cells. The local guanine content

within 100 bp window around 5hmC increase to 29.9%, higher than 25.6% around random 5mC in the genome (Yu et al., 2012b). This leads us to hypothesise that TET may prefer to oxidising the DNA strand with high guanine content. To study the strand preference of TET enzymes on full methylated CpG sites, a 36-bp model DNA containing one methylated CpG site was used. In the model DNA, the forward strand has higher guanine content than the reverse strand (28% in the forward strand and 22% in the reverse strand). To access the reaction on each strand separately, one strand was end-labelled with biotin. After reaction with TET, the biotinylated strand was separated from the non-biotinylated one. The nucleic acid composition of each strand was quantified separately with LC-MS/MS. As the biotin strand and the non-biotin strand were hydrolysed with different methods before LC-MS/MS, to control for the potential bias caused by the hydrolysis methods, DNA with biotin on either forward or reverse strand was used (Figure 4-2 a). In the reaction with TET1, there was no difference between the reaction on two strands (Figure 4-2 b). In the case of TET2, there was a difference in the reaction rate between two strands, which was not affected when the biotin label was moved to a different strand. The difference was at the step from 5hmC to 5fC and 5caC but not the step from 5mC to 5hmC (Figure 4-2 c). In summary TET2 but not TET1 had different reaction rates on two strands.


a
5mC
Biotin
5mC
TET
Biotin
F_Bio
R
5mC
5mC
Biotin
TET
F
Biotin
R_Bio


b
TET1
F_Bio R
Relative amount
100%
80%
60%
40%
20%
0%
0 min
2 min
5 min
10 min
20 min
40 min
Time
caC
fC
hmC
mC
TET1
F R_Bio
F_Bio_mC
R_Bio_mC
F_Bio_hmC
R_Bio_hmC
F_Bio_fC
R_Bio_fC
F_Bio_caC
R_Bio_caC
5mC percentage
5hmC percentage
5fC percentage
5caC percentage
Time/min
F_bio
F_Bio
R
F
R_Bio
**
*

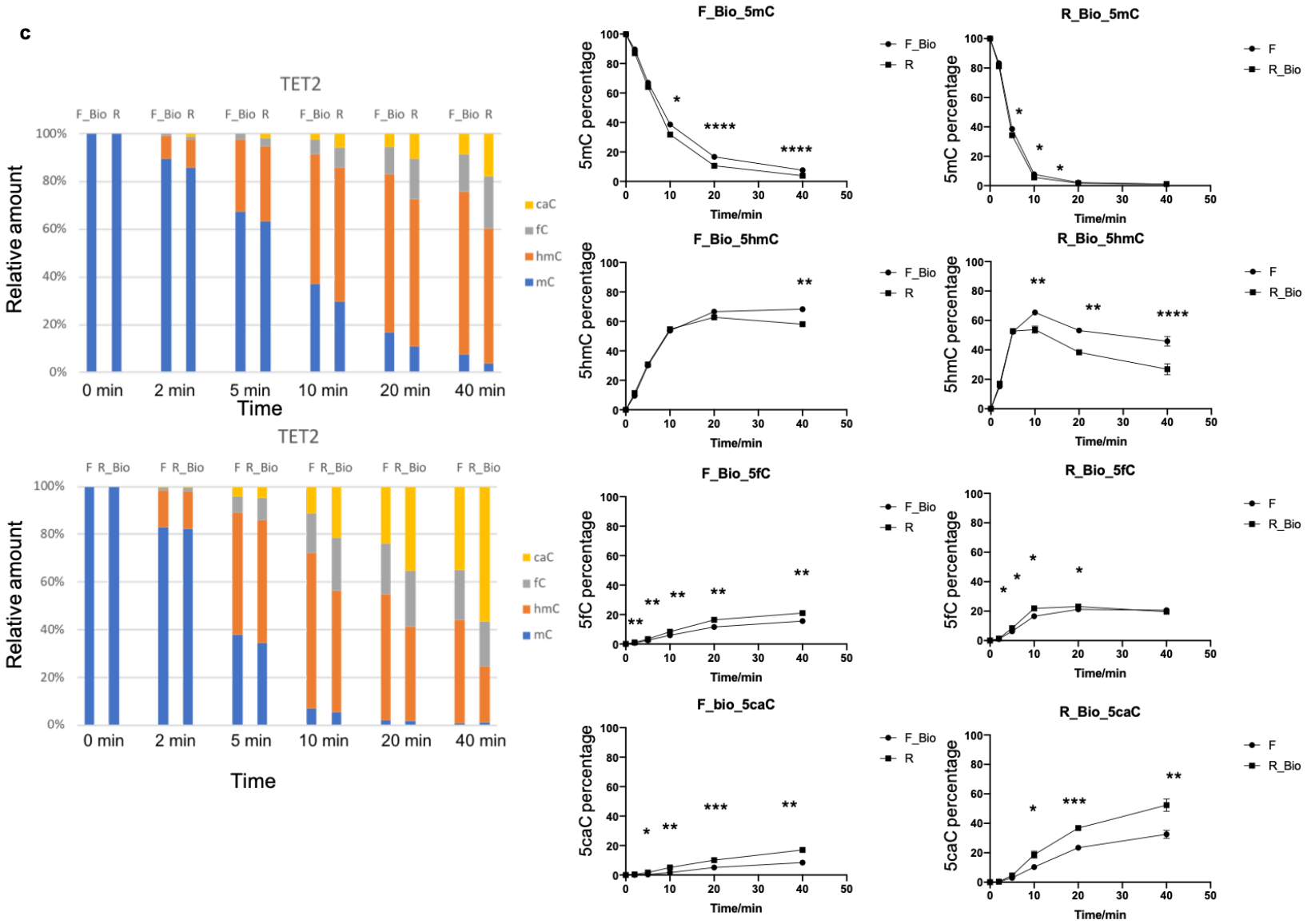

c
TET2
F_Bio R
Relative amount
100%
80%
60%
40%
20%
0%
0 min
2 min
5 min
10 min
20 min
40 min
Time
caC
fC
hmC
mC
TET2
F R_Bio
F_Bio_5mC
R_Bio_5mC
F_Bio_5hmC
R_Bio_5hmC
F_Bio_5fC
R_Bio_5fC
F_bio_5caC
R_Bio_5caC
5mC percentage
5hmC percentage
5fC percentage
5caC percentage
Time/min
F_Bio
R
F
R_Bio

Figure 4-2 Strand-specific activity of TET enzymes. a) A schematic diagram of the experimental design. DNA containing M/M sites was biotin-labelled on either forward stand or reverse strand. After the TET1 reaction, the biotin strand was separated from the non-biotin strand and analysed with LC-MS/MS separately. b, c) Comparison of TET1 (c) and TET2 (d) activities on the two strands from M/M sites. Left panel: representative results of LC-MS/MS quantification of TET1 (b) and TET2 (c) oxidation products, comparing the two strands from the same DNA; right panel, a summary of three replicates. Error bars indicate SD (n = 3). t-test with correction for multiple testing with Holm-Sidak methods was used to compare between two strands at each time point. Adjusted *p*-value was marked with asterisks where the test showed significant differences. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$

4.3 Discussion

The current method measured the reaction on each strand in bulk. The observation that there were similar reaction rates on the two strands can be explained by the conversion of M/M sites to H/H sites, but it can also be explained if similar amounts of M/H and H/M sites were formed. To determine which was the case, the single molecular imaging methods which quantify H/M site directly could be used (Song et al., 2016).

The mechanism of the different activity of TET2 on two strands requires further study. The difference observed in the current model could be caused by guanine content or sequence context. To further study the effect of guanine context, model DNAs with

various sequence context but the same guanine content could be used. To study the effect of sequence context, model DNAs with swapped local sequence context on forward and reverse strand could be used.

If the observation that TET2 reacts faster from 5hmC to 5fC and 5caC on one strand than the other can be generalised after studying more model DNA, it means TET2 turns symmetrical 5hmC/5hmC sites into asymmetrical 5hmC/5fC, 5hmC/5caC or 5hmC/C sites. To further understand the biological relevance of the biochemical observation, the effect of TET2 enzymes on symmetrical 5hmC/5hmC site could be studied in knockout cells.

5 Chapter 5 5hmC detection methods

5.1 Cell-free DNA 5hmC signature and cancer detection

Cell-free DNA (cfDNA) is DNA found in the bloodstream, which is mainly released by cell death. Tumour DNA also contributes to cfDNA, which provides a non-invasive window for cancer detection and monitoring (Wan et al., 2017). The fraction of tumour DNA in cell-free DNA is correlated with tumour size in nude mice injected human cancer cells (Kamat et al., 2006; Thierry et al., 2010). Tumour DNA is detectable in cfDNA from more than 75% of patients with advanced pancreatic, ovarian, colorectal, bladder, gastroesophageal, breast, melanoma, hepatocellular, and head and neck cancers, but less than 50% of primary brain, renal, prostate, or thyroid cancers. In case of localised tumours, tumour DNA was detected in cfDNA from 73, 57, 48, and 50% of patients with colorectal cancer, gastroesophageal cancer, pancreatic cancer, and breast adenocarcinoma, respectively (Bettegowda et al., 2014). In bladder cancer patients, mutations were shown to be detectable before the diagnose of cancer (Gormally et al., 2006). Therefore, cell-free DNA could potentially be used for diagnostic and early detection of cancer.

In cancer cells, there is a global loss and focal gain of 5mC. DNA hypermethylation at promoters and enhancers of tumour-suppressor genes was observed in various cancers (Vidal et al., 2017). 5mC has tissue-specific distribution in the genome (Dor and Cedar, 2018). The alteration of 5mC profiles could be used as a biomarker for cancer and tissue of origin. Studies have shown the potential of cell-free DNA

methylation as biomarkers for hepatocellular carcinoma (Chan et al., 2013; Sun et al., 2015) and for cancer detection and classification (Shen et al., 2018).

5hmC can also be used as a biomarker for cancer. Low levels of 5hmC are observed in haematological malignancies (Ko et al., 2010) and a wide range of solid tumours (Vasanthakumar and Godley, 2015), including lung cancers, pancreatic cancers (Yang et al., 2013), melanoma (Lian et al., 2012), hepatocellular carcinoma (Liu et al., 2013a; Yang et al., 2013), adenocarcinoma of colon (Haffner et al., 2011), breast cancers, prostate cancers (Haffner et al., 2011; Yang et al., 2013) and gliomas (Muller et al., 2012; Orr et al., 2012). Several studies of cell-free DNA 5hmC signature have shown that it could be a biomarker for various cancers such as oesophageal cancer (Tian et al., 2018) and non-small-cell lung cancer (Zhang et al., 2018). 5hmC exhibits a tissue-specific distribution (Globisch et al., 2010). Recent studies showed that genome-wide profiling of 5hmC in cell-free DNA could also be used to distinguish different cancer types and stages (Li et al., 2017b; Song et al., 2017). Therefore, 5hmC in cell-free DNA can be potentially used as a biomarker to identify cancer and tissue of origin.

5.2 Base-resolution 5mC and 5hmC mapping methods

The gold standard for base-resolution mapping is bisulfite sequencing, in which unmodified cytosine is converted to uracil and further recognised as T by DNA polymerase in the PCR (Frommer et al., 1992). Both 5mC and 5hmC are resistant to bisulfite conversion (Huang et al., 2010; Jin et al., 2010; Nestor et al., 2010). Therefore, in bisulfite sequencing, both 5mC and 5hmC are mapped as C while

unmodified cytosine is mapped as T. Modified bisulfite sequencing methods are developed to map 5mC and 5hmC separately, including oxidative bisulfite sequencing (oxBS-Seq) (Booth et al., 2012) and TET-assisted bisulfite sequencing (TAB-Seq) (Yu et al., 2012a; Yu et al., 2012b). In oxBS-Seq, 5hmC is specifically oxidised to 5fC by potassium perruthenate ($KRuO_4$), and converted to uracil after bisulfite treatment, allowing for discrimination of 5hmC and 5mC (Table 5-1) (Booth et al., 2012). With this method, 5mC is mapped directly, and the mapping of 5hmC is achieved by subtracting readout of oxBS-Seq (5mC) from those of conventional bisulfite sequencing (5mC plus 5hmC). Another method, TAB-Seq was developed to map 5hmC directly. In TAB-Seq, 5hmC is protected by glycosylation via β-glucosyltransferase (βGT) and 5mC is oxidised to 5caC by recombinant TET1 protein. In subsequent bisulfite treatment, both unmodified cytosine and 5caC are converted while 5hmC remains unchanged. Thus, in the sequencing, only 5hmC is read as C (Table 5-1) (Yu et al., 2012a; Yu et al., 2012b).

Previous studies on cell-free DNA 5hmC profiles were done with enrichment methods, where 5hmC containing DNA fragments were enriched and analysed by next-generation sequencing. Such methods do not provide base resolution and quantitative information. 5mC mapping in cell-free DNA was done with both enrichment method (Shen et al., 2018) and bisulfite sequencing (Chan et al., 2013; Sun et al., 2015). However, bisulfite sequencing causes significant DNA degradation (Grunau et al., 2001). The amount of cell-free DNA materials is limited, which is of 1 to 10 ng per mL plasma (Wan et al., 2017). To maximise the 5mC and 5hmC

information that can be obtained from cell-free DNA, a base-resolution mapping method that causes less DNA degradation than bisulfite sequencing is needed.

Table 5-1 Base conversion of bisulfite sequencing-based methods

	C	**5mC**	**5hmC**	**5fC**	**5caC**
Bisulfite sequencing	T	C	C	T	T
oxBS-seq	T	C	T	T	T
TAB-seq	T	T	C	T	T

5.3 Peroxotungstate selective 5hmC oxidation

The strategy to map 5hmC at base-resolution is to use a chemical reaction that selectively reacts with 5hmC to make it distinguishable from other forms of cytosine in subsequent sequencing. It is reported that a metal oxidant, dinuclear peroxotungstate selectively oxidises 5hmC to trihydroxylated thymine while 5mC and C remain unchanged (Okamoto et al., 2011). During PCR, the oxidised product is recognised as thymine, making it distinguishable from 5mC and C and enabling the base-resolution mapping of 5hmC (Figure 5-1).

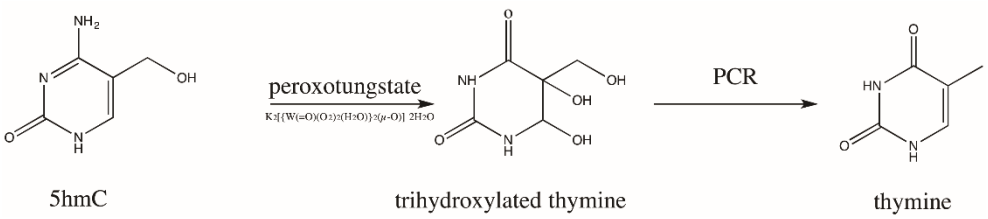


Figure 5-1 Selective conversion of 5hmC to thymine. 5hmC could be selectively oxidised to trihydroxylated thymine and further converted to thymine after PCR, allowing 5hmC to be distinguished from other forms of cytosine.

5.4 Reaction with single-strand DNA

Two model DNA fragments were used to determine the conversion rate of the reaction on single-strand DNA. To quantify the conversion rate, two strategies were employed. One was to use restriction enzyme digestion to detect the change of 5hmC to thymine. The synthetic DNA sequence contains a restriction enzyme cutting site, which is hydroxymethylated on cytosine. Before the reaction, this site could be recognised and cut by the restriction enzyme. After the oxidation reaction and PCR, if the 5hmC is successfully converted to thymine, the restriction enzyme cutting site will be disrupted and become resistant to the restriction enzyme digestion. The conversion rate could be semi-quantified from the band density of digested and undigested DNA fragments from an agarose gel. Both model DNA showed similar conversion rates of about 80% (Figure 5-2). A more quantitative way to detect the change of 5hmC to thymine was to clone the DNA fragment into a vector and analyse it by Sanger sequencing. The overall conversion rate is 81% on DNA1 and 62.5% on DNA2 (Figure 5-3). In the restriction enzyme assay, the sites analysed were the middle site in DNA1 and the first site in DNA2. The Sanger sequencing showed consistent quantification with restriction enzyme assay. The lower overall conversion rate of DNA2 was mainly resulted from the middle site (Figure 5-3), suggesting that certain sites were more difficult to be converted than others. In summary, the conversion of 5hmC by peroxotungstate can be achieved on the single-strand DNA models.

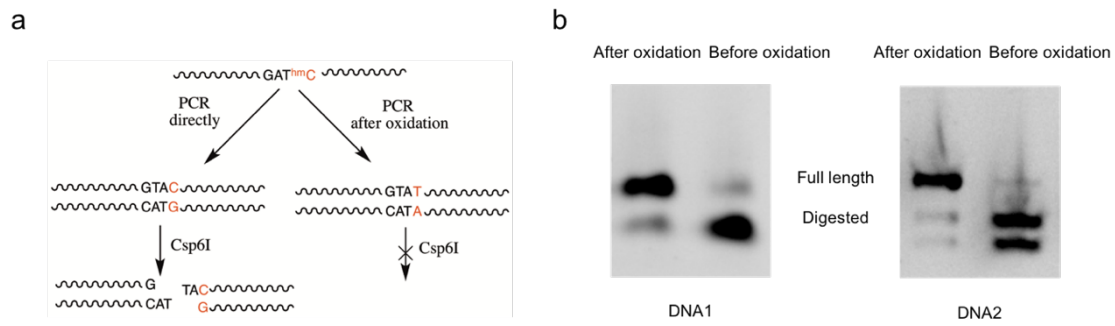


Figure 5-2 Semi-quantification of the conversion rate of 5hmC to thymine. a) A schematic diagram of the method. If 5hmC is converted to thymine, the recognition site of restriction enzyme Csp6I will be disrupted and resist to digestion. b) The conversion rate on single-strand DNA. Two DNA sequences were tested. The top bands are undigested DNA and the bottom bands are DNA fragments after digestion. In DNA1, the two fragments after digestion are of the same size.

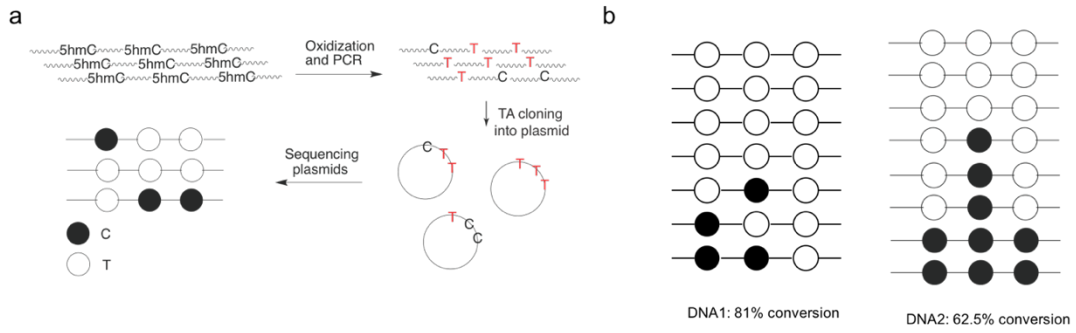


Figure 5-3 Quantification of the conversion rate using Sanger sequencing. a) A schematic diagram of the method. The DNA fragments after oxidation and PCR are cloned into sequencing plasmids and analysed by Sanger sequencing. b) Quantifications of the conversion rate on single-strand DNA. Two DNA sequences are tested. Each row represents the sequencing results from a DNA fragment with three 5hmC sites. Black dots indicate C in the sequencing, whereas white dots indicate T.

5.5 Reaction optimisation on double-strand DNA

After the selective conversion of 5hmC to thymine were verified on the single-stranded DNA models, the reaction conditions were optimised to achieve a high conversion rate on double-stranded DNA. Studies have shown the peroxotungstate reaction has a lower conversion rate of double-strand DNA compared to single-stranded DNA (Hayashi et al., 2016; Okamoto et al., 2011). Given the high conversion rate of single-stranded DNA, we hypothesised that increasing the conversion rate of double-stranded DNA could be achieved by using a reaction condition that could keep the DNA in a single-strand form as much as possible. Different strategies were tested to keep the DNA single-stranded. The DNA fragments used for testing are synthetic DNA fragments of the same sequence, which makes it easier to anneal back to each other than the heterogeneous genomic DNA sample. To better mimic genomic DNA, the synthetic DNA fragments were spiked into fragmented lambda phage DNA. The conversion rate was slightly improved. Different reaction conditions were tested to improve the denaturation of DNA. It is reported that the renaturation of DNA is facilitated by metal ions and prevented by DMSO and low pH (Wang et al., 2014). Therefore, several reaction conditions were tested, including decreasing the salt concentration of the buffer, increasing buffer pH, increasing reaction temperature, different buffers (such as phosphate buffer, MES buffer, Tris buffer), reaction with DMSO or betaine and reaction with crown ether that chelates metal ions. Zn^{2+} is reported to block the renaturation of DNA (Lu, 2014), so the addition of Zn^{2+} to the reaction was also tested. However, none of these substantially increased the conversion rate. The buffers screened and conversion rates were summarised in Figure

5-4. As the experiments were performed for screening propose, each condition was repeated only once. To reduce the rate of renaturation, DNA fragments were end-labelled with biotin and immobilised on beads after denaturation. This increased the conversion rate to approximately 30% (Figure 5-5). Prof. Colin Goding kindly suggested that dilution can be used to prevent DNA renaturation. To test this idea, DNA was end labelled with biotin, diluted 100 times before denaturation and kept diluted during the reaction. After the reaction, DNA was recovered on streptavidin beads. With this method, the conversion rate was comparable with reactions on single-strand DNA (Figure 5-6 a).

The other challenge with this reaction was that the conversion rate was not uniform. In the synthetic DNA model, there are three 5hmC sites. The conversion rate on one of the sites was roughly half that of the other two, as shown with both restriction enzyme digestion and Sanger sequencing (Figure 5-6 b, c). The causes of such differences are not clear. One reason could be the possible sequence or structure preference of DNA polymerases, so several DNA polymerases were compared. However, the difference in conversion rate persisted regardless of the polymerases used (Figure 5-6 d). Another possibility might be interference from DNA secondary structure. To reduce secondary structure, betaine was added to the oxidation reaction, but the difference between the two sites was not reduced (Figure 5-6 e). To further understand the cause of the uneven conversion rate, another synthetic DNA sequence was designed based on the sequence of DNA2. In the new DNA sequence (DNA3), the positions of the high conversion rate site and the low conversion rate site in DNA2

were swapped and sites with different sequence contexts were added. The uneven conversion rate was also observed in DNA3 (Figure 5-6 f).

a

Single strand DNA

Reaction Buffer	Conversion Rate
H_2O	72%
0.05 M NaOH	10%
50% DMSO	9%
2.5 M betaine	37%
50 mM phosphate buffer	86%
50 mM phosphate buffer, 100 mM NaCl	90%

b

Double strand DNA, reduce salt

Reaction Buffer	Conversion Rate
H_2O	35%
50mM phosphate buffer, 100mM NaCl	25%

c

Double strand DNA, pH

Reaction Buffer	Convresion Rate
10 mM Tris pH 7.5	48%
10 mM Tris pH 7	27%
10 mM Tris pH 8	17%
10 mM Tris pH 9	16%

d

Double strand DNA, Buffer and Zn^{2+}

Reaction Buffer	Conversion Rate
10mM PIPES PH 7	9%
10mM PIPES PH 7 + 1mM $ZnSO_4$	8%
10mM HEPES PH 7	7%
10mM MES PH 6	27%
10mM Tris PH 7.5	61%
10mM Tris PH 7.5 + 1mM $ZnSO_4$	22%
10mM Tris PH 7.5 + 10mM NaCl	45%
10mM Tris PH 7.5 + 10mM NaCl + 1mM $ZnSO_4$	7%

Figure 5-4 Summary of buffers screened for the reaction. The conversion rate was quantified from gel images of restriction enzyme digestion tests. Each condition was repeated only once. a) The effect of buffer to the conversion on single-strand DNA. b) Reactions in water gave a better conversion than the reaction in Na^+ containing buffer. c) The effect of buffer pH. d) The effect of buffer type and the addition of Zn^{2+}.

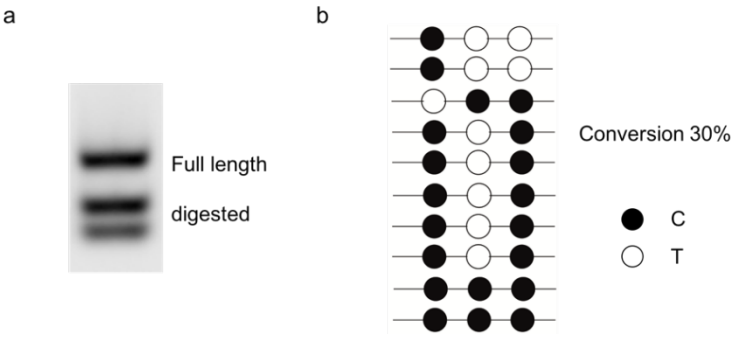


Figure 5-5 Conversion rate of reaction after immobilising denatured DNA on beads. a) Semi-quantification with restriction enzyme digestion. b) Quantification with Sanger sequencing. Each row represents the sequencing results from one DNA fragment with three 5hmC sites. Black dots indicate C in the sequencing, whereas white dots indicate T.

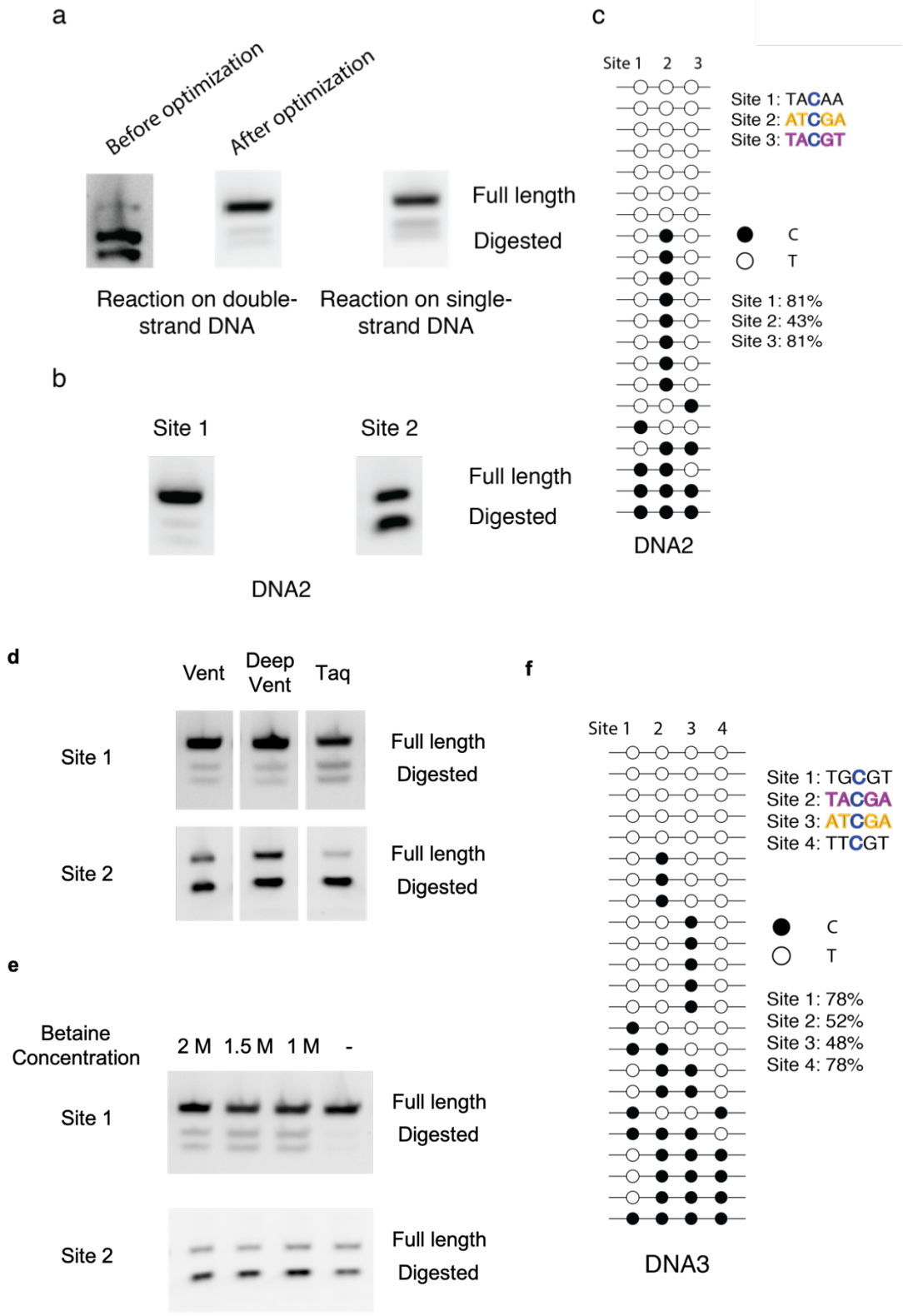
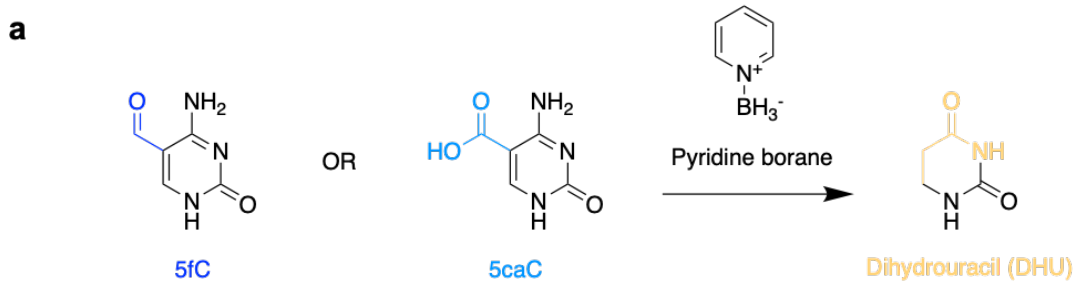

a
Before optimization
After optimization
Full length
Digested
Reaction on double-strand DNA
Reaction on single-strand DNA
b
Site 1
Site 2
Full length
Digested
DNA2
c
Site 1 2 3
Site 1: TACAA
Site 2: ATCGA
Site 3: TACGT
C
T
Site 1: 81%
Site 2: 43%
Site 3: 81%
DNA2
d
Vent
Deep Vent
Taq
Site 1
Site 2
Full length
Digested
e
Betaine Concentration
2 M 1.5 M 1 M -
Site 1
Site 2
Full length
Digested
f
Site 1 2 3 4
Site 1: TGCGT
Site 2: TACGA
Site 3: ATCGA
Site 4: TTCGT
C
T
Site 1: 78%
Site 2: 52%
Site 3: 48%
Site 4: 78%
DNA3

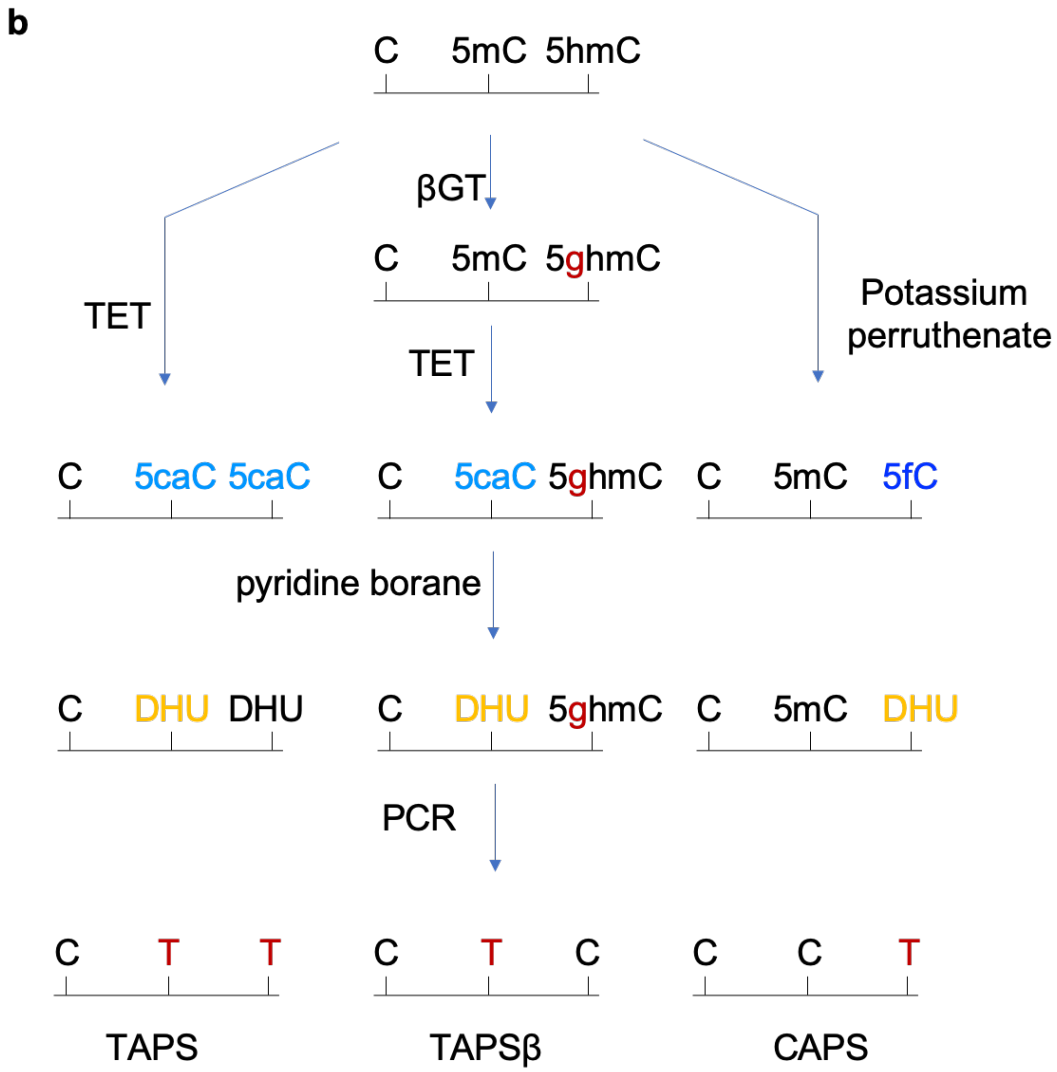
Figure 5-6 Optimisation of the reaction on double-strand DNA. a) Improved conversion rate is shown from restriction enzyme analysis. b, c) An uneven conversion rate observed on DNA2 demonstrated by a restriction enzyme analysis (b) and Sanger sequencing (c). d) Conversion rates of PCR with different DNA polymerases: Vent® (exo-) DNA polymerase, Deep Vent® (exo-) DNA polymerase and Taq DNA polymerases. e) Conversion rates with the addition of betaine to the oxidation reaction. f) An uneven conversion rate observed on DNA3 shown by Sanger sequencing. c, f) Each row represents the sequencing results from a DNA fragment. Black dots indicate C in the sequencing, whereas white dots indicate T.

5.6 TET-assisted pyridine borane sequencing

To map 5mC at base-resolution without using bisulfite reaction, Yibin Liu in our lab developed a chemical reaction using pyridine borane to reduce 5fC and 5caC to dihydrouracil (DHU) (Figure 5-7 a), which would be recognised as T by polymerase during PCR. When coupling with TET oxidation of 5mC and 5hmC to 5caC, the pyridine borane reaction can be used for 5mC and 5hmC mapping in a procedure termed TET-assisted pyridine borane sequencing (TAPS) (Figure 5-7 b). I purified FLAG-tagged mouse TET1 catalytic domain (mTET1CD) from 293T cells (Figure 5-7 c). The activity of mTET1CD was tested on an 11 bp double-strand model DNA with one 5mC. After the reaction with mTET1CD, a 30 Da increase of molecular weight was observed on the model DNA, indicating the complete oxidation of 5mC to 5caC (Figure 5-7 d). Thus, the mTET1CD can be used for TABS in 5mC sequencing. The expression and purification of mTET1CD were further optimised in Exp293F™ Cells for large-scale production by Paulina Siejka-Zielinska. The TAPS method can also be extended to sequence 5mC or 5hmC specifically. To sequence

5mC, 5hmC was protected by glucose with βGT, enabling for specific TET oxidation and subsequent mapping of 5mC. This method is named TAPSβ (Figure 5-7 b). 5hmC sites can be deduced from comparing the readout from TAPS and TAPSβ. Alternatively, 5hmC can also be mapped with chemical-assisted pyridine borane sequencing (CAPS). Using potassium perruthenate, 5hmC can be selectively oxidised to 5fC (Booth et al., 2012), which can be further converted to DHU by pyridine borane and read as thymine in the subsequent sequencing (Figure 5-7 b).




b
C 5mC 5hmC
βGT
C 5mC 5ghmC
TET
Potassium perruthenate
TET
C 5caC 5caC
C 5caC 5ghmC
C 5mC 5fC
pyridine borane
C DHU DHU
C DHU 5ghmC
C 5mC DHU
PCR
C T T
C T C
C C T
TAPS
TAPSβ
CAPS

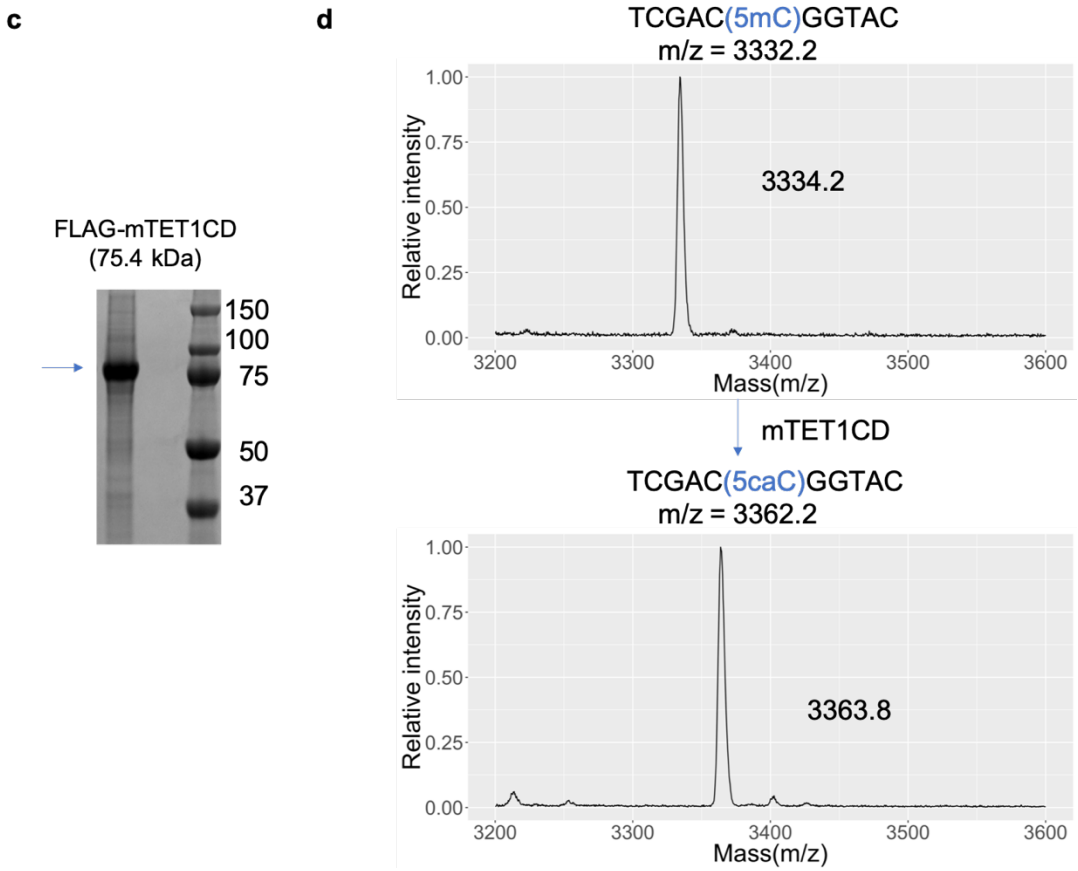


Figure 5-7 TET-assisted pyridine borane sequencing. a) Reduction of 5fC and 5caC to dihydrouracil (DHU) with pyridine borane. b) A schematic illustration of the strategy of TET-assisted pyridine borane sequencing (TAPS), TAPSβ and chemical-assisted pyridine borane sequencing (CAPS), modified from (Liu et al., 2019) c) SDS-PAGE showing purified FLAG-mTET1CD. Arrow pointing at the band of FLAG-mTET1CD (75.4 kDa). d) MALDI-MS results of a 5mC containing model DNA. 5mC was oxidised 5caC by mTET1CD.

5.7 Comparison of commonly used methods for DNA hydrolysis

To examine the oxidation of 5mC to 5caC by mTET1CD, DNA was enzymatically hydrolysed to single nucleoside for LC-MS/MS quantification. Surprisingly, the level of 5caC was much lower than expected in TET oxidised genomic DNA. Similarly, the 5caC level detected in the synthetic oligo was two to three orders of magnitude

lower than expected. Fang Yuan in our lab found that cause of the low 5caC signal in the LC-MS/MS quantification is one of the enzymes used in the hydrolysis, PDE1. 5caC is resistant to PDE1 digestion, resulting in incomplete digestion product which cannot be detected by LC-MS/MS (Yuan et al., 2019). The 5caC digestion efficiencies of several commonly used DNA hydrolysis protocols were compared on genomic DNA from mTET1CD overexpressed 293T cells. PDE1 method performed similarly to other methods in digesting 5mC, 5hmC and 5fC (Figure 5-8 a, b, c), but digested 5caC less efficiently than other methods (Figure 5-8 d). The addition of other nucleases such as DNase I and nuclease P1 improved the digestion of 5caC (Figure 5-8 d). Benzonase did not improve the detection of 5caC because Benzonase digests DNA into three to five base oligonucleotides instead of single nucleotides. Besides, the protocol with nuclease P1 hydrolysed DNA less efficiently compared with other protocols (Figure 5-8). In summary, PDE1 cannot be used for hydrolysis involving the quantification of 5caC unless in combination with other nucleases that digest DNA into single nucleotides.

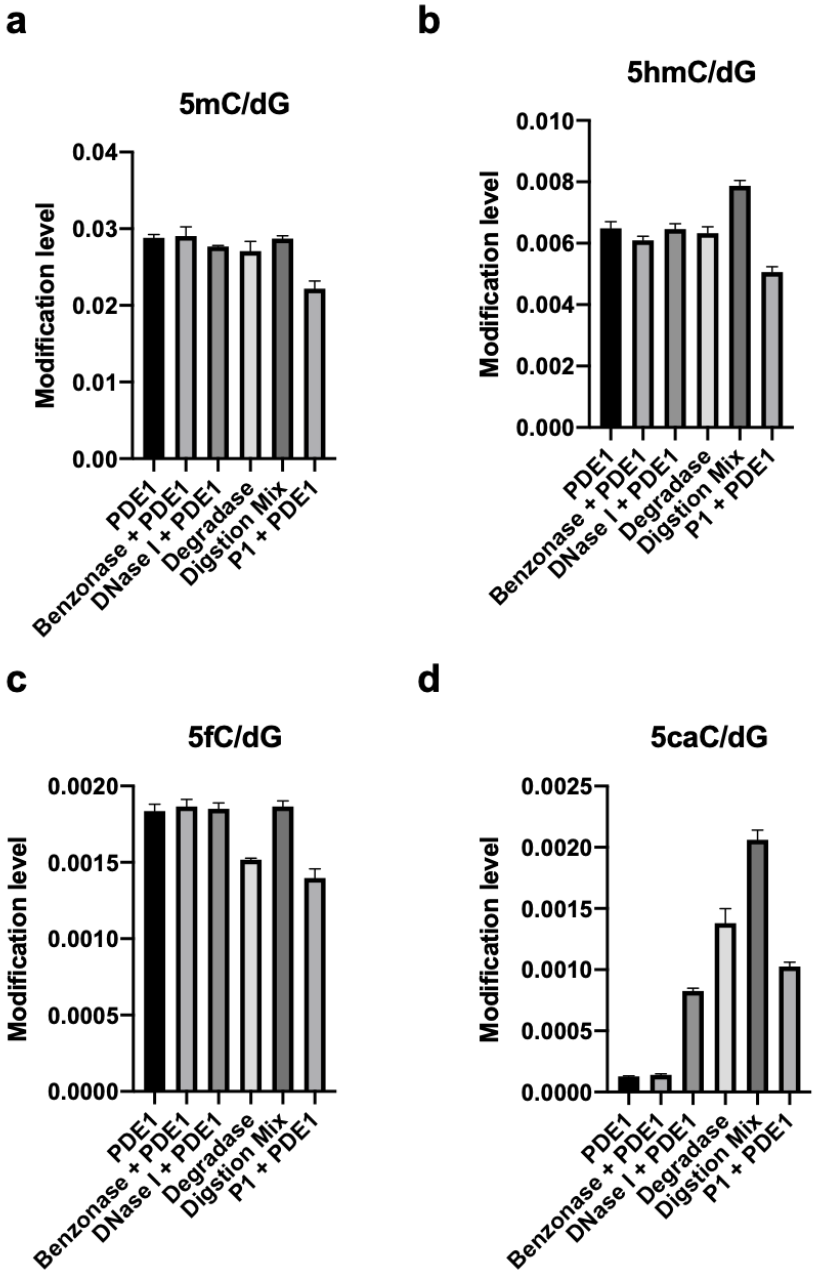
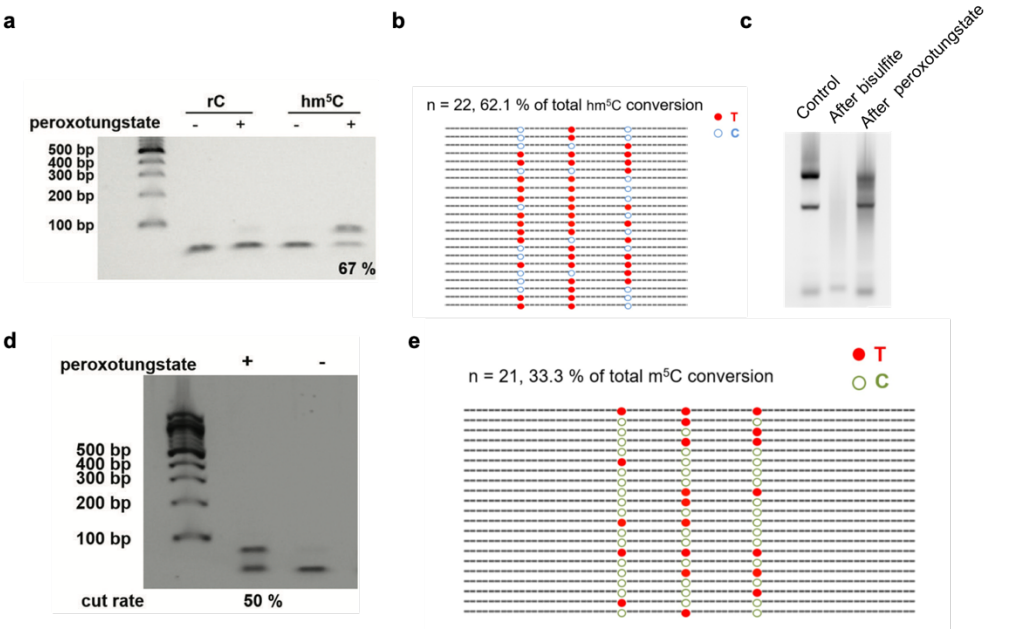
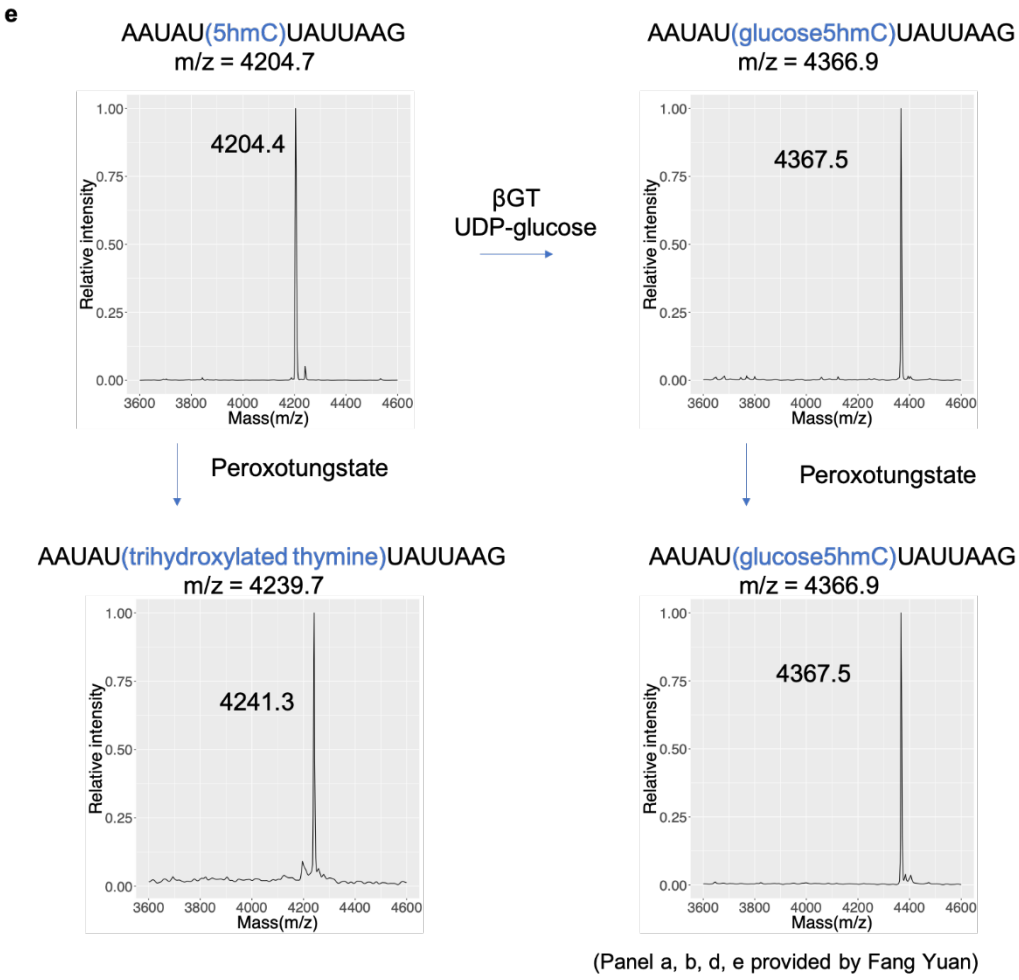


Figure 5-8 Comparisons of hydrolysis methods for LC-MS/MS quantification of modified cytosines. Genomic DNA from mTET1CD overexpressed 293T cells was hydrolysed with different DNA hydrolysis enzymes. (A) 5mC, (B) 5hmC, (C) 5fC, (D) 5caC levels were quantified with LC-MS/MS after the hydrolysis. Degradase: DNA Degradase Plus™ (Zymo), Digestion Mix: Nucleoside Digestion Mix (NEB). Error bars indicate SD (n = 3).

5.8 5hmC and 5mC mapping in RNA

The peroxotungstate selective oxidation reaction can also be used for mapping 5hmC in RNA. Working together with Fang, we developed peroxotungstate oxidation sequencing (WO-Seq). After reverse transcription and PCR of peroxotungstate oxidised RNA, 5hmC was converted to thymine. The conversion rate is 67% based on restriction digestion (Figure 5-9 a) and 62% quantified by Sanger sequencing (Figure 5-9 b). This method caused less RNA degradation than bisulfite treatment (Figure 5-9 c). The method was further extended to detect 5mC in RNA. Mammalian TET proteins were reported to oxidise 5mC to 5hmC (Fu et al., 2014). We tested *Naegleria* Tet-like oxygenase (NgTET) and found it also oxidise 5mC to 5hmC. Based on this, TET-Assisted WO-Seq (TAWO-Seq) was developed to detected 5mC in RNA. 5mC was converted to 5hmC by NgTET. 5hmC was then oxidised by peroxotungstate reaction to trihydroxylated thymine, which was recognised as T by reverse transcriptase during RT-PCR, allowing for subsequent mapping of 5mC in RNA. The reaction was performed on model RNA and the conversion rate was 50% by restriction digestion (Figure 5-9 d) and 33% by Sanger sequencing (Figure 5-9 e). To map 5mC specifically, before the NgTET oxidation reaction, glucose can be transferred onto 5hmC by βGT, which protects it from peroxotungstate oxidation (Figure 5-9 f). In summary, we demonstrated that the peroxotungstate reaction can be used for mapping 5hmC and 5mC in RNA.





(Panel a, b, d, e provided by Fang Yuan)

Figure 5-9 5hmC and 5mC mapping in RNA. a, d) Semi-quantification of 5hmC (a) or 5mC (d) to T conversion by WO-seq (a) or TAWO-seq (d) with restriction enzyme digestion assay. b) Quantification of 5hmC (b) or 5mC (e) to T conversion by WO-seq (b) or TAWO-seq (e) with Sanger sequencing. Each row represents the sequencing results from a DNA fragment with three modified sites. White dots indicate C in the sequencing, whereas red dots indicate T. c) Agarose gel showing total RNA after bisulfite treatment or peroxotungstate treatment. e) MALDI-MS results of a 5hmC containing RNA fragment. 5hmC was oxidised to trihydroxylated thymine by peroxotungstate while glucose labelled 5hmC was resistant to the oxidation.

5.9 Discussion

The uneven conversion rate of the peroxotungstate oxidation reaction on DNA limits the application of this method. Site 3 in DNA2 and site 2 in DNA3 have the same sequence context but differed in conversion rate (81% and 52% respectively) (Figure 5-6 c, f), indicating the difference was caused by other factors than sequence context, for example, distance to the end of the DNA fragment. To make the method applicable to quantify the distribution of 5hmC in the genome, more model oligos need to be tested to understand the cause of the difference in conversation rate to be able to quantify the 5hmC levels accordingly.

The CAPS and TAPSβ methods allowed for the detection 5hmC with a mild reaction, which could be potentially used for base-resolution 5hmC mapping in cell-free DNA. Several other bisulfite-free 5hmC base-resolution mapping methods have been published recently, including hmC-CATCH (Zeng et al., 2018) and ACE-seq (Schutsky et al., 2018). hmC-CATCH selectively oxidises 5hmC to 5fC and label 5fC

with azido derivative of 1,3-indandione, which enables enrichment as well as base-pairing change for base-resolution mapping. In ACE-seq, 5hmC is protected with glucose by T4 β-glucosyltransferase and read as C while unmodified cytosine and 5mC are deaminated by APOBEC3A and read as T in the sequencing. hmC-CATCH has been applied to 5hmC base-resolution mapping in cell-free DNA. It was done with enrichment therefore there was no quantitative information, but in principle, the method can be used without the enrichment step to get quantitative information.

The application of the peroxotungstate reaction to RNA enables base resolution mapping of 5hmC and 5mC in RNA. The conversion rate is around 50% which limits its application in detecting low abundance sites. Similar to the reaction in DNA, this method has uneven conversion rates among different sites in RNA (Figure 5-9 b, e). The currently available methods for 5hmC is antibody-based enrichment methods (Delatte et al., 2016). As there is no other base-resolution 5hmC mapping method available, WO-seq, despite its limitations, will still provide insight about hydroxymethylation in RNA when applied to RNA samples from cells. Bisulfite sequencing has been modified to be applicable for RNA (Schaefer et al., 2008) and was used to analyse methylation in non-coding and coding RNA (Squires et al., 2012). Bisulfite treatment causes RNA damage (Figure 5-9 c) and cannot distinguish 5mC from other modifications that are also resistant to bisulfite conversion (Shafik et al., 2016). Having TAWO-Seq as an alternative method for 5mC mapping could provide better insight into 5mC distribution in RNA.

6 Chapter 6 Conclusion and Discussion

In Chapter 3 and Chapter 4, 5hmC in the context of H/M sites was studied. The H/M site was found to account for 60% of all 5hmC in various cell types (mouse cerebellum, kidney, heart, liver and ESCs) (Song et al., 2016), but little is known about the biological functions of H/M sites. The H/M sites were studied from two aspects: the proteins that recognise the sites (Chapter 3) and enzymatic pathways that generate the sites (Chapter 4).

6.1 The recognition of H/M sites

DNA pull-down followed by LC-MS/MS was performed to identify potential binding proteins that recognise H/M sites specifically. GO analysis of all proteins identified from the pull-down showed enrichment of nuclear and DNA binding proteins, but there was also an enrichment of RNA binding proteins (Figure 3-7), which might result from the impurity of DNA probes used. The DNA probes were generated by annealing complementary single-strand DNA. There might be single-strand DNA left in the probe because of incomplete annealing. To improve the purity, Exonuclease I, which digest all single-strand DNA and leaves double-strand DNA intact, can be used.

The MS was done with two replicates for brain samples and three replicates for ESC samples. The small sample size hindered statistical power. As a result, the positive control MECP2, though had the trend of enrichment by the M/M probe, did not pass the statistical test in the brain samples (Figure 3-21). Increasing number replicates would increase statistical power and allow for the detection of more potential targets.

Alternatively, to get more information from the current data, protein interaction analysis with less stringent filtered MS results can be performed. For example, BRPF1 and its interaction protein KAT6A were enriched by H/M probe in brain samples (Figure 6-1). BRPF1 is the scaffold protein of histone acetyltransferase MOZ/MORF complexes (Yang, 2015) and has a PHD-zinc-knuckle-PHD module which binds to DNA (Klein et al., 2019). KAT6A (also known as MOZ) is part of the MOZ/MORF complex and has histone acetyltransferase activity (Voss et al., 2009). The identification of two proteins from the same complex increased the confidence of the result. Though there was a trend of enrichment in the MS results, there were variations between replicates (Figure 6-1). The results need to be validated by DNA pull-down followed by western blot before taken forward.

There was no overlap between the H/M site binding proteins identified from the brain and from the ESCs. A comparison of the MS results of identified proteins was summarized in Table 6-1. UHRF1, TET2, TET3 and YTHDF3 were detected in only one of the two tissues, whereas YBX1, WRAD complex and TET1 were detected in both tissues but only show enrichment by the H/M probe in one tissue.

Table 6-1 Comparison of MS results from brains and from ESCs

Protein name		MS result from brain samples	MS result from brain samples
YBX1		Enriched by H/M probe	Detected, not enriched by H/M probe
WRAD complex	RBBP5	Enriched by H/M probe	Detected, not enriched by H/M probe
	ASH2L	Enriched by H/M probe	Detected, not enriched by H/M probe
	DPY30	Detected, not enriched by H/M probe	Detected, not enriched by H/M probe
	WDR5	Detected, not enriched by H/M probe	Detected, not enriched by H/M probe

UHRF1	Not detected	Enriched by H/M probe
TET1	Detected, not enriched by H/M probe	Enriched by H/M probe
TET2	Not detected	Enriched by H/M probe
TET3	Detected, not enriched by H/M probe	Not detected
YTHDF3	Not detected	Enriched by H/M probe

a

	Unique peptides	ANOVA (p)
BRPF1	2	0.64
KAT6A	3	0.075
KAT8	2	0.29

b

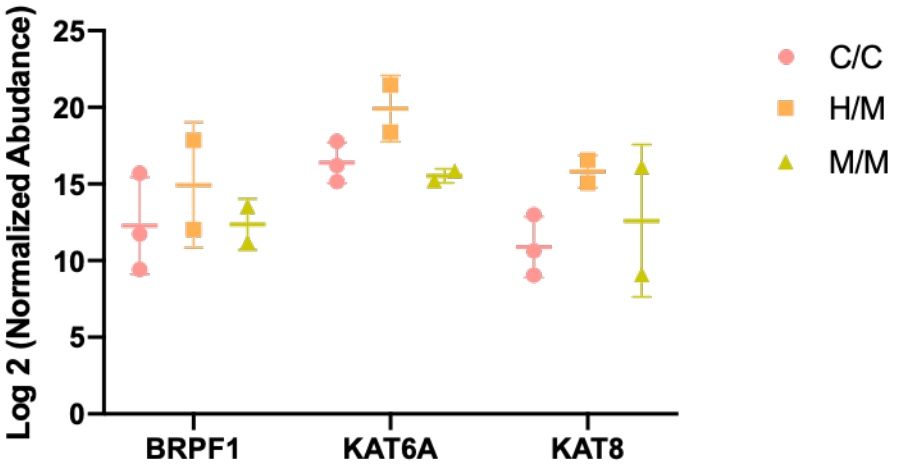


Figure 6-1 Summary of MS results of BRPF1 and KAT6A. a) Summary of the number of unique peptides and *p*-values of the ANOVA test among three probes. b) Log_2 abundance of protein pull-downed by DNA probes containing H/M, M/M or C/C sites.

Biochemical studies were performed to validate the candidates identified from LC-MS/MS. However, none of the five candidates has been confirmed to have the binding specificity to H/M sites. Based on the MS results, there are several high confidence hits especially UHRF1 from mESCs and YBX1 from mouse brains, which showed consistently high fold enrichment in all replicates (Figure 3-9, Figure 3-14). Comparing the experimental conditions used in DNA pull-down and in the EMSA, the buffers were the same but the DNA probes were different. In the pull-down, the DNA probe was 90 bp long with three modified sites whereas the probe in EMSA was 36 bp with one modified site. Besides, the sequence context of the two probes was different. If sequence context or the number of modified sites would also contribute to the binding affinity, the same probe used in the pull-down should be used in the EMSA to validate the binding. Moreover, in the pull-down, the DNA probe was incubated with all proteins from nuclear extract while in the EMSA only the recombinant target protein was used. The binding preference of the target protein may be affected by post-translation modifications or require interaction with other proteins. For example, in *Drosophila* CLOCK and CYCLE proteins forms heterodimer and interact with E-box DNA elements. In EMSA, the band shift was only observed when DNA was incubated with two proteins together but not with either of the proteins separately (Lee et al., 1999). Therefore, for proteins that are known to be part of a complex, EMSA could be done by incubating together with the interaction partners from the complex. Alternatively, EMSA could also be done with whole nuclear proteins extract and the target proteins could be identified by adding specific antibody to see if a further shift of the band could be observed. Alternative

assays could also be used in addition to EMSA, for example, fluorescence polarisation assay or isothermal titration calorimetry.

For YBX1, ASH2L and YTHDF3, band-shift was observed but there was no difference between M/M, H/M and C/C probes (Figure 3-10 c, Figure 3-13 b, Figure 3-20 c). It was not clear if the shift was a result of non-specific binding. There were several possible reasons for the non-specific binding. Firstly, the purity of the proteins was not perfect. There were additional bands on the Coomassie-stained gel of purified proteins (Figure 3-10 a, Figure 3-12, Figure 3-20 a). The shift could result from binding by the contaminants other than protein of interest. EMSA with DNA binding domain mutation of the protein could be done to see if the mutation could abolish DNA binding as expected. To improve protein purity, several aspects could be further optimised: reducing expression time, moving tag from N-terminus to C-terminus to remove incompletely translated proteins, purification with multiple affinity columns, gel filtration and/or ion exchange. Secondly, the non-specific binding could be from protein interaction with the fluorescence label on the DNA. The non-specific binding could be tested with competition assay using an excess of non-labelled DNA with the same modification. The problem could be solved by using another fluorescence label or P^{32} radio label.

After validating the binding of target protein to H/M sites biochemically, the binding needs to be further studied in the cellular context using techniques like reporter assays: the target protein could be fused with a transcription activator; the H/M sites, M/M sites or C/C sites could be put into a plasmid followed by a weak promoter and GFP

gene. When the fused proteins and the plasmid are put together into cells, the binding could be monitored by GFP expression. Functionally, how the protein affects the abundance of H/M sites, as well as other possible functions related to the known function of the target proteins, needs to be studied.

Another possibility there is no H/M sites specific binding proteins. H/M sites can function by reducing the binding of 5mC binding proteins. Biochemical studies showed that H/M sites reduced the binding of MeCP2, MBD1, MBD2, MBD4 (Hashimoto et al., 2012; Song et al., 2016). If in the cells H/M sites are sufficient to inhibit 5mC binding proteins and have similar functions to H/H sites, the existence of H/M sites might be due to a lack of selection pressure for completely oxidising H/M sites to H/H sites.

6.2 The generation of H/M sites

H/M sites could be generated from the oxidation of M/M sites by TET enzymes or methylation of H/C sites by DNMTs. The method used in Chapter 4 separated two strands of a double-strand DNA and analyse each strand individually. This method would detect H/M sites when all H/M sites have 5hmC on the same strand and 5mC on the other strand, but it was not able to distinguish a mixture of M/M sites and H/H sites from a mixture or H/M sites and M/H sites. To address this problem, a single molecular imaging method which measures H/M sites could be used (Song et al., 2016). Besides TET1 and TET2, whether TET3 has similar preference awaits further study.

Alternatively, it is possible that TET enzymes do not have strand preferences on genomic DNA in the cells. Base-resolution mapping of 5hmC showed that, in a population of cells, 5mC and 5hmC often coexist at the same cytosine (Yu et al., 2012b). H/M sites of 5hmC on the forward strand and on the reverse strand may both exist.

The other pathway to generate H/M site is the methylation of H/C sites by DNMTs. DNMT3A and DNMT3B can methylate H/C sites (Hashimoto et al., 2012; Otani et al., 2013). DNMT1 has lower activity on H/C sites than on M/C sites (Hashimoto et al., 2012; Otani et al., 2013; Seiler et al., 2018; Valinluck and Sowers, 2007). UHRF1 is known to increase DNMT1 activity in ethylating M/C sites (Bostick et al., 2007; Sharif et al., 2007). The observation that UHRF1 also binds to H/M and H/C sites (Figure 3-16) leads us to hypothesize that UHRF1 may also facilitate DNMT1 in methylating H/C sites. A method has been established to quantify the reaction of DNMT1 on model DNA with M/C or H/C sites by bisulfite conversion followed by next-generation sequencing. Further studies need to be done by including UHRF1 in the reaction to see if UHRF1 would enhance the activity of DNMT1 on H/C sites. If in the presence of UHRF1, the activity of DNMT1 on the H/C sites is similar to that on the M/C site, it not only provides a possible mechanism for H/M sites maintenance but also has an implication on the models of passive demethylation. Current model of role of 5hmC in replication mediated passive demethylation is based on the biochemical observation that DNMT1 has reduced activity on H/C sites compared to M/C sites (Hashimoto et al., 2012; Otani et al., 2013; Seiler et al., 2018; Valinluck

and Sowers, 2007). If with the addition of UHRF1, the activity of DNMT1 on H/C sites and M/C sites are similar, 5hmC will not lead to passive demethylation.

After all the biochemical studies, the effect of TET enzymes and DNMTs on the H/M sites needs to be studied in the cellular context. The change of global level of H/M site can be measured by a single molecular imaging method (Song et al., 2016). Cell-cycle arrested cells would be a good model to study the effect of TET enzymes on H/M sites. In Cell-cycle arrested cells, H/M sites would not be disrupted by DNA replication. By introducing the TET enzyme into cell-cycle arrested cells, the net effect of TET enzymes in generating H/M sites can be studied. The role of UHRF1 in facilitating DNMT1 to generate H/M sites can be studied by comparing the levels of H/M sites in UHRF1 knockout cells and wild type cells.

6.3 Detection of 5hmC

In Chapter 5, the 5hmC selective oxidation reaction (Okamoto et al., 2011) was applied to mapping 5hmC in DNA, 5hmC in RNA and when combined with TET oxidation, mapping 5mC in RNA. Using this method, 5hmC can be mapped at base-resolution with Sanger sequencing or next-generation sequencing. However, the inconsistent conversion rates among different sites make the quantification of 5hmC abundance at each site challenging. In the DNA model tested, the conversion rate was higher when the 5hmC site is closer to the end of DNA fragments, regardless of the sequence context (Figure 5-6). Interestingly, in the RNA model, the conversion rate was higher when the site is in the middle of the fragment (Figure 5-9). The RNA model has the same sequence as the DNA2 from five nucleotides upstream of the first

modified site to five nucleotides downstream of the last modified sites. The difference in the conversion rate of the DNA and RNA model indicates that the difference is not likely a result of secondary structures as the same sequence in RNA and DNA would lead to similar secondary structure. Based on the observation on the models used, although the cause is not clear, on DNA the conversion is better on the fragment end while on RNA it is better in the middle of the fragment. However, more model nucleotides need to be studied before the conclusion can be generalised. If the rule applies after testing more models, computational methods can be used to adjust for the differences accordingly when calculating the 5hmC abundance at each site.

The conversion rate on RNA was around 50% (Figure 5-9) and can be further improved to allow for the detection of low abundance sites. The conversion requires two steps: the chemical oxidation of 5-hydroxymethylcytosine to trihydroxylated thymine and the recognition of trihydroxylated thymine as thymine by reverse transcriptase. Both steps contribute to the final conversion rate. The oxidation step has over 80% conversion rate based on LC-MS/MS quantification but the final conversion rate was around only 50%, suggesting a reverse transcriptase that has better ability to recognise and read-through trihydroxylated thymine would improve the conversion rate.

For 5hmC mapping in the DNA, there are two recent methods that selectively convert 5hmC to T for sequencing and do not have the problem of uneven conversion rate among different sites: hmC-CATCH (Zeng et al., 2018) and CAPS (Liu et al., 2019).

5hmC was identified as a modification in mammalian RNA by mass spectrometry (Huang et al., 2016; Huber et al., 2015; Zhang et al., 2016) and by dot blot (Miao et al., 2016). The oxidation of the 5mC to 5hmC in RNA by TET was observed *in vitro* and the overexpression of TET leads to increased 5hmC levels in RNA in HEK293T cells (Huang et al., 2016). However, the genomic mapping of 5hmC has only been done with antibody-based enrichment methods (Delatte et al., 2016). There is no base-resolution method available for 5hmC mapping. Application of our WO-seq method to whole-genome 5hmC mapping in RNA would enable us to learn the distribution of 5hmC in RNA and have insights into the biological role of 5hmC in RNA.

Bisulfite conversion was used for 5mC mapping in RNA (Schaefer et al., 2008; Squires et al., 2012) but has the limitation of RNA degradation (Figure 5-9) and cannot distinguish 5mC from other modifications that are also resistant to bisulfite conversion (Shafik et al., 2016). Our TAWO-seq methods would allow for the detection of novel sites that would otherwise be missed with bisulfite sequencing.

7 References

Advani, S.M., Advani, P., DeSantis, S.M., Brown, D., VonVille, H.M., Lam, M., Loree, J.M., Mehrvarz Sarshekeh, A., Bressler, J., Lopez, D.S.*, et al.* (2018). Clinical, Pathological, and Molecular Characteristics of CpG Island Methylator Phenotype in Colorectal Cancer: A Systematic Review and Meta-analysis. Translational Oncology *11*, 1188-1201.

Altman, D.G., and Bland, J.M. (2005). Standard deviations and standard errors. BMJ *331*, 903-903.

Amouroux, R., Nashun, B., Shirane, K., Nakagawa, S., Hill, P.W.S., D'Souza, Z., Nakayama, M., Matsuda, M., Turp, A., Ndjetehe, E.*, et al.* (2016). De novo DNA methylation drives 5hmC accumulation in mouse zygotes. Nature Cell Biology *18*, 225.

Antequera, F., and Bird, A. (1993). Number of CpG islands and genes in human and mouse. Proceedings of the National Academy of Sciences *90*, 11995.

Arai, E., Chiku, S., Mori, T., Gotoh, M., Nakagawa, T., Fujimoto, H., and Kanai, Y. (2012). Single-CpG-resolution methylome analysis identifies clinicopathologically aggressive CpG island methylator phenotype clear cell renal cell carcinomas. Carcinogenesis *33*, 1487-1493.

Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y., and Shirakawa, M. (2008). Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. Nature *455*, 818.

Bachman, M., Uribe-Lewis, S., Yang, X., Williams, M., Murrell, A., and Balasubramanian, S. (2014). 5-Hydroxymethylcytosine is a predominantly stable DNA modification. Nature Chemistry *6*, 1049.

Barau, J., Teissandier, A., Zamudio, N., Roy, S., Nalesso, V., Hérault, Y., Guillou, F., and Bourc'his, D. (2016). The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. Science *354*, 909.

Baubec, T., Ivánek, R., Lienert, F., and Schübeler, D. (2013). Methylation-Dependent and -Independent Genomic Targeting Principles of the MBD Protein Family. Cell *153*, 480-492.

Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. (1988). Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells: The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. Journal of Molecular Biology *203*, 971-983.

Bestor, T.H., and Ingram, V.M. (1983). Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. Proceedings of the National Academy of Sciences *80*, 5559.

Bettegowda, C., Sausen, M., Leary, R.J., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B.R., Wang, H., Luber, B., Alani, R.M.*, et al.* (2014). Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies. Science Translational Medicine *6*, 224ra224.

Bhattacharya, S.K., Ramchandani, S., Cervoni, N., and Szyf, M. (1999). A mammalian protein with specific demethylase activity for mCpG DNA. Nature *397*, 579-583.

Bian, K., Lenz, S.A.P., Tang, Q., Chen, F., Qi, R., Jost, M., Drennan, C.L., Essigmann, J.M., Wetmore, S.D., and Li, D. (2019). DNA repair enzymes ALKBH2, ALKBH3, and AlkB oxidize 5-methylcytosine to 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine in vitro. Nucleic Acids Research *47*, 5522-5529.

Bienvenu, T., Carrié, A., de Roux, N., Vinet, M.-C., Jonveaux, P., Couvert, P., Villard, L., Arzimanoglou, A., Beldjord, C., Fontes, M.*, et al.* (2000). MECP2 mutations account for most cases of typical forms of Rett syndrome. Human Molecular Genetics *9*, 1377-1384.

Bird, A., Taggart, M., Frommer, M., Miller, O.J., and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell *40*, 91-99.

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic acids research *8*, 1499-1504.

Bird, A.P. (1986). CpG-rich islands and the function of DNA methylation. Nature *321*, 209-213.

Blattler, A., Yao, L., Wang, Y., Ye, Z., Jin, V.X., and Farnham, P.J. (2013). ZBTB33 binds unmethylated regions of the genome associated with actively expressed genes. Epigenetics & Chromatin *6*, 13.

Booth, M.J., Branco, M.R., Ficz, G., Oxley, D., Krueger, F., Reik, W., and Balasubramanian, S. (2012). Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. Science *336*, 934-937.

Booth, M.J., Ost, T.W.B., Beraldi, D., Bell, N.M., Branco, M.R., Reik, W., and Balasubramanian, S. (2013). Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. Nature Protocols *8*, 1841.

Bornelöv, S., Komorowski, J., and Wadelius, C. (2015). Different distribution of histone modifications in genes with unidirectional and bidirectional transcription and a role of CTCF and cohesin in directing transcription. BMC genomics *16*, 300-300.

Bostick, M., Kim, J.K., Estève, P.-O., Clark, A., Pradhan, S., and Jacobsen, S.E. (2007). UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells. Science *317*, 1760.

Campbell, N.J., Dengel, A.C., Edwards, C.J., and Griffith, W.P. (1989). Studies on transition metal peroxo complexes. Part 8. The nature of peroxomolybdates and peroxotungstates in aqueous solution. Journal of the Chemical Society, Dalton Transactions, 1203-1208.

Carey, M.F., Peterson, C.L., and Smale, S.T. (2012). Experimental Strategies for the Identification of DNA-Binding Proteins. Cold Spring Harbor Protocols *2012*, pdb.top067470.

Chan, K.C.A., Jiang, P., Chan, C.W.M., Sun, K., Wong, J., Hui, E.P., Chan, S.L., Chan, W.C., Hui, D.S.C., Ng, S.S.M.*, et al.* (2013). Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. Proceedings of the National Academy of Sciences *110*, 18761.

Chedin, F., Lieber, M.R., and Hsieh, C.L. (2002). The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. Proc Natl Acad Sci U S A *99*, 16916-16921.

Chen, D., Wang, Y., Mo, M., Zhang, J., Zhang, Y., Xu, Y., Liu, S.-Y., Chen, J., Ma, Y., Zhang, L.*, et al.* (2019a). Polymerization retardation isothermal amplification (PRIA): a strategy enables sensitively quantify genome-wide 5-methylcytosine oxides rapidly on handy instruments with nanoscale sample input. Nucleic Acids Research.

Chen, R., Zhang, Q., Duan, X., York, P., Chen, G.-D., Yin, P., Zhu, H., Xu, M., Chen, P., Wu, Q.*, et al.* (2017). The 5-Hydroxymethylcytosine (5hmC) Reader UHRF2 Is Required for Normal Levels of 5hmC in Mouse Adult Brain and Spatial Learning and Memory. J Biol Chem *292*, 4533-4543.

Chen, R.Z., Akbarian, S., Tudor, M., and Jaenisch, R. (2001). Deficiency of methyl-CpG binding protein-2 in CNS neurons results in a Rett-like phenotype in mice. Nature Genetics *27*, 327-331.

Chen, X., Li, A., Sun, B.-F., Yang, Y., Han, Y.-N., Yuan, X., Chen, R.-X., Wei, W.-S., Liu, Y., Gao, C.-C*., et al.* (2019b). 5-methylcytosine promotes pathogenesis of bladder cancer through stabilizing mRNAs. Nature Cell Biology *21*, 978-990.

Cheng, X., and Roberts, R.J. (2001). AdoMet-dependent methylation, DNA methyltransferases and base flipping. Nucleic Acids Research *29*, 3784-3795.

Chuang, L.S., Ian, H.I., Koh, T.W., Ng, H.H., Xu, G., and Li, B.F. (1997). Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21WAF1. Science *277*, 1996-2000.

Cramer, J.M., Scarsdale, J.N., Walavalkar, N.M., Buchwald, W.A., Ginder, G.D., and Williams, D.C. (2014). Probing the Dynamic Distribution of Bound States for Methylcytosine-binding Domains on DNA. J Biol Chem *289*, 1294-1302.

Crawford, D.J., Liu, M.Y., Nabel, C.S., Cao, X.-J., Garcia, B.A., and Kohli, R.M. (2016). Tet2 Catalyzes Stepwise 5-Methylcytosine Oxidation by an Iterative and de novo Mechanism. Journal of the American Chemical Society *138*, 730-733.

Cui, Y., and Irudayaraj, J. (2015). Dissecting the behavior and function of MBD3 in DNA methylation homeostasis by single-molecule spectroscopy and microscopy. Nucleic acids research *43*, 3046-3055.

Daniel, J.M., Spring, C.M., Crawford, H.C., Reynolds, A.B., and Baig, A. (2002). The p120 ctn -binding partner Kaiso is a bi-modal DNA-binding protein that recognizes both a sequence-specific consensus and methylated CpG dinucleotides. Nucleic Acids Research *30*, 2911-2919.

Dawlaty, Meelad M., Breiling, A., Le, T., Barrasa, M.I., Raddatz, G., Gao, Q., Powell, Benjamin E., Cheng, Albert W., Faull, Kym F., Lyko, F., *et al.* (2014). Loss of Tet Enzymes Compromises Proper Differentiation of Embryonic Stem Cells. Developmental Cell *29*, 102-111.

Delatte, B., Wang, F., Ngoc, L.V., Collignon, E., Bonvin, E., Deplus, R., Calonne, E., Hassabi, B., Putmans, P., Awe, S., *et al.* (2016). Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. Science *351*, 282-285.

Dor, Y., and Cedar, H. (2018). Principles of DNA methylation and their implications for biology and medicine. The Lancet *392*, 777-786.

Eden, A., Gaudet, F., Waghmare, A., and Jaenisch, R. (2003). Chromosomal Instability and Tumors Promoted by DNA Hypomethylation. Science *300*, 455.

Ehrlich, M., Gama-Sosa, M.A., Huang, L.H., Midgett, R.M., Kuo, K.C., McCune, R.A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucleic acids research *10*, 2709-2721.

Eliseeva, I.A., Kim, E.R., Guryanov, S.G., Ovchinnikov, L.P., and Lyabin, D.N. (2011). Y-box-binding protein 1 (YB-1) and its functions. Biochemistry (Moscow) *76*, 1402-1433.

Ernst, P., and Vakoc, C.R. (2012). WRAD: enabler of the SET1-family of H3K4 methyltransferases. Briefings in Functional Genomics *11*, 217-226.

Fang, F., Turcan, S., Rimner, A., Kaufman, A., Giri, D., Morris, L.G.T., Shen, R., Seshan, V., Mo, Q., Heguy, A., *et al.* (2011). Breast Cancer Methylomes Establish an Epigenomic Foundation for Metastasis. Science Translational Medicine *3*, 75ra25.

Fang, M., Ou, J., Hutchinson, L., and Green, Michael R. (2014). The BRAF Oncoprotein Functions through the Transcriptional Repressor MAFG to Mediate the CpG Island Methylator Phenotype. Molecular Cell *55*, 904-915.

Feng, Q., and Zhang, Y. (2001). The MeCP1 complex represses transcription through preferential binding, remodeling, and deacetylating methylated nucleosomes. Genes & Development *15*, 827-832.

Ficz, G., Branco, M.R., Seisenberger, S., Santos, F., Krueger, F., Hore, T.A., Marques, C.J., Andrews, S., and Reik, W. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature *473*, 398-402.

Figueroa, M.E., Abdel-Wahab, O., Lu, C., Ward, P.S., Patel, J., Shih, A., Li, Y.S., Bhagwat, N., Vasanthakumar, A., Fernandez, H.F*., et al.* (2010). Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation. Cancer Cell *18*, 553-567.

Filion, G.J.P., Zhenilo, S., Salozhin, S., Yamada, D., Prokhortchouk, E., and Defossez, P.-A. (2006). A Family of Human Zinc Finger Proteins That Bind Methylated DNA and Repress Transcription. Molecular and Cellular Biology *26*, 169.

Fournier, A., Sasai, N., Nakao, M., and Defossez, P.-A. (2011). The role of methyl-binding proteins in chromatin organization and epigenome maintenance. Briefings in Functional Genomics *11*, 251-264.

Fraga, M.F., Uriol, E., Diego, L.B., Berdasco, M., Esteller, M., Cañal, M.J., and Rodríguez, R. (2002). High-performance capillary electrophoretic method for the quantification of 5-methyl 2'-deoxycytidine in genomic DNA: Application to plant, animal and human cancer tissues. ELECTROPHORESIS *23*, 1677-1681.

Frauer, C., Hoffmann, T., Bultmann, S., Casa, V., Cardoso, M.C., Antes, I., and Leonhardt, H. (2011). Recognition of 5-Hydroxymethylcytosine by the Uhrf1 SRA Domain. PLOS ONE *6*, e21306.

Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., and Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proceedings of the National Academy of Sciences *89*, 1827.

Fu, L., Guerrero, C.R., Zhong, N., Amato, N.J., Liu, Y., Liu, S., Cai, Q., Ji, D., Jin, S.G., Niedernhofer, L.J., *et al.* (2014). Tet-mediated formation of 5-hydroxymethylcytosine in RNA. J Am Chem Soc *136*, 11582-11585.

Fujita, N., Takebayashi, S.-i., Okumura, K., Kudo, S., Chiba, T., Saya, H., and Nakao, M. (1999). Methylation-Mediated Transcriptional Silencing in Euchromatin by Methyl-CpG Binding Protein MBD1 Isoforms. Molecular and Cellular Biology *19*, 6415.

Gabel, H.W., Kinde, B., Stroud, H., Gilbert, C.S., Harmin, D.A., Kastan, N.R., Hemberg, M., Ebert, D.H., and Greenberg, M.E. (2015). Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature *522*, 89.

Gaudet, F., Hodgson, J.G., Eden, A., Jackson-Grusby, L., Dausman, J., Gray, J.W., Leonhardt, H., and Jaenisch, R. (2003). Induction of Tumors in Mice by Genomic Hypomethylation. Science *300*, 489.

Gaudreault, I., Guay, D., and Lebel, M. (2004). YB‐1 promotes strand separation in vitro of duplex DNA containing either mispaired bases or cisplatin modifications, exhibits endonucleolytic activities and binds several DNA repair proteins. Nucleic Acids Research *32*, 316-327.

Globisch, D., Münzel, M., Müller, M., Michalakis, S., Wagner, M., Koch, S., Brückl, T., Biel, M., and Carell, T. (2010). Tissue Distribution of 5-Hydroxymethylcytosine and Search for Active Demethylation Intermediates. PLOS ONE *5*, e15367.

Goll, M.G., Kirpekar, F., Maggert, K.A., Yoder, J.A., Hsieh, C.L., Zhang, X., Golic, K.G., Jacobsen, S.E., and Bestor, T.H. (2006). Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2. Science *311*, 395-398.

Gonzalez-Zulueta, M., Bender, C.M., Yang, A.S., Nguyen, T., Beart, R.W., Van Tornout, J.M., and Jones, P.A. (1995). Methylation of the 5' CpG island of the p16/CDKN2 tumor suppressor gene in normal and transformed human tissues correlates with gene silencing. Cancer Res *55*, 4531-4535.

Gormally, E., Vineis, P., Matullo, G., Veglia, F., Caboux, E., Le Roux, E., Peluso, M., Garte, S., Guarrera, S., Munnia, A.*, et al.* (2006). TP53 and KRAS2 mutations in plasma DNA of healthy subjects and subsequent cancer occurrence: a prospective study. Cancer Res *66*, 6871-6876.

Gowher, H., Liebert, K., Hermann, A., Xu, G., and Jeltsch, A. (2005). Mechanism of Stimulation of Catalytic Activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L. J Biol Chem *280*, 13341-13348.

Gruenbaum, Y., Cedar, H., and Razin, A. (1982). Substrate and sequence specificity of a eukaryotic DNA methylase. Nature *295*, 620-622.

Grunau, C., Clark, S.J., and Rosenthal, A. (2001). Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. Nucleic Acids Research *29*, e65-e65.

Gu, T.-P., Guo, F., Yang, H., Wu, H.-P., Xu, G.-F., Liu, W., Xie, Z.-G., Shi, L., He, X., Jin, S.-g.*, et al.* (2011). The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. Nature *477*, 606.

Guo, J.U., Su, Y., Shin, J.H., Shin, J., Li, H., Xie, B., Zhong, C., Hu, S., Le, T., Fan, G.*, et al.* (2013). Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. Nature Neuroscience *17*, 215.

Guy, J., Gan, J., Selfridge, J., Cobb, S., and Bird, A. (2007). Reversal of Neurological Defects in a Mouse Model of Rett Syndrome. Science *315*, 1143.

Guy, J., Hendrich, B., Holmes, M., Martin, J.E., and Bird, A. (2001). A mouse Mecp2-null mutation causes neurological symptoms that mimic Rett syndrome. Nature Genetics *27*, 322-326.

Hackett, J.A., Sengupta, R., Zylicz, J.J., Murakami, K., Lee, C., Down, T.A., and Surani, M.A. (2013). Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. Science (New York, NY) *339*, 448-452.

Haffner, M.C., Chaux, A., Meeker, A.K., Esopi, D.M., Gerber, J., Pellakuru, L.G., Toubaji, A., Argani, P., Iacobuzio-Donahue, C., Nelson, W.G.*, et al.* (2011). Global 5-hydroxymethylcytosine content is significantly reduced in tissue stem/progenitor cell compartments and in human cancers. Oncotarget *2*, 627-637.

Hainer, S.J., McCannell, K.N., Yu, J., Ee, L.-S., Zhu, L.J., Rando, O.J., and Fazzio, T.G. (2016). DNA methylation directs genomic localization of Mbd2 and Mbd3 in embryonic stem cells. eLife *5*, e21964.

Hasegawa, S.L., Doetsch, P.W., Hamilton, K.K., Martin, A.M., Okenquist, S.A., Lenz, J., and Boss, J.M. (1991). DNA binding properties of YB-1 and dbpA: binding to double-stranded, single-stranded, and abasic site containing DNAs. Nucleic acids research *19*, 4915-4920.

Hashimoto, H., Liu, Y., Upadhyay, A.K., Chang, Y., Howerton, S.B., Vertino, P.M., Zhang, X., and Cheng, X. (2012). Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. Nucleic Acids Research *40*, 4841-4849.

Haun, R.S., Serventi, I.M., and Moss, J. (1992). Rapid, Reliable Ligation-Independent Cloning of Pcr Products Using Modified Plasmid Vectors. Biotechniques *13*, 515-518.

Hayashi, G., Koyama, K., Shiota, H., Kamio, A., Umeda, T., Nagae, G., Aburatani, H., and Okamoto, A. (2016). Base-Resolution Analysis of 5-Hydroxymethylcytosine by One-Pot Bisulfite-Free Chemical Conversion with Peroxotungstate. Journal of the American Chemical Society *138*, 14178-14181.

He, Y.F., Li, B.Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L.*, et al.* (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science *333*, 1303-1307.

Hendrich, B., and Bird, A. (1998). Identification and Characterization of a Family of Mammalian Methyl-CpG Binding Proteins. Molecular and Cellular Biology *18*, 6538.

Hendrich, B., Guy, J., Ramsahoye, B., Wilson, V.A., and Bird, A. (2001). Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development. Genes & Development *15*, 710-723.

Hendrich, B., Hardeland, U., Ng, H.-H., Jiricny, J., and Bird, A. (1999). The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. Nature *401*, 301-304.

Herman, J.G., Latif, F., Weng, Y., Lerman, M.I., Zbar, B., Liu, S., Samid, D., Duan, D.S., Gnarra, J.R., and Linehan, W.M. (1994). Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. Proceedings of the National Academy of Sciences of the United States of America *91*, 9700-9704.

Hiltunen, M.O., Alhonen, L., Koistinaho, J., Myöhänen, S., Pääkkönen, M., Marin, S., Kosma, V.-M., and Jänne, J. (1997). Hypermethylation of the APC (adenomatous

Polyposis Coli) gene promoter region in human colorectal carcinoma. International Journal of Cancer *70*, 644-648.

Hon, Gary C., Song, C.-X., Du, T., Jin, F., Selvaraj, S., Lee, Ah Y., Yen, C.-a., Ye, Z., Mao, S.-Q., Wang, B.-A.*, et al.* (2014). 5mC Oxidation by Tet2 Modulates Enhancer Activity and Timing of Transcriptome Reprogramming during Differentiation. Molecular Cell *56*, 286-297.

Hotchkiss, R.D. (1948). THE QUANTITATIVE SEPARATION OF PURINES, PYRIMIDINES, AND NUCLEOSIDES BY PAPER CHROMATOGRAPHY. J Biol Chem *175*, 315-332.

Hsu, C.H., Peng, K.L., Kang, M.L., Chen, Y.R., Yang, Y.C., Tsai, C.H., Chu, C.S., Jeng, Y.M., Chen, Y.T., Lin, F.M.*, et al.* (2012). TET1 suppresses cancer invasion by activating the tissue inhibitors of metalloproteinases. Cell Rep *2*, 568-579.

Hu, L., Li, Z., Cheng, J., Rao, Q., Gong, W., Liu, M., Shi, Y.G., Zhu, J., Wang, P., and Xu, Y. (2013). Crystal Structure of TET2-DNA Complex: Insight into TET-Mediated 5mC Oxidation. Cell *155*, 1545-1555.

Hu, L., Lu, J., Cheng, J., Rao, Q., Li, Z., Hou, H., Lou, Z., Zhang, L., Li, W., Gong, W.*, et al.* (2015). Structural insight into substrate preference for TET-mediated oxidation. Nature *527*, 118.

Huang, H., Jiang, X., Li, Z., Li, Y., Song, C.X., He, C., Sun, M., Chen, P., Gurbuxani, S., Wang, J.*, et al.* (2013). TET1 plays an essential oncogenic role in MLL-rearranged leukemia. Proceedings of the National Academy of Sciences *110*, 11994-11999.

Huang, W., Lan, M.-D., Qi, C.-B., Zheng, S.-J., Wei, S.-Z., Yuan, B.-F., and Feng, Y.-Q. (2016). Formation and determination of the oxidation products of 5-methylcytosine in RNA. Chemical Science *7*, 5495-5502.

Huang, Y., Chavez, L., Chang, X., Wang, X., Pastor, W.A., Kang, J., Zepeda-Martínez, J.A., Pape, U.J., Jacobsen, S.E., Peters, B.*, et al.* (2014). Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells. Proceedings of the National Academy of Sciences *111*, 1361.

Huang, Y., Pastor, W.A., Shen, Y., Tahiliani, M., Liu, D.R., and Rao, A. (2010). The Behaviour of 5-Hydroxymethylcytosine in Bisulfite Sequencing. PLOS ONE *5*, e8888.

Huber, S.M., van Delft, P., Mendil, L., Bachman, M., Smollett, K., Werner, F., Miska, E.A., and Balasubramanian, S. (2015). Formation and Abundance of 5-Hydroxymethylcytosine in RNA. ChemBioChem *16*, 752-755.

Iida, T., Suetake, I., Tajima, S., Morioka, H., Ohta, S., Obuse, C., and Tsurimoto, T. (2002). PCNA clamp facilitates action of DNA cytosine methyltransferase 1 on hemimethylated DNA. Genes to Cells *7*, 997-1007.

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

Inoue, A., and Zhang, Y. (2011). Replication-Dependent Loss of 5-Hydroxymethylcytosine in Mouse Preimplantation Embryos. Science *334*, 194.

Iqbal, K., Jin, S.-G., Pfeifer, G.P., and Szabó, P.E. (2011). Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. Proceedings of the National Academy of Sciences *108*, 3642.

Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466*, 1129-1133.

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

Iurlaro, M., Ficz, G., Oxley, D., Raiber, E.-A., Bachman, M., Booth, M.J., Andrews, S., Balasubramanian, S., and Reik, W. (2013). A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. Genome Biology *14*, R119.

Iwan, K., Rahimoff, R., Kirchner, A., Spada, F., Schröder, A.S., Kosmatchev, O., Ferizaj, S., Steinbacher, J., Parsa, E., Müller, M.*, et al.* (2017). 5-Formylcytosine to cytosine conversion by C–C bond cleavage in vivo. Nature Chemical Biology *14*, 72.

Iyer, L.M., Tahiliani, M., Rao, A., and Aravind, L. (2009). Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. Cell Cycle *8*, 1698-1710.

Ji, D., Lin, K., Song, J., and Wang, Y. (2014). Effects of Tet-induced oxidation products of 5-methylcytosine on Dnmt1- and DNMT3a-mediated cytosine methylation. Molecular BioSystems *10*, 1749-1752.

Jin, C., Lu, Y., Jelinek, J., Liang, S., Estecio, M.R.H., Barton, M.C., and Issa, J.-P.J. (2014). TET1 is a maintenance DNA demethylase that prevents methylation spreading in differentiated cells. Nucleic acids research *42*, 6956-6971.

Jin, S.-G., Kadam, S., and Pfeifer, G.P. (2010). Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. Nucleic Acids Research *38*, e125-e125.

Jin, S.-G., Wu, X., Li, A.X., and Pfeifer, G.P. (2011). Genomic mapping of 5-hydroxymethylcytosine in the human brain. Nucleic Acids Research *39*, 5015-5024.

Jin, S.-G., Zhang, Z.-M., Dunwell, Thomas L., Harter, Matthew R., Wu, X., Johnson, J., Li, Z., Liu, J., Szabó, Piroska E., Lu, Q., *et al.* (2016). Tet3 Reads 5-Carboxylcytosine through Its CXXC Domain and Is a Potential Guardian against Neurodegeneration. Cell Reports *14*, 493-505.

Jørgensen, H.F., Ben-Porath, I., and Bird, A.P. (2004). Mbd1 Is Recruited to both Methylated and Nonmethylated CpGs via Distinct DNA Binding Domains. Molecular and Cellular Biology *24*, 3387.

Kafri, T., Ariel, M., Brandeis, M., Shemer, R., Urven, L., McCarrey, J., Cedar, H., and Razin, A. (1992). Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. Genes Dev *6*, 705-714.

Kamat, A.A., Bischoff, F.Z., Dang, D., Baldwin, M., Han, L.Y., Lin, Y.G., Merritt, W.M., Landen, C.N., Lu, C., Gershenson, D.M., *et al.* (2006). Circulating cell-free DNA: A novel biomarker for response to therapy in ovarian carcinoma. Cancer Biology & Therapy *5*, 1369-1374.

Karpf, A.R., and Matsui, S.-i. (2005). Genetic Disruption of Cytosine DNA Methyltransferase Enzymes Induces Chromosomal Instability in Human Cancer Cells. Cancer Research *65*, 8635.

Keshet, I., Lieman-Hurwitz, J., and Cedar, H. (1986). DNA methylation affects the formation of active chromatin. Cell *44*, 535-543.

Khare, T., Pai, S., Koncevicius, K., Pal, M., Kriukiene, E., Liutkeviciute, Z., Irimia, M., Jia, P., Ptak, C., Xia, M., *et al.* (2012). 5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary. Nature Structural &Amp; Molecular Biology *19*, 1037.

Klein, B.J., Cox, K.L., Jang, S.M., Côté, J., Poirier, M.G., and Kutateladze, T.G. (2019). Molecular Basis for the PZP Domain of BRPF1 Association with Chromatin. Structure.

Klimasauskas, S., Kumar, S., Roberts, R.J., and Cheng, X. (1994). Hhal methyltransferase flips its target base out of the DNA helix. Cell *76*, 357-369.

Ko, M., An, J., Bandukwala, H.S., Chavez, L., Äijö, T., Pastor, W.A., Segal, M.F., Li, H., Koh, K.P., Lähdesmäki, H.*, et al.* (2013). Modulation of TET2 expression and 5-methylcytosine oxidation by the CXXC domain protein IDAX. Nature *497*, 122.

Ko, M., Bandukwala, H.S., An, J., Lamperti, E.D., Thompson, E.C., Hastie, R., Tsangaratou, A., Rajewsky, K., Koralov, S.B., and Rao, A. (2011). Ten-Eleven-Translocation 2 (TET2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice. Proceedings of the National Academy of Sciences *108*, 14566.

Ko, M., Huang, Y., Jankowska, A.M., Pape, U.J., Tahiliani, M., Bandukwala, H.S., An, J., Lamperti, E.D., Koh, K.P., Ganetzky, R.*, et al.* (2010). Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. Nature *468*, 839-843.

Koh, K.P., Yabuuchi, A., Rao, S., Huang, Y., Cunniff, K., Nardone, J., Laiho, A., Tahiliani, M., Sommer, C.A., Mostoslavsky, G.*, et al.* (2011). Tet1 and Tet2 Regulate 5-Hydroxymethylcytosine Production and Cell Lineage Specification in Mouse Embryonic Stem Cells. Cell Stem Cell *8*, 200-213.

Kothari, R., and Shankar, V. (1976). 5-Methylcytosine content in the vertebrate deoxyribonucleic acids: species specificity. Journal of molecular evolution *7*, 325-329.

Kriaucionis, S., and Heintz, N. (2009). The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. Science *324*, 929-930.

Kubosaki, A., Tomaru, Y., Furuhata, E., Suzuki, T., Shin, J.W., Simon, C., Ando, Y., Hasegawa, R., Hayashizaki, Y., and Suzuki, H. (2012). CpG site-specific alteration of hydroxymethylcytosine to methylcytosine beyond DNA replication. Biochemical and Biophysical Research Communications *426*, 141-147.

Kuo, K.C., McCune, R.A., Gehrke, C.W., Midgett, R., and Ehrlich, M. (1980). Quantitative reversed-phase high performance liquid chromatographic determination of major and modified deoxyribonucleosides in DNA. Nucleic acids research *8*, 4763-4776.

Laget, S., Joulie, M., Le Masson, F., Sasai, N., Christians, E., Pradhan, S., Roberts, R.J., and Defossez, P.-A. (2010). The Human Proteins MBD5 and MBD6 Associate with Heterochromatin but They Do Not Bind Methylated DNA. PLOS ONE *5*, e11982.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lee, C., Bae, K., and Edery, I. (1999). PER and TIM Inhibit the DNA Binding Activity of aDrosophila CLOCK-CYC/dBMAL1 Heterodimer without Disrupting Formation of the Heterodimer: a Basis for Circadian Transcription. Molecular and Cellular Biology *19*, 5316.

Lemonnier, F., Couronne, L., Parrens, M., Jais, J.P., Travert, M., Lamant, L., Tournillac, O., Rousset, T., Fabiani, B., Cairns, R.A.*, et al.* (2012). Recurrent TET2 mutations in peripheral T-cell lymphomas correlate with TFH-like features and adverse clinical parameters. Blood *120*, 1466-1469.

Leonhardt, H., Page, A.W., Weier, H.-U., and Bestor, T.H. (1992). A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. Cell *71*, 865-873.

Lewis, J.D., Meehan, R.R., Henzel, W.J., Maurer-Fogy, I., Jeppesen, P., Klein, F., and Bird, A. (1992). Purification, sequence, and cellular localization of a novel chromosomal protein that binds to Methylated DNA. Cell *69*, 905-914.

Li, A., Chen, Y.-S., Ping, X.-L., Yang, X., Xiao, W., Yang, Y., Sun, H.-Y., Zhu, Q., Baidya, P., Wang, X.*, et al.* (2017a). Cytoplasmic m6A reader YTHDF3 promotes mRNA translation. Cell Res *27*, 444-447.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915-926.

Li, W., and Liu, M. (2011). Distribution of 5-Hydroxymethylcytosine in Different Human Tissues. Journal of Nucleic Acids *2011*.

Li, W., Zhang, X., Lu, X., You, L., Song, Y., Luo, Z., Zhang, J., Nie, J., Zheng, W., Xu, D.*, et al.* (2017b). 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. Cell Research *27*, 1243.

Li, X., Wei, W., Zhao, Q.-Y., Widagdo, J., Baker-Andresen, D., Flavell, C.R., D'Alessio, A., Zhang, Y., and Bredy, T.W. (2014). Neocortical Tet3-mediated accumulation of 5-hydroxymethylcytosine promotes rapid behavioral adaptation. Proceedings of the National Academy of Sciences *111*, 7120.

Li, Z., Cai, X., Cai, C.-L., Wang, J., Zhang, W., Petersen, B.E., Yang, F.-C., and Xu, M. (2011a). Deletion of Tet2 in mice leads to dysregulated hematopoietic stem cells and subsequent development of myeloid malignancies. Blood *118*, 4509-4518.

Li, Z., Cai, X., Cai, C.L., Wang, J., Zhang, W., Petersen, B.E., Yang, F.C., and Xu, M. (2011b). Deletion of Tet2 in mice leads to dysregulated hematopoietic stem cells and subsequent development of myeloid malignancies. Blood *118*, 4509-4518.

Lian, C.G., Xu, Y., Ceol, C., Wu, F., Larson, A., Dresser, K., Xu, W., Tan, L., Hu, Y., Zhan, Q.*, et al.* (2012). Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma. Cell *150*, 1135-1146.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. Nature *362*, 709-715.

Lindahl, T., and Nyberg, B. (1974). Heat-induced deamination of cytosine residues in deoxyribonucleic acid. Biochemistry *13*, 3405-3410.

Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D.*, et al.* (2013). Global epigenomic reconfiguration during mammalian brain development. Science (New York, NY) *341*, 1237905-1237905.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M.*, et al.* (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315.

Liu, C., Liu, L., Chen, X., Shen, J., Shan, J., Xu, Y., Yang, Z., Wu, L., Xia, F., Bie, P.*, et al.* (2013a). Decrease of 5-hydroxymethylcytosine is associated with progression of hepatocellular carcinoma through downregulation of TET1. PLoS One *8*, e62828.

Liu, S., Wang, J., Su, Y., Guerrero, C., Zeng, Y., Mitra, D., Brooks, P.J., Fisher, D.E., Song, H., and Wang, Y. (2013b). Quantitative assessment of Tet-induced oxidation products of 5-methylcytosine in cellular and tissue DNA. Nucleic Acids Research *41*, 6421-6429.

Liu, Y., Siejka-Zielinska, P., Velikova, G., Bi, Y., Yuan, F., Tomkova, M., Bai, C., Chen, L., Schuster-Bockler, B., and Song, C.X. (2019). Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. Nat Biotechnol *37*, 424-429.

Liutkevičiūtė, Z., Kriukienė, E., Ličytė, J., Rudytė, M., Urbanavičiūtė, G., and Klimašauskas, S. (2014). Direct Decarboxylation of 5-Carboxylcytosine by DNA C5-Methyltransferases. Journal of the American Chemical Society *136*, 5884-5887.

Liutkevičiūtė, Z., Lukinavičius, G., Masevičius, V., Daujotytė, D., and Klimašauskas, S. (2009). Cytosine-5-methyltransferases add aldehydes to DNA. Nature Chemical Biology *5*, 400-402.

Lock, L.F., Takagi, N., and Martin, G.R. (1987). Methylation of the Hprt gene on the inactive X occurs after chromosome inactivation. Cell *48*, 39-46.

Lorsbach, R.B., Moore, J., Mathew, S., Raimondi, S.C., Mukatira, S.T., and Downing, J.R. (2003). TET1, a member of a novel protein family, is fused to MLL in acute myeloid leukemia containing the t(10;11)(q22;q23). Leukemia *17*, 637-641.

Lu, F., Liu, Y., Jiang, L., Yamaguchi, S., and Zhang, Y. (2014). Role of Tet proteins in enhancer activity and telomere elongation. Genes & Development *28*, 2103-2119.

Lu, S. (2014). Zn2+ Blocks Annealing of Complementary Single-Stranded DNA in a Sequence-Selective Manner. Scientific Reports *4*, 5464.

Lyst, M.J., Ekiert, R., Ebert, D.H., Merusi, C., Nowak, J., Selfridge, J., Guy, J., Kastan, N.R., Robinson, N.D., de Lima Alves, F.*, et al.* (2013). Rett syndrome mutations abolish the interaction of MeCP2 with the NCoR/SMRT co-repressor. Nature Neuroscience *16*, 898.

Maiti, A., and Drohat, A.C. (2011). Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine POTENTIAL IMPLICATIONS FOR ACTIVE DEMETHYLATION OF CpG SITES. J Biol Chem *286*, 35334-35338.

Mayer, W., Niveleau, A., Walter, J., Fundele, R., and Haaf, T. (2000). Demethylation of the zygotic paternal genome. Nature *403*, 501-502.

Meehan, R.R., Lewis, J.D., and Bird, A.P. (1992). Characterization of MeCP2, a vertebrate DNA binding protein with affinity for methylated DNA. Nucleic acids research *20*, 5085-5092.

Mellen, M., Ayata, P., Dewell, S., Kriaucionis, S., and Heintz, N. (2012). MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. Cell *151*, 1417-1430.

Miao, Z., Xin, N., Wei, B., Hua, X., Zhang, G., Leng, C., Zhao, C., Wu, D., Li, J., Ge, W.*, et al.* (2016). 5-hydroxymethylcytosine is detected in RNA from mouse brain tissues. Brain Research *1642*, 546-552.

Millar, C.B., Guy, J., Sansom, O.J., Selfridge, J., MacDougall, E., Hendrich, B., Keightley, P.D., Bishop, S.M., Clarke, A.R., and Bird, A. (2002). Enhanced CpG Mutability and Tumorigenesis in MBD4-Deficient Mice. Science *297*, 403.

Misawa, K., Imai, A., Mochizuki, D., Mima, M., Endo, S., Misawa, Y., Kanazawa, T., and Mineta, H. (2018). Association of TET3 epigenetic inactivation with head and neck cancer. Oncotarget *9*, 24480-24493.

Mohandas, T., Sparkes, R.S., and Shapiro, L.J. (1981). Reactivation of an Inactive Human X Chromosome: Evidence for X Inactivation by DNA Methylation. Science *211*, 393-396.

Monk, M., Boubelik, M., and Lehnert, S. (1987). Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development *99*, 371.

Moran-Crusio, K., Reavie, L., Shih, A., Abdel-Wahab, O., Ndiaye-Lobry, D., Lobry, C., Figueroa, M.E., Vasanthakumar, A., Patel, J., Zhao, X.*, et al.* (2011). Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. Cancer Cell *20*, 11-24.

Morgan, H.D., Dean, W., Coker, H.A., Reik, W., and Petersen-Mahrt, S.K. (2004). Activation-induced Cytidine Deaminase Deaminates 5-Methylcytosine in DNA and Is Expressed in Pluripotent Tissues: IMPLICATIONS FOR EPIGENETIC REPROGRAMMING. J Biol Chem *279*, 52353-52360.

Muller, T., Gessi, M., Waha, A., Isselstein, L.J., Luxen, D., Freihoff, D., Freihoff, J., Becker, A., Simon, M., Hammes, J.*, et al.* (2012). Nuclear Exclusion of TET1 Is Associated with Loss of 5-Hydroxymethylcytosine in IDH1 Wild-Type Gliomas. Am J Pathol *181*, 675-683.

Nan, X., Campoy, F.J., and Bird, A. (1997). MeCP2 Is a Transcriptional Repressor with Abundant Binding Sites in Genomic Chromatin. Cell *88*, 471-481.

Nan, X., Meehan, R.R., and Bird, A. (1993). Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. Nucleic Acids Research *21*, 4886-4892.

Nestor, C., Ruzov, A., Meehan, R.R., and Dunican, D.S. (2010). Enzymatic approaches and bisulfite sequencing cannot distinguish between 5-methylcytosine and 5-hydroxymethylcytosine in DNA. Biotechniques *48*, 317-319.

Nestor, C.E., and Meehan, R.R. (2014). Hydroxymethylated DNA Immunoprecipitation (hmeDIP). In Functional Analysis of DNA and Chromatin, J.C.

Stockert, J. Espada, and A. Blázquez-Castro, eds. (Totowa, NJ: Humana Press), pp. 259-267.

Nestor, C.E., Ottaviano, R., Reddington, J., Sproul, D., Reinhardt, D., Dunican, D., Katz, E., Dixon, J.M., Harrison, D.J., and Meehan, R.R. (2012). Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. Genome Research *22*, 467-477.

Noushmehr, H., Weisenberger, D.J., Diefes, K., Phillips, H.S., Pujara, K., Berman, B.P., Pan, F., Pelloski, C.E., Sulman, E.P., Bhat, K.P.*, et al.* (2010). Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. Cancer Cell *17*, 510-522.

Okamoto, A., Sugizaki, K., Nakamura, A., Yanagisawa, H., and Ikeda, S. (2011). 5-Hydroxymethylcytosine-selective oxidation with peroxotungstate. Chem Commun (Camb) *47*, 11231-11233.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. Cell *99*, 247-257.

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

Okano, M., Xie, S., and Li, E. (1998b). Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells. Nucleic Acids Research *26*, 2536-2540.

Ono, R., Taki, T., Taketani, T., Taniwaki, M., Kobayashi, H., and Hayashi, Y. (2002). Leukemia-associated Protein with a CXXC Domain, Is Fused to in Acute Myeloid

Leukemia with Trilineage Dysplasia Having t(10;11)(q22;q23). Cancer Research *62*, 4075.

Orr, B.A., Haffner, M.C., Nelson, W.G., Yegnasubramanian, S., and Eberhart, C.G. (2012). Decreased 5-Hydroxymethylcytosine Is Associated with Neural Progenitor Phenotype in Normal Brain and Shorter Survival in Malignant Glioma. PLOS ONE *7*, e41036.

Oshimo, Y., Nakayama, H., Ito, R., Kitadai, Y., Yoshida, K., Chayama, K., and Yasui, W. (2003). Promoter methylation of cyclin D2 gene in gastric carcinoma. Int J Oncol *23*, 1663-1670.

Oswald, J., Engemann, S., Lane, N., Mayer, W., Olek, A., Fundele, R., Dean, W., Reik, W., and Walter, J. (2000). Active demethylation of the paternal genome in the mouse zygote. Current Biology *10*, 475-478.

Otani, J., Kimura, H., Sharif, J., Endo, T.A., Mishima, Y., Kawakami, T., Koseki, H., Shirakawa, M., Suetake, I., and Tajima, S. (2013). Cell Cycle-Dependent Turnover of 5-Hydroxymethyl Cytosine in Mouse Embryonic Stem Cells. PLOS ONE *8*, e82961.

Pastor, W.A., Pape, U.J., Huang, Y., Henderson, H.R., Lister, R., Ko, M., McLoughlin, E.M., Brudno, Y., Mahapatra, S., Kapranov, P*., et al.* (2011). Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature *473*, 394.

Penn, N., Suwalski, R., O'riley, C., Bojanowski, K., and Yura, R. (1972). The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. Biochemical Journal *126*, 781-790.

Petronzelli, F., Riccio, A., Markham, G.D., Seeholzer, S.H., Genuardi, M., Karbowski, M., Yeung, A.T., Matsumoto, Y., and Bellacosa, A. (2000). Investigation of the substrate spectrum of the human mismatch-specific DNA N-glycosylase MED1 (MBD4): fundamental role of the catalytic domain. J Cell Physiol *185*, 473-480.

Popp, C., Dean, W., Feng, S., Cokus, S.J., Andrews, S., Pellegrini, M., Jacobsen, S.E., and Reik, W. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature *463*, 1101.

Privat, E., and Sowers, L.C. (1996). Photochemical Deamination and Demethylation of 5-Methylcytosine. Chemical Research in Toxicology *9*, 745-750.

Prokhortchouk, A., Hendrich, B., Jorgensen, H., Ruzov, A., Wilm, M., Georgiev, G., Bird, A., and Prokhortchouk, E. (2001). The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor. Genes & Development *15*, 1613-1618.

Quivoron, C., Couronne, L., Della Valle, V., Lopez, C.K., Plo, I., Wagner-Ballon, O., Do Cruzeiro, M., Delhommeau, F., Arnulf, B., Stern, M.H.*, et al.* (2011). TET2 inactivation results in pleiotropic hematopoietic abnormalities in mouse and is a recurrent event during human lymphomagenesis. Cancer Cell *20*, 25-38.

Ramsahoye, B.H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A.P., and Jaenisch, R. (2000). Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proceedings of the National Academy of Sciences *97*, 5237.

Robinson, L., Guy, J., McKay, L., Brockett, E., Spike, R.C., Selfridge, J., De Sousa, D., Merusi, C., Riedel, G., Bird, A.*, et al.* (2012). Morphological and functional reversal of phenotypes in a mouse model of Rett syndrome. Brain *135*, 2699-2710.

Roloff, T.C., Ropers, H.H., and Nuber, U.A. (2003). Comparative study of methyl-CpG-binding domain proteins. BMC Genomics *4*, 1.

Rudenko, A., Dawlaty, Meelad M., Seo, J., Cheng, Albert W., Meng, J., Le, T., Faull, Kym F., Jaenisch, R., and Tsai, L.-H. (2013). Tet1 Is Critical for Neuronal Activity-Regulated Gene Expression and Memory Extinction. Neuron *79*, 1109-1122.

Saito, M., and Ishikawa, F. (2002). The mCpG-binding Domain of Human MBD3 Does Not Bind to mCpG but Interacts with NuRD/Mi2 Components HDAC1 and MTA2. J Biol Chem *277*, 35434-35439.

Santos, F., Hendrich, B., Reik, W., and Dean, W. (2002). Dynamic Reprogramming of DNA Methylation in the Early Mouse Embryo. Developmental Biology *241*, 172-182.

Schaefer, M., Pollex, T., Hanna, K., and Lyko, F. (2008). RNA cytosine methylation analysis by bisulfite sequencing. Nucleic Acids Research *37*, e12-e12.

Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T.*, et al.* (2000). A gene expression database for the molecular pharmacology of cancer. Nature Genetics *24*, 236-244.

Schutsky, E.K., DeNizio, J.E., Hu, P., Liu, M.Y., Nabel, C.S., Fabyanic, E.B., Hwang, Y., Bushman, F.D., Wu, H., and Kohli, R.M. (2018). Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. Nature Biotechnology *36*, 1083.

Seiler, C.L., Fernandez, J., Koerperich, Z., Andersen, M.P., Kotandeniya, D., Nguyen, M.E., Sham, Y.Y., and Tretyakova, N.Y. (2018). Maintenance DNA Methyltransferase Activity in the Presence of Oxidized Forms of 5-Methylcytosine: Structural Basis for Ten Eleven Translocation-Mediated DNA Demethylation. Biochemistry *57*, 6061-6069.

Selivanova, O.M., Guryanov, S.G., Enin, G.A., Skabkin, M.A., Ovchinnikov, L.P., and Serdyuk, I.N. (2010). YB-1 is capable of forming extended nanofibrils. Biochemistry (Moscow) *75*, 115-120.

Shafik, A., Schumann, U., Evers, M., Sibbritt, T., and Preiss, T. (2016). The emerging epitranscriptomics of long noncoding RNAs. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms *1859*, 59-70.

Sharif, J., Muto, M., Takebayashi, S.-i., Suetake, I., Iwamatsu, A., Endo, T.A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., Okamura, K.*, et al.* (2007). The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. Nature *450*, 908.

Shen, J.C., Rideout, W.M., 3rd, and Jones, P.A. (1994). The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic acids research *22*, 972-976.

Shen, L., Wu, H., Diep, D., Yamaguchi, S., D'Alessio, Ana C., Fung, H.-L., Zhang, K., and Zhang, Y. (2013). Genome-wide Analysis Reveals TET- and TDG-Dependent 5-Methylcytosine Oxidation Dynamics. Cell *153*, 692-706.

Shen, S.Y., Singhania, R., Fehringer, G., Chakravarthy, A., Roehrl, M.H.A., Chadwick, D., Zuzarte, P.C., Borgida, A., Wang, T.T., Li, T.*, et al.* (2018). Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature *563*, 579-583.

Shi, H., Wang, X., Lu, Z., Zhao, B.S., Ma, H., Hsu, P.J., Liu, C., and He, C. (2017). YTHDF3 facilitates translation and decay of N6-methyladenosine-modified RNA. Cell Res *27*, 315-328.

Siegfried, Z., Eden, S., Mendelsohn, M., Feng, X., Tsuberi, B.-Z., and Cedar, H. (1999). DNA methylation represses transcription in vivo. Nature Genetics *22*, 203-206.

Song, C.-X., Diao, J., Brunger, A.T., and Quake, S.R. (2016). Simultaneous single-molecule epigenetic imaging of DNA methylation and hydroxymethylation. Proceedings of the National Academy of Sciences *113*, 4338-4343.

Song, C.-X., Szulwach, Keith E., Dai, Q., Fu, Y., Mao, S.-Q., Lin, L., Street, C., Li, Y., Poidevin, M., Wu, H.*, et al.* (2013). Genome-wide Profiling of 5-Formylcytosine Reveals Its Roles in Epigenetic Priming. Cell *153*, 678-691.

Song, C.-X., Szulwach, K.E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.-H., Zhang, W., Jian, X.*, et al.* (2011). Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat Biotech *29*, 68-72.

Song, C.-X., Yin, S., Ma, L., Wheeler, A., Chen, Y., Zhang, Y., Liu, B., Xiong, J., Zhang, W., Hu, J.*, et al.* (2017). 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. Cell Research *27*, 1231.

Song, J., Teplova, M., Ishibe-Murakami, S., and Patel, D.J. (2012). Structure-Based Mechanistic Insights into DNMT1-Mediated Maintenance DNA Methylation. Science *335*, 709.

Sperlazza, M.J., Bilinovich, S.M., Sinanan, L.M., Javier, F.R., and Williams, D.C. (2017). Structural Basis of MeCP2 Distribution on Non-CpG Methylated and Hydroxymethylated DNA. Journal of Molecular Biology *429*, 1581-1594.

Spruijt, C.G., Gnerlich, F., Smits, A.H., Pfaffeneder, T., Jansen, P.W., Bauer, C., Munzel, M., Wagner, M., Muller, M., Khan, F.*, et al.* (2013). Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. Cell *152*, 1146-1159.

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

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., Nimwegen, E.v., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D.*, et al.* (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490.

Stein, R., Gruenbaum, Y., Pollack, Y., Razin, A., and Cedar, H. (1982). Clonal inheritance of the pattern of DNA methylation in mouse cells. Proceedings of the National Academy of Sciences *79*, 61.

Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S., and Jacobsen, S.E. (2011). 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. Genome Biology *12*, R54.

Stroynowska-Czerwinska, A., Piasecka, A., and Bochtler, M. (2018). Specificity of MLL1 and TET3 CXXC domains towards naturally occurring cytosine modifications. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms *1861*, 1093-1101.

Suetake, I., Shinozaki, F., Miyagawa, J., Takeshima, H., and Tajima, S. (2004). DNMT3L Stimulates the DNA Methylation Activity of Dnmt3a and Dnmt3b through a Direct Interaction. J Biol Chem *279*, 27816-27823.

Sun, K., Jiang, P., Chan, K.C.A., Wong, J., Cheng, Y.K.Y., Liang, R.H.S., Chan, W.-k., Ma, E.S.K., Chan, S.L., Cheng, S.H.*, et al.* (2015). Plasma DNA tissue mapping by genome-wide methylation sequencing for noninvasive prenatal, cancer, and transplantation assessments. Proceedings of the National Academy of Sciences *112*, E5503.

Szulwach, K.E., Li, X., Li, Y., Song, C.-X., Han, J.W., Kim, S., Namburi, S., Hermetz, K., Kim, J.J., Rudd, M.K.*, et al.* (2011a). Integrating 5-Hydroxymethylcytosine into the Epigenomic Landscape of Human Embryonic Stem Cells. PLOS Genetics *7*, e1002154.

Szulwach, K.E., Li, X., Li, Y., Song, C.-X., Wu, H., Dai, Q., Irier, H., Upadhyay, A.K., Gearing, M., Levey, A.I.*, et al.* (2011b). 5-hmC–mediated epigenetic dynamics during postnatal neurodevelopment and aging. Nature Neuroscience *14*, 1607.

Szwagierczak, A., Bultmann, S., Schmidt, C.S., Spada, F., and Leonhardt, H. (2010). Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic DNA. Nucleic Acids Research *38*, e181-e181.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L.*, et al.* (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science *324*, 930-935.

Takai, H., Masuda, K., Sato, T., Sakaguchi, Y., Suzuki, T., Suzuki, T., Koyama-Nasu, R., Nasu-Nishimura, Y., Katou, Y., Ogawa, H.*, et al.* (2014). 5-Hydroxymethylcytosine plays a critical role in glioblastomagenesis by recruiting the CHTOP-methylosome complex. Cell Rep *9*, 48-60.

Tamanaha, E., Guan, S., Marks, K., and Saleh, L. (2016). Distributive Processing by the Iron(II)/α-Ketoglutarate-Dependent Catalytic Domains of the TET Enzymes Is Consistent with Epigenetic Roles for Oxidized 5-Methylcytosine Bases. Journal of the American Chemical Society *138*, 9345-9348.

Thierry, A.R., Mouliere, F., Gongora, C., Ollier, J., Robert, B., Ychou, M., Del Rio, M., and Molina, F. (2010). Origin and quantification of circulating DNA in mice with human colorectal cancer xenografts. Nucleic Acids Research *38*, 6159-6175.

Thorvaldsen, J.L., Duran, K.L., and Bartolomei, M.S. (1998). Deletion of the H19 differentially methylated domain results in loss of imprinted expression of H19 and Igf2. Genes & development *12*, 3693-3702.

Thul, P.J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L.M.*, et al.* (2017). A subcellular map of the human proteome. Science *356*, eaal3321.

Tian, X., Sun, B., Chen, C., Gao, C., Zhang, J., Lu, X., Wang, L., Li, X., Xing, Y., Liu, R.*, et al.* (2018). Circulating tumor DNA 5-hydroxymethylcytosine as a novel diagnostic biomarker for esophageal cancer. Cell Research *28*, 597-600.

Tomkova, M., McClellan, M., Kriaucionis, S., and Schuster-Boeckler, B. (2016). 5-hydroxymethylcytosine marks regions with reduced mutation frequency in human DNA. eLife *5*, e17082.

Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J.G., Baylin, S.B., and Issa, J.P. (1999). CpG island methylator phenotype in colorectal cancer. Proceedings of the National Academy of Sciences of the United States of America *96*, 8681-8686.

Tremblay, K.D., Saam, J.R., Ingram, R.S., Tilghman, S.M., and Bartolomei, M.S. (1995). A paternal–specific methylation imprint marks the alleles of the mouse H19 gene. Nature Genetics *9*, 407-413.

Unoki, M., Nishidate, T., and Nakamura, Y. (2004). ICBP90, an E2F-1 target, recruits HDAC1 and binds to methyl-CpG through its SRA domain. Oncogene *23*, 7601-7610.

Valinluck, V., and Sowers, L.C. (2007). Endogenous Cytosine Damage Products Alter the Site Selectivity of Human DNA Maintenance Methyltransferase DNMT1. Cancer Research *67*, 946.

Vasanthakumar, A., and Godley, L.A. (2015). 5-hydroxymethylcytosine in cancer: significance in diagnosis and therapy. Cancer Genetics *208*, 167-177.

Vaughan, R.M., Dickson, B.M., Cornett, E.M., Harrison, J.S., Kuhlman, B., and Rothbart, S.B. (2018). Comparative biochemical analysis of UHRF proteins reveals

molecular mechanisms that uncouple UHRF2 from DNA methylation maintenance. Nucleic Acids Research *46*, 4405-4416.

Venolia, L., Gartler, S.M., Wassman, E.R., Yen, P., Mohandas, T., and Shapiro, L.J. (1982). Transformation with DNA from 5-azacytidine-reactivated X chromosomes. Proceedings of the National Academy of Sciences of the United States of America *79*, 2352-2354.

Vidal, E., Sayols, S., Moran, S., Guillaumet-Adkins, A., Schroeder, M.P., Royo, R., Orozco, M., Gut, M., Gut, I., Lopez-Bigas, N.*, et al.* (2017). A DNA methylation map of human cancer at single base-pair resolution. Oncogene *36*, 5648.

Viguié, F., Aboura, A., Bouscary, D., Ramond, S., Delmer, A., Tachdjian, G., Marie, J.P., and Casadevall, N. (2005). Common 4q24 deletion in four cases of hematopoietic malignancy: early stem cell involvement? Leukemia *19*, 1411-1415.

Voss, A.K., Collin, C., Dixon, M.P., and Thomas, T. (2009). Moz and Retinoic Acid Coordinately Regulate H3K9 Acetylation, Hox Gene Expression, and Segment Identity. Developmental Cell *17*, 674-686.

Walsh, C.P., Chaillet, J.R., and Bestor, T.H. (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nature Genetics *20*, 116-117.

Wan, J.C.M., Massie, C., Garcia-Corbacho, J., Mouliere, F., Brenton, J.D., Caldas, C., Pacey, S., Baird, R., and Rosenfeld, N. (2017). Liquid biopsies come of age: towards implementation of circulating tumour DNA. Nature Reviews Cancer *17*, 223.

Wang, R., Gao, X., and Yu, L. (2019). The prognostic impact of tet oncogene family member 2 mutations in patients with acute myeloid leukemia: a systematic-review and meta-analysis. BMC Cancer *19*, 389.

Wang, T., Pan, Q., Lin, L., Szulwach, K.E., Song, C.-X., He, C., Wu, H., Warren, S.T., Jin, P., Duan, R.*, et al.* (2012). Genome-wide DNA hydroxymethylation changes

are associated with neurodevelopmental genes in the developing human cerebellum. Human Molecular Genetics *21*, 5500-5510.

Wang, X., Lim, H.J., and Son, A. (2014). Characterization of denaturation and renaturation of DNA for DNA hybridization. Environmental Health and Toxicology *29*, e2014007.

Weber, A.R., Krawczyk, C., Robertson, A.B., Kuśnierczyk, A., Vågbø, C.B., Schuermann, D., Klungland, A., and Schär, P. (2016). Biochemical reconstitution of TET1–TDG–BER-dependent active DNA demethylation reveals a highly coordinated mechanism. Nature Communications *7*, 10806.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Pääbo, S., Rebhan, M., and Schübeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nature Genetics *39*, 457-466.

Weissmann, S., Alpermann, T., Grossmann, V., Kowarsch, A., Nadarajah, N., Eder, C., Dicker, F., Fasan, A., Haferlach, C., Haferlach, T.*, et al.* (2011). Landscape of TET2 mutations in acute myeloid leukemia. Leukemia *26*, 934.

Wen, L., Li, X., Yan, L., Tan, Y., Li, R., Zhao, Y., Wang, Y., Xie, J., Zhang, Y., Song, C.*, et al.* (2014). Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. Genome Biology *15*, R49.

Wigler, M., Levy, D., and Perucho, M. (1981). The somatic replication of DNA methylation. Cell *24*, 33-40.

Williams, K., Christensen, J., Pedersen, M.T., Johansen, J.V., Cloos, P.A.C., Rappsilber, J., and Helin, K. (2011). TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature *473*, 343.

Wilson, V.L., Smith, R.A., Autrup, H., Krokan, H., Musci, D.E., Le, N.-N.-T., Longoria, J., Ziska, D., and Harris, C.C. (1986). Genomic 5-methylcytosine determination by 32P-postlabeling analysis. Analytical Biochemistry *152*, 275-284.

Wossidlo, M., Nakamura, T., Lepikhov, K., Marques, C.J., Zakhartchenko, V., Boiani, M., Arand, J., Nakano, T., Reik, W., and Walter, J. (2011). 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. Nature Communications *2*, 241.

Wu, H., D'Alessio, A.C., Ito, S., Wang, Z., Cui, K., Zhao, K., Sun, Y.E., and Zhang, Y. (2011). Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. Genes & Development *25*, 679-684.

Wu, S.C., and Zhang, Y. (2010). Active DNA demethylation: many roads lead to Rome. Nature Reviews Molecular Cell Biology *11*, 607.

Wyatt, G.R. (1951). Recognition and estimation of 5-methylcytosine in nucleic acids. The Biochemical journal *48*, 581-584.

Wyatt, G.R., and Markham, R. (1950). The quantitative analysis of deoxypentose nucleic acids. Biochem J *47*, vii.

Xiong, J., Zhang, Z., Chen, J., Huang, H., Xu, Y., Ding, X., Zheng, Y., Nishinakamura, R., Xu, G.L., Wang, H.*, et al.* (2016). Cooperative Action between SALL4A and TET Proteins in Stepwise Oxidation of 5-Methylcytosine. Mol Cell *64*, 913-925.

Xu, Y., Wu, F., Tan, L., Kong, L., Xiong, L., Deng, J., Barbera, A.J., Zheng, L., Zhang, H., Huang, S.*, et al.* (2011). Genome-wide Regulation of 5hmC, 5mC, and Gene Expression by Tet1 Hydroxylase in Mouse Embryonic Stem Cells. Molecular Cell *42*, 451-464.

Yamaguchi, S., Shen, L., Liu, Y., Sendler, D., and Zhang, Y. (2013). Role of Tet1 in erasure of genomic imprinting. Nature *504*, 460.

Yang, H., Liu, Y., Bai, F., Zhang, J.Y., Ma, S.H., Liu, J., Xu, Z.D., Zhu, H.G., Ling, Z.Q., Ye, D.*, et al.* (2013). Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. Oncogene *32*, 663-669.

Yang, X.-J. (2015). MOZ and MORF acetyltransferases: Molecular interaction, animal development and human disease. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research *1853*, 1818-1826.

Yang, Y., Fan, X., Mao, M., Song, X., Wu, P., Zhang, Y., Jin, Y., Yang, Y., Chen, L.-L., Wang, Y.*, et al.* (2017). Extensive translation of circular RNAs driven by N6-methyladenosine. Cell Res *27*, 626-641.

Yang, Y., Wang, L., Han, X., Yang, W.-L., Zhang, M., Ma, H.-L., Sun, B.-F., Li, A., Xia, J., Chen, J.*, et al.* (2019). RNA 5-Methylcytosine Facilitates the Maternal-to-Zygotic Transition by Preventing Maternal mRNA Decay. Molecular Cell.

Ye, Z., Li, J., Han, X., Hou, H., Chen, H., Zheng, X., Lu, J., Wang, L., Chen, W., Li, X.*, et al.* (2016). TET3 inhibits TGF-β1-induced epithelial-mesenchymal transition by demethylating miR-30d precursor gene in ovarian cancer cells. Journal of Experimental & Clinical Cancer Research *35*, 72.

Yildirim, O., Li, R., Hung, J.H., Chen, P.B., Dong, X., Ee, L.S., Weng, Z., Rando, O.J., and Fazzio, T.G. (2011). Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. Cell *147*, 1498-1510.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F.*, et al.* (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*, eaaj2239.

Yoder, J.A., Soman, N.S., Verdine, G.L., and Bestor, T.H. (1997). DNA (cytosine-5)-methyltransferases in mouse cells and tissues. Studies with a mechanism-based probe. J Mol Biol *270*, 385-395.

Yu, H., Su, Y., Shin, J., Zhong, C., Guo, J.U., Weng, Y.-L., Gao, F., Geschwind, D.H., Coppola, G., Ming, G.-l.*, et al.* (2015). Tet3 regulates synaptic transmission and homeostatic plasticity via DNA oxidation and repair. Nature Neuroscience *18*, 836.

Yu, M., Hon, G.C., Szulwach, K.E., Song, C.-X., Jin, P., Ren, B., and He, C. (2012a). Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. Nat Protocols *7*, 2159-2170.

Yu, M., Hon, G.C., Szulwach, K.E., Song, C.X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B.*, et al.* (2012b). Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell *149*, 1368-1380.

Yuan, F., Bi, Y., Zhang, J.-Y., Zhou, Y.-L., Zhang, X.-X., and Song, C.-X. (2019). 5-Carboxylcytosine is resistant towards phosphodiesterase I digestion: implications for epigenetic modification quantification by mass spectrometry. RSC Advances *9*, 29010-29014.

Zeng, H., He, B., Xia, B., Bai, D., Lu, X., Cai, J., Chen, L., Zhou, A., Zhu, C., Meng, H.*, et al.* (2018). Bisulfite-Free, Nanoscale Analysis of 5-Hydroxymethylcytosine at Single Base Resolution. Journal of the American Chemical Society *140*, 13190-13194.

Zhang, H., Zhang, X., Clark, E., Mulcahey, M., Huang, S., and Shi, Y.G. (2010). TET1 is a DNA-binding protein that modulates DNA methylation and gene transcription via hydroxylation of 5-methylcytosine. Cell Research *20*, 1390.

Zhang, H.-Y., Xiong, J., Qi, B.-L., Feng, Y.-Q., and Yuan, B.-F. (2016). The existence of 5-hydroxymethylcytosine and 5-formylcytosine in both DNA and RNA in mammals. Chemical Communications *52*, 737-740.

Zhang, J., Han, X., Gao, C., Xing, Y., Qi, Z., Liu, R., Wang, Y., Zhang, X., Yang, Y.-G., Li, X.*, et al.* (2018). 5-Hydroxymethylome in Circulating Cell-free DNA as A Potential Biomarker for Non-small-cell Lung Cancer. Genomics, proteomics & bioinformatics *16*, 187-199.

Zhang, R.-R., Cui, Q.-Y., Murai, K., Lim, Yen C., Smith, Zachary D., Jin, S., Ye, P., Rosa, L., Lee, Yew K., Wu, H.-P.*, et al.* (2013). Tet1 Regulates Adult Hippocampal Neurogenesis and Cognition. Cell Stem Cell *13*, 237-245.

Zhang, Y., Ng, H.-H., Erdjument-Bromage, H., Tempst, P., Bird, A., and Reinberg, D. (1999). Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. Genes & Development *13*, 1924-1935.

Zhou, T., Xiong, J., Wang, M., Yang, N., Wong, J., Zhu, B., and Xu, R.-M. (2014). Structural Basis for Hydroxymethylcytosine Recognition by the SRA Domain of UHRF2. Molecular Cell *54*, 879-886.

Zhu, B., Zheng, Y., Hess, D., Angliker, H., Schwarz, S., Siegmann, M., Thiry, S., and Jost, J.-P. (2000). 5-Methylcytosine-DNA glycosylase activity is present in a cloned G/T mismatch DNA glycosylase associated with the chicken embryo DNA demethylation complex. Proceedings of the National Academy of Sciences *97*, 5135.