

The Biological Role of DNA Modifications Deposited Over Gene Bodies

Pijus Brazauskas

Green Templeton College



This thesis is submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Ludwig Institute for Cancer Research

Nuffield Department of Clinical Medicine

University of Oxford

Trinity 2015

This thesis is the result of my own work unless otherwise stated and does not constitute part of any other thesis. The work herein described was carried out at the Ludwig Institute for Cancer Research, under the supervision of Dr Skirmantas Kriaucionis.

Pijus Brazauskas
Green Templeton College
2015

The Biological Role of DNA Modifications Deposited Over Gene Bodies

Doctor of Philosophy in Clinical Medicine

Pijus Brazauskas, Green Templeton College

Trinity 2015

Abstract

5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) are the most common DNA modifications in vertebrate genomes. Promoter-residing 5mC is a well-established transcriptional repression mark. However, the role of gene body DNA modifications, especially those involving 5hmC, remains elusive and controversial. In this study, the role of genic 5mC and 5hmC in transcriptional regulation is investigated. By engineering a cell line with inducible 5hmC it is shown that deposition of 5hmC causes elevated expression of genes with the highest accumulation of this modification. The transcriptional elevating potential of 5hmC is also confirmed via expression analysis of integrated and transfected artificial reporters. Furthermore, it is demonstrated that promoter proximal gene body DNA modifications can modulate transcription initiation *in cis*. Finally, the ability of methyl-CpG-binding domain protein 1 (MBD1) to interact with genic DNA methylation and repress transcription is analysed. Upon oxidation of 5mC to 5hmC or demethylation, MBD1 binding is abolished and 5mC-mediated repression is relieved. Altogether, this study suggests that 5hmC plays a de-repressive role and proposes a mechanism by which this is achieved.

To my loving parents...

Acknowledgements

First of all, I would like to express my gratitude to my supervisor Dr. Skirmantas Kriaucionis for his guidance, support and patience with a sometimes slow and unfocussed scholar. I am particularly grateful for his encouragement to take risks, try new things and explore the unknown because, at the end of the day, this is what science and life is about. I would like to thank him, the Ludwig Institute for Cancer Research and the Nuffield Department of Medicine for the opportunity to pursue a Doctoral degree, which has been a truly extraordinary experience.

The next thanks go to my colleagues Martin Cusack, Dr. Marie Gerus Durand, Dr. Michael McClellan, Dr. Melania Zauri, Martin Fritzsche, Dr. Zhe Zhao, Norbert Volkmar, Dr. Leonid Nikitenko and others for their support and excellent discussions, not only on science but also on many other topics. I would like to thank all my students: Caroline Oedekoven, Kathrin Jansen, Fidel Lopez-Verdugo and Souradeep Basu. All of them, I believe, will become excellent scientists. I would also like to express many thanks to Dr. Claire Beveridge for critical reading of this thesis.

Last but not least, I would like to thank my parents Laima and Zigmantas, my siblings Indre and Justas, my grandmother, my girlfriend Nilufar, and my best friends Ieva, Vytautas and Paulius for all their support and encouragement during the ups and downs of this D.Phil.

Table of Contents

List of Abbreviations	9
1. Introduction	12
1.1. The Evolution and Biological Significance of DNA Modifications in Vertebrates.....	13
1.2. Writers of DNA Modifications	19
1.2.1. DNMT Family Enzymes Establish and Maintain DNA Methylation Patterns.....	20
1.2.2. TET Family Enzymes Mediate 5mC Oxidation.....	24
1.3. The Role of 5mC Oxidation in DNA Demethylation.....	27
1.4. Genomic Distribution of 5-methylcytosine and 5-hydroxymethylcytosine..	30
1.4.1. Amount, Genomic Context and Distribution of 5mC and 5hmC in ESCs.....	31
1.4.2. Distribution of 5mC and 5hmC During Brain Development	35
1.4.3. Genomic Distribution and the Role of DNA Modifications in Cancer	38
1.5. Readers of Methylated Cytosine and its Oxidised Derivatives	42
1.5.1. MBD Family Proteins.....	43
1.5.2. Other Readers of Cytosine Modifications.....	46

1.5.3. CXXC Domain-Containing Proteins Bind to Unmodified CpG Dinucleotides	47
1.5.4. Proteins Binding to Unmodified DNA.....	48
1.6. Promoter DNA Modifications and Transcriptional Regulation	49
1.7. Gene Body DNA Modifications and Transcriptional Modulation.....	53
1.7.1. Gene Bodies DNA Modifications and Regulation of Transcription Associated Processes.....	53
1.7.2. Gene Body DNA Modification Effects on Transcriptional Elongation and Promoter Activity.....	58
2. Aims of the Study	61
3. Materials and Methods.....	63
3.1. Molecular Cloning.....	64
3.1.1. Plasmids Used in the Study.....	64
3.1.2. Polymerase Chain Reaction (PCR)	66
3.1.3. Restriction Digestion and Dephosphorylation of Plasmids.....	67
3.1.4. Ligation of PCR Product into Vector	68
3.1.5. Bacterial Transformation	69
3.1.6. Colony PCR.....	70

3.1.7.	Plasmid Amplification and Sequencing of the Insert	70
3.2.	Genomic DNA Extraction.....	71
3.3.	DNA Hydrolysis for HPLC analysis.....	71
3.4.	Determination of Relative Extinction Coefficients for Deoxycytidine and its Derivatives.....	72
3.5.	RNA Extraction	73
3.6.	RT-qPCR and data analysis.....	73
3.7.	Adapter Ligation for MeDIP, hmSeal and EpiJET DNA Modification Enrichment.....	75
3.8.	Methylated DNA Immunoprecipitation (MeDIP)	76
3.9.	5-hydroxymethylcytosine Selective Chemical Labelling and Pulldown (hMe- Seal).....	76
3.10.	5hmC Enrichment Using the EpiJET Kit	77
3.11.	PCR Amplification and Molecular Barcoding of Enriched Material.....	78
3.12.	Size selection and Quantification of Sequencing Libraries.....	79
3.13.	Enrichment Quality Control using Spiked-In Modified Oligonucleotides	80
3.14.	Electrophoretic Mobility Shift Assay (EMSA)	81
3.15.	Integration of Differential Amounts of Modified Bases into DNA	83

3.16.	<i>In vitro</i> Transcription Assays.....	85
3.16.1.	T7 RNA Polymerase <i>in vitro</i> Transcription Assay	85
3.16.2.	Transcription Assay in HeLa Nuclear Extracts	86
3.17.	Transfected Reporter Assays.....	87
3.17.1.	Production of Differentially Modified Transcription Constructs.....	87
3.17.2.	Transfection of Differentially Modified Constructs into Cells	89
3.17.3.	Recording of Luciferase Readings and Data Analysis	89
3.18.	Construction of Integrated Luciferase Reporter System.....	90
3.18.1.	<i>In vitro</i> Methylation of the Luciferase Reporter System by M.SssI.....	91
3.18.2.	Integration of the Methylated and Unmethylated Reporter Systems into the Genome of the HEK293 TRex Cell Line.....	92
3.19.	Protein methods.....	93
3.19.1.	Protein Extraction	93
3.19.2.	Protein Quantification.....	93
3.19.3.	Protein Analysis using SDS-PAGE Electrophoresis and Immuno- blotting.....	94
3.20.	CRISPR-Cas9-Mediated Knockouts	96
3.21.	Oxidative Bisulfite Sequencing (oxBS).....	98

3.22.	Cell Culture Techniques	100
3.22.1.	Growth Conditions for Cell Lines Used in This Study	100
3.22.2.	Cell Counting.....	102
3.23.	Construction of TET1 Catalytic Domain-Expressing Cell Lines.....	102
3.24.	Recovery of Transfected Constructs.....	103
3.25.	Combined Bisulfite Restriction Analysis (COBRA)	104
3.26.	Cell Treatments	105
3.26.1.	L-Ascorbic acid.....	105
3.26.2.	5-Aza-2'-deoxycytidine	105
3.26.3.	Trichostatin A.....	106
3.27.	Genomic Data Analysis (DNA Modification Enrichment)	106
3.27.1.	Quality Control, Mapping and Duplicate Removal	106
3.27.2.	Constructing Bigwig Files.....	106
3.27.3.	Calculating DNA Modification Enrichment over Genic Regions.....	107
3.27.4.	Read Saturation Analysis.....	107
3.27.5.	Hierarchical Cluster Analysis.....	107
3.27.6.	Enrichment Analysis over Genomic Elements and Peak Calling.....	108

3.28.	Differential Gene Expression Analysis.....	108
4.	Accumulation of Genomic 5hmC and Gene Regulation in Cancer Cells.....	109
4.1.	Introduction.....	110
4.2.	Results.....	111
4.2.1.	Construction of Cell Lines Overexpressing TET-CD	111
4.2.2.	Dynamic Changes of DNA Modifications after Overexpression of TET Catalytic Domain.....	114
4.2.3.	Genome-wide Mapping of 5hmC in TRex HEK239 Cells.....	118
4.2.4.	Genic Distribution of 5hmC Upon TET1 Catalytic Domain Overexpre- ssion.....	127
4.2.5.	Genome-Wide Mapping of DNA Methylation	129
4.2.6.	Differential Gene Expression Analysis in a TET-CD Overexpressing Cell Line.....	135
4.2.7.	Genic DNA Modifications Mediated Transcriptional Regulation: Mechanistic Insights from Genomic Data.....	145
4.3.	Discussion.....	151
5.	Gene Body Hydroxymethylation and Demethylation Can Reverse 5mC- Mediated Repression	153
5.1.	Introduction.....	154

5.2.	Results.....	155
5.2.1.	Engineering and Integration of a Differentially Methylated Reporter System into the Genome of the TRex HEK293 TET-CD Cell Line	155
5.2.2.	Deposition of 5hmC Alleviates Genic 5mC-Mediated Repression	158
5.2.3.	Transfected Reporter Assays: DNA Methylation but not Hydroxy-methylation over Gene Bodies can Repress Transcription in a Concentration-Dependant Manner	163
5.2.4.	CMV, SV40 and p16, but not EEf1A1, Promoters are Sensitive to Gene Body Methylation	167
5.2.5.	CTCF Does Not Play a Role in Desensitising the EEf1A1 Promoter to Gene Body Methylation.....	170
5.2.6.	Promoter Proximal DNA Modifications Affect Transcription Initiation.....	173
5.3.	Discussion.....	175
6.	The Molecular Mechanism of Gene Body DNA Modification-Mediated Transcriptional Regulation	178
6.1.	Introduction.....	179
6.2.	Results.....	181
6.2.1.	T7 RNAP Elongation Rates through Modified Templates	181

6.2.2. <i>In vitro</i> Transcription in HeLa Nuclear Extracts Using Modified DNA Templates	184
6.2.3. MeCP2 and Modulation of Gene Expression via Gene Body DNA Modifications.....	188
6.2.4. MBD1 and Modulation of Gene Expression via Gene Body DNA Modifications.....	193
6.3. Discussion.....	197
7. Discussion.....	202
7.1. Gene Body DNA Modifications: Potential Effects on Transcription.....	203
7.2. Genic 5hmC Acts as Transcriptional Activator.....	204
7.3. Genic DNA Modifications Affect Transcription via MBD Family Proteins ..	206
7.4. 5mC Hydroxylation Versus Demethylation	207
7.5. The Model: Genic DNA Modifications Affect Transcription Efficiency	208
7.6. Future Perspectives.....	209
Appendix.....	212
Bibliography	223

List of Abbreviations

(In alphabetical order)

5' or 3' UTR	5' or 3' Untranslated region
5azadC	5-azadeoxycytidine
5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethylcytosine
5hmdC	5-hydroxymethyldeoxycytidine
5hmU	5-hydroxymethyluracil
5mC	5-methylcytosine
5mdC	5-methyldeoxycytidine
α KG	α -Ketoglutarate
Base J	β -D-Glucopyranosyloxymethyluracil
BER	Base excision repair
bp	Base pair
BS	Bisulfite sequencing
$^{\circ}\text{C}$	Degrees Celsius
CAGE	Cap analysis of gene expression
CGI	CpG island
CIMP	CpG island methylator phenotype
CNS	Central nervous system
CpA	Genomic sequence with C followed by A in a 5' to 3' direction
CpG	Genomic sequence with C followed by G in a 5' to 3' direction
CXXC	CpG binding domain

List of Abbreviations

dA	2'- deoxyadenosine
dC	2'-deoxycytidine
dG	2'-deoxyguanosine
DMR	Differentially methylated region
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
dT	2'- deoxythymidine
ESCs	Embryonic stem cells
FDR	False discovery rate
FPKM	Fragments per kilobase per million reads
GRO-seq	Global nuclear run-on sequencing
h	Hour
H3K4	Histone 3 lysine 4
H3K27	Histone 3 lysine 27
H3K36	Histone 3 lysine 36
HDAC	Histone deacetylase
hMeDIP	Hydroxymethylated DNA immunoprecipitation
hmSeal	5hmC Selective chemical labelling
HPLC	High performance liquid chromatography
HTF	HpaII tiny fragment
IAP	Intracisternal A-particle retrotransposon
ICF anomalies	Immunodeficiency, centromeric instability and facial
kb	Kilobase (1000 bases)
kDa	Kilo Dalton
KO	Knockout

List of Abbreviations

L	Litre
MBD	Methyl-CpG binding domain
MeDIP	Methylated DNA immunoprecipitation
μ M	Micro molar
mol	Mole
oxBS	Oxidative bisulfite sequencing
PWWP	Domain named after a conserved Pro-Trp-Trp-Pro motif
RNA	Ribonucleic acid
RNAPII	RNA polymerase 2
s	Seconds
SAM	S-adenosyl-L-methionine
TAB-seq	TET assisted bisulfite sequencing
TRD	Transcriptional repression domain
TSA	Trichostatin A
TSS	Transcription start site
TTS	Transcription termination site

Chapter 1

Introduction

1.1. The Evolution and Biological Significance of DNA Modifications in Vertebrates

All known cellular organisms code their genetic information in a macromolecular polymer called deoxyribonucleic acid (DNA), first isolated by Miescher in 1865 (Dahm, 2005). DNA is mainly composed of four different nucleosides: 2'-deoxyguanosine (dG), 2'-deoxycytidine (dC), 2'-deoxyadenosine (dA) and 2'-deoxythymidine (dT). In DNA, the nucleosides are phosphorylated to form nucleotides, and the phosphate part of each nucleotide molecule is used to bridge between the 3' and 5' oxygens of ribose sugars to form oligonucleotides (Neidle, 2008). Although Miescher isolated nuclein in the late 1800s, DNA wasn't demonstrated to act as the cell's genetic material until 1944 (Avery et al., 1944). Furthermore, the structure of DNA remained a mystery until a breakthrough came from the work of Watson and Crick (Watson and Crick, 1953). We now understand that two oligonucleotide strings can be assembled in a DNA double helix via hydrogen bonds formed between nitrogenous bases (Neidle, 2008).

DNA is a product of billions of years of evolution, an extraordinary macromolecule that is stable enough to maintain and pass genetic information through generations and yet is prone to mutations (small alterations in DNA sequence), which lead to the vast numbers of diverse organisms on earth (Holman, 2000). Interestingly, DNA does not seem to have been life's first molecule of choice for coding genetic material. Ribonucleic acid (RNA) has been proposed to have been used initially and replaced later by DNA (Spirin, 2010). This theory is strongly supported by the fact that deoxyribonucleotides are made from ribonucleotide precursors that are the main building blocks of RNA (Lundin et al., 2010), and that

some organisms such as viruses use RNA as their genetic information carrier (Patton and Spencer, 2000). Importantly, the shift from RNA to more stable and easily repairable DNA probably allowed the appearance of complex cellular organisms (Lazcano et al., 1988).

The complete set of DNA in an organism including all the genes is called a genome (Holman, 2000). Interestingly, the correlation of genome size with the complexity of a biological organism is not particularly strong. However, there is a clear trend towards genomic expansion as organisms acquire more genes and develop more complex traits (Gregory, 2001). The expansion of the number of genes in genomes and the emergence of multicellularity led to the evolution of various mechanisms for gene regulation, including complex transcription factor networks, DNA modifications and chromatin (Romero et al., 2012). This gene regulatory information together with genetic material is transferred during somatic cell division. The study of genome regulatory mechanisms that are not coded by the primary DNA sequence, and their heritability, is called epigenetics (Berger et al., 2009).

The best-studied epigenetic DNA modification is cytosine methylation. However, numerous covalent DNA modifications exist across living organisms (Figure 1.1). The most abundant modified base in vertebrates, 5-methylcytosine, has been shown to be important in many diverse biological processes in all domains of life (Schubeler, 2015). In bacteria, for example, DNA methylation is used as a defence mechanism against bacteriophages (Arber, 1974). Bacterial genomes contain substantial amount of 5-methylcytosine, N6-methyladenine and, some strains also have N4-methylcytosine (Janulaitis et al., 1983) (Figure 1.1).

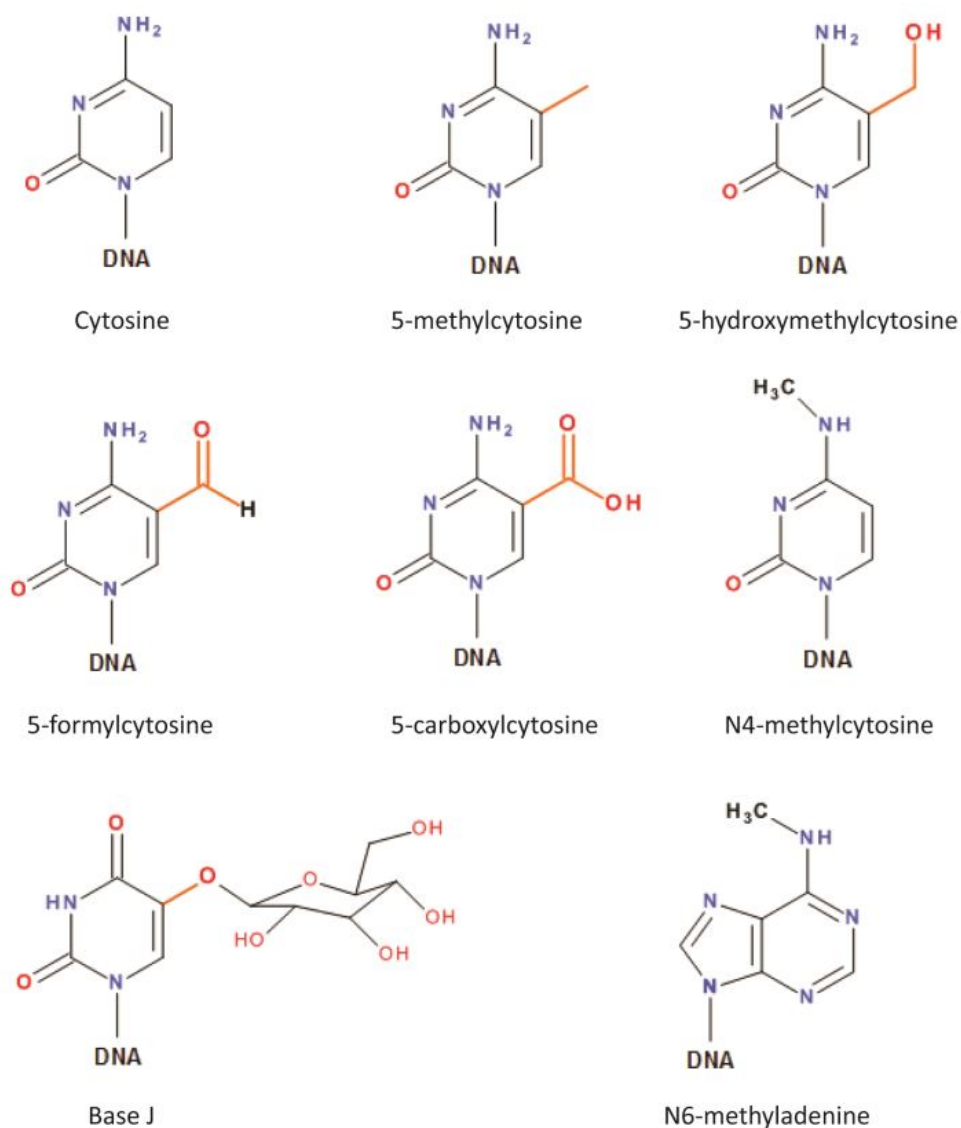


Figure 1.1 Molecular structures of common DNA modifications

All cytosine modifications with an exception of N4-methylcytosine have been reported in vertebrates. Base J has been found in *Trypanosoma brucei* (Gommers-Ampt et al., 1993). N4-methylcytosine and N6-methyladenine are primarily found in bacterial genomes.

These DNA modifications are found in specific DNA sequence contexts and protect bacterial genomes from the activity of restriction endonucleases (RE) (Arber, 1974). While vertebrates lack a RE system, DNA modifications are important for other biological processes in these organisms. The main epigenetic DNA modifications in

vertebrates are cytosine methylation and hydroxymethylation (Figure 1.1) (Hotchkiss, 1948, Kriaucionis and Heintz, 2009). Although the biological role of 5-hydroxymethylcytosine remains largely unknown, 5-methylcytosine has been reported to be essential in processes including development (Li et al., 1992), genomic imprinting (Reik et al., 1987) and X-chromosome inactivation (Mohandas et al., 1981).

The importance of DNA methylation in vertebrate development and cell differentiation is highlighted by the early embryonic lethality of DNA methylation-deficient animals that lack DNA methyltransferases (DNMTs), the enzymes that deposit DNA methylation over the genome (Okano et al., 1999, Li et al., 1992). However, while DNA methylation is not essential for the establishment and self-renewal of pluripotent cells (Tsumura et al., 2006), its absence results in failure to differentiate into three germ layers. This failure has been shown to be caused by their inability to inactivate pluripotency genes such as *Elf5* (Ng et al., 2008). The molecular mechanisms of DNA methylation-mediated gene silencing are discussed in depth in section 1.6. In addition to early development, DNA methylation has been shown to be important in the self-renewal and differentiation of haematopoietic stem cells (Hodges et al., 2011). The importance of DNA methylation in haematopoiesis is further supported by the frequent *DNMT3A* gene mutations observed in leukaemia (Ley and Grant, 2011). While the importance of DNA methylation in mammalian development has been demonstrated, the biological function of 5hmC is only starting to emerge: Deletion of Tet-Eleven Translocase 3 (*Tet3*), responsible for 5mC oxidation, (discussed in section 1.1.2), causes a failure to establish pluripotent cells (Gu et al., 2011). In addition, *Tet1* knockout (KO), *Tet2* KO or double KO animals seem to exhibit marginal abnormalities and are born in highly skewed Mendelian ratios

(Dawlaty et al., 2013, Dawlaty et al., 2011). Furthermore, *Tet2* KO animals show expansion of some haematopoietic compartments and subsequently develop myelodysplastic syndromes (Li et al., 2011). Taken together, these observations indicate that DNA modifications play an important role in early development and cell differentiation.

DNA methylation has been long known to silence imprinted and germ line-specific genes (Reik et al., 1987). A classic example of imprinting is the *IGF2-H19* regulation demonstrated in mice. The maternal allele of *H19* gene harbours a CTCF binding site. CTCF keeps the locus unmethylated until the CTCF-DNA interaction is abolished and the locus acquires DNA methylation in male gametogenesis (Lee et al., 2010). Also, DNA methylation-mediated imprint silencing is strongly supported by the evidence of substantial upregulation of imprints in DNA methylation-deficient embryonic stem cells (ESCs) (Karimi et al., 2011). In addition to imprinted genes, germline-specific genes (e.g. the MAGE family) are also significantly upregulated in DNA methylation-deficient ESCs. This indicates that DNA methylation is a commonly used mechanism for gene silencing.

Yet another important role of DNA methylation in silencing is the process called X chromosome inactivation (XCI). XCI is a dosage compensation mechanism first reported in female mice, where one X chromosome is inactivated to adjust the expression of sex-linked genes (Lyon, 1961). Although XCI is achieved by a number of mechanisms such as deposition of histone macroH2A1 (Costanzi and Pehrson, 1998) and expression of *XIST* RNA (Clemson et al., 1996), defects in X chromosome inactivation in the *Dnmt1* KO animals strongly suggests the involvement of DNA methylation (Panning and Jaenisch, 1996). Furthermore, promoters of the genes on the inactive X are more often methylated compared to the active X and the opposite is

true for the gene bodies (Sharp et al., 2011). Therefore, XCI mechanism further supports the repressive role of DNA methylation in vertebrates.

Importantly, DNA methylation has been reported to play a role in human disease. Upon malignant transformation, for example, cells tend to methylate the promoters (discussed in sections 1.4.3 and 1.6) of tumour suppressor genes such as *CDKN2A* (p16) (Merlo et al., 1995). Mutations in *DNMT3B* cause Immunodeficiency, Centromeric instability and Facial anomalies (ICF) syndrome (Hansen et al., 1999). DNA methylation is also a mechanism to suppress the expression of various retroelements which, upon activation, can cause genomic instability or increase susceptibility to cancer (Kaer and Speek, 2013). Hence, DNA methylation plays a number of roles in various biological processes, mainly through transcriptional regulation.

Although DNA methylation is vital for vertebrates, a large number of living organisms exist without, or with little, DNA methylation. These include multicellular organisms such as *Drosophila melanogaster* and *Caenorhabditis elegans* and more simple eukaryotes such as *Saccharomyces cerevisiae* (Raddatz et al., 2013).

In addition to the control of various biological processes, DNA methylation is thought to have shaped the genomic landscape of vertebrates and other species. CpG dinucleotides (the main targets of DNMTs) have been largely depleted from mammalian genomes (Bird, 1980). However, certain regions of DNA have maintained close-to-expected CpG densities; these regions were initially called HTF islands and later renamed CpG islands (Illingworth and Bird, 2009, Cooper et al., 1983). These islands are proposed to have evolved due to the fact that 5mC undergoes spontaneous deamination to thymidine in the genome (J. C. Shen, Rideout, & Jones,

1994). This is thought to be the reason why mammalian genomes are depleted of CpG dinucleotides where methylation occurs, but not in CpG islands, which remain unmethylated. This suggests that DNA methylation has played an important role in vertebrate genome evolution. Importantly, CpG islands often overlap with gene promoters, further supporting the role of DNA methylation in transcriptional regulation. Interestingly, oxidised forms of 5-methylcytosine, such as 5-hydroxymethylcytosine (Kriaucionis and Heintz, 2009, Tahiliani et al., 2009b) and others (Ito et al., 2011), have also been reported in mammalian genomes (Figure 1.1). They have similarly been proposed to play a role in gene regulation and other biological processes (Mellen et al., 2012). The molecular mechanisms behind how and where DNA modifications are deposited, how these marks are read and how they directly or indirectly influence gene expression are discussed later in this chapter.

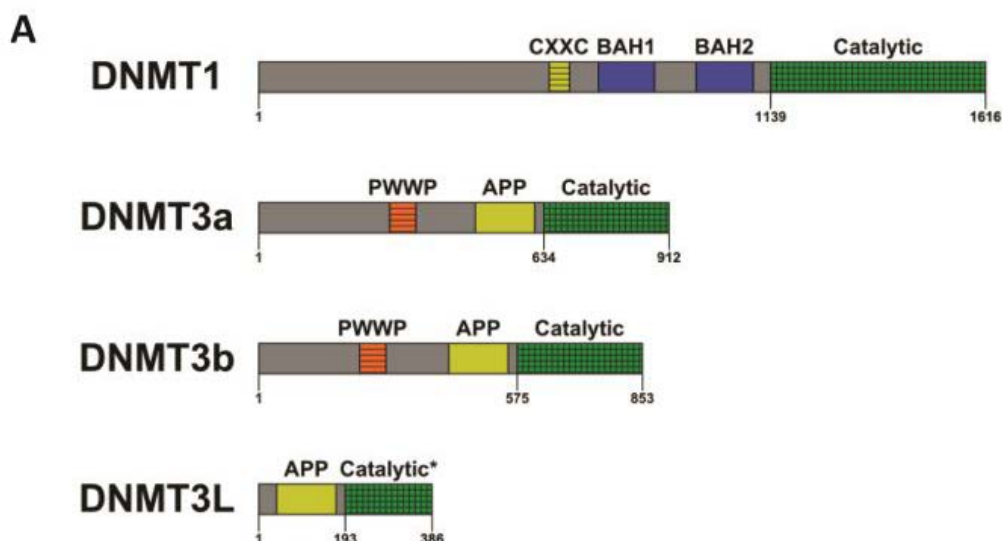
1.2. Writers of DNA Modifications

In mammals, DNA methylation occurs via the covalent addition of a methyl group to the 5th position of the cytosine pyrimidine ring. This reaction is catalysed by the DNMT enzyme family (Bestor, 2000). There are three catalytically active mammalian DNMTs: Dnmt1, Dnmt3a and Dnmt3b. 5-hydroxymethylcytosine (5hmC) is another modified cytosine variant found in mammalian DNA. The TET family of enzymes establish this modification (Tahiliani et al., 2009b). Three TET family members are encoded in mammalian genomes: Tet1, Tet2 and Tet3 (Shen and Zhang, 2012). It has also been shown that these enzymes can further oxidise 5hmC to yield 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) in the DNA (Ito et al., 2011).

However, due to the low amount of 5fC and 5caC in the genome, their functional significance as *bona fide* epigenetic marks remains highly debatable.

1.2.1. DNMT Family Enzymes Establish and Maintain DNA Methylation Patterns

DNA methylation in mammalian genomes was observed as early as 1948 by Rollin Hotchkiss (Hotchkiss, 1948), who discovered the cytosine derivative using paper chromatography. Three catalytically active DNMTs have been characterized to date (Figure 1.2A). Dnmt1 was the first DNMT identified in mammals. Subsequent cloning of its cDNA and comparison of its sequence with bacterial methyltransferases allowed a better understanding of the structure of the enzyme (Bestor et al., 1988). As depicted in Figure 1.2A, DNMT1's regulatory part is located at the N-terminus and the catalytic domain at the C-terminus. The catalytic



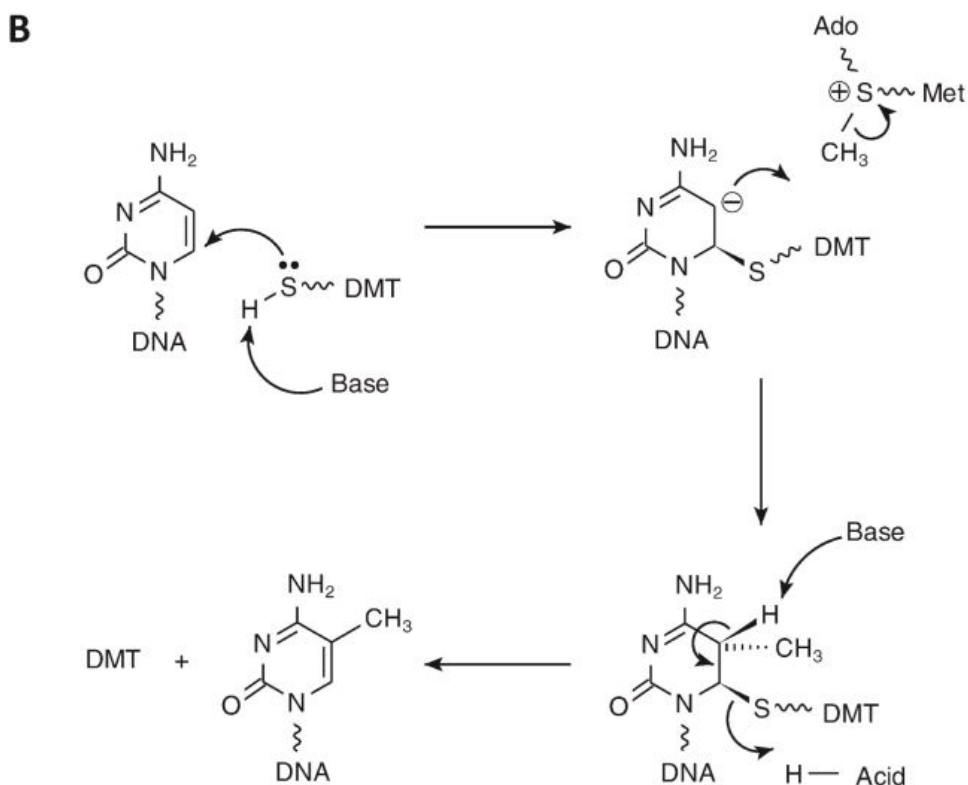


Figure 1.2 Schematic representation and catalytic mechanism of mammalian DNMTs

A) Schematic representation of mammalian DNA methyltransferases. Protein regulatory and catalytic domains are indicated. B) The catalytic mechanism in the active site of DNA methyltransferases is shown.

domains of DNMTs are highly conserved from bacteria to mammals and exhibit the same general mechanism of catalysis using *S*-adenosyl-L-methionine (SAM) as a cofactor. The catalytic mechanism is initiated by the nucleophilic attack of the 6th carbon of the pyrimidine ring by a methionine residue of DNMTs, and the subsequent transfer of a methyl group from SAM to the 5th carbon of the pyrimidine ring (Figure 1.2B) (Chen et al., 1991). Despite the similarities in the catalytic mechanism and domain architecture of mammalian DNMTs, the biological functions of these enzymes are rather different. Dnmt1 is commonly called the maintenance

DNMT, while Dnmt3a and b were named *de novo* methyltransferases (Goll and Bestor, 2005). In addition to the three major DNMTs, *DNMT3L* is another member of the DNMT family that has lost its catalytic activity (Goll and Bestor, 2005). Nevertheless, Dnmt3l plays an important role in establishing DNA methylation patterns (Chedin et al., 2002). Yet another member of this family is *DNMT2*, which catalytic domain exhibits significant sequence similarity to other members of the family. However, it has not been shown to methylate DNA but RNA instead (Goll et al., 2006).

Dnmt3a and b enzymes are thought to be primarily responsible for establishing DNA methylation *de novo* and this hypothesis is supported by several lines of evidence. Firstly, several studies revealed that, early after fertilisation, mammalian genomes undergo nearly complete demethylation and remethylation (Guo et al., 2014, Smith et al., 2014). The major role of the DNMT3 enzymes in the reestablishment of DNA methylation patterns has been convincingly demonstrated by interrogating enzyme expression patterns during these processes (Guo et al., 2014). Secondly, DNMT3A is capable of methylating cytosine not only in a CpG but also a CpA context, which exhibits asymmetrical structure and, therefore, cannot be easily propagated during DNA replication (Ramsahoye et al., 2000).

The enzyme that faithfully copies DNA methylation during DNA replication is DNMT1. This role for DNMT1 has been assigned because of its direct interactions with PCNA (Proliferating Cell Nuclear Antigen) (Chuang et al., 1997) and UHRF1 (Ubiquitin-like, containing PHD and RING Finger domains 1) (Bostick et al., 2007). These two enzymes are components of the DNA replication machinery. Furthermore, DNMT1 has been shown to be more efficient when catalysing the reaction on the hemi-methylated DNA substrate *in vitro* (Hermann et al., 2004).

Recently, several lines of evidence have challenged the understanding of the way DNMTs work. Firstly, it has been shown that DNMTs can interact with each other via N-terminal domains and might exist as a complex (Jin et al., 2009) (Figure 1.3). Secondly, substantial *de novo* methylation activity has been demonstrated for Dnmt1 (Biniszkiewicz et al., 2002). Finally, Dnmt3a and b have been shown to interact with H3K36me2/3 via PWWP (a domain with conserved PWWP motif) domains (Dhayalan et al., 2010), and all DNMTs were demonstrated to interact with HDAC (Histone deacetylase) complexes (Figure 1.3) (Fuks et al., 2000). Although most of these interactions were confirmed experimentally, relatively weak affinities between separate components indicate that these complexes are possibly transient rather than robust.

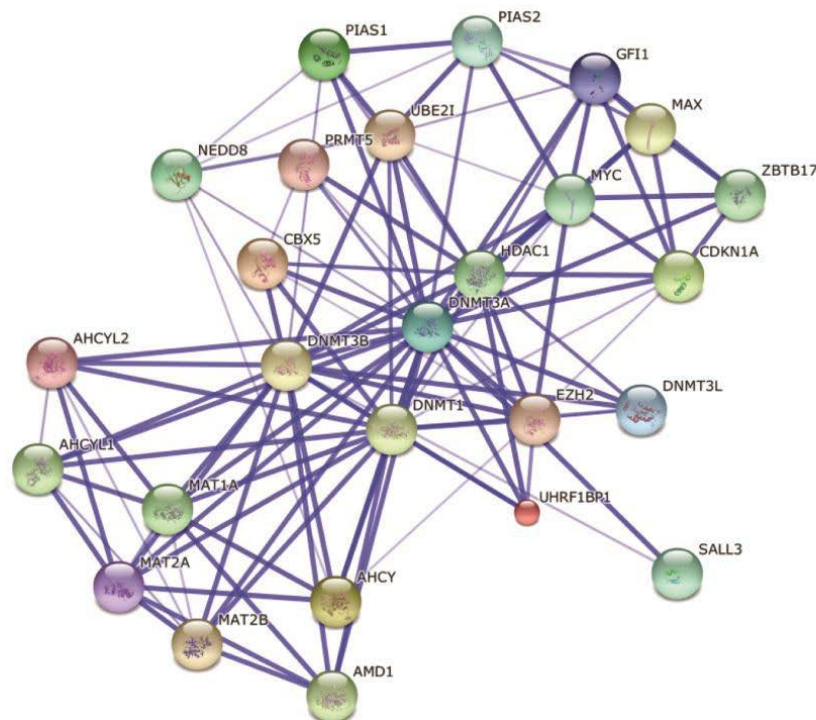


Figure 1.3 Schematic diagram of the interaction partners of the mammalian DNA methyltransferases.

Taken together, the evidence indicates that DNA methylation is targeted to broad genomic regions by the coordinated activity of all three DNMTs, rather than methylation being established on specific CpG sites by Dnmt3a/b and then faithfully maintained via DNMT1 activity throughout the organism's life.

1.2.2. TET Family Enzymes Mediate 5mC Oxidation

The existence of DNA methylation in mammalian genomes has been known for around six decades. In contrast, the presence of 5-hydroxymethylcytosine (5hmC) in mammalian cells has only recently been convincingly demonstrated (Kriaucionis and Heintz, 2009, Tahiliani et al., 2009b). The family of TET enzymes, which catalyse the oxidation of 5mC to 5hmC and further, was discovered by a homology search for sequences similar to the thymine oxidizing enzymes JBP1 and 2 (Base J binding protein 1 and 2) found in *Trypanosoma brucei* (Tahiliani et al., 2009b, Iyer et al., 2009). Part of the *TET1* gene has previously been reported to translocate and form a fusion with myeloid/lymphoid or mixed-lineage leukemia protein 1 (MLL1) in human leukaemias (Ono et al., 2002). This is where the name TET (Ten-Eleven Translocation) comes from. TET family members share significant homology of the C-terminal catalytic domains, however, the N-terminal parts of the proteins are different and largely unstructured (Figure 1.4A). TET1 and TET3 also contain a CXXC domain, which can bind to unmodified CpG dinucleotides (Xu et al., 2012). Interestingly, TET2 is thought to have lost its CXXC domain during the course of evolution; it seems that it has been translocated and expressed as part of the protein IDAX (Ko et al., 2013). Idax directly interacts with Tet2's catalytic domain, indicating that the CXXC domain of the TET family members is functionally important (Ko et al.,

2013), and it may also play a role in their biological function (discussed in section 1.5.3).

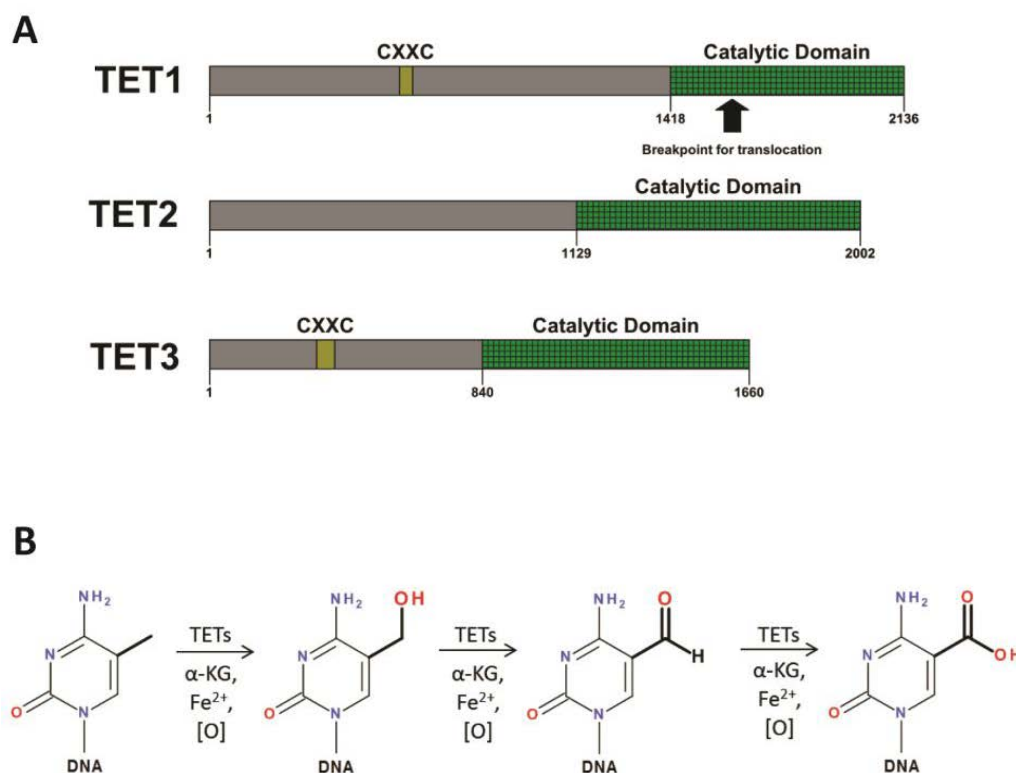


Figure 1.4 TET family enzymes and oxidation of 5-methylcytosine

A) Schematic representation of TET family enzymes (*H. sapiens*). Catalytic and CXXC domains are represented. Translocation site of TET1 is indicated by black arrow. B) 5mC oxidation by TET family enzymes. 5hmC and the further oxidation products 5fC and 5caC are shown.

Soon after the discovery of 5hmC and the TET enzymes, it was reported that TETs can further oxidise 5hmC to 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Figure 1.4B) (Ito et al., 2011, Pfaffeneder et al., 2011). Another study reported that 5fC and 5caC can be efficiently excised by Thymidine DNA Glycosylase (TDG) (Maiti and Drohat, 2011). This study opened up the possibility that 5mC oxidation could be the pathway for active DNA demethylation (discussed in section 1.3). However, the relatively low amounts of 5fC and 5caC in the genome (even in TDG

knockouts) suggest that the extent of active demethylation via this mechanism in pluripotent and differentiated cells is probably very low (Ito et al., 2011).

Numerous studies have sought the interaction partners of the TET family members and several interesting patterns are emerging. Several recent studies reported that OGT (O-linked N-acetylglucosamine [O-GlcNAc] transferase) is a direct interaction partner of all TET family enzymes (Chen et al., 2013, Deplus et al., 2013, Vella et al., 2013). Its ability to interact with all the members of the TET family suggests that this interaction dates back to before the amplification of TET gene and that it has been maintained through protein diversification. This interaction results in several biological outcomes. Firstly, TET proteins are O-GlcNAcylated, but the biological function of this post-translational modification remains unknown. Secondly, TET proteins co-localise to the same genomic regions as OGT. Interestingly, these regions are mainly promoter associated CpG islands where TETs and/or OGT have been shown to facilitate transcription (Vella et al., 2013). Surprisingly, depletion of OGT does not affect the recruitment of the TET enzymes to the CpG island promoters, in fact the opposite is the case. In addition, TET2 and TET3 have been shown to interact with the SET1/COMPASS complex (Deplus et al., 2013), which deposits H3K4 (histone 3 lysine 4) methylation over the promoter regions. All these studies have reported the role of TET proteins as transcriptional activators. However, one study reported that TET1 could also act as transcriptional co-repressor via the recruitment of Sin3A (a transcriptional repressor) protein to CpG-rich promoters (Williams et al., 2011). Based on interaction partner analysis of the TET enzymes, it seems that they are likely to act as transcriptional regulators independently of their catalytic function. However, the exact details of this regulation remain elusive.

1.3. The Role of 5mC Oxidation in DNA Demethylation

Since the discovery of DNA methylation, scientists have been curious about the reverse process of this reaction. It was quickly realised that DNA methylation can be diluted during the course of DNA replication because the new DNA strand is synthesised in an unmethylated state. However, DNMT1 which interacts with DNA replication machinery quickly re-established DNA methylation on a newly synthesized strand (Figure 1.5A). Alternatively, in the absence or inhibition of DNMT1, newly synthesized strand loses parental DNA methylation pattern (Figure 1.5B). This process has been named the 'passive DNA demethylation pathway' (Ooi and Bestor, 2008). Shortly after discovery of 5hmC it became evident that 5mC oxidation might play a role in DNA demethylation. One study proposed that 5hmC may mediate the process of demethylation by reducing the binding affinity of DNMT1 to hemi-hydroxymethylated DNA compared to hemi-methylated DNA (Hashimoto et al., 2012). However, these conclusions were based mainly on affinity measurements *in vitro*. In further support of this hypothesis, it has been shown that this 'semi-passive' DNA demethylation mechanism (Figure 1.5C) is used to erase DNA methylation in the paternal pronucleus after the fusion of gametes (Inoue & Zhang, 2011). According to this study, Tet3, which is present in the oocyte, oxidises 5mC in the paternal pronucleus leading to the loss of cytosine modification during cell division. Several other studies suggested that a similar mechanism is utilised during imprint erasure in primordial germ cell development (Hackett et al., 2013; Piccolo et al., 2013). This theory appears to be supported by stronger evidence than that of modulation of DNMT1 levels during cell division, which has been also proposed (Figure 1.5B) (Kar et al., 2012).

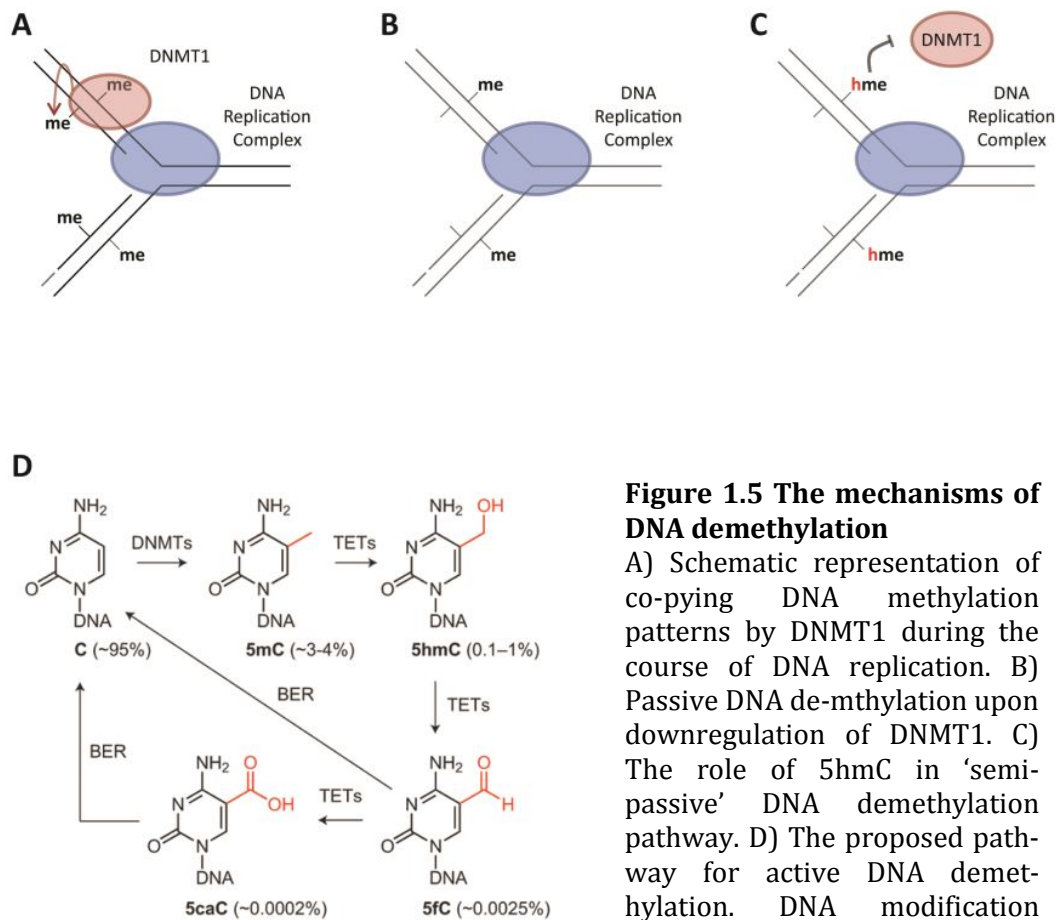


Figure 1.5 The mechanisms of DNA demethylation

A) Schematic representation of co-copying DNA methylation patterns by DNMT1 during the course of DNA replication. B) Passive DNA de-methylation upon downregulation of DNMT1. C) The role of 5hmC in 'semi-passive' DNA demethylation pathway. D) The proposed pathway for active DNA demethylation. DNA modification abundance in ESCs is shown.

Nevertheless, the rapid loss of DNA methylation in the paternal nucleus after gamete fusion hinted that an active DNA demethylation pathway, which does not require cell division, should also exist (Mayer et al., 2000). However, the high stability of carbon-carbon covalent bonds makes direct enzymatic removal of the methyl group from the 5th position of 5-methylcytosine very unlikely. Several early studies suggested the involvement of RNA or Mbd2 (Methyl-CpG binding protein 2) in active DNA demethylation, however, these findings were never reproduced (Swisher et al., 1998, Bhattacharya et al., 1999). Some useful insights regarding DNA demethylation

can be gained from the mechanisms in plants, where DNA demethylating enzymes have been identified. In *Arabidopsis thaliana*, DNA methylation is established by Drm2 and maintained by Met1 and Cmt3 (Law and Jacobsen, 2010). In order to demethylate DNA, plants have evolved glycosylases such as Dme, Dml2, Dml3 and Ros1 to efficiently excise 5mC (Penterman et al., 2007, Gong et al., 2002). The mechanism in plants implies that the evolutionarily preferred mechanism for DNA demethylation is mediated by base excision rather than by direct removal of the methyl residue.

Soon after its discovery, TET enzyme-mediated oxidation of 5mC has been also shown to be involved in active DNA demethylation (i.e. without cell division) processes. Several lines of evidence indicate that endogenous cytosine deaminases, such as APOBEC or AID, could act as DNA demethylases by deaminating 5hmC to 5hmU (5-hydroxymethyluracil). 5hmU is recognised as an unnatural DNA base and subsequently excised by DNA glycosylases such as TDG or SMUGs (Guo et al., 2011). However, these activities could only be documented upon overexpression of both TET1 and AID enzymes. Although this mechanism could play a role in DNA demethylation, another more convincing mechanism was proposed following the discovery of 5fC and 5caC (Ito et al., 2011). It has been shown that 5caC is a very good TDG substrate and can be very efficiently excised by the base excision repair (BER) pathway (Figure 1.5D) (Zhang et al., 2012). This mechanism is also supported by the evidence that *Tdg* KO animals die during the early stages of development. Interestingly, this early lethality can not be explained by accumulation of mutational load because MEFs isolated from the TDG KO animals do not show increased sensitivity to DNA damaging agents or more mutations in the BigBlue assay (Cortazar et al., 2011). In addition, depletion of TDG leads to increased DNA methylation and

the accumulation of 5fC and 5caC in genomic DNA (He et al., 2011). Finally, oxidised forms of 5hmC accumulate over regulatory elements in *Tdg*-deficient mouse ESCs (Shen et al., 2013). These findings indicate that oxidised forms of 5hmC, together with *Tdg*, constitute the active DNA demethylation pathway in mammals.

1.4. Genomic Distribution of 5-methylcytosine and 5-hydroxy-methylcytosine

The earliest evidence about genomic vertebrate DNA methylation patterns dates back to pre-sequencing days. One study showed that the digestion of genomic DNA with methylation-sensitive restriction enzyme HpaII gave bimodal DNA fragment distribution (Cooper et al., 1983). This was the first evidence that methylation is not uniformly distributed in mammalian genomic DNA. It became evident that genomes have numerous short unmethylated regions called CpG islands (Bird et al., 1985). Subsequent rapid advances in DNA sequencing technology and the development of numerous methods to determine the genomic location of DNA modifications have made genome-wide DNA methylation mapping possible (Laird, 2010). These methods range from low resolution antibody-based methods such as methylated DNA immunoprecipitation (MeDIP) (Weber et al., 2005) to single base resolution methods like bisulfite sequencing (Frommer et al., 1992). These methods have been widely applied to determine DNA methylation patterns in various cell lines and tissues.

Soon after the discovery of 5hmC in the vertebrate DNA, a number of methods were developed to determine its location in the genome. Methods based on antibody pulldown of cytosine 5-methylenesulfonane (CMS) (Pastor et al., 2011) or 5hmC itself

(hMeDIP) (Tan et al., 2013), have been widely used in addition to enzymatic/chemical methods such as 5hmC selective chemical labeling (hmSeal) (Song et al., 2011a). Single base resolution methods such as oxidative bisulfite sequencing (oxBS) (Booth et al., 2012) and TET assisted bisulfite sequencing (TAB-seq) (Yu et al., 2012) have also been developed. All these methods have made it possible to determine the localisation of 5mC and 5hmC in various biological systems. Here, genome-wide studies in ESCs, brain tissues and cancer cells will be discussed in detail.

1.4.1. Amount, Genomic Context and Distribution of 5mC and 5hmC in ESCs

ESCs are derived from 5-day preimplantation embryos. They are capable of dividing without differentiating in culture, and, upon implantation, are able to differentiate into three primary germ layers (de Wert and Mummery, 2003). ESCs have become a widely used model system in molecular biology and epigenetics for several reasons. They are considered to be normal cells, meaning that they have normal karyotypes and can give rise to the animal if implanted into the blastocyst. Yet, these cells are relatively easy to culture, enabling the generation of a large amount of material for biochemical studies. In addition, genetic manipulations in these cells are relatively straightforward. For these reasons, many scientists use ESCs for studying epigenetic mechanisms.

Total methylation and hydroxymethylation amounts in cells or tissues vary. However, they can be accurately quantified using analytical methods such as mass spectrometry (Ehrich et al., 2005). In ESCs, 5mC levels range between 4–5% of total

dC, while 5hmC levels are around 0.1% of total dC in the genome (Ficz et al., 2013, Dawlaty et al., 2013). The 5hmC level in ESCs seems low compared to that of neuronal tissues (1–2% of total dC) but, taking into consideration that ESCs cycle very rapidly, deposition of 5hmC in these cells is very efficient. However, absolute numbers of DNA modifications do not provide much insight into the biological functions of 5mC or 5hmC. Therefore, there is a need to determine the exact location of these modifications over the genome.

Several excellent studies have investigated DNA modification patterns in ESCs using MethylC-seq (Lister et al., 2009) and TAB-seq (Yu et al., 2012) for DNA methylation and hydroxymethylation, respectively. These and other genome-wide studies have confirmed that DNA methylation, in a CpG dinucleotide context in ESCs, has a bimodal distribution. This means that certain genomic CpG sites are completely demethylated, however, the rest of the CpG sites in the genome are highly methylated (Lister et al., 2009). These studies have also confirmed that most of genomic CpGs are highly methylated with the exception of those residing in CpG islands, the majority of which are kept unmethylated genome-wide. In addition, the lack of 5hmC over CpG island promoters has been reported. This is as expected, as 5hmC is the product of 5mC and, therefore, can only be found on cytosines that have previously been methylated. Interestingly, unlike DNA methylation, the distribution of 5hmC is continuous across the genome (Yu et al., 2012).

In ESCs, around 25% of DNA methylation is found in a non-CpG context (Lister et al., 2009). After CpG, the second most prominently methylated dinucleotide sequence is CpA. Interestingly, CpA methylation is lost upon differentiation into a fibroblast lineage (Lister et al., 2009). In addition, despite the high levels of CpA methylation in ESCs found genome-wide, the probability of any particular CpA

position in the genome being methylated is low. This implies that, unlike for CpG methylation, there is likely no biological mechanism to maintain CpA methylation through cell division. In addition, 5hmC levels in a CpA context have been shown to be negligible in ESCs (Yu et al., 2012). However, it has been reported that the TET catalytic domain can convert 5mC to 5hmC in any genomic context *in vitro*, although the kinetics of this process remain unknown (Yu et al., 2012). For now, the reasons for the very low levels of 5hmC in the CpA context remain unidentified.

In ESCs, CpG methylation has been shown to be depleted from transcription start sites and distal regulatory elements (Figure 1.6) (Lister et al., 2009). Indeed, one study proposed that, in addition to highly methylated and unmethylated regions, a third group called intermediately methylated regions exists, which overlap with distal regulatory elements to large extent (Stadler et al., 2011). In particular, it has been shown that DNA methylation is reduced in the binding sites of transcription factors such as p300, KLF4, OCT4 and TAF1. This depletion has been reported to be true for CpG as well as non-CpG methylation, suggesting that DNA methylation is regulated at the distal regulatory elements. Furthermore, 5hmC has been shown to be highly enriched in distal regulatory elements in ESCs (Yu et al., 2012). 5hmC enrichment techniques, as well as sequencing at single nucleotide resolution, has revealed that it is deposited in enhancer regions, p300 binding sites in proximity to expressed genes compared to non-expressed genes (Figure 1.6) (Huang et al., 2014). Taking into consideration that 5hmC has been proposed to be an intermediate in the DNA demethylation process (Shen et al., 2013), it seems likely that it plays a role in the process of enhancer activation.

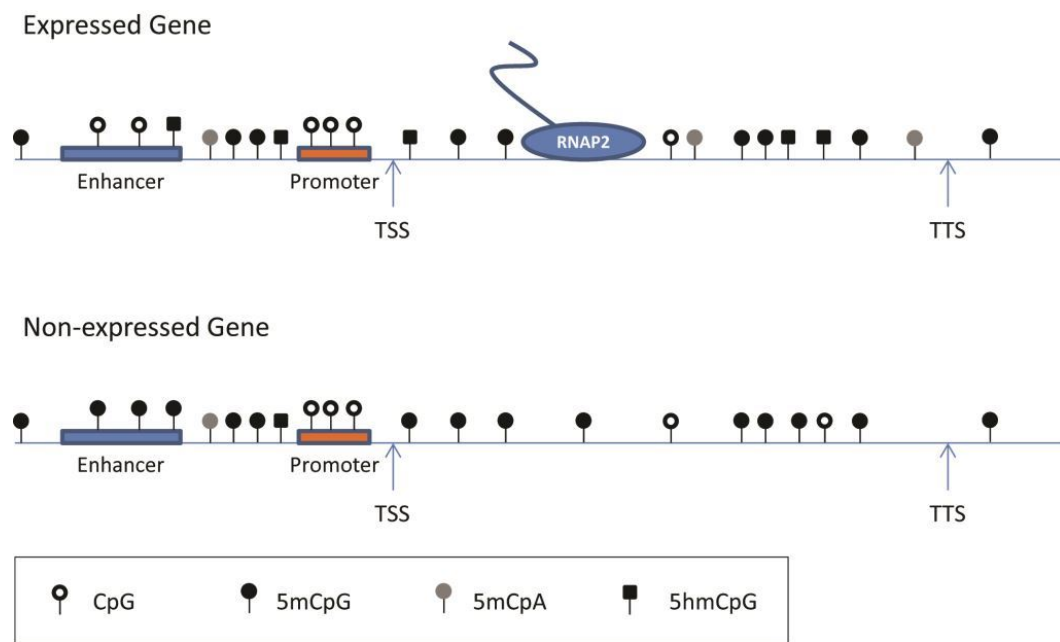


Figure 1.6 The DNA modification landscape in genic and proximal regions in ESCs

Schematic representation of DNA modifications in expressed and non-expressed genes. Unmethylated, methylated and hydroxymethylated CpG dinucleotides, and methylated CpA sites are represented. Transcription start sites (TSS) and transcription termination sites (TTS) are indicated by arrows.

Substantial amounts of DNA methylation have been found over gene bodies in ESCs (Lister et al., 2009). Interestingly, while CpG methylation present in promoter regions shows a negative correlation with gene expression (Weber et al., 2007), CpG methylation over gene bodies does not correlate with gene expression in ESCs. However, genic non-CpG methylation positively correlates with gene expression (Figure 1.6) (Lister et al., 2009). In addition, non-CpG methylation is slightly more enriched on the antisense strand, indicating that this modification could play a role in transcription processes. 5hmC in ESCs has also been shown to be enriched in the gene bodies of highly expressed genes compared to non-expressed ones (Huang et al.,

2014). However, the exact role of gene body DNA modifications in ESCs remains largely unknown.

1.4.2. Distribution of 5mC and 5hmC During Brain Development

The central nervous system (CNS) is one of the most complex organ systems in mammals. It consists of multiple neuronal cell types, glia and other cells. Most of the cells in the brain exit the cell cycle early in development but their identity continues to be refined during processes such as memory formation (Zhao et al., 2013). This implies that healthy neuronal cells undergo minor changes during their lifetime, which have been shown to be partially regulated by various epigenetic mechanisms (Graff and Tsai, 2013). DNA methylation in particular has been shown to play a role in brain development, maturation and learning processes (Feng et al., 2010a, Miller and Sweatt, 2007). The importance of epigenetic modifications in the CNS is also emphasised by various diseases such as Rett syndrome, which is caused by mutations in the methyl-CpG binding protein MeCP2 (Amir et al., 1999). Importantly, 5hmC has also been shown to be present at high levels in the brain, especially in Purkinje and other neuronal cells compared to other tissues (Kriaucionis and Heintz, 2009). Such findings highlight the importance of DNA modifications in the brain.

Neuronal cells permanently exit the cell cycle upon differentiation, as opposed to stem cells or other cycling cell types. For this reason, passive dilution mechanisms can no longer be utilised to reshape DNA modification patterns. This suggests that the DNA methylation dynamics in these cells could be mediated by the active demethylation pathways. High active demethylation of DNA in the brain has been proposed because of the high amounts of 5hmC present (Kriaucionis and Heintz,

2009). However, elegant work from the Balasubramanian laboratory (Bachman et al., 2014) rejected this hypothesis by showing that DNA modifications in the brain are long lived. Based on this evidence, it seems that 5hmC is a stable epigenetic mark in the brain. Lister *et al.* have studied DNA modifications during brain development at single nucleotide resolution (Lister et al., 2013). Their research found that global DNA methylation amount in the CpG context remain constant from birth to old age, both in humans and in mice. However, significant gains in hydroxymethylation and non-CpG methylation have been observed during the maturation of the brain and synapse formation (Lister et al., 2013). This suggests that DNA methylation patterns in neuronal cell types are established during early development and are possibly fine-tuned later by hydroxylation of 5mC and deposition of non-CpG methylation.

To gain better insight into the functions of DNA modifications, it is important to investigate their genomic distributions. CpG methylation is the most abundant DNA modification in the brain as well as in other tissues in humans (Leung et al., 2015). In addition, the general pattern of CpG DNA methylation is quite similar among different cell types, where the majority of the genome is methylated with the exception of CpG islands (CGIs) that overlap with active transcription start sites (TSS) (Leung et al., 2015). Differentially methylated region (DMR) is a commonly used name to determine the extent of the change in DNA methylation upon different biological processes. Interestingly, upon brain development, no DMRs are found in the CGIs indicating that CpG island methylation does not play a role in the process of development from the foetal to adult brain. This is surprising because gene expression is upregulated and downregulated during brain development (Lister et al., 2013). These findings agree with previous observations that promoter CpG island methylation status is a good predictor of repressed genes. However, only a fraction of

genes that are downregulated acquire CpG island methylation (Weber et al., 2007). Intriguingly, the majority of DMRs (~60%) occur in gene bodies and ~10% in enhancer regions. Furthermore, during the maturation of the brain, gene bodies accumulate non-CpG methylation and hydroxymethylation (Lister et al., 2013). Interestingly, non-CpG methylation is negatively correlated with gene expression in the brain while 5hmC shows strong positive correlation (Figure 1.7) (Lister et al., 2013, Mellen et al., 2012, Song et al., 2011a). In addition, 5hmC enrichment on the sense strand compared to the antisense strand has been reported in neuronal cell types (Wen et al., 2014). Interestingly, 5hmC accumulation on highly expressed genes has been reported in other biological processes such as the course of T-cell development and differentiation (Tsagaratou et al., 2014). Thus, the evidence to date

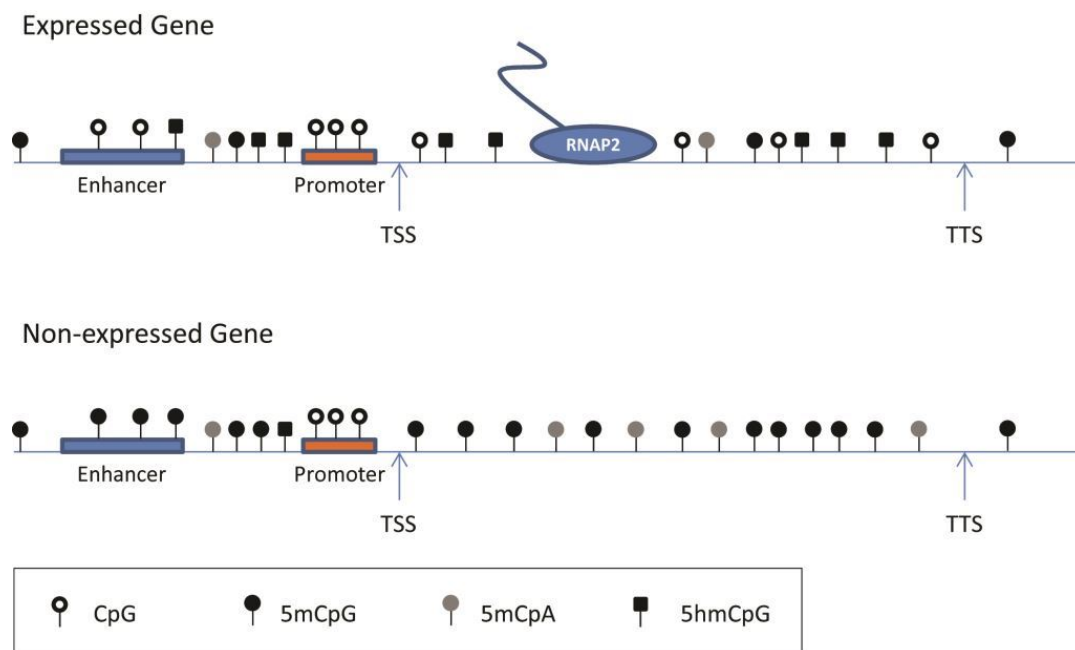


Figure 1.7 The DNA modification landscape in genic and proximal regions in neuronal cells.

Schematic representation of DNA modifications in expressed and non-expressed genes. Unmethylated cytosines are marked as hollow lollipops, methylated cytosines are marked as filled lollipops, and hydroxymethylated cytosines are marked as black squares. All black DNA lollipops represent CpG dinucleotides. Methylated CpA sites

are represented as grey lollipops. Transcription start sites (TSS) and transcription termination sites (TTS) are indicated by arrows.

indicates that DNA modifications are likely to be important for the function and differentiation of neuronal and possibly other cell types. However, whether changes in gene body DNA modifications actually aid transcriptional regulation or are only the consequence of changing transcriptional landscapes is currently unknown.

1.4.3. Genomic Distribution and the Role of DNA Modifications in Cancer

Studies are accumulating that highlight the importance of DNA modifications in the processes of development, ageing and disease. However, the vast majority of the research publications about DNA methylation comes from cancer research (Kulis and Esteller, 2010). To date, more than 5000 cancer methylomes have been interrogated using microarrays, enrichment methods followed by sequencing or whole-genome bisulfite sequencing (WGBS) (Stirzaker et al., 2014). These studies have greatly enhanced our understanding of the role of DNA methylation in cancer. Importantly, a CpG island methylator phenotype (CIMP) has been described in colorectal cancer (Toyota et al., 1999). The study found that, despite the global demethylation of the cancer genome, CpG island promoters of a specific set of genes are hypermethylated. Interestingly, CIMP has now been shown to be a feature of many neoplasias including glioblastomas, neuroblastomas and leukaemias (Hughes et al., 2013). Furthermore, the mechanistic details of CIMP cancers are starting to emerge from studies of glioblastomas (Noushmehr et al., 2010, Turcan et al., 2012) and leukaemias (Figuerola

et al., 2010). Glioblastomas with a CIMP phenotype have been shown to contain mutations in isocitrate dehydrogenase (IDH) family enzymes (Turcan et al., 2012). IDH family enzymes produce α -ketoglutarate, which is a cofactor for the TET enzymes (Tahiliani et al., 2009b). In addition, TET2 mutations have been shown to be associated with genomic DNA hypermethylation in leukaemia (Figueroa et al., 2010). Although, this data indicates that a loss of 5hmC might be important in the formation of a CIMP phenotype, the exact role of this DNA modification remains to be elucidated.

The loss of 5hmC is now considered by some as a universal feature of cancer (Ficz and Gribben, 2014, Ko et al., 2010) (Kudo et al., 2012, Muller et al., 2012). While 5hmC loss can be explained by mutations in the TET and IDH proteins, these mutations are limited to only a small fraction of cancer types (Figueroa et al., 2010). Recently, it has been proposed that 5hmC cannot be accumulated due to the rapid cycling of cancer cells (Bachman et al., 2014). However, it is known that actively dividing ESCs maintain relatively large amounts of 5hmC. Although the cause of the loss of 5hmC observed in cancer is not currently clear, one study has reported that the deposition of 5hmC in a melanoma model could stop and even reverse cancer progression (Lian et al., 2012). These findings strongly suggest that a loss of 5hmC is required for malignant transformation.

While promoter hypermethylation is a prominent feature of cancer genomes, a number of studies have reported global hypomethylation in neoplastic malignancies (Ehrlich, 2009). Whereas it has been suggested that DNA demethylation of centromeric repeats and retrotransposons results in genomic instability and cancer progression (Eden et al., 2003, Daskalos et al., 2009), the mechanistic details of these phenomena are currently unknown. Interestingly, the vast majority of globally

demethylated regions in cancer are packed into heterochromatin (Ehrlich, 2002), which implies that DNA methylation is not essential for the formation of heterochromatin in cancer cells. However, some studies argue that global DNA demethylation occurs in late replicating regions and that it is merely the consequence of increased replication rates in cancer (Hon et al., 2012).

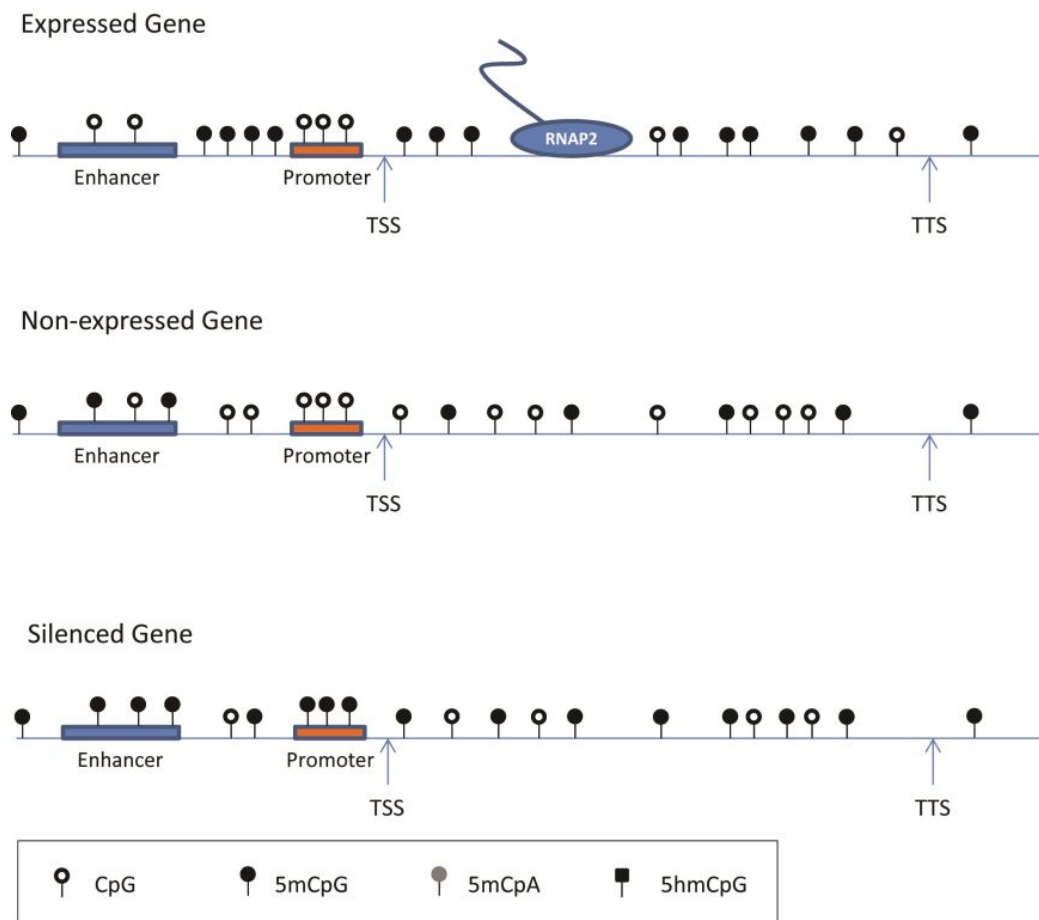


Figure 1.8 The DNA modification landscape in genic and proximal regions in cancer cells

Schematic representation of DNA modifications in expressed, non-expressed or silenced genes. Unmethylated, methylated and hydroxymethylated CpG dinucleotides, and methylated CpA sites are represented. Transcription start sites (TSS) and transcription termination sites (TTS) are indicated by arrows.

Furthermore, loss of non-CpG methylation is another phenomenon common to all cancers (Figure 1.8) (Stirzaker et al., 2014). Because the role of non-CpG methylation is currently unclear, it is very hard to speculate whether or not this loss actually contributes to neoplastic transformation.

In addition to CpG islands, the methylation patterns of other genomic elements such as gene bodies have been investigated in cancer cells (Yang et al., 2014). One study, using bisulfite padlock probes, estimated that the most highly expressed genes exhibit the highest methylation levels, while non-expressed genes have little DNA methylation, in the GM06990 cell line (Figure 1.8) (Ball et al., 2009). In another study, a more comprehensive meta-analysis showed that genic DNA methylation in cancer cell lines looks like a parabola curve when plotted against gene expression (Jjingo et al., 2012). This means that non-expressed genes have little gene body DNA methylation, which increases with expression levels, while the gene bodies on the most highly expressed genes start to lose DNA methylation. However, these conclusions were drawn based on reduced representation bisulfite sequencing experiments, which provide little data about gene body DNA methylation and should be considered with caution. Nonetheless, it seems that cancer cell DNA modification landscapes do not resemble those of ESCs. However, even though the methylomes of cancer cells are, to some extent, remodelled they still resemble the methylation patterns of the parental cell types before malignant transformation (Kulis et al., 2012). In addition, although cancer cells can be efficiently killed using DNA demethylating drugs such as 5-azacitidine *in vitro*, only very few cancer types respond to this therapy in the clinic. Furthermore, recent studies interrogating the mechanism of action of 5-azacitidine have proposed that the main reason for its ability to induce cell death might be its activation of retroelements and induction of

the cellular inflammatory response (Roulois et al., 2015, Chiappinelli et al., 2015). Thus, while it is probable that DNA modifications play a role in cancer, they do not seem to be the main driver of malignant transformation.

1.5. Readers of Methylated Cytosine and its Oxidised Derivatives

It is thought that DNA modifications and chromatin evolved to facilitate the complex gene regulation mechanisms required in higher organisms. DNA modifications in vertebrates expanded the regulatory potential of the genome by increasing or reducing DNA's binding affinities to a number of transcription factors and chromatin remodellers. These factors, whose binding is affected by a modification in the DNA sequence, can be referred to as readers of DNA modifications. Readers of CpG DNA methylation have been studied for a long time and can be subdivided in several categories (Jeltsch, 2008):

1) MBD (Methyl-CpG binding domain) family proteins that bind methylated cytosine via MBD domains. The main members of this family are MeCP2, MBD1, MBD2, MBD3 and MBD4.

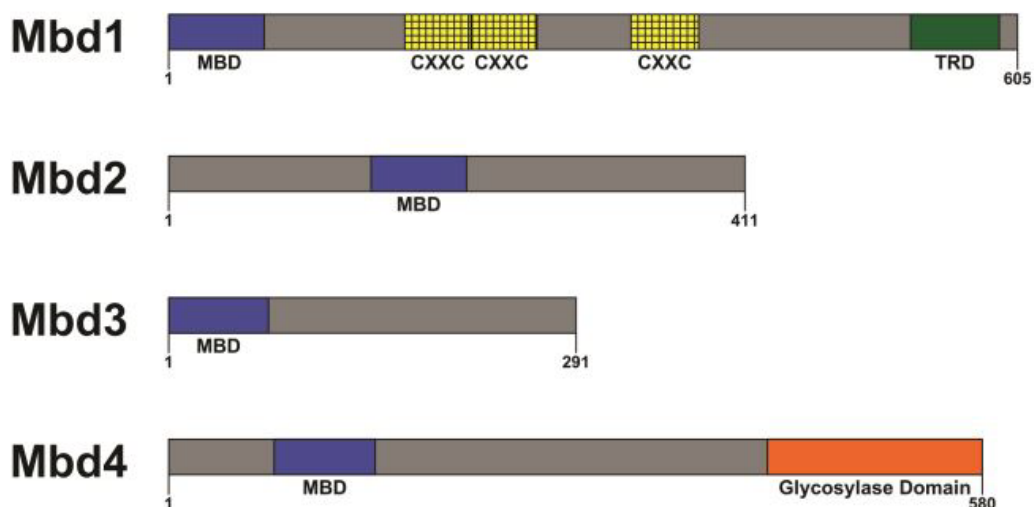
2) SRA family proteins that utilise an SRA domain to bind methylated DNA. This family has two members UHRF1 and UHRF2.

3) Other transcription factors that use zinc finger domains to bind methylated DNA, such as the Kaiso protein family.

All of these proteins have higher affinities to methylated DNA compared to unmethylated DNA, and integrate DNA methylation signals to facilitate downstream processes.

1.5.1. MBD Family Proteins

The MBD protein family is probably the most studied group of proteins capable of binding to modified DNA and consists of five main homologues: MBD1 (Cross et al., 1997), MBD2, MBD3, MBD4 (Hendrich and Bird, 1998) and MeCP2 (Meehan et al., 1992). All family members contain a MBD (Figure 1.8), the modular domain which binds to methylated DNA (Ohki et al., 2001). The exception is MBD3 protein, which MBD domain does not seem to bind to methylated DNA, although several reports have shown that it has weak affinity for 5hmC-containing DNA instead (Roloff et al., 2003, Mellen et al., 2012). All MBD family members act as transcriptional repressors, with the exception of MBD4 which has glycosylase domain (Figure 1.9) and is involved in DNA damage repair (Hendrich et al., 1999).



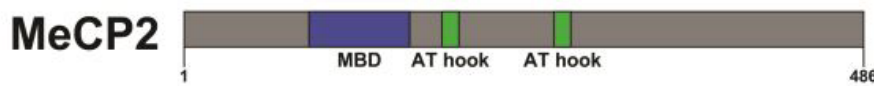


Figure 1.9 Schematic representation of canonical isoforms of MBD family proteins.

Mutations in the *MECP2* gene cause Rett syndrome, an autism spectrum disorder in females (Amir et al., 1999). MeCP2 is expressed at high levels in the brain and binds symmetrically methylated CpG, as well as methylated CpA dinucleotides (Ho et al., 2008, Chen et al., 2015). In addition, it has been reported to bind hydroxymethylated CpA dinucleotides (Chen et al., 2015). Interestingly, a large amount of non-CpG methylation exist in the brain and MeCP2 binding to this mark seems to play a role in transcriptional regulation (Chen et al., 2015). However, little or no 5hmC in a CpA context is found in the brain or ESCs (Lister et al., 2013, Yu et al., 2012) and the importance of the binding to this dinucleotide remains unknown. MeCP2 and other MBD family proteins contain TRDs (Transcriptional Repression Domains) which are responsible for transcriptional repression. In addition, MeCP2 has been shown to interact with histone deacetylases (HDACs) and other transcriptional repressors (Nan et al., 1998), and MeCP2's interaction with the NCoR/SMRT co-repressor complex was recently reported (Lyst et al., 2013). This interaction can be abolished by the mutations in MeCP2 found in patients with Rett syndrome, strongly supporting the role of MeCP2 as a bridging molecule between DNA methylation and transcriptional regulators. However, in *Mecp2* KO mouse models very few genes are deregulated in the brain (Smrt et al., 2007). In addition, MeCP2 has been proposed to be a global activator of gene expression (Li et al., 2013) yet the evidence for this is highly controversial. Finally, using a *Mecp2* KO animal

model that has a severe neuronal, Rett syndrome-like phenotype, it has been convincingly shown that MeCP2 plays a repressive role on long genes that are enriched in non-CpG methylation (Gabel et al., 2015).

MBD1, which is expressed in the majority of tissues, has three CXXC domains in addition to a MBD domain, and one of the CXXC domains (CXXC-3) has been shown to be able to bind unmethylated CpG dinucleotides (Jorgensen et al., 2004). The ability of MBD1 to bind methylated as well as unmethylated DNA seems to be counterintuitive. However, the presence of multiple targeting domains can be observed in the MeCP2 protein, which can bind to AT-rich DNA regions as well as methylated CpGs (Baker et al., 2013). Alternatively, CXXC might act as autoinhibitory module as has been proposed for the CXXC domain of DNMT1 (Song et al., 2011b). *Mbd1* KO animals show a neuronal phenotype that resembles autism spectrum disorders (Allan et al., 2008). However, this phenotype is not as pronounced as that of *Mecp2* KO animals. Although the exact function of Mbd1 is not entirely clear, in HeLa cells, MBD1 has been shown to interact with and recruit histone H3K9 methyltransferase SETDB1 to methylated DNA and establish repressive chromatin specifically in S phase (Sarraf and Stancheva, 2004).

MBD2 is another member of the MBD family that has been shown to act as a transcriptional repressor (Ng et al., 1999). MBD2, as well as MBD3, is part of the NuRD transcriptional repressor complex, however, the presence of these proteins in the complex has been shown to be mutually exclusive (Le Guezennec et al., 2006). MBD2 has been shown to recruit the NuRD complex to methylated DNA (Zhang et al., 1999), which subsequently mediates transcriptional repression via chromatin remodelling and HDAC activity (Ng et al., 1999). *Mbd2* KO animals are fertile and appear to be normal despite some mild behaviour abnormalities. However, *Mbd3* KO

animals show early embryonic lethality (Hendrich et al., 2001), the reason for which is unknown.

In summary, the MECP2, MBD1 and MBD2 proteins seem to be important for neuronal functions, and possibly others that are not easily detected in KO animal models. In addition, MBD family proteins seem to link DNA modifications to chromatin modifying enzymes, as can be seen from interaction data. However, linking this activity to global transcriptional regulation is difficult, probably due to the fact that these proteins play a modulatory role in transcriptional regulation.

1.5.2. Other Readers of Cytosine Modifications

Recently, a comprehensive study has interrogated transcription factor binding to DNA containing oxidised forms of 5mC using a mass spectrometry-based approach (Spruijt et al., 2013). The authors reported that the majority of 5mC binding proteins have reduced or diminished ability to bind oxidised forms of 5mC. One example is the Kaiso family of proteins, which have been shown to bind specifically to methylated DNA and act as transcriptional repressors via their interactions with nuclear receptor co-repressor (NCoR) and HDAC3 (Yoon et al., 2003). The Kaiso family does not show binding to oxidised forms of 5mC (Spruijt et al., 2013). However, binding to only one specific modification is not a universal feature of the readers of DNA modifications. UHRF1, for example, has been shown to interact with both 5mC and 5hmC in a hemi-modified form via its SRA domain (Spruijt et al., 2013, Frauer et al., 2011). UHRF1 has been shown to stimulate DNMT1 activity and has been speculated to be involved in the propagation of DNA modifications during cell division (Bashtrykov et al., 2014). UHRF2 is another SRA domain-containing protein that has been demonstrated to

bind 5hmC, 5caC and 5fC, although weekly (Spruijt et al., 2013). Taking into consideration that UHRF2 has been shown to play a role in the DNA damage response (Luo et al., 2013), it seems that it could be involved in the active DNA demethylation pathway. Indeed, many proteins that are reported to bind 5fC and 5caC, such as N1, have been shown to be DNA glycosylases (Spruijt et al., 2013). This further supports the role of these bases in active demethylation.

1.5.3. CXXC Domain-Containing Proteins Bind to Unmodified CpG Dinucleotides

One important group of DNA binding proteins is the group that contains the CXXC domain. The CXXC is zinc finger DNA binding domain with a binding preference to unmodified CpG dinucleotides (Voo et al., 1999). This family is not only important for its ability to bind CpG sites, which are the most common site for DNA modifications, but also because many epigenetic modifiers contain this particular domain. Lysine demethylases KDM2A and B, transcription factors FBXL19, CFP1, IDAX and CXXC5, and also DNMT1, MLL1, MLL2, MBD1, TET1, TET3 all have CXXC domains, although not all of them are functional (Long et al., 2013). Binding of CFP1, KDM2A or B to CpG dinucleotides in DNA can be efficiently abolished by DNA methylation, which ensures targeting of these chromatin modifiers to unmethylated CpG islands (Xu et al., 2011a, Farcas et al., 2012). Interestingly, the CXXC domain present in TET enzymes has been proposed to target the proteins to CpG islands and eliminate methylation deposited by mistargeted DNMT family enzymes (Xu et al., 2011b). However, some CXXC domains have more complex functions. The CXXC

domain present in DNMT1 has been shown to act as an autoinhibitory module that suppresses the transfer of the methyl group to the DNA (Song et al., 2011b).

1.5.4. Proteins Binding to Unmodified DNA

While a number of proteins exist that specifically bind to modified DNA, the vast majority of proteins bind to their consensus DNA sequences in an unmodified state. The effect of DNA methylation on transcription factor binding has been investigated for ETS (Gaston and Fried, 1995), AP-2 (Comb and Goodman, 1990), c-Myc (Prendergast and Ziff, 1991), Ah receptor (Shen and Whitlock, 1989), NF- κ B (Bednarik et al., 1991), CREB/ATF (Becker et al., 1987), E2F (Kovesdi et al., 1987), GABP (Yokomori et al., 1995) and many others. The resulting general consensus is that transcription factor binding affinities to DNA are reduced by DNA methylation. However, it is important to stress that cytosine must be present in the binding motif in order for the factor to be affected. In addition, it seems that the discovery of abundant non-CpG methylation in the brain and ESCs should fuel more effort to determine whether it plays a role in transcription factor binding. Interestingly, a number of transcription factors exist that are not affected by DNA methylation. A few notable examples would be Sp1 and YY1 (Tate and Bird, 1993).

As discussed above, there is a strong interplay between DNA sequence, DNA modifications, histone marks and DNA binding proteins. This complex system plays a role in many cellular processes, most notably the transcriptional regulation of gene expression.

1.6. Promoter DNA Modifications and Transcriptional Regulation

During the evolution of simple single cell organisms to complex multicellular life forms, multiple layers of gene regulation have emerged (Chen and Rajewsky, 2007). Transcription factors, chromatin, DNA modifications and other regulatory elements comprise a complex system, with multiple feedback and feed-forward loops, where different elements act in tandem to drive cell-specific transcription programmes.

In terms of DNA methylation, the most studied genomic regulatory element is the CpG island (CGI) (Bird et al., 1985). CGIs are found in the genomes of species with genomic DNA methylation such as vertebrates. CpG island definition varies slightly from study to study but, in general, they are defined as genomic regions of at least 200 bp with a GC content of more than 50% and observed to expected CpG ratio of more than 0.6 (Illingworth and Bird, 2009). CGIs often overlap with transcription start sites but intra- and intergenic CpG islands also exist. These islands are thought to represent alternative or cryptic promoters (Illingworth et al., 2010). Methylation of CpG islands harbouring TSSs has been linked to inactivation of gene expression. This has been observed in numerous single gene (Merlo et al., 1995) and genome-wide expression studies (Hon et al., 2012, Lister et al., 2013, Pastor et al., 2011). In addition, demethylation of methylated promoter CpG islands by treatment with DNMT inhibitors or downregulation of DNMT enzymes leads to re-expression of some genes (Lund et al., 2014, Yang et al., 2014), but not all of them. Therefore, an alternative hypothesis has been proposed that CGI methylation follows gene inactivation. This model has been supported by several studies of gene inactivation by Polycomb proteins (Vire et al., 2006). In general, a substantial amount of evidence exists that CpG island methylation is the mechanism of gene repression regardless of the timing of its deposition. This repression is thought to be mediated by the blocking

of transcription factor binding sites on the DNA (discussed in sections 1.5.3 and 1.5.4) and recruitment of mCpG binding proteins, which establish the repressive chromatin (section 1.5.1) (Klose and Bird, 2006). Importantly, more than 60% of promoters are marked by CpG islands that meet the formal definition. However, genome-wide DNA methylation analysis in somatic and germline cells has revealed that the CpG density of all the promoters can vary from high to intermediate to low, indicating that some promoters can be methylated and depleted of CpGs and yet be fully functional and active (Weber et al., 2007). Furthermore, the same study argued that methylation could enhance the transcription from low CpG density promoters. Most promoters that reside within CGIs lack a TATA box and have rather broad transcription initiation patterns, as indicated by RNA cap pulldown analysis (CAGE) experiments. Interestingly, developmentally important genes tend to have wider CGIs compared to housekeeping genes that contain rather narrow promoter CGIs (Carninci et al., 2006). The evidence, therefore, suggests that CpG islands are important gene regulatory elements that provide the platform for transcription initiation. However, while genes with methylated promoter CGIs are silenced, a large fraction of genes are also silenced despite the fact that TSSs harbouring CpG islands are unmethylated. Furthermore, only around 3% of CGI promoters have been found to be methylated in terminally differentiated cells (Weber et al., 2007). Taken together, these data indicate that promoter DNA methylation is one of several mechanisms used to repress mammalian gene expression.

Recent advances in the field of transcriptional regulation and the development of genome-wide RNA polymerase 2 (RNAPII) maps indicate that transcriptional machinery is recruited to both active and inactive genes. This suggests that the recruitment of RNAPII is not a rate limiting step in the initiation of transcription

(Core et al., 2008). It is now becoming clear that transcription is controlled by the so-called negative transcription elongation factors DSIF (Wada et al., 1998) and NELF (Yamaguchi et al., 1999), which cause promoter proximal pausing of RNAPII. This promoter proximal pausing of the transcriptional machinery can be reversed by the positive transcription elongation factor P-TEFb (Cheng and Price, 2007). One of the proposed mechanisms suggests that P-TEFb is recruited to the desired promoters via BRD4 protein which binds to acetylated lysine (Hargreaves et al., 2009). While there is no evidence that DNA modifications play a role in this mechanism of transcription regulation, transcriptional machinery recruited to the promoters of inactive genes could protect them from being methylated.

Another mechanism for gene repression where CpG islands play an important role is Polycomb-mediated gene silencing (Schuettengruber et al., 2007). Polycomb-mediated silencing is initiated by PRC2 (Polycomb Repressive Complex 2), which mediates histone 3 lysine 27 methylation (H3K27me3). H3K27me3 is subsequently recognised by the PRC1 complex, which ubiquitinates H2A and assembles compact chromatin, leading to transcriptional repression. Interestingly, gene promoters repressed by Polycomb can be methylated, which is thought to act as an additional layer of repression (Vire et al., 2006). Furthermore, this double repressive mechanism probably does not allow genes to be upregulated upon demethylation.

While promoter methylation represses transcription, unmethylated CpG islands allow binding of initiating transcription factors and the general transcription machinery, which would otherwise be blocked by DNA methylation (discussed in section 1.5.4). Furthermore, CpG islands aid transcription initiation by recruiting various chromatin modifiers via so-called CpG signalling (described in section 1.5.3).

Early chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) experiments have shown that CpG islands are associated with rather distinct chromatin signatures that make these regions different from the rest of the genome. CGIs have been linked to high levels of histone H3 and H4 acetylation (Wang et al., 2008), as well as histone H3K4me3 (Mikkelsen et al., 2007). H3K4me3 is established via CfpI protein, which binds to CGIs via its CXXC domain and recruits the Setd1 which is a part of histone methyltransferase complex (Lee and Skalnik, 2005). H3K4 methylation seems to act as a platform and recruits chromatin remodelling complexes and histone acetyltransferases to enhance transcription (Saksouk et al., 2009, Wysocka et al., 2006). Histone acetylation puts a negative charge on the histone tails and is thought to loosen the interactions between histones and negatively charged DNA, which aids transcription initiation (Li et al., 2007). This indicates that promoter CpG islands act as a platform with open chromatin where transcription initiation events are executed.

Based on all the evidence discussed so far, it seems that CpG islands are evolutionarily conserved regions which, together with co-evolved molecular machinery, regulate gene expression. It appears that they are important elements for majority of the promoters where various cellular triggers are integrated and converted into a transcriptional output. In addition to promoters, dynamic changes in DNA modifications in development and disease have also been observed in regulatory elements and gene bodies. However, little is known about how these DNA modifications affect transcription.

1.7. Gene Body DNA Modifications and Transcriptional Modulation

Gene bodies are very different genomic elements compared to promoters. By definition, a gene body is a genomic region between a transcription start site (TSS) and a transcription termination site (TTS) (Nanty et al., 2011). From a bioinformatics perspective, this definition is very convenient because the coordinates of the TSS and, to some extent, the TTS can be determined in RNA-seq experiments. In contrast, the genomic boundaries of promoters or enhancers are quite complicated to estimate. In addition, genomic elements can overlap. For instance, promoters extend into gene bodies and enhancers (also alternative promoters) can be found within gene bodies (Deaton and Bird, 2011). Interestingly, genic DNA is highly modified compared to promoters (Ball et al., 2009, Mellen et al., 2012). Furthermore, high conservation of gene body DNA methylation in numerous organisms such as mammals, honeybees, ants and even plants suggests that gene body DNA methylation has been selected for over the course of evolution (Feng et al., 2010b). In addition, the recently discovered novel epigenetic modification 5hmC has been shown to predominantly exist in the genic regions of terminally differentiated cells (Mellen et al., 2012). This suggests that distinct genic DNA modifications could differently affect various biological processes associated with genic regions. These effects on transcriptional regulation are discussed further.

1.7.1. Gene Bodies DNA Modifications and Regulation of Transcription Associated Processes

Promoter DNA methylation is associated with the complete shutdown of transcription. However, gene body methylation seems be the opposite; the general consensus to date is that expressed genes tend to have highly modified gene body

DNA (Mellen et al., 2012, Ball et al., 2009). This contradiction is known as the DNA methylation paradox (Jones, 1999). It is thought that gene body methylation plays an important role in the repression of various retroelements, including murine intracisternal A-particle retrotransposons or IAPs (Walsh et al., 1998), also in the silencing of tissue-specific intragenic promoters (Illingworth et al., 2010) and possibly enhancers (Stone et al., 2015), and also regulating alternative splicing (Maor, Yearim, & Ast, 2015). How changes in the genic DNA modification landscape affect these processes and transcription have been addressed in previous studies but a lot of details remain unknown.

1.7.1.1. Genic DNA Methylation and Silencing of Repetitive Elements

In the modern human genome, repetitive elements are thought to be efficiently silenced via epigenetic silencing mechanisms and the accumulation of a large number of inactivating mutations (Cordaux and Batzer, 2009). Although the mutational load might have inactivated the vast majority of retroelements, several studies comparing the genomes of different human individuals have established that one retrotransposition event occurs approximately once in every 200 births (Xing et al., 2009), and some of these events have been reported to cause human disease (Kaer and Speek, 2013). DNA methylation as a silencing mechanism for endogenous retroelements is supported by a number of observations. Firstly, the promoter of the L1 retroelements is a CpG island and it is typically highly methylated (Hata and Sakaki, 1997). Secondly, other repetitive elements such as Alu and SVA also have a high density of CpG dinucleotides (Xing et al., 2004, Wang et al., 2005). Finally, loss of DNA methylation in ESCs causes moderate upregulation of various retroelements (Karimi et al., 2011). As shown in figure 1.10 reactivation of retroelements not only

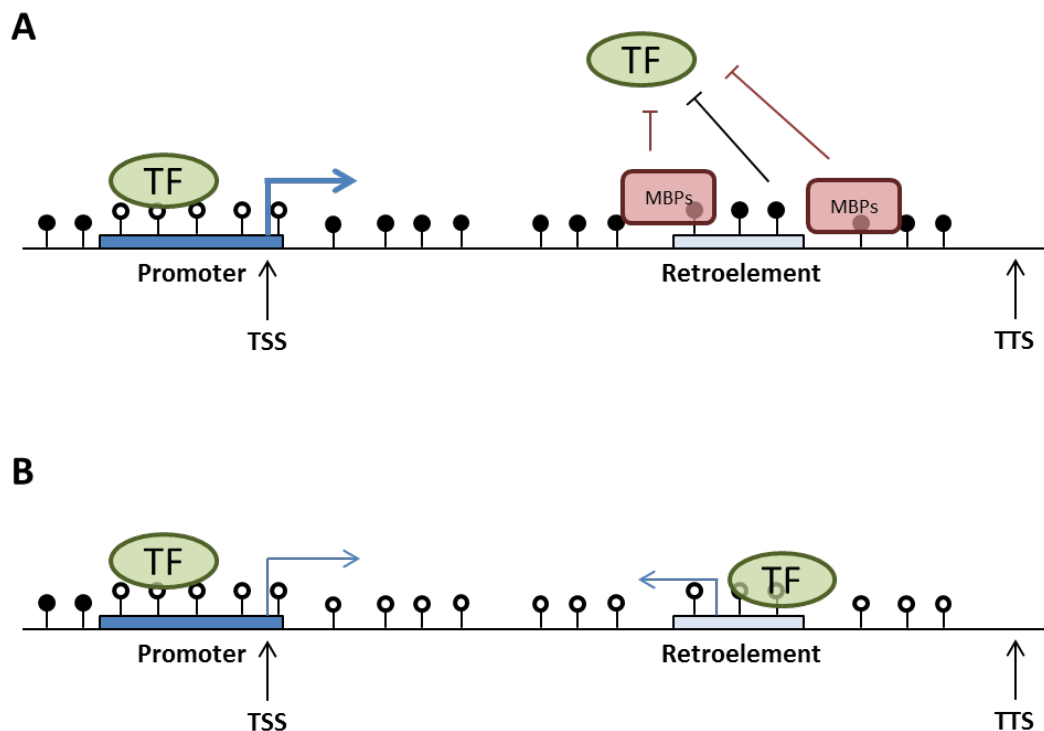


Figure 1.10 DNA methylation mediated repression of retroelements

A) Retroelements in the genome are thought to be partially repressed by blocking transcription factor binding by DNA methylation and recruiting repressive chromatin via MBD family proteins. B) Demethylation of gene body might cause activation of spurious transcription and reduction of expression of adjacent gene.

can cause deleterious genomic disruptions by integration but also downregulate the expression of surrounding genes (S. Kriaucionis unpublished data). Interestingly, other mechanisms such as methylation of histone H3 lysine 9 has also been shown to aid DNA methylation-mediated retroelement silencing in mouse ESCs (Karimi et al., 2011). In addition, it does not seem that DNA methylation-mediated silencing is required for all organisms. For example, *D. melanogaster* genome has little and *C. elegans* completely lack DNA methylation and yet manage to silence retrotransposons via various mechanisms such as piRNAs (Patil and Kai, 2010). Taken together the evidence published so far indicates that DNA methylation plays at least partial role in

suppressing activity of endogenous retroviruses. Loss of DNA methylation from the repeats can also downregulate gene expression. However, the role of 5hmC in these processes remains unknown.

1.7.1.2. Gene Body 5mC and Silencing of Intragenic Promoters

Recent mapping of transcription start sites by an RNA cap enrichment technique (CAGE) has identified more than a million transcription start sites in the human genome, while the number of used TSSs in any particular cell type is around 50,000 (Forrest et al., 2014). These findings indicate that the majority of the promoters in any cell type must be inactivated. Interestingly, a large fraction of inactivated promoters fall into the genic regions and are methylated and have, therefore, been proposed to be inactivated by genic DNA methylation. Although activation of intragenic promoters upon loss of DNA methylation has not been thoroughly addressed, evidence so far suggest that gene body methylation aids silencing of the elements capable of transcription initiation in order to facilitate the desired transcriptional programmes.

1.7.1.3. The Role of Genic DNA Modifications in Regulation of Intragenic Enhancers

Enhancers are another important type of regulatory element that can often reside in the gene bodies (Gonzalez-Vallinas et al., 2015). Enhancers are normally CpG poor regions, but a number of genome-wide methylation studies have shown that enhancer regions have dynamic and rather low levels of DNA methylation (Ziller et al., 2013). Another study has reported that demethylation of the intragenic enhancers can induce their activity (Blattler et al., 2014). Interestingly, 5hmC

mapping in ESCs and the brain has indicated that 5hmC is highly enriched over enhancer regions (Lister et al., 2013). However, the precise role of 5hmC in enhancer regulation is still unknown. In addition, it remains unclear how transcriptional elongation is affected by active intragenic enhancers, which attract transcription factors and can even produce short transcripts called enhancer RNAs (Kim et al., 2010). Although, large number of regulatory elements found within gene bodies suggests that the presence of them is not deleterious or even selected for.

1.7.1.4. DNA Methylation and Alternative Splicing

Alternative splicing is a biological mechanism that enables increased proteome complexity by generating multiple mRNA products from a single gene (Ast, 2004). It is an important process in higher organisms with large complex genomes (Keren et al., 2010) and seems to be at least partially regulated by DNA modifications (Maor et al., 2015). Alternative splicing is a co-transcriptional event that has been proposed to be affected by the transcriptional elongation rate of RNA polymerase (Brody et al., 2011, Brzyzek and Swiezewski, 2015). Gene body DNA and histone modifications have been shown to modulate RNAPII rates and affect gene splicing (Luco et al., 2011). In addition, the effects of DNA methylation on alternative splicing have been demonstrated both *in vivo* (Wan et al., 2013) and *in vitro* (Yearim et al., 2015). Interestingly, several different mechanisms for this regulation have been proposed. Firstly, higher levels of DNA methylation over some exons have been linked to higher nucleosome occupancy and higher exon inclusion rates mediated by RNAPII slowdown (Hodges et al., 2009). In addition, MeCP2 protein has also been shown to bind to more highly methylated exons and facilitate their inclusion into the mRNA (Maunakea et al., 2013). These findings indicate that DNA methylation facilitates alternative splicing by slowing down the elongation rate of RNAPII. However, recent

study has shown that alternative splicing might be regulated by alternative mechanism. CTCF can modulate alternative splicing by binding to DNA and establishing a 'roadblock' for RNAPII (Shukla et al., 2011). Interestingly, CTCF binding to DNA is abolished by DNA methylation, which in this case seems to promote exon exclusion. Altogether, DNA methylation has been suggested to regulate around 22% of splicing events (Maor et al., 2015).

Taken together, all the available data indicate that in higher organisms gene body DNA modifications affect or actively modulate various biological processes such as enhancer activation or pre-RNA splicing. In addition, changes in genic DNA modification patterns will affect these elements and mechanisms and relay the effect on the expression of the gene. How genic DNA modifications affect transcriptional initiation and elongation directly is discussed further.

1.7.2. Gene Body DNA Modification Effects on Transcriptional Elongation and Promoter Activity

To date, several attempts have been made to measure gene body DNA methylation effects on transcription genome-wide (Blattler et al., 2014, Yang et al., 2014). The authors used DNA demethylating drug to assess the role of gene body DNA methylation in transcriptional regulation. However, these studies have yielded hard-to-interpret results, with many genes being upregulated as well as downregulated. The drug used in these studies was 5-azacytidine, which has high toxicity and has been shown to kill cells via various different molecular mechanisms (Roulois et al., 2015, Chiappinelli et al., 2015). However, it has been concluded that genic DNA methylation is a transcription activating mark.

Gene body DNA methylation effect on transcription has also been investigated using artificial reporter approach. Hsieh established that gene body DNA methylation can downregulate transcriptional activity using a patch methylated episome assay (Hsieh, 1994, Hsieh, 1997). Later, similar results were produced by another study, which used a differentially methylated GFP reporter gene stably integrated into the genome of immortalised cells (Lorincz et al., 2004). This study also found that the highly methylated reporter acquired high levels of H3K9me3 over the gene body and that the unmethylated reporter gene body was hyperacetylated. The authors concluded that histone modifications and DNA methylation reduced the elongation rate of RNAPII leading to gene downregulation. However, how endogenous gene body DNA methylation affects transcriptional elongation rates genome-wide have not been addressed to date. This is probably because elongation rates are hard to measure *in vivo*. Although, new techniques such as global run-on sequencing (GRO-seq) are emerging that are capable of measuring the travelling speed of RNA polymerases (Core et al., 2008). This technique has been already used to demonstrate that genic 5caC effectively slows down the elongating RNAPII (Wang et al., 2015). However, the role of other DNA modifications in transcriptional elongation process remains to be seen.

Recently, Appanah *et al.* published a study (using the same system as in (Lorincz et al., 2004)) where DNA methylation spreading towards the promoter reduced promoter activity (Appanah et al., 2007). Although this study has provided some interesting insights, whether the same holds true for endogenous genes remains to be seen. Despite the importance of gene body methylation in development and disease, its precise role remains elusive and requires further investigation.

Recently, high levels of 5hmC deposited over gene bodies have been observed in various tissues and cell types. It has been convincingly demonstrated that highly expressed genes contain more 5hmC compared to low- or non-expressed genes (Mellen et al., 2012). However, its function and significance with regard to transcriptional regulation remained unknown and therefore has been addressed in this study.

Chapter 2

Aims of the Study

The main aim of this study is to better understand the role of gene body DNA modifications in transcriptional regulation. Particular emphasis is put on the contrasting roles of 5-methylcytosine and 5-hydroxymethylcytosine (5hmC). It is important to note that this study is focussed on transcription rather than gene expression. This is mainly for convenience, to eliminate additional complexity added by RNA splicing events. The study aims to explore the role of genic 5hmC in transcription (Chapter 4). In addition, attempts are made to go beyond simple correlations and use a reductionist approach to decouple the influence of gene body DNA modifications on transcriptional initiation and elongation (Chapter 5). To better understand the mechanistic details of genic DNA modification-mediated transcriptional fine-tuning, a candidate based approach is employed to identify the binders of the gene body DNA modifications that affect transcription (Chapter 6).

Chapter 3

Materials and Methods

3.1. Molecular Cloning

3.1.1. Plasmids Used in the Study

Expression plasmids used in this study (Table 3.1) were cloned using the strategy outlined below. Other plasmids were obtained from various different laboratories as indicated in Table 3.1.

Name	Vector	Insert	Source	RE Sites	Selection
pcDNA5/FRT/TO	pcDNA5/FRT/TO	N/A	J Christianson Lab	N/A	Amp
pBDNF1	pCR-Blunt II-TOPO	Bdnf promoter fragment + AT-rich oligo (46 bp)	S Kriaucionis Lab	N/A	KAN
pBDNF1.5	pCR-Blunt II-TOPO	Bdnf promoter fragment + AT-rich oligo (69 bp)	S Kriaucionis Lab	N/A	KAN
pBDNF3	pCR-Blunt II-TOPO	Bdnf promoter fragment + AT-rich oligo (92 bp)	S Kriaucionis Lab	N/A	KAN
pCSII-EF-MCS	pCSII-EF-MCS	N/A	S Kriaucionis Lab	N/A	Amp
pCAGGS-flpE-puro	pCAGGS-flpE-puro	N/A	J Christianson Lab	N/A	Amp
pCpGfree-	pCpGfree-	N/A	InVivoGEN	N/A	KAN

vitroNmcs	vitroNmcs				
pCpG-CpGless Renilla	pCpGfree-vitroNmcs	CpGless Renilla	Cloned	ScaI, NheI	KAN
pCpG-EGFP	pCpGfree-vitroNmcs	EGFP	Cloned	ScaI, NcoI	KAN
phRL-CMV	phRL-CMV	N/A	S Kriaucionis Lab	N/A	Amp
pMX TET1 CD	pMX	Human TET1 (1417-2136 aa)	S Kriaucionis Lab	N/A	Amp
pCDNA3	pCDNA3	N/A	Xin Lu Lab	N/A	Amp
pCDNA3 MBD1	pCDNA3	Human MBD1 FL	Cloned	KpnI, EcoRI	Amp
pCDNA3 MBD2	pCDNA3	Human MBD2 FL	Cloned	BamHI, EcoRI	Amp
pCDNA3 MeCP2	pCDNA3	Human MeCP2 FL	Cloned	BamHI, EcoRI	Amp
pCDNA3 MBD3	pCDNA3	Human MBD3 FL	Cloned	BamHI, EcoRI	Amp
pCDNA5/FRT /TO TET-CD-GFP	pCDNA5/FRT /TO	Human 3Xflag-TET1-CD-EGFP	Cloned	KpnI, NotI	Amp

Table 3.1 Plasmids used in the study

3.1.2. Polymerase Chain Reaction (PCR)

Q5 High Fidelity Hot Start DNA Polymerase (#M0493L, NEB) was used for amplification of the inserts to be cloned into plasmids. The reactions were assembled on ice according to the manufacturer's instructions. The primers used for construction of the mammalian expression vectors are shown in Table 3.2. In order to confirm that amplification was successful, 5 μ L of PCR reaction products were mixed with 1 μ L of 6 $\times$ DNA loading dye (R0611, Thermo Fisher Scientific) and subjected to agarose gel electrophoresis together with a DNA size ladder (SM1331, Thermo Fisher Scientific). Typically, 0.8–2% w/v agarose gels in 1 $\times$ TBS (89 mM Tris pH 7.6, 89 mM Boric acid, 2 mM EDTA) were used, supplemented with 1 $\times$ peqGREEN DNA Dye (PEQL37-5000, VWR). Electrophoresis was performed using a horizontal electrophoresis system (1704401, Biorad) at 100 V for 40 min. The bands were visualised using a GelDoc Imager (UVP). Provided that a single band of correct size was visible, PCR products were purified using a Q-Spin Gel Extraction/PCR Purification Kit (K1-0040, Geneflow) according to the manufacturer's instructions. Purified PCR products were subjected to restriction digestion or stored at -20 °C.

PP	Sequence	T _m	Target	Template	Plasmid Cloned	RE
SK511	5'GCTAGGTACCATGGACTAT AAGGACCACGACGGAGACTA CAAGGATCATGATATTGACT ACAAGGATGACGATGACAAG TCTGAACTGCCACCTGCAG CT3'	60 °C	TET1-CD (1417- 2136 aa)	pMX TET1 CD	pCDNA5 /FRT/T O TET- CD-GFP	KpnI, NotI

SK513	5'GCTAGCGGCCGCTTACTTG TACAGCTCGTCCATGCCG3'					
SK435	5'GCTACGGATCCATGGTAGC TGGGATGTTAGGGC3'	67 °C	Human MeCP2 FL	pCSMThME CP2 (R Klose Lab)	pcDNA3 MeCP2	BamHI , EcoRI
SK436	5'GCTACGAATTCCTAGCTAA CTCTCTCGGTCACGGG3'					
SK534	5'GCTAGGTACCATGGCTGAG GACTGGCTGGACTG3'	64 °C	Human MBD1 FL	Hematopoie tic stem cell cDNA (D Higgs Lab)	pcDNA3 MBD1	KpnI, EcoRI
SK535	5'GCTACGAATTCCTACTGCT TTCTAGCTCCAGG3'					
SK437	5'GCTACGGATCCAGGGCTCG GTTACGGAAGAAG3'	65 °C	Human MBD2 FL	cDNA clone MGC:45085 IMAGE:549 6721	pcDNA3 MBD2	BamHI , EcoRI
SK438	5'GCTACGAATTCATTCTTAG GCTTCATCTCCACTGTC3'					
SK439	5'GCTACGGATCCATGGAGCG GAAGAGGTGGGAGTGCC3'	65 °C	Human MBD3 FL	cDNA clone MGC:50958 IMAGE:615 6267	pcDNA3 MBD3	BamHI , EcoRI
SK440	5'GCTACGAATTCCTAGACGT GCTCCATCTCCGGGTC3'					

Table 3.2 Primers used in molecular cloning**3.1.3. Restriction Digestion and Dephosphorylation of Plasmids**

1 µg of purified PCR product and 1 µg of target vector were digested with the desired restriction enzyme (Table 3.1) for 1 h at 37 °C. Restriction enzymes used in this study are listed in Table 3.3. Following digestion, enzymes were heat inactivated following manufacturers' instructions and digestion products purified using a Q-Spin Gel Extraction/PCR Purification Kit (K1-0040, Geneflow). All plasmids were subjected

to dephosphorylation by calf intestine phosphatase (CIP) (M0290L, NEB) as stated in the manufacturer's instructions, followed by purification using a PCR purification kit.

Restriction Enzyme	Manufacturer	Temperature
KpnI-HF	NEB # R3142S	37 °C
NotI-HF	NEB # R3189S	37 °C
BamHI-HF	NEB # R3136S	37 °C
EcoRI-HF	NEB # R3101S	37 °C
NheI-HF	NEB # R3131S	37 °C
ScaI-HF	NEB # R3122S	37 °C
NcoI	NEB # R0193S	37 °C

Table 3.3 Restriction Endonucleases used in Cloning

3.1.4. Ligation of PCR Product into Vector

A 1:10 plasmid to insert ratio was used in the ligation reaction. The formula for the calculation of the amount of insert needed is shown below:

$$\text{Insert mass (ng)} = \frac{\text{vector mass (ng)} \times \text{insert size (kb)}}{\text{vector size (kb)}} \times \frac{\text{Insert amount (mol)}}{\text{Vector amount (mol)}}$$

Normally, 100 ng of vector would be sufficient for the reaction. Ligations were performed using T4 DNA Ligase (M0202S, NEB) according to the manufacturer's recommendations. The reactions were incubated in a water bath (Thermo Fisher Scientific) at 16 °C overnight.

3.1.5. Bacterial Transformation

Insert ligated into a vector (the desired construct) and vector only (negative control) were transformed into chemicompetent bacterial cells (XL10 Gold, Agilent Technologies) using the following protocol. Bacteria taken from a -80 °C freezer were thawed on ice for approximately 20 min. Next, ligation products were added to the cell suspension, mixed gently and incubated on ice for 20 min. Following the incubation, bacteria were heat-shocked at 42 °C in a water bath (Thermo Fisher Scientific) for 1 min and transferred back on ice for 2 min. After transformation, cells were resuspended in 1 mL of supplemented SOB media (0.5% (w/v) yeast extract, 2% (w/v) tryptone, 10 mM NaCl, 10mM MgSO₄, 10mM MgCl, 20mM Glucose) and incubated for 1 h at 37 °C. Following this, bacteria were centrifuged at 10,000 g for 30 sec in a microcentrifuge (5424R, Eppendorf), resuspended in 50 µL supplemented SOB and plated on pre-warmed LB agar plates with appropriate antibiotics. 100 µg/mL were used for Ampicillin plates and 50 µg/mL were for Kanamycin plates. Bacteria were grown overnight at 37 °C in an incubator (IGS60, Thermo Fisher Scientific). The following morning, bacterial plates were analysed by colony PCR.

3.1.6. Colony PCR

Colony PCR reactions were assembled as follows. 10 μ L of ddH₂O was dispensed into thin wall PCR tubes (A1402-3700, StarLab). A fresh bacterial colony was touched with a sterile p2 pipette tip, making sure that there was no excess of bacteria on the tip. The tip was subsequently transferred to a pre-prepared PCR tube. The PCR mix was assembled using DreamTaq DNA Polymerase (EP0701, Thermo Fisher Scientific). Briefly, 1 $\times$ Reaction Buffer, 0.2 μ M of dNTP mix (R0191, Thermo Fisher Scientific), 1 μ M of forward and reverse primers (plasmid specific) and 0.5 U of DNA polymerase were assembled in a total reaction volume of 20 μ L. The forward primer was designed to anneal on the plasmid and the reverse one to anneal on the insert. The reactions were run in a Piko Thermal Cyclor (Thermo Fisher Scientific) using the following cycling conditions: Initial denaturation for 5 min at 95 °C, denaturation for 95 °C for 30 sec, annealing reaction (primer specific, normally the lower melting temperature of the two primers minus 3 °C), extension at 72 °C for 1 min/kb, final extension at 72 °C for 7 min. Routinely, 24 cycles were used. Following amplification, 5 μ L of the product were loaded onto an appropriate percentage agarose gel and screened for the amplicon of desired size (as described previously).

3.1.7. Plasmid Amplification and Sequencing of the Insert

At least four positive colonies, based on colony PCR results, were grown in 5mL LB media at 37 °C overnight. The plasmids were isolated using a QIAprep Spin Miniprep Kit (27104, Qiagen) following the manufacturer's instructions. The plasmids were quantified using a NanoDrop (Thermo Fisher Scientific), diluted to a concentration of 100 ng/ μ L and sent to a Sanger sequencing service (Source

Bioscience, UK). Sequences were aligned to reference sequences using ApE Plasmid software. Quality of base calls was assessed and the final assembled DNA sequence was translated *in silico*. The obtained amino acid sequence was blasted against the NCBI protein database to ensure that no mutations were present. The desired plasmids were amplified using a Plasmid Maxi Kit (12162, Qiagen) or Plasmid Midi Kit (12143, Qiagen), depending on the amount of plasmid required. Plasmid concentrations were quantified using a Nanodrop instrument (Thermo Fisher Scientific).

3.2. Genomic DNA Extraction

After harvesting, genomic DNA from cultured cells was extracted using a Gene Jet Genomic DNA Extraction Kit (K0721, Thermo Fisher Scientific) according to the manufacturer's instructions.

3.3. DNA Hydrolysis for HPLC analysis

Extracted gDNA was hydrolysed as described by (Kriaucionis and Heintz, 2009). Proteins were removed by centrifugation at 14,000 g for 20 min in an Amicon Ultra-0.5 (3 kDa) column (UFC5003BK, EMD Millipore). The flow-through was lyophilised and resuspended in 50 µL of Buffer A (100 mM ammonium acetate pH 6). 40 µL of the solution was injected into an Agilent 1290 Infinity Series HPLC system. The flow rate was set at 0.400 mL/min; nucleosides were separated using an Eclipse Plus C18, 2.1×150 mm, 1.8 µm column (Agilent) at 40 °C. The run was performed using a 1.8–

30% linear gradient of 40% acetonitrile. The acquired data was analysed using Agilent 1290 Infinity Series software.

3.4. Determination of Relative Extinction Coefficients for Deoxycytidine and its Derivatives

HPLC-UV analysis allows the comparison of concentrations of different nucleosides in a sample based on absorbance measurements. Absorbance is determined by the Bear-Lambert Law and described by the formula:

$$ABS = \varepsilon \times b \times c$$

Where ABS is absorbance, ε is the extinction coefficient, b is the path length of the chamber and c is the concentration of the substance. Extinction coefficient has a specific value for specific conditions. In order to compare relative cytosine modification amounts in different samples, equal amounts of dC, 5mdC and 5hmdC were run on a HPLC instrument to obtain the ABS peak areas ($A(C)$, $A(5mc)$ and $A(5hmc)$) for a specific cytosine modification. The experiment was done in triplicate and concentrations of all cytosine modifications were kept constant ($c(C)=c(5mC)=c(5hmC)$). From the Bear-Lambert Law, the following was derived:

$$A_c/\varepsilon_c=A_{mc}/\varepsilon_{mc}=A_{hmc}/\varepsilon_{hmc}$$

Absorbance values (A) were experimentally defined and the relative extinction coefficients calculated.

3.5. RNA Extraction

RNA from adherent cell lines was extracted using an RNeasy Mini kit (74104, QIAGEN) according to the manufacturer's instructions. Briefly, cells were lysed by passing them seven times through a 21 G needle (GS351, Appleton Woods) using a 1 mL syringe (SS01T1, TERUMO). On-column genomic DNA digestion was performed according to the manufacturer's instructions. RNA was eluted in 30–50 μ L of elution buffer and the concentration determined by NanoDrop (Thermo Fisher Scientific).

3.6. RT-qPCR and data analysis

For RT-qPCR analysis, the GoTaq 1-Step RT-qPCR System (A6020, Promega) was used. 20 μ L reactions were assembled on ice in 0.1 mL strip tubes (981103, QIAGEN). Primers (Table 3.4) were used at a concentration of 250 nM each. The primer sequences were either taken from published studies or designed to be 75–150 bp and span exon–exon junctions, unless otherwise stated. The amplification efficiencies were determined for the custom designed primers using a four point serial dilution for RNA (25 ng, 50 ng, 100 ng, 200 ng per reaction) and the size of the product was confirmed by agarose gel electrophoresis. Normally, 100 ng of total RNA was used as a template. The reactions were run using a 72 reaction rotor in a Rotor-Gene Q Thermal Cycler (QIAGEN). Data analysis was performed using the comparative C_T method (Schmittgen and Livak, 2008). All the measurements were made using three biological replicates and three technical replicates for each biological replicate. Genes deregulated with $p < 0.05$ (Student's *t*-test) were considered to be significantly differentially expressed.

PP	Sequence	T _m	Target
SK1346	5'-CTTATCCGTGGTCTGGAGCAGG-3'	60 °C	SOGA1 mRNA
SK1347	5'-CGTTTCTTCTCCACCACCTCG-3'		
SK1114	5'-AACATTGTCGTCATTGGACA-3'	60 °C	EEF1A1 mRNA
SK1115	5'-ACTTGCTGGTCTCAAATTTTC-3'		
SK1116	5'-TGGAGGCGGGGGCGCTGCCCA-3'	60 °C	CDKN2A mRNA
SK1117	5'-TCGTGCACGGGTCGGGTGAGA-3'		
SK1118	5'-CTACCAGGCCAGCGTGTC-3'	60 °C	DLL1 mRNA
SK1119	5'-GCGGCTGATGAGTCTTTCT-3'		
SK1120	5'-AAGAGTCGCAGCGTGCTGCGC-3'	60 °C	SIX2 mRNA
SK1121	5'-CACCGACTTGCCGCTGCCATTTCAG-3'		
SK1122	5'-AGCCTGAGGGAGGAGGAGGAG-3'	60 °C	KCNG1 mRNA
SK1123	5'-TCCACCAGCAGCGTGATGTAG-3'		
SK1124	5'-AGGAGGAGGTCATCCGGCTC-3'	60 °C	MAFA mRNA
SK1125	5'-CTTCTCGCTCTCCAGAATGTG-3'		
SK1126	5'-AGAGACGTGCTCAGCCCCAGG-3'	60 °C	OSR1 mRNA
SK1127	5'-GGGCAGATGGTCCGAAGGCA-3'		
SK1128	5'-TGCTGCAGCCAGAAGGTGC-3'	60 °C	ZFYVE21 mRNA
SK1554	5'-CTGAGTTTCCAAACGTGACGAGG-3'		

SK1553	5'-AAGGCTCGTTCCTGTTTCAGA-3'	60 °C	GATA2 mRNA
SK1129	5'-GGCATTGCACAGGTAGTGG-3'		

Table 3.4 RT-qPCR primers used in the study

3.7. Adapter Ligation for MeDIP, hmSeal and EpiJET DNA Modification Enrichment

5 µg of extracted genomic DNA in 120 µL of ddH₂O was sonicated to the size of 200–300 bp using a Covaris System (Covaris S220). The DNA was sheared for 55 sec at 10% duty cycle, 200 cycles per burst and intensity level 4. Next generation sequencing libraries were prepared prior to subjecting DNA to enrichment as follows. Ends of sonicated DNA were repaired using a NEBNext End Repair Module (E6050S, NEB) following the manufacturer's instructions. The reaction was then purified using a Q-Spin Gel Extraction/PCR Purification Kit (K1-0040, Geneflow) according to the manufacturer's instructions. The recovered DNA was subjected to dA tailing using a NEBNext dA-Tailing Module (E6053S, NEB) followed by purification as indicated above. Purified DNA was subjected to Illumina adapter ligation using a NEBNext Quick Ligation Module (E6056S, NEB). The DNA was once again purified using a PCR Purification kit and stored at -20 °C until further use.

3.8. Methylated DNA Immunoprecipitation (MeDIP)

Prior to incubation with antibody, 10 μ L magnetic Dynabeads (10007D, Thermo Fisher Scientific) were washed twice with 50 μ L of 1 $\times$ PBS (Phosphate Buffered Saline, pH 7.4). The bead-antibody complex was prepared by incubating washed beads with 2 μ g of anti-5-methylcytosine antibody (Bi-MECY-1000, Eurogentec) in 200 μ L of PBS on a rotating platform for 2 h at 4 $^{\circ}$ C. The bead-antibody complex was then washed with IP buffer (10 mM Sodium Phosphate, 140 mM NaCl, 0.05% Triton X-100). The DNA was diluted with water to a volume of 180 μ L, denatured on a heating block for 10 min at 100 $^{\circ}$ C and immediately transferred to on an ice slurry for 10 min. 20 μ L of 10 $\times$ IP buffer was added to the denatured DNA and the whole 200 μ L mix was transferred to magnetic beads conjugated to the 5mC antibody. The DNA and bead mix was then incubated for 2 h on a rotating platform at 4 $^{\circ}$ C. The beads were washed twice for 10 min with 500 μ L of 1 $\times$ IP buffer on a rotating platform at 4 $^{\circ}$ C. Finally, the beads were resuspended in 100 μ L of Proteinase K Digestion Buffer (50 mM Tris pH 8.0, 10 mM EDTA, 0.5% SDS (sodium dodecyl sulphate)) and 3 μ L of Proteinase K (19131, Qiagen) were added. The mixture was then incubated for 2 h at 50 $^{\circ}$ C with constant agitation. The DNA was purified using a DNA Clean and Concentrator 5 Kit (D4004, Zymo Research) and eluted in a 20 μ L volume of ddH₂O.

3.9. 5-hydroxymethylcytosine Selective Chemical Labelling and Pulldown (hMe-Seal)

1 μ g of DNA with ligated Illumina adapters was incubated with 0.2 pg/ μ L β -glucosyltransferase, 20 nM UDP-6-N-glucose (C Schofield Lab), 50 mM HEPES buffer

(pH 7.9), and 25 mM MgCl₂ in a reaction volume of 20 µL for 1 h at 37 °C. After incubation, DNA was purified using a Q-Spin Gel Extraction/PCR Purification Kit (K1-0040, GeneFlow) according to the manufacturer's instructions and DNA was eluted in a 30 µL volume of ddH₂O. For the biotinylation, EZ-Link™ Sulfo-NHS-SS-Biotin linker (21328, Thermo Fisher Scientific) was added to the DNA at a final concentration of 150 µM, gently mixed and incubated on a heating platform at 37°C for 2 h. The biotin-labelled DNA was purified using a PCR Purification Kit with three washes instead of two to get rid of all residual biotin. The DNA was eluted in 50 µL of water. Pull-down was performed using MyOne™ Streptavidin C1 beads (65001, Invitrogen). Prior to pull-down, 50 µL beads were washed three times with 50µL 2 × B&W Buffer (1 × B&W Buffer: 10 mM Tris-HCl pH 7.5, 1 mM EDTA, 1 M NaCl, 0.02% Tween). After the last wash, beads were resuspended in 50 µL 2 × B&W Buffer and mixed with 50 µL of DNA from the previous step. The mix was then incubated on a rotating platform for 15 min at room temperature and washed three times with 1 × B&W Buffer. Finally, DNA was eluted by adding 100 µL of fresh 50 mM DTT (646563, Sigma) and incubating the beads on a rotating platform for 2 h at the room temperature. Eluted DNA was subsequently purified using a Zymo DNA Clean and Concentrator 5 Kit (D4004, Zymo Research) and eluted in 20 µL of ddH₂O.

3.10. 5hmC Enrichment Using the EpiJET Kit

5hmC-containing DNA was enriched using an EpiJET 5hmC Enrichment Kit (K1491BID, Thermo Scientific) according to the manufacturer's instructions.

3.11. PCR Amplification and Molecular Barcoding of Enriched Material

DNA libraries were amplified and barcoded using Kappa HiFi DNA Polymerase (KK2501, Kappa Bioscience). Briefly, DNA recovered from enrichment reactions was amplified in a total volume of 50 μ L using 0.3 mM dNTP mix, 0.2 μ M of forward primer (Table 3.5), 1 μ L of reverse primer 1.2 (Illumina) and 1 U of Kappa HiFi DNA polymerase. The cycling conditions were used as suggested in the Kappa HiFi manual. Eighteen amplification cycles were used for hmSeal and EpiJET enriched samples, and 22 cycles for MeDIP enriched DNA. The barcodes of each primer are underlined (Table 3.5).

Primer	Sequence
SK925	5'CAAGCAGAAGACGGCATAACGAGAT <u>CGTGAT</u> GTGACTGGAGTTCAGACGTGTGCTCTCCGATCT3'
SK941	5'CAAGCAGAAGACGGCATAACGAGAT <u>ACATCGG</u> TGACTGGAGTTCAGACGTGTGCTCTCCGATCT3'
SK942	5'CAAGCAGAAGACGGCATAACGAGAT <u>GCCTAAG</u> TGACTGGAGTTCAGACGTGTGCTCTCCGATCT3'
SK943	5'CAAGCAGAAGACGGCATAACGAGAT <u>TGGTCA</u> GTGACTGGAGTTCAGACGTGTGCTCTCCGATCT3'
SK944	5'CAAGCAGAAGACGGCATAACGAGAT <u>CACTGT</u> GTGACTGGAGTTCAGACGTGTGCTCTCCGATCT3'

SK945	5'CAAGCAGAAGACGGCATACGAGAT <u>ATTGGC</u> GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT3'
SK946	5'CAAGCAGAAGACGGCATACGAGAT <u>GATCTG</u> GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT3'
SK947	5'CAAGCAGAAGACGGCATACGAGAT <u>TCAAGT</u> GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT3'
SK948	5'CAAGCAGAAGACGGCATACGAGAT <u>CTGATC</u> GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT3'
SK949	5'CAAGCAGAAGACGGCATACGAGAT <u>AAGCTA</u> GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT3'
SK950	5'CAAGCAGAAGACGGCATACGAGAT <u>GTAGCC</u> GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT3'
SK951	5'CAAGCAGAAGACGGCATACGAGAT <u>TACAAG</u> GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT3'
SK952	5'CAAGCAGAAGACGGCATACGAGAT <u>CGTACA</u> GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT3'

Table 3.5 Barcoding primers for Illumina sequencing**3.12. Size selection and Quantification of Sequencing Libraries**

Amplified libraries were run on a 1% Agarose gel and gel extracted to get the average size of 350 bp. Gel extraction was performed using a Q-Spin Gel Extraction/PCR Purification Kit (K1-0040, Geneflow) according to the manufacturer's instructions. The extracted DNA was then quantified using PicoGreen reagent (P11496, Life Technologies) and a GlowMax instrument (Promega). An eight point standard curve was used to determine the DNA amount of each library. Libraries then

were pooled together to achieve a concentration of 10 nM of DNA fragments and sequenced on an Illumina HiSeq2500 Instrument (Illumina).

3.13. Enrichment Quality Control using Spiked-In Modified Oligonucleotides

To ensure the good efficiency of the enrichment experiments, 5ng of three differentially modified oligonucleotides (Appendix) were spiked in the genomic DNA. The oligonucleotides were made using different templates which gave different size PCR products when amplified with SK166 (5'-CCGCCAGTGTGATGGATATC-3') and SK525 (5'-TCAAGCTTGGTACCGAGCTC-3') as shown in Figure 3.1. Different DNA modifications were introduced by substituting dC with either 5mdC or 5hmdC in the

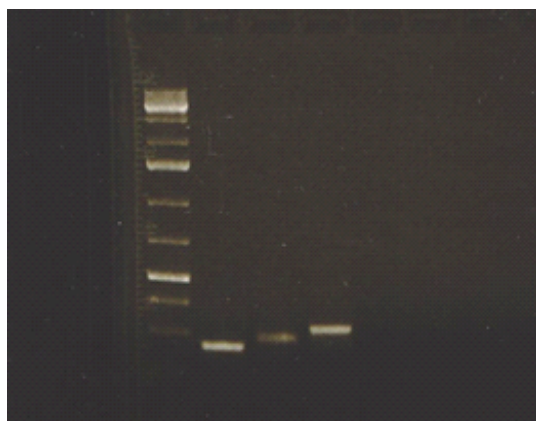


Figure 3.1 Agarose gel image of amplification products of three differentially modified oligonucleotides

DNA Ladder is shown in lane 1. Lanes 2, 3 and 4 represent unmodified, methylated and hydroxymethylated oligonucleotides, respectively.

dNTP mix. DreamTaq DNA Polymerase (EP0701, Thermo Fisher Scientific) was used to amplify the constructs. Following the enrichment procedure, pulled-down

oligonucleotides were amplified again using the same primers and enrichment of particular modification-carrying oligonucleotides was assessed by running them on an 8% acrylamide gel for 1.5 h at 90 V in a Mini Protean Tetra cell (1658004EDU, Biorad). The gels were subsequently stained with SYBR® Green I Nucleic Acid Gel Stain (S-7563, Life Technologies) and imaged using a ChemiDOC (Biorad).

3.14. Electrophoretic Mobility Shift Assay (EMSA)

Oligonucleotides used as EMSA probes (Table 3.6) were ordered from ATDbio UK and annealed by incubating relevant pairs (Table 3.7) on a heating block for 10 min prior to letting them cool down to room temperature over 3 h. The probes were labelled using γ -³²P-ATP (BLU002A001MC, Alfa Perkin) and T4 Polynucleotide Kinase (M0201S, NEB) for 1 h at 37 °C and purified using Oligo Spin Columns (11814397001, Roche). EMSA buffer (8% ficoll, 20 mM HEPES pH 7.9, 150 mM KCl, 1 mM EDTA and 0.5 mM DTT) was supplemented with 50 nM of purified MeCP2 MBD domain, 20 ng/μL polyIdC competitor DNA (P4929, Sigma) and incubated for 10 min at room temperature. Finally, 2 fmol of radioactive probe was added to 20 μL of the mixture and incubated at room temperature for 25 min. The samples were run on a 5% acrylamide, 0.5 × TBE gel at 240 V for 3 h at 4 °C. The gel was then dried for 1 h at 80 °C and exposed to imaging screed (170-7841, Biorad) overnight. The screen was imaged using the Personal Molecular Image System (170-9400, Biorad) and data analysed with QuantityOne software (Biorad).

Name	Sequence	Modification
B2785	5'-TCTGGAA C AGAATTCTTCTA-3'	5hmC

B2786	5'-TAGAAGAATTCTGTTCCAGA-3'	5hmC
B2787	5'-TCTGGAAACAGAATTCTTCTA-3'	5mC
B2788	5'-TAGAAGAATTCTGTTCCAGA-3'	5mC
B2789	5'-TCTGGAAACAGAATTCTTCTA-3'	None
B2790	5'-TAGAAGAATTCTGTTCCAGA-3'	None
B2791	5'-TCTGGAAACACAATTCTTCTA-3'	5hmC
B2792	5'-TAGAAGAATTGTGTTCCAGA-3'	None
B2793	5'-TCTGGAAACACAATTCTTCTA-3'	5mC
B2794	5'-TCTGGAAACATAATTCTTCTA-3'	5hmC
B2795	5'-TCTGGAAACATAATTCTTCTA-3'	5mC
B2796	5'-TAGAAGAATTATGTTCCAGA-3'	None
B0975	5'-TCTGGAAACGGAATTCTTCTA-3'	5hmC
B0976	5'-ATAGAAGAATTCGTTCCAG-3'	5hmC
BDMF	5'-TCTGGAAACGGAATTCTTCTA-3'	5mC
BDMR	5'-ATAGAAGAATTCGTTCCAG-3'	5mC
SK93	5'-TCTGGAAACGGAATTCTTCTA-3'	None
SK94	5'-ATAGAAGAATTCGTTCCAG-3'	None

Table 3.6 Modified oligonucleotides used in EMSA the experiment

Probe Name	Oligo 1	Oligo 2	Cytosine modification (forward-reverse)	Context
1	B0975	B0976	5hmC-5hmC	CpG
2	BDMF	BDMR	5mC-5mC	CpG
3	SK93	SK94	C-C	CpG
4	B0975	BDMR	5hmC-5mC	CpG
5	B0975	SK94	5hmC-C	CpG

6	BDMF	B0976	5mC-5hmC	CpG
7	SK93	B0976	C-5hmC	CpG
8	BDMF	SK94	5mC-C	CpG
9	SK93	BDMR	C-5mC	CpG
10	B2785	B2786	5hmC-5hmC	CpApG
11	B2787	B2788	5mC-5mC	CpApG
12	B2789	B2790	C-C	CpApG
13	B2785	B2788	5hmC-5mC	CpApG
14	B2785	B2790	5hmC-C	CpApG
15	B2786	B2787	5mC-5hmC	CpApG
16	B2786	B2789	C-5hmC	CpApG
17	B2787	B2790	5mC-C	CpApG
18	B2789	B2788	C-5mC	CpApG
19	B2791	B2792	5hmC-n.a.	CpApC
20	B2793	B2792	5mC-n.a.	CpApC
21	B2794	B2796	5hmC-n.a.	CpApT
22	B2795	B2796	5mC-n.a.	CpApT

Table 3.7 Probes used in EMSA experiment

3.15. Integration of Differential Amounts of Modified Bases into DNA

5mC- and 5hmC-containing PCR products for the transcription assays were synthesised using DreamTaq DNA Polymerase (EP0701, Thermo Fisher Scientific). Different ratios of dC to 5mdC or 5hmdC in the dNTP mix were used to synthesise

PCR products with various DNA modification densities (Table 3.8). The amounts of differentially modified cytosines in PCR products were estimated by HPLC analysis and quantification (Figure 3.2).

Fraction of DNA Modification in PCR Product	5mdC (% of Total dC in dNTP mix)	5hmdC (% of Total dC in dNTP mix)
10%	11%	35.6%
20%	28.6%	66.7%
40%	47.7%	75%
60%	75%	91.7%
100%	100%	100%

Table 3.8 The amounts of modified dC needed to obtain the relevant modification amounts in the PCR product

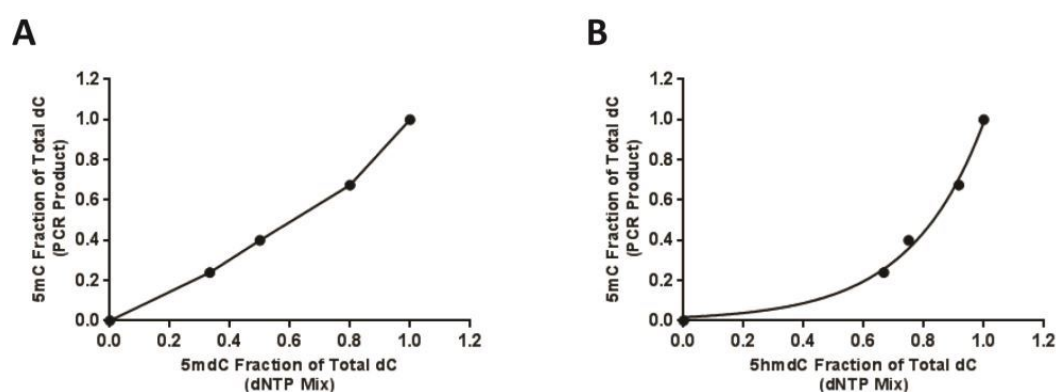


Figure 3.2 Estimation of DNA modifications in the PCR products.

A. The dependence of 5mC in the PCR product on the fraction of 5mdC in the dNTP

mix. B. The fraction of the 5hmC in the PCR product depending on the fraction of 5hmdC in the dNTP mix.

3.16. *In vitro* Transcription Assays

3.16.1. T7 RNA Polymerase *in vitro* Transcription Assay

Templates for *in vitro* transcription assays were designed as follows. The T7 promoter was amplified from the pJet1.2 plasmid (K1231, Fermentas) using DreamTaq DNA Polymerase (EP0701, Thermo Fisher Scientific) and primers 5'-GCCTGAACACCATATCCATCCGGC-3' and 5'-CCTTAATTAATGCTGAAAACTCGAGCC-ATCCG-3' (Integrated DNA Technologies). The reverse primer carries a PacI restriction site (underlined). The 315 bp ORF was amplified from pBDNF3 (Table 3.1), which contains an artificial sequence based on the Bdnf gene using primers 5'-CCTTAATTAAGCGTG-AATTTGCTAGGACTGG-3' and 5'-CCGCCAGTGTGATGGATATC-3'. DNA modifications were introduced by replacing 2'-deoxycytidine with either 2'-deoxy-5-methylcytidine or 2'-deoxy-5-hydroxymethylcytidine in the dNTP mix. PCR products were purified using a QIAquick PCR Purification Kit (28104, QIAGEN) and 3 µg digested with PacI (R0547S, NEB) restriction endonuclease. PacI digested products were purified as above and ligated using T4 DNA Ligase (M0202S, NEB) at 16 °C overnight. The correct size ligation products were gel extracted (QIAGEN Gel Extraction Kit) and DNA concentrations were determined using NanoDrop (Thermo Fisher Scientific). To confirm the amounts and quality, templates were run on an 8% acrylamide gel and imaged using ChemiDoc (BioRad). Experiments were reproduced with different sets of (550 bp) templates, prepared as indicated above. Primers 5'-CTCAAATTCGTTCTCCGGCCGCTTGGGTG-3' and 5'-CGCCTTGAGCCTGGCGAACA-3', 5'-GCCTGAACACCATATCCATCCGGC-3' and 3'-CTCAAATTCGTGCTGAAAACTCGAG-

CCATCCG-5' were used to amplify the ORF and T7 promoter, respectively. However, here *ApoI* (R0566S, NEB) restriction enzyme was used (site underlined). The ORF was amplified from the sequence of the mouse *Zdhh2* gene. Transcription experiments were performed in a Rotor-Gene Q Thermal Cycler (QIAGEN) monitoring RNA levels as an increase in fluorescence. More specifically, a High Yield T7 Transcription Kit (K0441, Fermentas) was used according to the manufacturer's instructions and the reaction solution was supplemented with 1 × SybrGreen II RNA binding dye (S-7564, Life Technologies). The reaction was performed using 20 ng of the DNA template in a 20 µL total reaction volume. The ability to detect different rates of transcription was validated by incubation of different amounts of DNA template. It was concluded that 2-fold differences in transcription rates could be confidently resolved using this method. In addition, reaction products were run on an denaturing acrylamide gel and visualised using ChemiDoc (Biorad).

3.16.2. Transcription Assay in HeLa Nuclear Extracts

HeLa nuclear extracts were purified as in Dignam et al. (Dignam et al., 1983) from 15 × 150mm confluent plates of HeLa cells, grown in DMEM High Glucose media (12-604Q, Lonza) supplemented with 10% FBS (FB-1001, Biosera), 1% Penicillin/Streptomycin (15140122, Life Technologies) and 1% L-Glutamine (25030-081, Life Technologies), or bought as a HeLaScribe *in vitro* Transcription Kit (E3110, Promega). Transcription experiments were performed as described in the HeLaScribe *in vitro* Transcription Kit manual. Briefly, 100 ng of template containing the pCMV promoter amplified from a pcDNA3.1 plasmid was ligated to the 550 bp ORF of the *Zdhh2* gene, *Renilla* Luciferase fragment as indicated above. Equal amounts of

templates were demonstrated by running them on a denaturing acrylamide gel for 90 min at 100 V. The gels were dried on a Vacuum and Heated Gel Dryer (Camlab) and subsequently exposed to imaging screed (170-7841, Biorad) overnight. The screen was imaged using a Personal Molecular Image System (170-9400, Biorad) and data analysed with QuantityOne software (Biorad).

3.17. Transfected Reporter Assays

3.17.1. Production of Differentially Modified Transcription Constructs

Modified Luciferase reporters were amplified using DreamTaq DNA Polymerase (EP0701, Thermo Fisher Scientific), primers and plasmids as indicated in Table 3.9. In order to ligate differentially modified reporters to unmodified promoters, restriction enzyme sites were introduced into primer sequences (Table 3.9). To avoid endonuclease digestion of linear constructs, primers at both ends of the construct were designed to have four thioester bonds (indicated by stars in the primer sequences) that are resistant to hydrolysis. After PCR amplification, the products were purified using a Q-Spin Gel Extraction/PCR Purification Kit (K1-0040, GeneFlow) according to the manufacturer's instructions. Following purification, PCR products were digested using AflIII (R0541S, NEB) or BamHI (R3136S, NEB) restriction enzymes for at least three hours at the temperatures recommended by the manufacturer. Next, digestion products were purified again and ligated using T4 DNA Ligase (M0202S, NEB) overnight at 16 °C. Next morning, T4 DNA Ligase was heat inactivated by incubating the reaction for 20 min at 65 °C. Ligation products were then loaded onto a 1–1.5% agarose gel and run until sufficient separation of the

products was achieved. The right molecular size fragments were then extracted from the gel and purified using a Q-Spin Gel Extraction/PCR Purification Kit (K1-0040, GeneFlow) according to the manufacturer's instructions. The concentration of each construct was estimated using NanoDrop (Thermo Fisher Scientific) and adjusted to 10 ng/ μ L with ultra-pure water. 50 ng of each construct were run on a 1–2% agarose gel to ensure the equal amounts and high quality of each product. Constructs were then stored at -20 °C until further use.

PCR Product	Source Plasmid	Primers	RE
<i>Renilla</i> Luc ORF	phRL-CMV (Promega)	5'CATCCTACACGTCCATGGCTTCCAAGGTGTACG3' 5'C*G*A*CACGGAAATGTTGAATACTCATAC3'	AflIII
Firefly Luc ORF	pGL4.11[luc2 P] (Promega)	5'CATCCTACACGTATCCGGTACTGTTGGTAAAG3' 5'G*C*T*GACTGGGTTGAAGGCTC3'	AflIII
CMV Promoter	phRL-CMV (Promega)	5'CATCCTACACGTAGAGCTGTAATTGAACTGGGAGTG3' 5'T*T*T*GCTCACATGGCTCGACAG3'	AflIII
EEF1A1 Promoter	CS2-EF-MCS (Riken Institute)	5'GCTACGGATCCGGTGGCTAGCCTATAGTGAGTCG3' 5'C*G*T*GAGGCTCCGGTGCCCGTC3'	Bam HI
SV40 Promoter	pcDNA3 (Xin Lu Lab)	5'GCTACGGATCCGGTGGCTAGCCTATAGTGAGTCG3' 5'T*G*T*CAGTTAGGGTGTGGAAAGTCC3'	Bam HI
CDKN2A Promoter	Genomic DNA**	5'GCTACGGATCCGGTGGCTAGCCTATAGTGAGTCG3' 5'A*T*T*CTCCCCGTCGGTATTAAATAAACCTC3'	Bam HI

Table 3.9 Primers and restriction enzymes used for the construction of differentially modified reporters

3.17.2. Transfection of Differentially Modified Constructs into Cells

24 h before transfection, cells (for three biological replicates) were seeded at 50% confluency on 24-well plates (3524, Costar) in DMEM Glutamax media (10566, Thermo Fisher Scientific) supplemented with 1% Penicillin/Streptomycin (15140122, Life Technologies), 1% L-Glutamine (25030-081, Life Technologies) and 10% FBS (FB-1001, Biosera). Cells were transfected with 40 ng of modified *Renilla* Luciferase construct and 40 ng of unmodified Firefly Luciferase construct (Appendix A) for normalisation. 3 µL of Lipofectamine 2000 (11668019, Life Technologies) was used per well in all transfection reactions. Transfections were carried out according to the manufacturer's instructions.

3.17.3. Recording of Luciferase Readings and Data Analysis

24 h after transfection, cells were lysed using a Dual-Luciferase Reporter Assay System (E1910, Promega) according to the manufacturer's instructions. Briefly, cells were lysed in 1 × PLB buffer for 15 min. 20 µL of lysate was transferred into a white 96-well plate (82050-726, VWR) and 100 µL of the substrate for Firefly luciferase was added to each well. Readings were taken using a 5 sec integration time on a Glowmax Luminometer (Promega). Subsequently, the substrate for *Renilla* luciferase was added and luminescence readings were recorded as above. Relative luminescence values were calculated according to the formula:

$$\text{Relative Luminescence} = \frac{\text{Modified Renilla Luminescence Value}}{\text{Firefly Luminescence Values (Control)}}$$

Relative luminescence values of at least 3 biological replicates were averaged and compared based on different DNA modifications.

3.18. Construction of Integrated Luciferase Reporter System

The construction of the integrated dual reporter system began with the pCpGfree-vitroNmcs plasmid (pcpgvtn-mcsg2, Invivogen). Firstly, the R6K replication origin was cut out using the *PacI* restriction endonuclease (R0547S, NEB). The *E. coli* replication origin was amplified from the pCR4 plasmid (K4575-02, Thermo Fisher Scientific) with primers SK652 (5'-CCTTAATTAATCATGACCAAAATCCCTTAA-CGTGAG-3') and SK653 (5'-CCTTAATTAACGTAAAAAGGCC-GCGTTGCTG-3') using Q5 DNA Polymerase (M0491S, NEB) according to the manufacturer's instructions. The amplified origin was subsequently digested with *PacI* (R0547S, NEB) and cloned into pCpGfree-vitroNmcs, which lacked a bacterial replication origin, to yield plasmid pCpGfree-vitroNmcs1. In order to separate the promoter from MCS to avoid genic DNA modifications being too close to the promoter, a human genomic region (chr21: 40,879,494-40,880,848) was amplified with primers SK654 (5'-CGATCTCGAGG-TTTCATAGAGATGGGGCGTC-3') and SK655 (5'-CGATCTCGAGCATACCAAAGATTGATAGTGCTATCTAGCG-3'), which had *XhoI* sites encoded at their overhangs. After digestion with *XhoI* (R0146S, NEB), the intron was cloned into pCpGfree-vitroNmcs1 to yield pCpGfree-vitroNmcs2.

The CpGless *Renilla* gene was designed by mutating out all of the CpG dinucleotides without changing the amino acid sequence. The construct was synthesised by the Life Technologies gene synthesis service (Appendix). The construct was amplified with primers SK604 (5'-CTGAGTACTATGGCTTCCA-

AGGTGTATG-3') and SK646 (5'-CTGGCTAGCTTACTGCTCATTCTTCAGCACTC-3'), which have Sall and NheI restriction sites engineered in their overhangs. CpGless *Renilla* reporter was then digested and ligated into the pCpGfree-vitroNmcs1 plasmid, which was digested with the same restriction enzymes. Next, the CpGless CMV promoter together with the CpGless *Renilla* ORF and polyA tail was amplified using Q5 High Fidelity DNA Polymerase (M0491S, NEB) and primers SK610 (5'-GCTGAATTCTGGGAAAAACCCATTGGAGCCAAG-3') and SK611 (5'-GCTGAATTCAAGATGAACACAGTGGGGCTCC-3'), which encode EcoRI RE sites at their overhangs. The amplicon was then cloned into pCpGfree-vitroNmcs2 plasmid digested with EcoRI-HF (R3101S, NEB). This process produced the plasmid pCpGfree-vitroNmcs3. Finally, Firefly luciferase gene was amplified from plasmid pGL4.10[luc2] (E6651, Promega) using primers SK606 (5'-GCTCAGATCTCATGGAAGATGCCA-AAAACATTAAGAAGG-3') and SK607 (5'-GCTCCCATGGGCTGACTGGGTGAAGGCTC-3'). Plasmid pCpGfree-vitroNmcs3 was digested with BglII (R0144S, NEB) and NcoI (R0193S, NEB) and amplified. Firefly luciferase gene was cloned into the multiple cloning site. This finally yielded the plasmid pCpGless-Ren-FF (Appendix A), which was used further in the study.

3.18.1. *In vitro* Methylation of the Luciferase Reporter System by M.SssI

10 µg of pCpGless-Ren-FF were linearised with PacI (R0547S, NEB) for 1 h at 37 °C and purified using a Q-Spin Gel Extraction/PCR Purification Kit (K1-0040, GeneFlow) according to the manufacturer's instructions. 3 µg of linear product were methylated *in vitro* using CpG methyltransferase M.SssI (EM0821, Thermo Fisher Scientific) following the manufacturer's instructions. The enzyme was heat-

inactivated at 65 °C for 20 min. After purification of the methylated DNA, the efficiency of the reaction was assessed by restriction digestion of constructs with Sall (R0138S, NEB) and EcoRI-HF (R3101S, NEB). Digested products were run on an agarose gel and visualised using a GelDoc Imager (UVP).

3.18.2. Integration of the Methylated and Unmethylated Reporter Systems into the Genome of the HEK293 TRex Cell Line

1 µg of methylated or unmethylated linearised pCpGless-Ren-FF construct was transfected into 80% confluent HEK293 T-Rex Flp-In TET-CD cells grown on 6 cm tissue culture plates (CLS430166, Sigma) using Lipofectamine 2000 (11668019, Life Technologies), according to the manufacturer's instructions. 24 h after transfection, cells were treated with 800 µg/mL of G418 (10131035, Life Technologies). The media and selection reagent were changed every two days. Nine days after transfection, all of the cells on the untransfected plate were dead and colonies on the transfected plates started to emerge. The concentration of the selection reagent for the transfected cells was reduced to 400 µg/mL and cells were maintained for another four passages.

Promoter methylation and integration has been done with the help of student Kathrin Jansen.

3.19. Protein methods

3.19.1. Protein Extraction

For protein analysis, cells were normally grown on 6-well plates (3516, Costar) until they reached confluency. Prior to protein extraction, cells were washed twice with 3 mL ice-cold PBS and scraped into 300 μ L of ice cold RIPA buffer (150 mM NaCl, 1.0% NP-40, 0.5% sodium deoxycholate, 0.1% SDS, 50 mM Tris-HCl pH 8.0) supplemented with Complete Protease Inhibitor Cocktail (11697498001, Roche). Cell lysates were transferred to pre-chilled Eppendorf tubes and transferred to a rotation platform at 4 °C for 30 min. Tubes with lysed cells were subsequently transferred to a pre-cooled microcentrifuge (5418R, Eppendorf) and centrifuged at maximum speed for 20 min. The supernatant containing extracted proteins was carefully transferred to new 1.5 mL tubes and kept at -20 °C until further use.

3.19.2. Protein Quantification

The concentration of the extracted proteins was determined using Bradford assay (B6916, Sigma). Briefly, the assay was performed in a 96-well flat bottom plate (3650, Costar) using 200 μ L of Bradford reagent per well. The standard curve was obtained by diluting BSA protein standard (82516, Fluka) in ddH₂O. The BSA concentration in each well ranged from 0.0025–0.05 μ g/ μ L. Prior to quantification, 96-well plates were incubated for 10 min at room temperature under cover. Measurements were made in triplicate using absorbance at 595 nm wavelength in a plate reader (Magellan, Tecan). The standard curve was fitted using Microsoft Excel software.

3.19.3. Protein Analysis using SDS-PAGE Electrophoresis and Immunoblotting

20 μ L of total protein extract (2 μ g/ μ L) was mixed with 5 μ L of 5 $\times$ SDS-PAGE loading buffer (200 mM Tris-HCl pH 6.8, 400 mM DTT, 8% SDS, 0.4% bromophenol blue, 40% glycerol) and incubated at 100 $^{\circ}$ C for 5 min, then immediately placed on ice until loading on the gel. SDS-PAGE electrophoresis was performed by loading 20 μ L of the sample along with Page Ruler Plus protein size standard (26619, Thermo Fisher Scientific) on a pre-cast 4–20% gradient polyacrylamide gel (4561093SEDU, Biorad). Electrophoresis was performed using a Mini Protean Tetra cell (1658004EDU, Biorad). Gels were run in 1 $\times$ SDS Running Buffer (25 mM Tris-base, 192 mM glycine, 0.1% SDS) at 200 V for 35 min to achieve good separation of proteins. After separation, proteins were transferred to a nitrocellulose membrane (10600002, GE Healthcare), which was activated in water and soaked in 1 $\times$ Transfer Buffer (25 mM Tris-HCl pH 7.6, 192 mM glycine, 20% methanol and 0.05% SDS) prior to transfer. Proteins were transferred in a 2-Gel Tetra Blotting Module (1660827EDU, Biorad) filled with 1 $\times$ Transfer Buffer for 1 h at 100 V. The successful transfer of the proteins was confirmed by briefly staining the membrane with Ponceau S (P7170, Sigma) and destaining with ddH₂O. The membrane was then blocked for 1 h in PBST-milk (PBS pH 7.4, 0.1% Tween-20, 5% w/v fat free milk powder (Marvel)). After blocking, the membrane was incubated with primary antibody diluted in PBST-milk for 2 h at room temperature with gentle shaking. Information about the antibodies and dilutions used in this study is provided in Table 3.11. The membrane was subsequently washed three times with PBS 0.1% Tween for 10 min each wash. Incubation with secondary antibody was performed at room temperature for 45 min. After incubation, the membrane was washed three times with at least 50 mL of PBS 0.1% Tween and

rinsed with water. Membranes were visualised using ECL Western Blotting Detection Reagent (RPN2209, GE Healthcare) and imaged using a ChemiDoc instrument (Biorad).

Antibody	Dilution	Species	Usage	Source
Anti-FLAG	1:3000	Mouse	Primary	#F3165, Sigma
Anti-MeCP2 (D4F3)	1:1000	Rabbit	Primary	#3456, Cell Signalling
Anti-Mbd1 (SK65)	1:1000	Rabbit	Primary	I Stancheva Lab
Anti-Actin (AC-15)	1:30000	Mouse	Primary HRP-conjugated	# ab49900, Abcam
Anti-PCNA	1:10000	Mouse	Primary HRP-conjugated	# sc-56, Santa Cruz
Anti-Mouse	1:20000	Goat	Secondary HRP	# sc-2005, Santa Cruz
Anti-Rabbit	1:20000	Goat	Secondary HRP	# sc-2004, Santa Cruz

Table 3.10 Antibodies used in this study

3.20. CRISPR-Cas9-Mediated Knockouts

CRISPR-Cas9 knockout technology is a recently developed technique that allows efficient and quick gene knockouts in animals, as well as in cell lines (Ran et al., 2013). CRISPR-Cas9 was used to knockout MECP2 and MBD1 in the HEK293 T-Rex Flip-In TET-CD cell line. Guiding RNA sequences (gRNAs) were designed using an online design tool (<http://www.e-crisp.org/E-CRISP/>). Two pairs of targeting gRNAs were designed for MECP2 and MBD1 (Table 3.12). Targeting oligonucleotides required for cloning were ordered from IDT Technologies, ligated into pSpCas9(BB)-2A-GFP (48138, Addgene) and transformed into bacteria as indicated above. (gRNAs were designed and cloned with help of summer student Sourdeep Basu). Plasmids

gRNA #	Targeting Oligonucleotides	Target Gene	Target Exon
62	5'-CACCGCCATGGTAGCTGGGATGTTA-3' 5'-AAACTAACATCCCAGCTACCATGGC-3'	MECP2	2
64	5'-CACCGTAGCTGGGATGTTAGGGCTC-3' 5'-AAACGAGCCCTAACATCCCAGCTAC-3'	MECP2	2
66	5'-CACCGACTTGAGGGGTTTGTCTTG-3' 5'-AAACCAAGGACAAACCCCTCAAGTC-3'	MECP2	3
78	5'-CACCGCGGAAGTCTTTCGCAAGTC-3' 5'-AAACGACTTGCGAAAGACTTCGCGC-3'	MBD1	2
80	5'-CACCGACTTCGCGGCGCTTCCAGCC-3'	MBD1	5

	5'-AAACGGCTGGAAGCGCCGGAAGTC-3'		
82	5'-CACCGTTCTCCCAGGCAGTAGACCC-3' 5'-AAACGGGTCTACTGCCTGGGAGAAC-3'	MBD1	15

Table 3.11 Cas9-targeting oligonucleotides

were then purified using a QIAprep Spin Miniprep Kit (27104, Qiagen) and sent for Sanger sequencing (Source Bioscience, UK). The plasmids with the right inserts were amplified and purified using a QIAGEN Plasmid Maxi Kit (12163, Qiagen) and transfected into cells. Cells were seeded onto 6-well plates (3516, Costar) at 70% confluency one day prior to transfection. 24 h later, HEK293 T-Rex Flip-In TET-CD cells were transfected with a mix of two plasmids (2 µg each) using 18 µL of Lipofectamine 2000 (11668019, Life Technologies). The transfections were carried out according to the manufacturer's instructions. 24 h after transfection, cells were inspected under a microscope for expression of GFP. Provided that more than 50% of the cells were expressing GFP, cells were harvested and seeded into 96-well plates (3650, Costar) using serial dilution. The cells were diluted to a density of 0.75 cells per 100 µL media, and 100 µL of suspension were dispensed into each well. 24 h later, wells were screened under a microscope and each well containing a single GFP-positive cell was marked. Cells were expanded for 15 days, with change of media every 3 days. The growing clones were then transferred to 24-well plates (3524, Costar) and subsequently into 6-well plates (3516, Costar). Once enough material was collected, clones were screened for successful deletions using immunoblotting. The positive clones were further expanded and frozen for long term storage.

3.21. Oxidative Bisulfite Sequencing (oxBS)

10 µg of genomic DNA resuspended in 50 µl of ddH₂O were sonicated for 55 seconds using a Covaris system (Covaris S220) at 10% duty cycle. 200 cycles per burst at intensity level 4 were set in order to achieve the size range of 200–500 bp. The oxidative bisulfite (oxBS) conversion was performed on 400 ng of gDNA using a True Methyl 24 Kit (Cambridge Epigenetix) following the manufacturer's instructions. Successful conversion of the methylated and hydroxymethylated DNA was confirmed by Taq α I digestion assay (Figure 3.2).

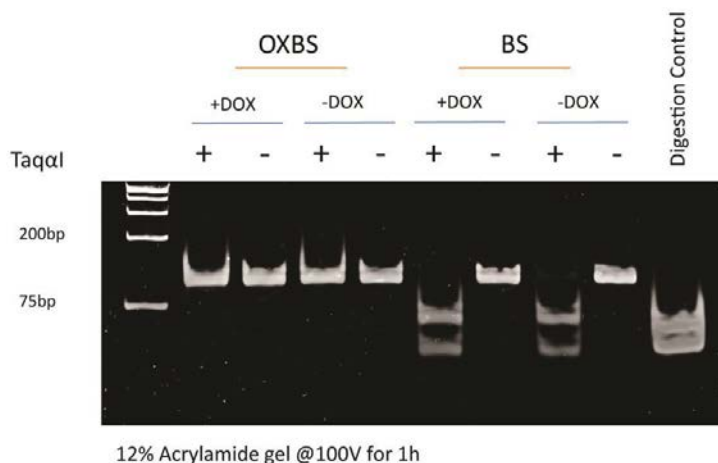


Figure 3.2 Acrylamide gel demonstrating the successful conversion of 5hmC to T in the oxBS treatment but not in BS treatment

The oligonucleotide contains a T(hm)CGA site for Taq α I that is eliminated upon oxBS treatment.

Subsequently, genic regions of interest were amplified using strand-specific primers (Table 3.12) designed using the MethPrimer tool (<http://www.urogene.org/cgi-bin/methprimer/methprimer.cgi>). The regions of interest were amplified using ZymoTaq Premix (E2003, Zymo Research) according to

the manufacturer's instructions. The right size of the amplicons was confirmed by running 10 µL of PCR product on a 2% agarose gel. Fresh PCR products (4 µl) were cl-

Target Gene	Primer Sequences	Amplicon Size (bp)
EEF1A1	(SK1112) 5'-TTGTTGTAGGGAGTTTAAAATGGAG-3' (SK1113) 5'-CCACCCACTCAATATAAAAAAACTC-3'	246
ZFYVE21	(SK1094) 5'-TATGTTTATGAGGGAGAGGATTAGG-3' (SK1095) 5'-AATTCCACAAAAAACAACAAAAC-3'	201
SIX2	(SK1104) 5'-GGTATTATGTTTATGTTGTTTATTTT-3' (SK1105) 5'-AACTCTCCAAAATCTTATAAACTC-3'	215
DLL1	(SK1106) 5'-GAAGTTTAAGATAAAATTAGGAAGT-3' (SK1107) 5'-AACCCAAAACCTTCTTTCTTTAAAAAC-3'	237
Firefly Luciferase	(SK706) 5'-GGATTATGTGGTTAGTTAGGTTATAAT-3' (SK707) 5'-AAACAACAAAATATCTATCCATACC-3'	254

Table 3.12 Amplification primers for bisulfite-converted DNA

oned into a pCR4 vector using a TOPO-TA Cloning Kit (K207020, Life technologies) according to the manufacturer's instructions. Ligation products were then transformed into XL10 Gold supercompetent cells and plated onto pre-warmed agar plates containing 100 µg/ml ampicillin. Prior to plating of the bacteria, plates were pre-treated with 40 µl of a 0.1 M IPTG solution and 40 µl of a 20 mg/ml X-Gal solution (B4252, Sigma) to allow efficient screening of the positive colonies. The white clones were picked and screened by colony-PCR in a 20 µL volume using primers SK418 (5'-CCCAGTCACGACGTTGTAAAACG-3') and SK419 (5'-AGCGGATAACAATTTACACAGG-

3') and DreamTaq DNA Polymerase (EP0701, Thermo Fisher Scientific) according to the manufacturer's instructions. PCR products (5 µL) for positive colonies were treated with 10 U of Exonuclease I (EN0581, Thermo Fisher Scientific) and 1 U of FastAP (Thermosensitive Alkaline Phosphatase) (EF0654, Thermo Fisher Scientific). The reactions were assembled on 96-well plates (E1403-1200, StarLab), incubated in a thermal cycler for 15 min at 37 °C and subsequently heat-inactivated for 15 min at 85 °C. 2 µL of the product were diluted in 8 µL of water and sent for the sequencing to the Department of Zoology sequencing facility (University of Oxford). The data was analysed using the BISMA tool (Rohde et al., 2010) available online (<http://services.ibc.uni-stuttgart.de/BDPC/BISMA/>).

3.22. Cell Culture Techniques

3.22.1. Growth Conditions for Cell Lines Used in This Study

All tissue culture procedures were carried out in a dedicated room inside a vertical flow hood (Model 1.2, Holton). Prior to handling of cells, all surfaces were sterilised with 70% industrial methylated liquid (16368, Sigma).

HEK293 T-Rex Flp-In (Life Technologies, R780-07), HeLa T-Rex Flp-In (a kind gift from Stephen Taylor's Lab) and HeLa S3 (Colin Goding's Lab) cell lines were grown in DMEM Glutamax (10566, Thermo Fisher Scientific) supplemented with 10% foetal bovine serum (FBS) (FB-1001, Biosera), 1% Penicillin/Streptomycin (15140122, Life Technologies) and 1% L-Glutamine (25030-081, Life Technologies) unless otherwise stated. J1 mouse embryonic stem cells (J1 ESCs) were grown in DMEM (BE12-604F, Lonza) supplemented with 10% FBS (FB-1001G-500, Biosera),

1% Penicillin/Streptomycin (15140122, Life Technologies), 1% L-Glutamine (25030-081, Life Technologies), 1% Non-Essential Amino Acids (11140-050, Thermo Scientific) and 1000 U/mL LIF (ESG1106, Millipore). Prior to seeding of the ESCs, all the plates and flasks were coated with a 0.1% gelatine solution in PBS.

For passaging, cells were washed with PBS and treated with Trypsin/EDTA (25300-096, Life Technologies). After cells detached, Trypsin was inactivated with cell culture media. Cells were collected by centrifuging them at 1000 rpm in a bench-top centrifuge (Multifuge 3SR+, Heraeus). Cells were normally split at a ratio of 1:4 and seeded onto an appropriate dish. All cell lines were grown at 37 °C in a 5% CO₂ atmosphere in a humidified incubator (Binder).

Cell lines were frozen according to the following procedure. Cells were grown in T75 flasks (431464U, Corning) until 70% confluency and subsequently harvested by trypsinisation. After pelleting, cells were resuspended in DMEM Glutamax (10566, Thermo Fisher Scientific) supplemented with 10% FBS (FB-1001, Biosera), 1% Penicillin/Streptomycin (15140122, Life Technologies), 1% L-Glutamine (25030-081, Life Technologies) and 10% DMSO (D2650, Sigma). 1 mL of cell suspension was aliquoted into 2 mL cryotubes (5001-0020, Thermo Fisher Scientific), which were wrapped in tissue paper to prevent instant freezing and placed in a -80 °C freezer for 24 h. The next day, cells were transferred to liquid nitrogen for long-term storage.

Frozen cells were thawed in a 37 °C water bath (SUB Aqua 26, Grant) and resuspended in 10 mL complete media. Next, cells were pelleted at 1000 rpm in a bench-top centrifuge (Multifuge 3SR+, Heraeus), resuspended in 10 mL complete medium, seeded in a T75 flask (431464U, Corning) and incubated at 37 °C in a 5% CO₂ atmosphere.

3.22.2. Cell Counting

Cell counting was performed using a TC10 Cell Counter (1450102, Biorad). Briefly, after the centrifugation and resuspension of cells, a 10 μ L aliquot was taken and mixed with 10 μ L of 0.4% Trypan blue solution (15250-061, Thermo Scientific) in a 1.5 mL Eppendorf tube. 10 μ L of the mix were transferred to a counting slide (1450011, Biorad) and counted in a cell counter. All the dilutions were made to take live cell counts into consideration. For the determination of proliferation efficiencies, cells were grown in triplicate and counted as described above. The average of the counts was determined and standard errors of the mean were calculated using Microsoft Excel software.

3.23. Construction of TET1 Catalytic Domain-Expressing Cell Lines

HEK293 T-Rex Flp-In (R780-07, Life Technologies) and HeLa T-Rex Flp-In (a kind gift from Stephen Taylor's Lab) cell lines were used for stable integration of the doxycycline inducible TET-CD gene. All the procedures were carried out following the pcDNA[™]5/FRT/TO vector (V6520-20, Life Technologies) manual with some minor alternations. The cloning of TET-CD into the pcDNA[™]5/FRT/TO vector is described in the molecular cloning section above. HeLa and HEK293 cells were transfected at 50% confluency on 10 cm tissue culture plates (430167, Corning) using Lipofectamine 2000 transfection reagent, according to the manufacturer's instructions. Cells were transfected with 10 μ g pcDNA[™]5/FRT/TO-TET-CD mixed with 1 μ g pCAGGS-flpE-puro plasmid (a kind gift from J. Christianson Lab). The next day, cells were subjected to selection with 200 μ g/mL of Hygromycin B (10687-010, Thermo Fisher Scientific).

The selection reagent was kept for 16 days, and the media changed every 3 days until all the cells on the untransfected plates had died and single colonies started to emerge on the plates transfected with pcDNATM5/FRT/TO-TET-CD. Single colonies were then picked and transferred to 96-well plates (3650, Costar). Colonies were subsequently expanded and the induction of TET-CD was confirmed by western blotting. Cells were induced by adding doxycycline (D9891-5G, Sigma) to a 1 µg/mL final concentration.

3.24. Recovery of Transfected Constructs

The protocol for the recovery of transfected constructs modified with 5hmC was developed based on work done in previous studies (Stenglein et al., 2010, Cortellino et al., 2011). HeLa cells were seeded at a density of 100,000 cells per single well of a 6-well plate (3516, Costar). Hydroxymethylated reporter construct was produced as described above. 100 ng of the construct DNA was transfected in a single well using TransIT-LT1 transfection reagent (MIR 2304, Mirus) following the manufacturer's instructions. 24 h later, cells were washed four times with PBS and harvested using a rotatable cell scraper (99003, TPP). Next, transfected constructs were purified using a QIAprep Spin Miniprep Kit (27104, Qiagen) according to the manufacturer's recommendations for bacterial plasmid extraction. In order to make sure that material from transfected cells was successfully recovered, 1 µL of the eluate was subjected to PCR reaction using primers SK700 and SK701, which were specific for the constructs. The amplified PCR products were run on a 2% agarose gel. As indicated in Figure 3.3, DNA from transfected wells, but not wells where no transfection reagent was used, showed the expected band. Following this, 10 µL of

eluate for the 5hmC modified construct were subjected to bisulfite conversion using an EpiTect Bisulfite Kit (59104, Qiagen). Following the conversion, DNA was amplified and sequenced in the same way as described for oxidative bisulfite sequencing above.

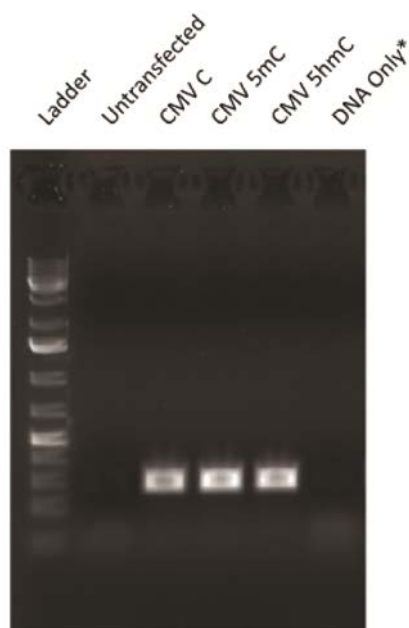


Figure 3.3 PCR products indicating the successful recovery of transfected reporter constructs

'DNA only' was a mock transfection without transfection reagent.

3.25. Combined Bisulfite Restriction Analysis (COBRA)

500 ng of genomic DNA, extracted from HEK293 T-Trex Flp-In TET-CD cells with integrated differentially modified reporter constructs, were subjected to bisulfite treatment using an EpiTect Bisulfite Kit (59104, Qiagen). Subsequently, the fragment of the firefly locus was amplified using Zymo-Taq Premix (Zymo Research) using the following primers (SK706) 5'-GGATTATGTGGTTAGTTAGGTTATAAT-3' and (SK707) 5'-AAACAACAAAATATCTATCCATACC-3'. Finally, PCR products were purified using a Q-Spin Gel Extraction/PCR Purification Kit (K1-0040, Geneflow) according to the

manufacturer's instructions and DNA concentrations were measured using NanoDrop (Thermo Fisher Scientific). 500 ng of DNA were digested using Taq^αI (R0149S, NEB) for 1 h at 65 °C and subjected to gel electrophoresis using a 2% agarose gel. (COBRA assay has been done with the help of student Kathrin Jansen).

3.26. Cell Treatments

3.26.1. L-Ascorbic acid

γ-irradiated L-ascorbate was purchased as a powder (A4403-100MG, Sigma). It was further resuspended in autoclaved ddH₂O to obtain a stock concentration of 1 mM and frozen at -20 °C in 1 mL aliquots. Cells were treated with L-Ascorbate to a final concentration of 10 μM in culture media. For longer treatments, media with L-ascorbate was changed every 24 h.

3.26.2. 5-Aza-2'-deoxycytidine

5-Aza-2'-deoxycytidine was purchased in a powder form (A3656-5MG, Sigma) and dissolved in DMSO (472301-100ML, Sigma) to obtain a stock concentration of 1 mM. The solution was aliquoted into 1.5 mL Eppendorf tubes and immediately frozen at -20 °C. Once aliquots were defrosted they were never frozen again. Cells were treated with 5-Aza-2'-deoxycytidine to obtain a final concentration of 5 μM in the media.

3.26.3. Trichostatin A

A 0.5 mM solution of Trichostatin A (T8552-1MG, Sigma) in DMSO was obtained from Colin Goding's lab and used at a final concentration of 200 nM for cell treatments.

3.27. Genomic Data Analysis (DNA Modification Enrichment)

3.27.1. Quality Control, Mapping and Duplicate Removal

The quality of sequencing data was assessed using FASTQC (v0.10.1) in a default mode. The short read files, which successfully past the QC, were aligned to the human genome (hg19) using bowtie2 (v 2.2.6) default parameters to obtain .bam files. Bam files for the same experimental conditions were then merged using `-cat` command in Samtools (v0.1.19) (<http://samtools.sourceforge.net/>). The reads in these files were then sorted according to genomic coordinates and indexed using Samtools default parameters. The duplicated reads, which are essentially PCR artefacts, were removed using the REMOVE_DUPLICATES function in Picard Tools (release 1.139).

3.27.2. Constructing Bigwig Files

Bigwig files were constructed for easy visualisation of the enrichment data in IGV and other genome browsers. The custom script which was run on R (release 3.3.2) using the following packages: `chipseq`, `rtracklayer`, `BSgenome.Hsapiens`, `UCSC.hg19`. The code for the script is provided in the Appendix.

3.27.3. Calculating DNA Modification Enrichment over Genic Regions

R package called CoverageView version 1.5.2, developed by Ernesto Lowy at EMBL-EBI (<https://www.bioconductor.org/packages/release/bioc/html/CoverageView.html>), was used to calculate the read coverage over genic regions. Briefly, genomic sequences of all protein coding genes (coordinates obtained from the UCSC Genome Browser) were binned into 100 equal size bins and the number of reads per bin was calculated. The number of reads in each bin was then normalised to the input and the values of all the bins across the gene averaged. It is important to point out that bam files for plus and minus doxycycline samples were subsampled to the same number of reads using Samtools prior to coverage analysis. Data output was a gene matrix with relative coverage values which were used in further analysis.

3.27.4. Read Saturation Analysis

The saturation analysis, which ensures that samples are not undersequenced, of the enrichment data was performed using MEDIPS package for R (Lienhard et al., 2014).

3.27.5. Hierarchical Cluster Analysis

Hierarchical clustering was performed based on gene body DNA modification enrichment values using the `hclust()` function in R.

3.27.6. Enrichment Analysis over Genomic Elements and Peak Calling

Enrichment of specific genomic elements was calculated using an overage command in Bedtools (version 2.21.0). The coordinates for genomic elements were extracted from the UCSC Genome Browser. Peak calling for enrichment data was performed using the MACS package (Feng et al., 2012) (version 1.4.2). The peak calling threshold was set at a significance level of $p < 0.05$, which is recommended for DNA modification enrichment data.

3.28. Differential Gene Expression Analysis

Reads obtained from the nascent RNA-seq experiment were aligned to the human genome (hg19) using Tophat2 version 2.0.12 (Trapnell et al., 2014). The read alignments were done using default parameters except that reads were aligned in a strand-specific manner. Next, Cuffdiff was used to quantify the reads over the genes annotated in a .gtf file available in the BS.genome package. Differentially expressed genes were called using the DEseq2 R package (Love et al., 2014). All the reads present over gene bodies (not only exonic reads) were used in the analysis because the nascent RNA enrichment step had been performed. Top 100 differentially expressed genes are shown in the Appendix.

Chapter 4

Accumulation of Genomic 5hmC and Gene Regulation in Cancer Cells

4.1. Introduction

Among the numerous covalent DNA modifications found in prokaryotes and eukaryotes, cytosine methylation (Ehrlich and Wang, 1981) and hydroxymethylation (Tahiliani et al., 2009a, Kriaucionis and Heintz, 2009) are the most abundant epigenetic marks in vertebrate genomes. A large body of evidence indicates that DNA modification signatures vary between different tissues and cell types (Leung et al., 2015). This complicates epigenetic research *in vivo* because the average modification maps of many cell types cannot easily be interpreted. To overcome this problem, several flow cytometry-based methods are available to purify specific cell populations (Mellen et al., 2012), but extensive sorting makes it hard to obtain a sufficient amount of material for biochemical studies. Cultured cancer cells are widely used as an experimental system. They are easy to culture and genetic manipulations are straightforward. Furthermore, cell populations tend to be rather homogeneous, which makes them an attractive system for dissecting basic epigenetic mechanisms. Interestingly, the vast majority of cultured cancer cells tend to lose 5hmC from their genomes, which does not necessarily occur as a consequence of TET or IDH family enzyme mutations (Ficz and Gribben, 2014). In addition, many cancer cells use promoter CpG island methylation of tumour suppressor genes such as CDKN2A as a binary mechanism to silence transcription (Merlo et al., 1995). Finally, gene bodies have been reported to undergo dynamic changes upon malignant transformation, indicating that these processes may be important in the course of cancer progression (Kulis et al., 2012). Several studies have aimed to elucidate the role of gene body modifications and have shown that expressed genes in cancer cells accumulate DNA methylation (Jjingo et al., 2012). However, they are descriptive and correlative as their observations are limited to steady state experiments. Only a few studies to date

have attempted to manipulate the DNA methylation landscape of cancer cells by treating them with demethylating agents (Yang et al., 2014, Lund et al., 2014). Interestingly, groups of genes that are upregulated as well as downregulated were reported but the mechanism of this regulation remains largely unknown. In addition, it was recently shown that DNA demethylating agents like 5-azacytidine kill cells via an intracellular viral response (Roulois et al., 2015, Chiappinelli et al., 2015). The role of 5hmC deposition over gene bodies is similarly understudied.

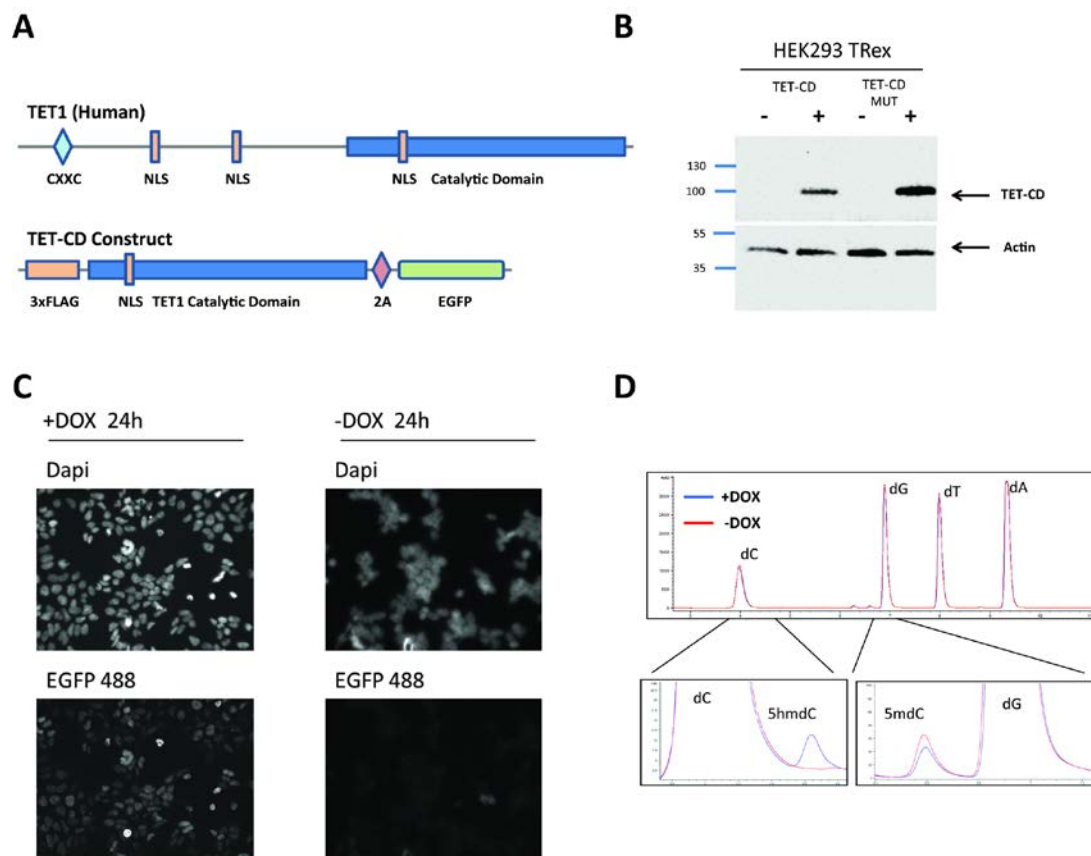
This chapter describes the design and characterisation of a system that allows the reproducible manipulation of DNA modification levels by overexpressing the TET1 catalytic domain and depositing 5hmC over the genome of HEK293 cells. This system not only allows us to gain insights into the function of DNA modifications but also investigate the role and dynamics of 5hmC deposition over the genome. Here, cancer cell lines were used because this enables a sufficient amount of material to be obtained for biochemical studies. In addition, this homogeneous system overcomes problems such as cell heterogeneity in animal tissues. Therefore, this novel experimental system can enhance our knowledge of the role of DNA modifications deposited over genic regions.

4.2. Results

4.2.1. Construction of Cell Lines Overexpressing TET-CD

In order to develop a robust system with which to investigate the role and dynamics of DNA modifications deposited over genic regions, 5hmC was generated over the genome using the TET1 catalytic domain (TET-CD). The catalytic domain was

specifically chosen to omit parts of the protein, such as the CXXC domain or other domains, that may mediate protein targeting. In addition I aimed to decouple certain regulatory functions, such as interaction with Sin3A (Vella et al., 2013), from the catalytic function of the enzyme. This allowed better insight into the biological role of 5hmC by reducing the possibility of indirect effects caused by overexpression of the enzyme. TET-CD is located at the C-terminus of the human TET1 protein and spans amino acids 1418-2136 (Figure 4.1A) (Tahiliani et al., 2009a). To induce the TET1 catalytic domain, a well-characterised tetracycline-controlled transcriptional activator system was chosen (Gossen and Bujard, 1992). In this system, transgene expression is induced by adding tetracycline or doxycycline (DOX) into the tissue



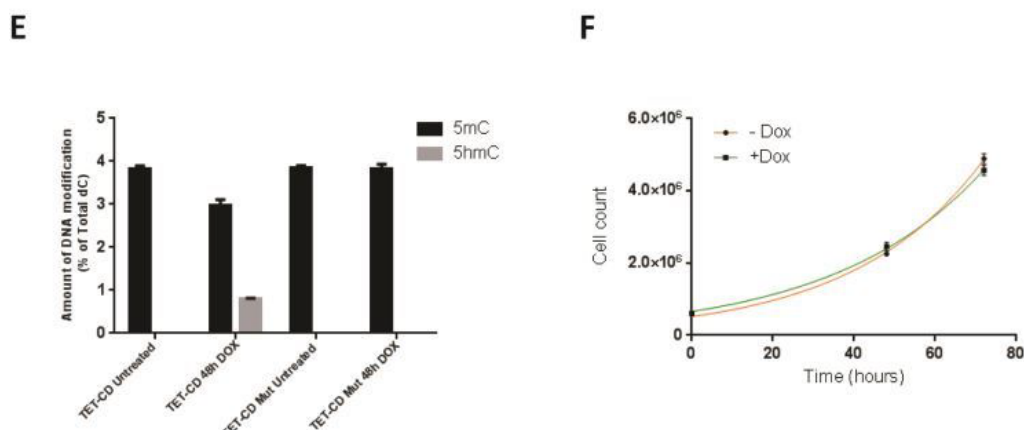


Figure 4.1 Characterisation of HEK293 cells overexpressing the TET1 catalytic domain

A) Schematic representation of full-length human TET1 and the TET catalytic domain construct integrated into cell lines. B) TET-CD is induced 24 h after DOX treatment, as depicted in the western blot. β -Actin was used as a loading control. C) Microscopy image showing homogeneous expression of TET1 catalytic domain 24 h after induction. D) HPLC traces showing accumulation of 5hmC in the genome with a corresponding decrease in 5mC. E) Quantification of modified cytosines in the genome. 5hmC accumulates to around 25% of 5mC levels 48 h after induction of the TET catalytic domain. The TET catalytic mutant does not accumulate any 5hmC. F) Cell growth curve after induction of TET-CD. Cells do not show any significant change in doubling time. All error bars represent the standard error of the mean, $n = 3$.

culture medium. In order to detect and monitor the overexpressed construct it was tagged with a 3xFLAG tag on the N-terminus and EGFP on the C-terminus (Figure 4.1A). To avoid EGFP interfering with its catalytic activity, a 2A cleavage site was introduced between TET-CD and EGFP. The TET-CD catalytic mutant was constructed by introducing H1671D and Y1673A mutations into the TET1 catalytic domain. These residues are predicted to bind cofactors and these particular mutations have been demonstrated to abolish enzymatic activity (Tahiliani et al., 2009a).

In order to establish a robust and highly regulated system, TRex Flip-In cell lines were used that already have a single locus transgene integration site and an inbuilt

tetracycline transactivator. Two cell lines were established: HEK293 and HeLa. Both of them showed robust expression of the TET1 catalytic domain upon DOX treatment (Figure 4.1B, S1). Single clones specifically selected and expanded in order to gain homogeneous populations. All of the cells in the population were able to express TET-CD upon induction and only around 1–2% of cells showed leakiness in the absence of DOX (Figure 4.1C). We further investigated whether the cells successfully deposit 5hmC over the genome. After 48 h treatment with DOX, a clear 5hmC peak could be detected by HPLC analysis, which was not present in untreated samples (Figure 4.1D). Subsequent quantitative analysis indicated that 22% of the 5mC were converted to 5hmC during 48 h treatment with DOX (Figure 4.1E). The equivalent reduction of 5mC indicates that 5hmC is a direct product of 5mC. As expected, catalytically dead constructs were unable to accumulate 5hmC and maintained initial levels of DNA methylation (Figure 4.1E). Finally, we investigated the effect of 5hmC induction on cell viability and cell cycle progression. As shown in Figure 4.1F, cell cycling and viability were not affected by the induction of TET-CD and deposition of 5hmC over the genome. Interestingly, HeLa cells failed to deposit 5hmC over the genome despite the successful integration of the construct (Appendix). TET-CD expression was reproducible as observed in the immunoblot analysis. Reasons for the absence of 5hmC deposition are not clear.

4.2.2. Dynamic Changes of DNA Modifications after Overexpression of TET Catalytic Domain

The ability to reproducibly deposit 5hmC over the genome of HEK293 cells enabled interrogation of the dynamic changes of the epigenetic marks over time.

Firstly, the dynamics of DNA modifications during the long-term induction of TET1 catalytic domain were analysed. Interestingly, 5hmC reached equilibrium around 48 h after induction, while 5mC decreased until it reached a steady state around 96 h after induction (Figure 4.2A). In addition to an accumulation of 5hmC, there was a significant decrease in the amount of modified cytosine, i.e. some of the 5mC marks were converted to 5hmC and a small fraction (around 5%) demethylated. Furthermore, DNA modifications in the cells grown with DOX reached equilibrium after around 100 h and this state was further maintained. As expected for the catalytically inactive mutant, 5mC levels remained stable throughout the course of DOX treatment (Figure 4.2A).

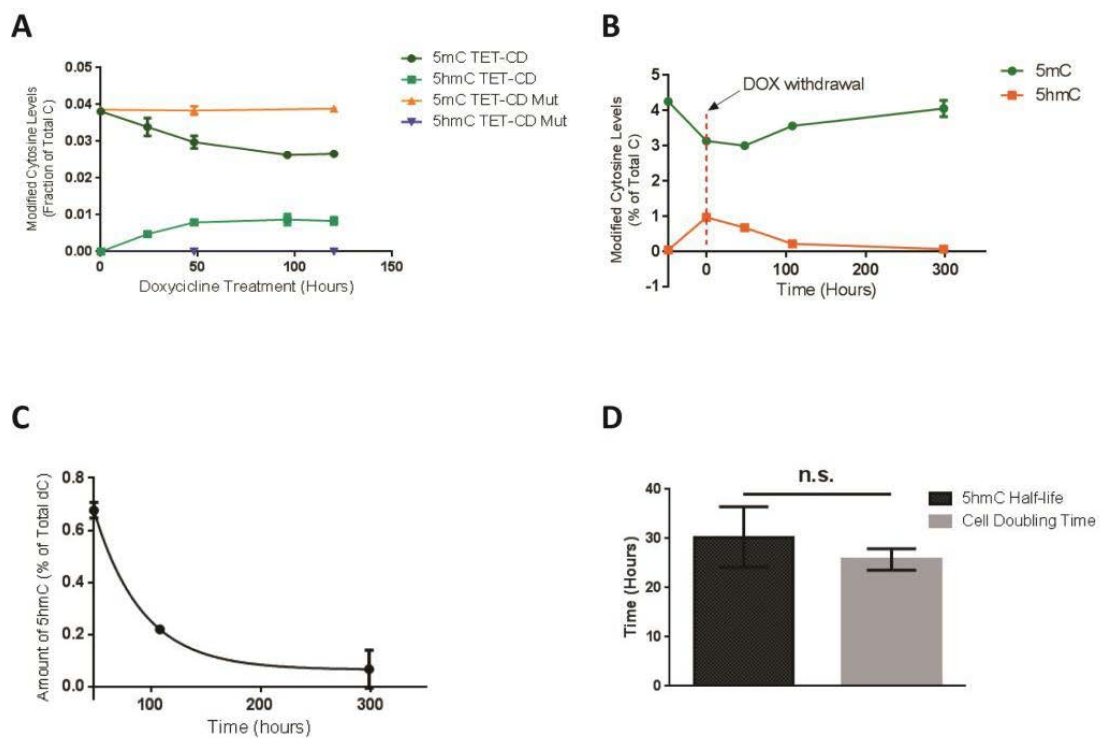


Figure 4.2 Dynamics of DNA modifications upon overexpression of the TET1 catalytic domain

A) Changes in DNA modifications upon long-term overexpression of the TET1 catalytic domain or the catalytically inactive mutant. B) Levels of 5mC and 5hmC upon 48 h induction of TET-CD followed by withdrawal of DOX. C) Exponential

decrease of 5hmC levels after removal of DOX. D) Comparison of 5hmC decay with cell doubling rate after removal of TET-CD. All error bars represent the standard error of the mean, $n = 3$.

We further investigated whether cells could re-establish 5-mC levels upon withdrawal of DOX. The TET1 catalytic domain was induced for 48 h after which DOX was withdrawn. Soon after withdrawal, 5hmC levels started to diminish (Figure 4.2B) and seemed to undergo substantial decay starting 48 h after withdrawal. 5mC, on the other hand, continued to decrease for at least 48 h after withdrawal of DOX. This is likely to be caused by the presence of the remaining TET1 catalytic domain. 96 h after removal of DOX, methylation levels started to increase and reached the initial levels of the system after 10 days (Figure 4.2B). Taken together, these data indicate that global DNA methylation levels are predetermined, at least in this cell type, and are robustly restored after perturbations.

To further investigate whether 5hmC is lost by passive dilution or is actively removed during cell division, the decay of 5hmC after withdrawal of DOX (Figure 4.2C) was modelled and compared to the exponential growth model of the cells. Superimposition of the models and comparison of the 5hmC half-life with the cell doubling rate (Figure 4.2D) gave no significant difference between two measures. This strongly suggests that, in this system, 5hmC is passively diluted during cell division and little active demethylation occurs.

Several studies have previously reported that supplementation of the growth media with ascorbate (vitamin C) can induce TET enzyme activity and increase the amount of genomic 5hmC (Blaschke et al., 2013). In addition, 5hmC levels are also modulated by the availability of other cofactors such as α -ketoglutarate (Carey et al., 2015). To explore how ascorbate affects DNA modifications in HEK293 cells and

whether it can further enhance the deposition of 5hmC, cells were treated with 10 μ M ascorbate for 48 h (Figure 4.3A, B). As shown in the bar graph, the amount of 5hmC doubled in HEK293 cells upon ascorbate treatment. This observation is in agreement with a previous report that ascorbate can enhance the activity of endogenous TET enzymes (Blaschke et al., 2013). However, upon overexpression of TET-CD plus ascorbate treatment (Figure 4.3B), there was only a marginal increase in 5hmC levels. This suggests that either ascorbate does not have much effect on the enhancing

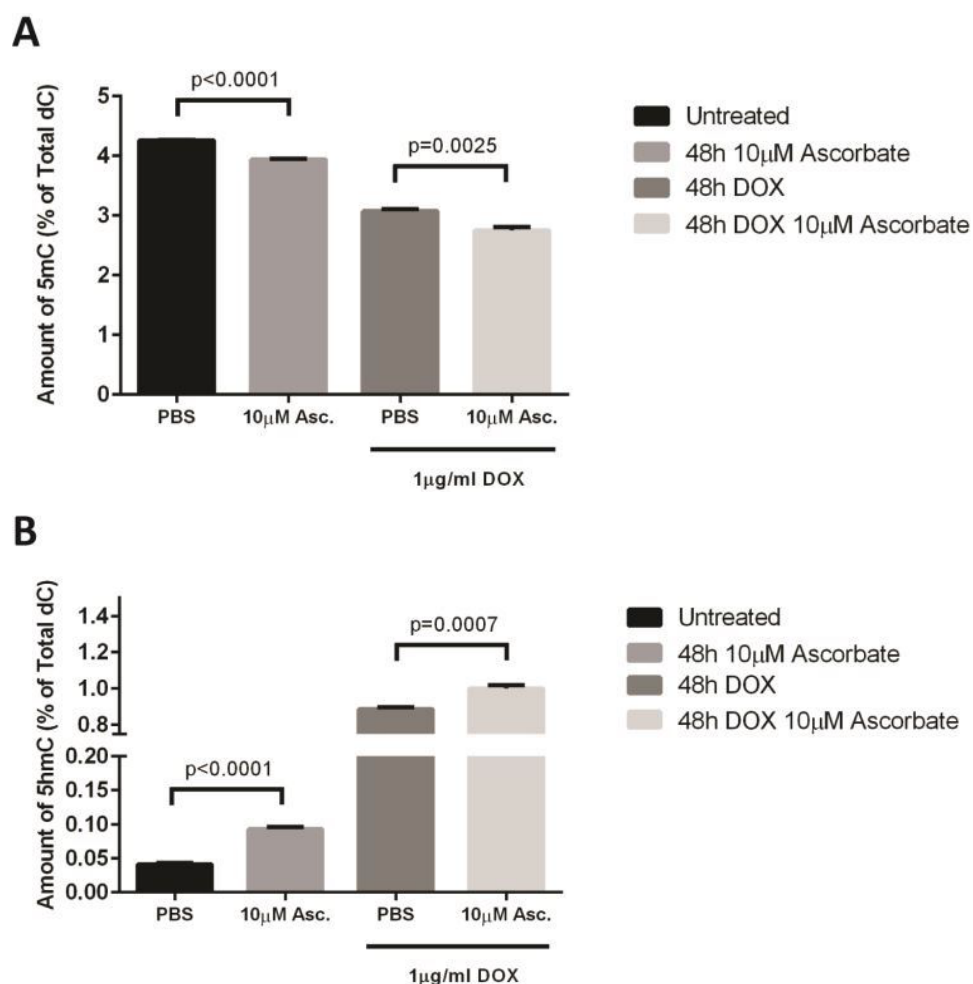


Figure 4.3 Changes in DNA methylation and hydroxymethylation upon ascorbate treatment

A) DNA methylation changes upon treatment with DOX, ascorbate or a combination

of the two. B) 5hmC increases significantly upon treatment with ascorbate in HEK293 TET-CD overexpressing cell lines. All error bars represent the standard error of the mean, n = 3.

activity of the TET catalytic domain or that 5hmC is saturated in all the accessible regions and this is the maximum capacity of this modification.

4.2.3. Genome-wide Mapping of 5hmC in TRex HEK239 Cells

In order to learn more about the function of 5hmC over the genome, the genomic location of the mark was mapped using two different methods. A well-established 5hmC enrichment method called hmSeal (Song et al., 2011a) was chosen in addition to EpiJET, a new technique developed by Thermo Fisher Scientific (Kaniusaite et al., 2015). The hmSeal method is based on treating DNA with β -glucosyltransferase enzyme, which specifically transfers a modified glucose moiety to the hydroxyl group of 5hmC (Figure 4.4A). The modified glucose has an azide linker which permits the use of click chemistry and the attachment of biotin to glycosylated 5hmC. Then 5hmC-containing DNA fragments are enriched using streptavidin beads and subsequently sequenced using massively parallel DNA sequencing. The method utilised by the EpiJET kit is based on the modification of 5hmC with bacterial DNA methyltransferase followed by the subsequent attachment of biotin, streptavidin pull-down and deep sequencing (Figure 4.4C).

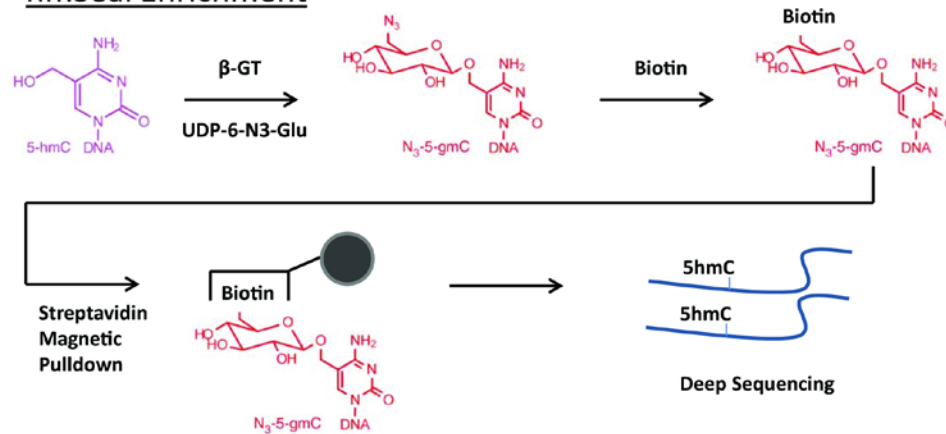
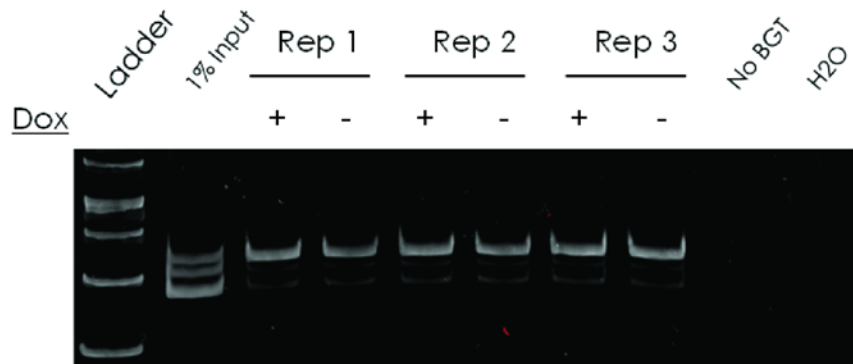
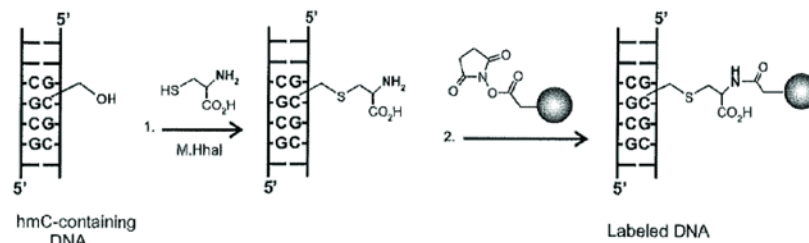
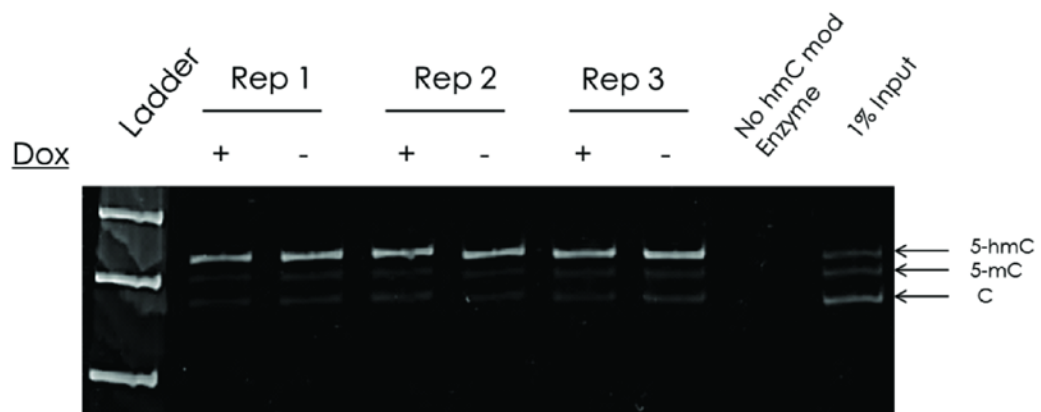
A hmSeal Enrichment**B****C** EpiJet Enrichment**D**

Figure 4.4 Genome-wide mapping of 5hmC

A) Schematic representation of the hmSeal method for mapping 5hmC. B) Acrylamide gel with modified probes spiked in the DNA representing efficient and reproducible pull-down of a 5hmC-containing probe in three biological replicas. C) Schematic representation of the EpiJET 5hmC enrichment procedure. D) Acrylamide gel representing efficient pull-down of the 5hmC probe in three biological replicas.

A quality control experiment was performed to assess the enrichment methods before committing to deep sequencing. Initially, genomic DNA was spiked with three differently modified oligonucleotides containing unmodified cytosine, methylated cytosine, and hydroxymethylated cytosine. PCR amplification of these oligonucleotides was performed after DNA enrichment and it was observed that 5hmC-containing oligonucleotides were consistently enriched in all samples (Figure 4.4B, D).

For the sequencing experiment, a 48 h induction time point was chosen because it was sufficient to allow the accumulation of a large amount of 5hmC with demethylation of only 5–7% of modified cytosines. In addition, it was reasoned that shorter induction would reduce the probability of indirect effects on the epigenomic landscape and transcription. 5hmC landscapes between induced and untreated cells were compared using the two different methods described above. Two biological replicas were sequenced per condition following the quality control assay. Each replica was sequenced to around 60 million reads (Table 4.1). Sequencing saturation analysis indicated capture of more than 80% of the unique DNA fragments present in each sample. This meant that deeper sequencing would not provide a substantial amount of new biological information. Interestingly, comparing unique reads for each replica, it was noticed that duplicate levels in untreated cells were much higher compared to cells treated with DOX (Table 4.1). This was expected because the

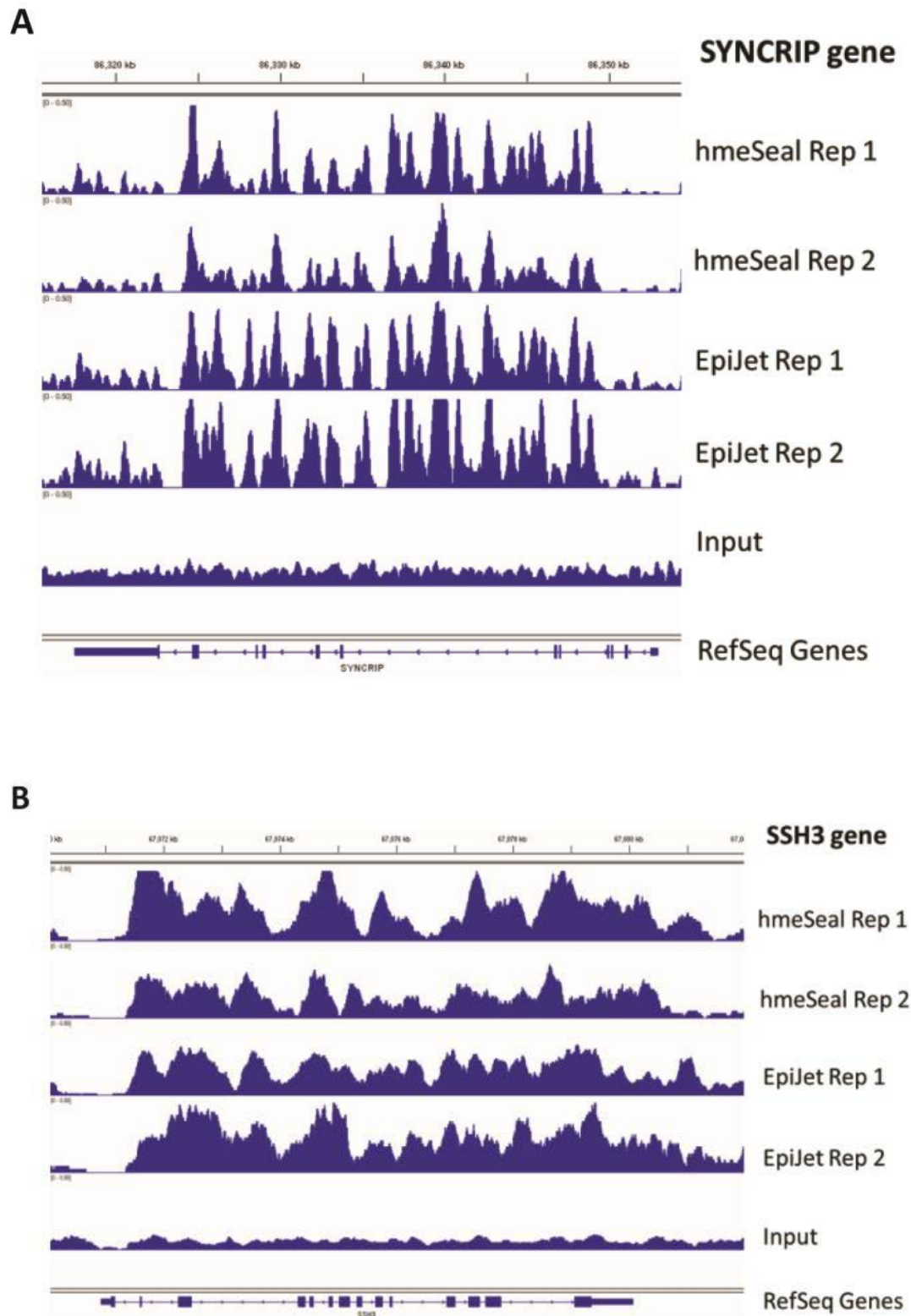
amount of 5hmC in untreated HEK293 cells is very low, leading to very few DNA fragments being pulled down in the reaction and these fragments being proportionally amplified in the PCR reaction. This results in sequencing of high numbers of PCR generated copies of the same fragment. To further analyse the data,

Sample	Total Aligned Reads	Total Unique Reads	Saturation Level
hmSeal1 +DOX	69470051	39739183	>80%
hmSeal1 -DOX	62106170	15982012	>80%
hmSeal2 +DOX	57263308	24938530	>80%
hmSeal2 -DOX	30175925	5915759	>80%
EpiJET1 +DOX	87379324	68074549	>80%
EpiJET1 -DOX	60127762	36337521	>80%
EpiJET2 +DOX	43443299	40603798	>80%
EpiJET2 -DOX	75936585	36644172	>80%

Table 4.1 5hmC enrichment deep sequencing metrics

two different enrichment methods for genome wide mapping of 5hmC were compared. The read data (Table 4.1) indicated that samples from both methods were sequenced to a similar depth, however, the number of unique reads obtained by the EpiJET method exceeded that obtained by hmSeal. As mentioned earlier, the number of unique reads is proportional to the number of DNA fragments obtained after pull-down. This indicates that the hmSeal technique is more stringent compared to EpiJET or, alternatively, that the hmSeal chemistry may be less efficient compared to EpiJET. This data is in agreement with the observations in the spiked-in oligonucleotide experiment. To further compare the two methods, 5hmC enrichment tracks of a few randomly selected genes were compared. The tracks were very highly reproducible between two biological replicas and the two different enrichment methods (Figure

4.5A). This suggests that 5hmC deposition is not entirely random. Genome-wide



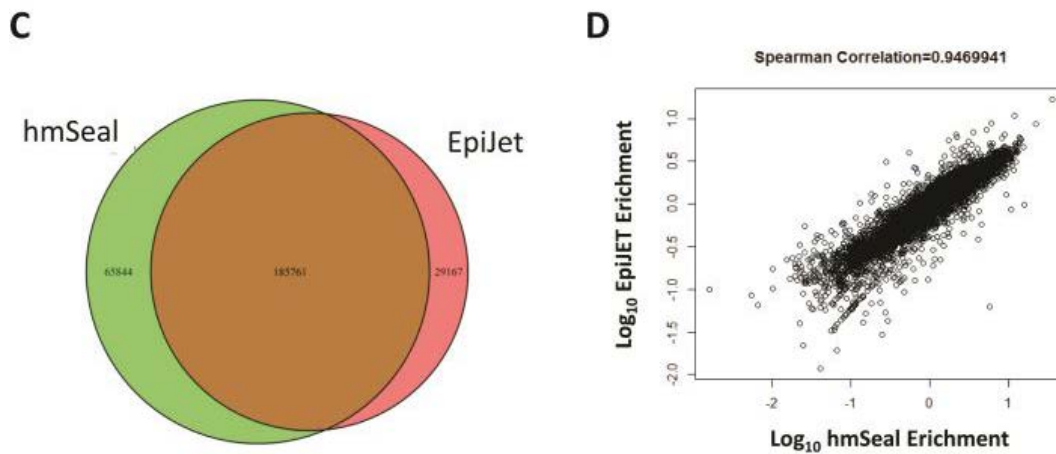
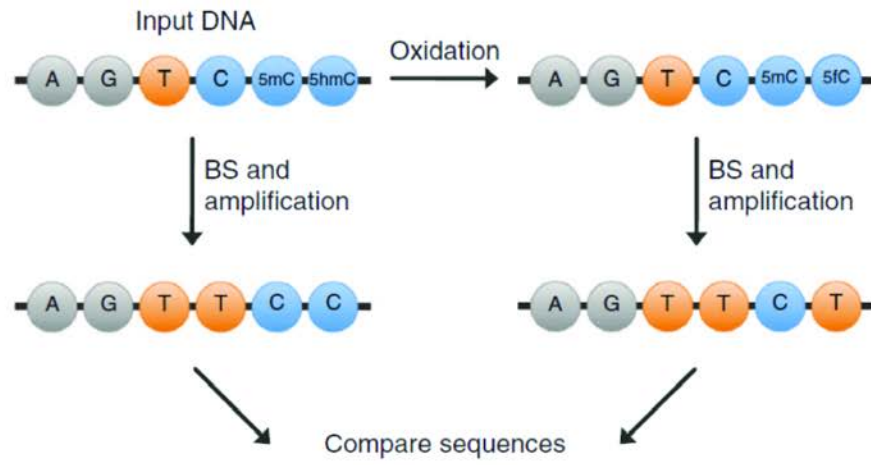
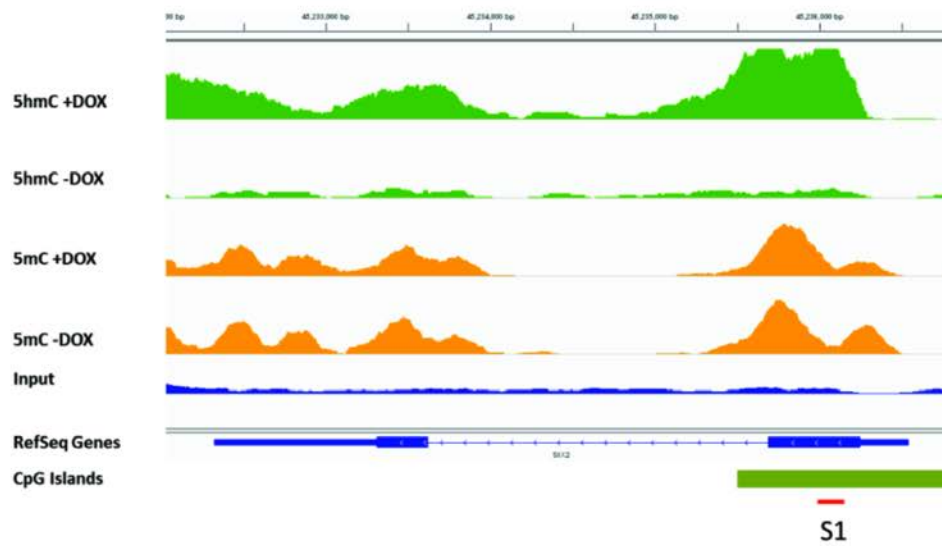
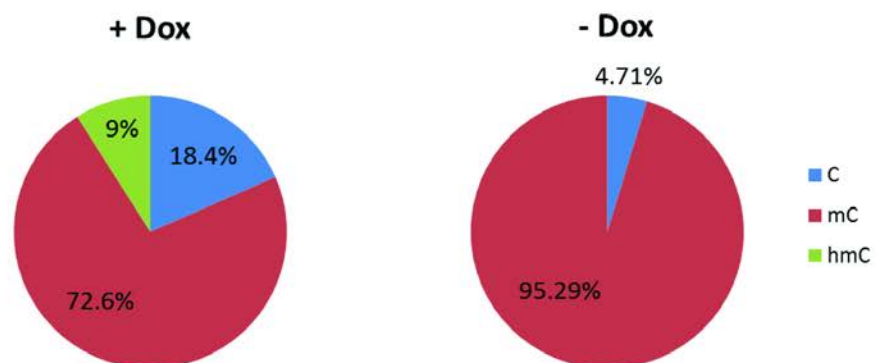


Figure 4.5 Comparison of hmSeal and EpiJET sequencing methods

A,B) Integrated genome viewer (IGV) tracks for SYNCRIP and SSH3 genes, respectively. Enrichment tracks are very similar between the two different methods and biological replicates. C) Numbers of overlapping MACS peaks between the two methods. D) Spearman's rank correlation of the amount of 5hmC per gene body as calculated from different enrichment methods.

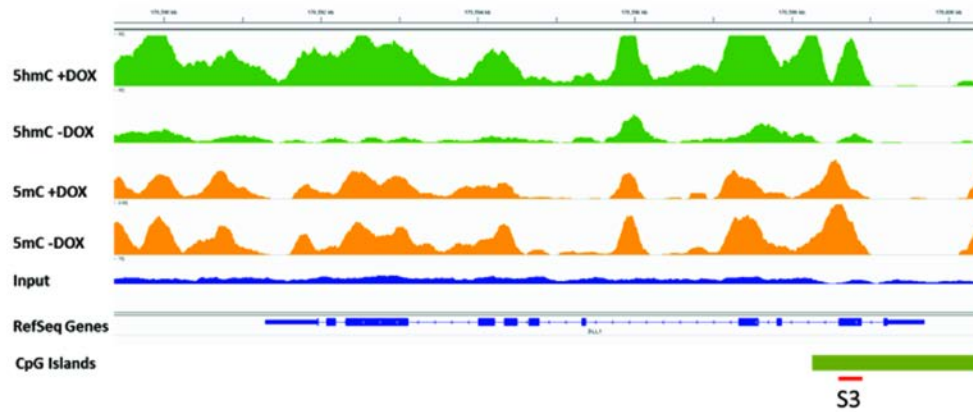
5hmC enrichment was further analysed in a more quantitative manner to better compare the biological replicates and different enrichment methods. To achieve this, 5hmC enrichment using both methods was compared over genic regions. As expected, high correlations were obtained, indicating that both methods yielded highly reproducible results (Figure 4.5D). MACS Peak Caller (Feng et al., 2012) is commonly used software to identify the significantly enriched regions compared to the input samples. Following the peak calling, global enrichment landscapes can be compared between samples. In order to compare enrichment profiles in the 5hmC enrichment data, MACS peaks were identified genome-wide, and the number of these peaks that overlapped was determined. As expected, the majority (86%) of the peaks overlapped (Figure 4.5C) confirming that the data generated by both methods were highly reproducible.

In order to further validate the results, the presence of 5hmC at several genomic regions was confirmed using Oxidative Bisulfite Sequencing (oxBS) (Booth et al., 2012). This technique can determine the presence of 5hmC at single nucleotide resolution, as well as giving the absolute quantity of DNA modifications per base. The chemistry of this method is based on the oxidation of 5hmC to 5fC. When subjected to bisulfite treatment, 5fC deaminates to 5-formyluracil and is converted to thymidine in a PCR reaction (Figure 4.6A). By complimenting standard bisulfite sequencing with oxBS and subtracting the data, 5hmC can be accurately quantified. Using oxBS, it was shown that 5hmC increases over the genic regions of the SIX2 and DLL1 genes upon treatment with DOX (Figure 4.6). Interestingly, the increase in 5hmC was closely linked to an increase in demethylated cytosines. In the first exon of the SIX2 gene, for example, no 5hmC could be detected in untreated cells. However, 5hmC levels increased to 9% in this particular region upon expression of the TET1 catalytic domain. Interestingly, the number of demethylated cytosines increased from 4.73% to 18.4% (Figure 4.6C). The regions that were sequenced using the oxBS protocol are in close proximity to promoters (Figure 4.6B,D). There is a lot of epigenetic heterogeneity in these regions, as can be seen from the single nucleotide resolution data. Previous studies have named these regions CpG island shores and demonstrated that they exhibit a substantial amount of variation in both cancer and normal cells (Irizarry et al., 2009).

A**B****SIX2****C****SIX2**

D

DLL1



E

DLL1

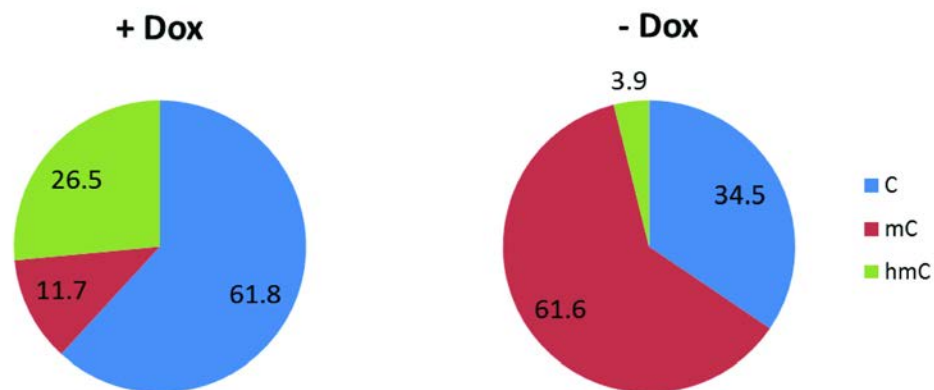


Figure 4.6 Confirmation of the presence of 5hmC in enriched regions by oxidative bisulfite (oxBS) sequencing

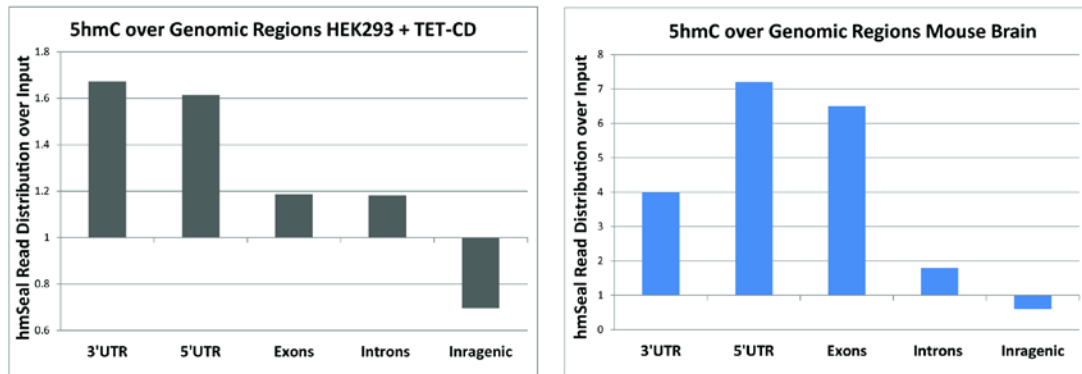
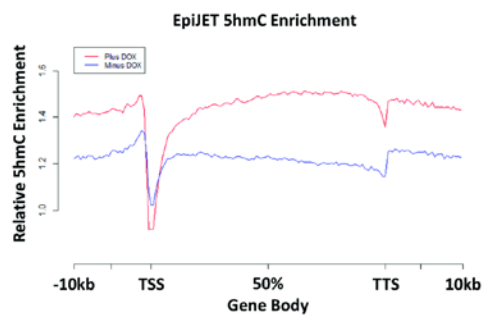
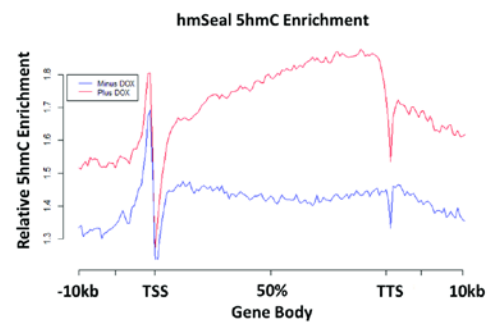
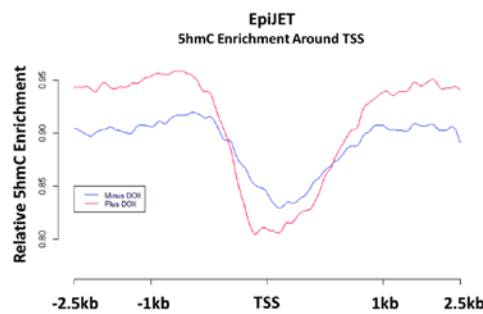
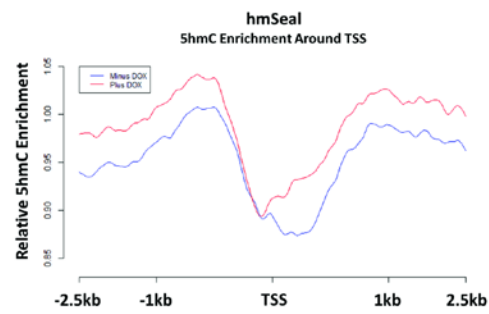
A) Schematic representation of the chemistry of the oxBS method. B) DNA methylation and hydroxymethylation profile for the *SIX2* gene. The region sequenced by the oxBS method is marked in red. C) Pie charts indicating the amount of cytosine modifications in the sequenced region. D) DNA modification profiles of the *DLL1* gene. E) Quantification of DNA modifications on the *DLL1* gene.

4.2.4. Genic Distribution of 5hmC Upon TET1 Catalytic Domain

Overexpression

Numerous papers have reported dynamic changes in gene body DNA modifications in development and pronounced differences among different cell types (Lister et al., 2013, Mellen et al., 2012). Therefore, I mainly focussed on the function of 5hmC deposited over gene bodies. Specifically, whether the expression of TET-CD could hydroxylate 5mC deposited over various genic regions and induce transcriptional changes was of interest. Firstly, 5hmC enrichment over gene bodies in induced and untreated cells was compared (Figure 4.7 B,C). High enrichment of 5hmC was observed over gene bodies regardless of the enrichment method used. In addition, a sharp dip was seen at transcription start sites (TSSs). This was expected due to the fact that the majority of CpG island promoters are unmethylated. Interestingly, a sharp peak upstream of the TSSs was noted, which has also been reported by previous studies (Mellen et al., 2012). To gain more insight into the deposition of 5hmC, enrichment of the mark in different genic elements was compared between TET catalytic domain overexpressing HEK293 cells and mouse brain (Figure 4.7A). Interestingly, strong 5hmC enrichment could be seen over transcribed elements in both systems with a higher enrichment over 3'- and 5'-untranslated regions or UTRs. Importantly, less 5hmC has been observed in intragenic regions in both TET-CD overexpressing HEK293 cells and mouse brain (Song et al., 2011a, Szulwach et al., 2011). These data indicate that, by overexpressing the TET1 catalytic domain, HEK293 cells can deposit 5hmC in a similar pattern as has been observed in mouse neuronal tissues.

Promoter demethylation has been thoroughly studied and linked to gene activation (Merlo et al., 1995). To investigate whether there was a significant level of

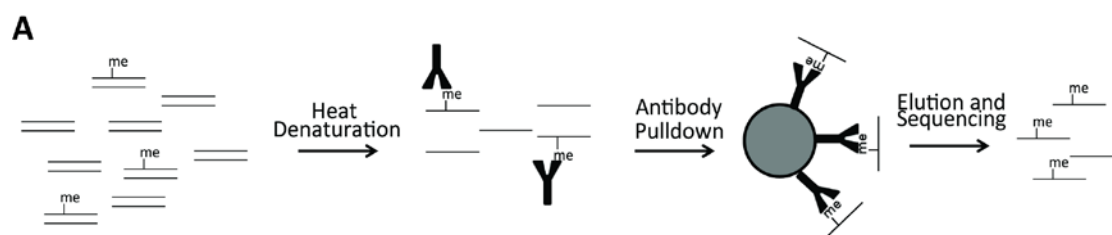
A**B****C****D****E****Figure 4.7 5hmC deposition over genomic regions**

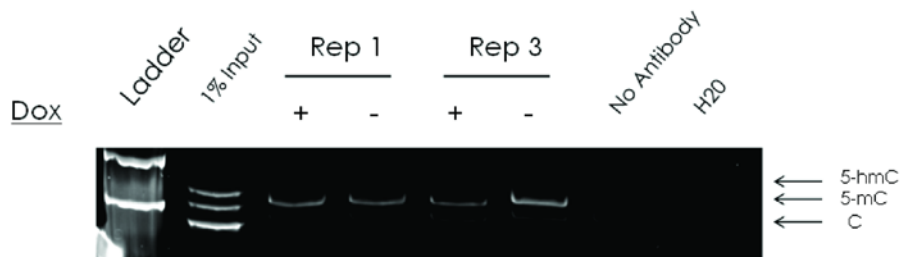
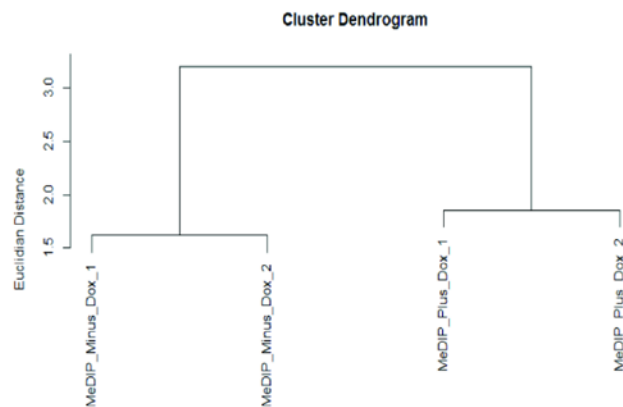
A) Bar chart representing enrichment of 5hmC over genomic features normalised to the input. Data is shown for HEK293 cells overexpressing the TET1 catalytic domain and mouse brain data obtained from (Szulwach et al., 2011). B,C) 5hmC enrichment profiles over gene bodies for the EpiJET and hmSeal methods, respectively. D,E) Cytosine hydroxymethylation profiles for TSS +/- 2.5 kb for the EpiJET and hmSeal methods, respectively. Relative enrichment values were calculated by dividing enrichment reads by the input.

5hmC deposition over promoter regions, 5hmC levels on promoters were compared between induced and untreated cell lines. Interestingly, no significant 5hmC enrichment was observed at TSSs (Figure 4.7D,E), although 5hmC enrichment in the CpG island shores and sites proximal to the promoter could be shown by both pull-down methods (Figure 4.7D,E). Taken together, these results indicate that most of the TSSs are demethylated, do not acquire many marks and that unmethylated promoters are unlikely to be affected by overexpression of the TET1 catalytic domain. This permitted the investigation of 5hmC function over genic regions as opposed to over promoters.

4.2.5. Genome-Wide Mapping of DNA Methylation

To fully characterise DNA modifications in the engineered cell line, DNA methylation was mapped genome-wide using the Methylated DNA Immunoprecipitation (MeDIP) technique (Weber et al., 2005). MeDIP is an enrichment method based on the antibody pull-down of methylated DNA sequences (Figure 4.8A). The enriched genomic DNA fragments are then subjected to deep DNA sequencing. Prior to submitting DNA samples for sequencing, enrichment efficiency was validated in a similar way to in 5hmC enrichment methods (Figure 4.8B).



B**MeDIP 5-mC enrichment****C****Figure 4.8 Mapping of DNA methylation in HEK293 cells genome-wide**

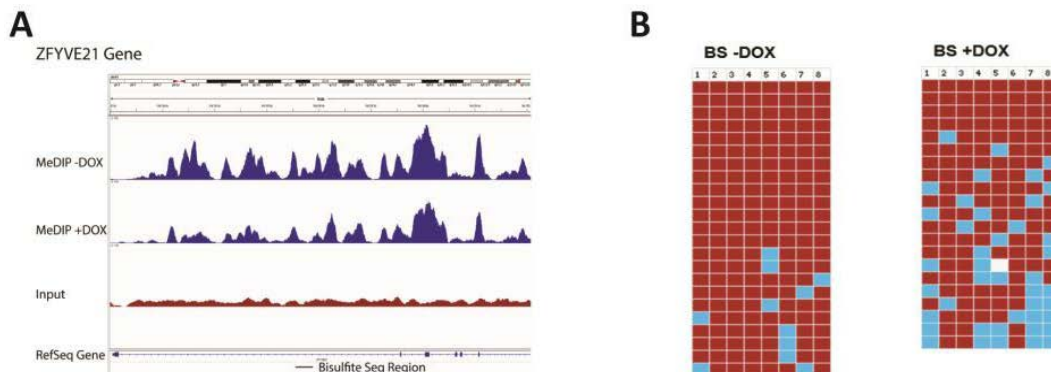
A) Schematic representation of the MeDIP enrichment method. B) An acrylamide gel representing the specific enrichment of a methylated probe spiked in genomic DNA. C) Hierarchical clustering dendrogram for DNA methylation samples.

Sample	Total Aligned Reads	Total Unique Reads	Saturation Level
MeDIP1 -DOX	78860664	50480592	>80%
MeDIP1 +DOX	68708180	46005754	>80%
MeDIP2 -DOX	79432967	58368710	>80%

MeDIP2 +DOX	81253872	52079021	>80%
-------------	----------	----------	------

Table 4.2 5mC enrichment deep sequencing metrics

Two biological replicates of HEK293 cells, either untreated or treated with DOX for 48 h, were sequenced to the average depth of 50 million unique reads (sequencing metrics are presented in Table 4.2). In order to check induction efficiency, samples were clustered according to the amount of DNA methylation deposited over gene bodies. As shown in Figure 4.8C, TET-CD-expressing samples and untreated samples clustered on the different branches of the dendrogram with little variation between biological replicates. The amount of DNA methylation in treated cells has been reduced by around 25% upon DOX treatment (HPLC data not shown). This reduction was also apparent over the body of the ZFYVE21 gene (Figure 4.9A). The reduction of DNA methylation was been further confirmed by bisulfite sequencing of the genic region of the ZFYVE21 gene (Figure 4.9B) and the EEF1A1 gene, which was confirmed by both methods to have no methylation over the gene body (Figure 4.9C,D). As reported previously, the highest expressed genes tend to lose DNA methylation over their gene bodies (Jjingo et al., 2012). This phenomenon was observed over the gene



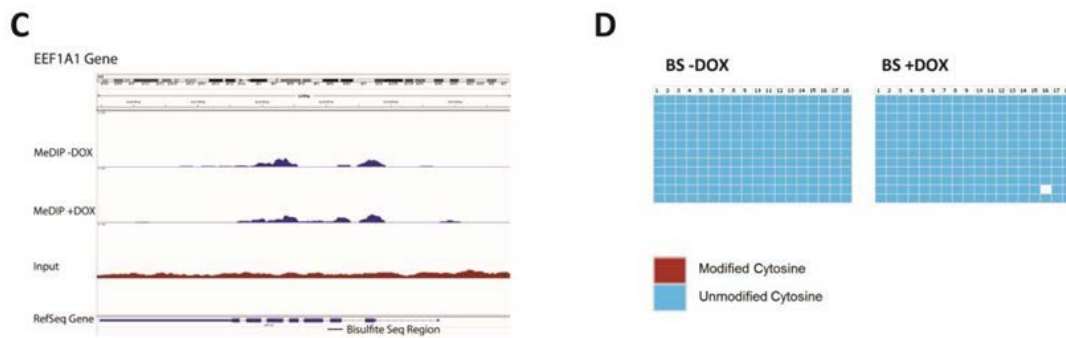


Figure 4.9 Mapping genome-wide DNA methylation in HEK293 cells.

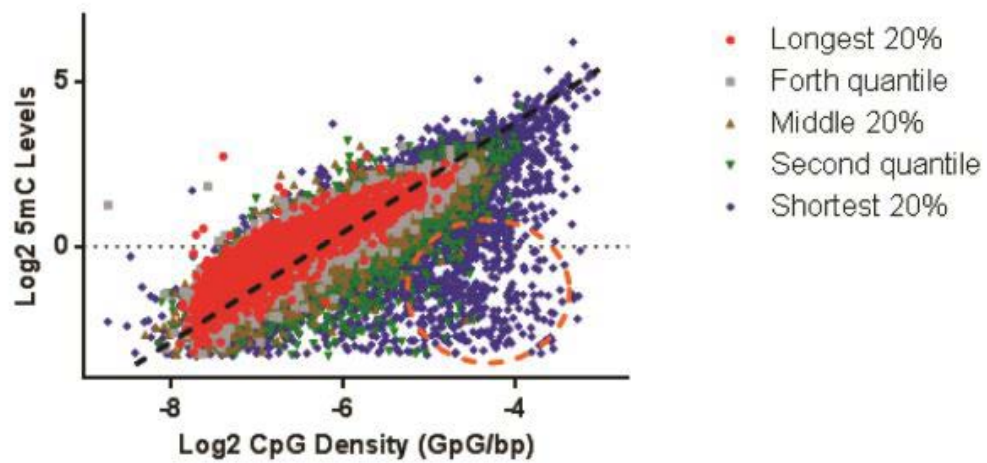
A) DNA methylation tracks for the ZFYVE21 gene and bisulfite sequencing of the region marked in black. B) Single base resolution DNA methylation analysis of the region in the body of the ZFYVE21 gene. C) IGV tracks for DNA methylation of the EEF1A1 gene. The region used in bisulfite sequencing is underlined in black. D) Cytosine modification profile for the EEF1A1 gene. Red squares = modified cytosine, blue squares = unmodified cytosine.

body of the EEF1A1 gene, which is very highly expressed in HEK293 cells. Taken together, these data indicate that MeDIP enrichment was reproducible and of high quality, enabling further investigation of the role of DNA modifications over the genome.

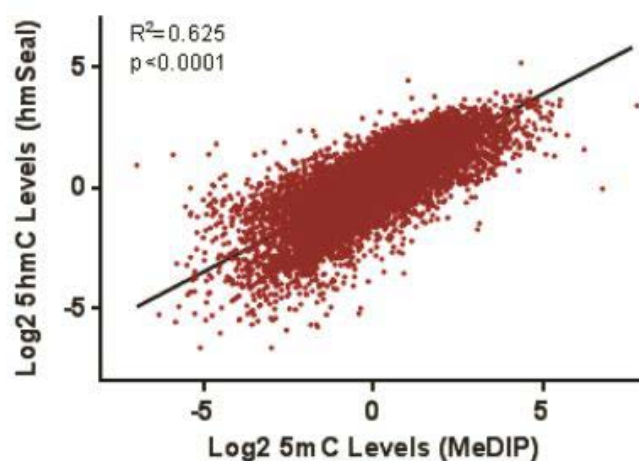
To investigate gene body DNA methylation, 5mC density values were correlated with the density of genic CpG dinucleotides, which are the primary target of DNA methylation (Figure 4.10A). As can be seen from Figure 4.10, DNA methylation density correlated well with CpG density. Interestingly, a group of outliers were observed on the bottom right corner of the graph (highlighted by an orange dashed circle). Binning genes according to gene length indicated that the bodies of the shortest genes with high CpG densities exhibited the highest variation of DNA methylation over genic regions (Figure 4.10A). These genes are likely to contain promoter CpG islands extending into the gene body. Whether highly methylated

regions were those where the most 5mC was converted to 5hmC upon the induction of TET-CD was of interest. Figure 4.10B shows the correlation between 5hmC and 5mC levels over genic regions genome-wide. Interestingly, the higher the methylation over a gene body, the more 5hmC is gained by the gene. This indicates that the level of 5hmC deposited over gene bodies depends on the amount of substrate present and

A



B



C

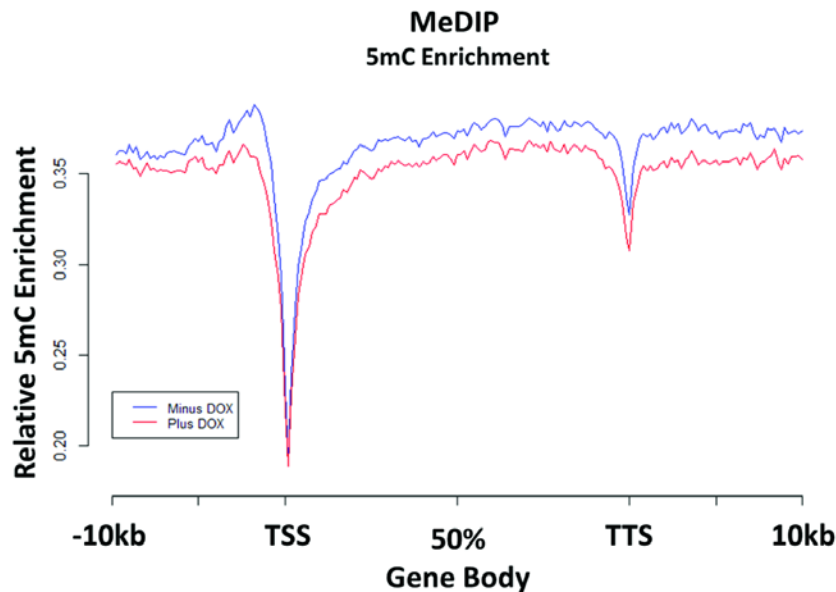


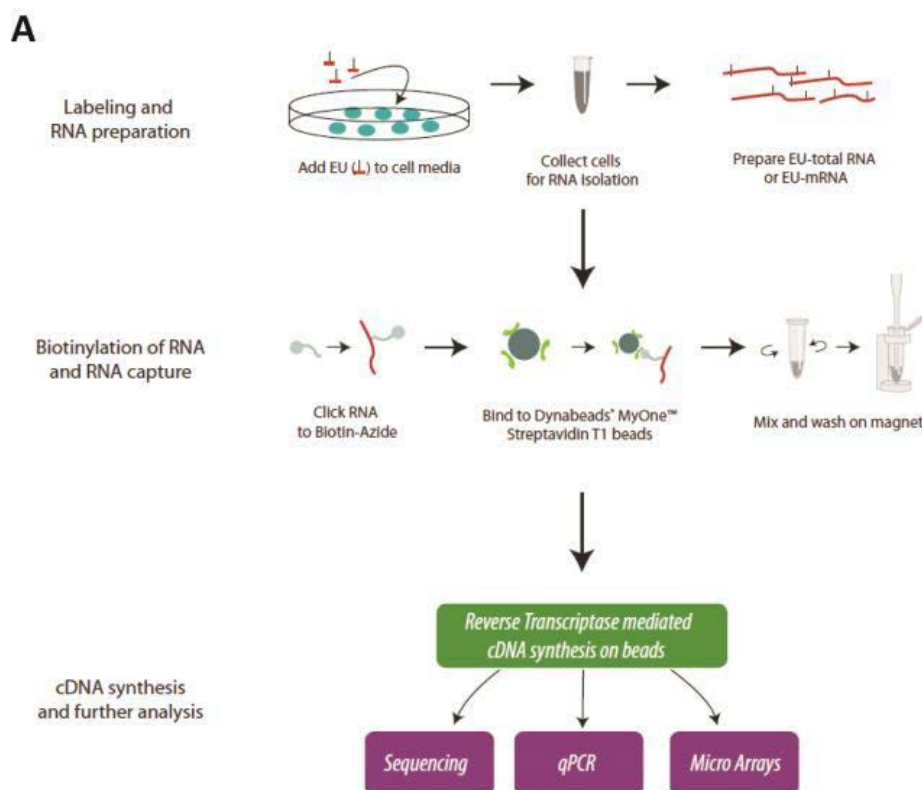
Figure 4.10 Genomic context of DNA methylation

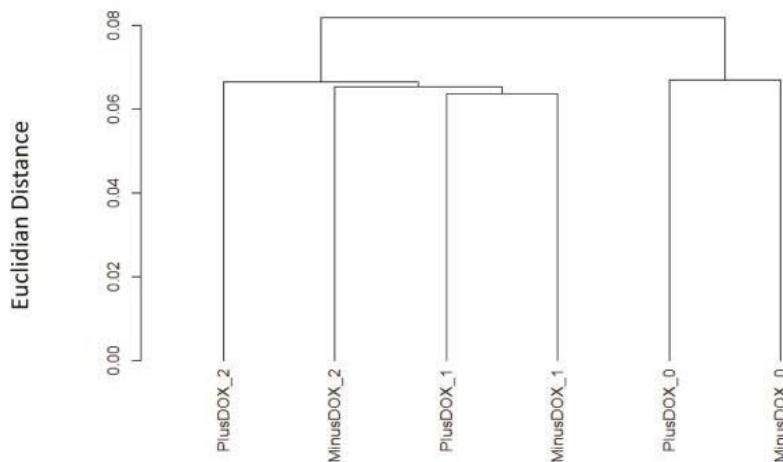
A) DNA methylation correlation with CpG density over genic regions. The group of outlier gene bodies is marked by an orange dashed circle. B) 5hmC levels correlate with 5mC levels. C) Gene body DNA methylation profile. Blue line represents DNA methylation enrichment in untreated cells, red line represents 5mC levels in cells treated with DOX.

that there was no specific targeting of TET-CD to specific genic regions. Finally, 5mC cumulative enrichment over gene bodies was calculated (Figure 4.10C). The plot shown indicates that there was depletion of DNA methylation over genes upon induction of the TET-CD the TET catalytic domain. A sharp dip around TSSs indicates that there is a large fraction of unmethylated CpG island promoters in HEK293 cells. Moreover, a peak of methylation was observed just upstream of TSSs which was largely reduced upon DOX treatment. In addition, the peak of 5hmC appeared at the exactly same position when TET-CD was expressed (Figure 4.7C). This indicates that this upstream region undergoes a high degree of change in DNA modifications, suggesting that it could be functionally important.

4.2.6. Differential Gene Expression Analysis in a TET-CD Overexpressing Cell Line

In order to investigate how the deposition of 5hmC affects gene expression in HEK293 cells, differential gene expression analysis was performed. Due to the short 48 h induction time, Nascent RNA-seq was undertaken to look at active transcription rather than explore transcriptome profiles, which could be done by conventional RNA-seq experiments. A Nascent RNA-seq protocol based on the spiking of a small amount of ethenouridine into the media, which is efficiently taken up by the cells and incorporated into the RNA, was used. Following RNA purification, biotin was attached to ethenouridine using click chemistry (Figure 4.11A). Nascent RNA was then enriched using streptavidin beads, reverse transcribed and subjected to deep sequencing. The samples were sequenced to a depth of around 40 million unique



B**Figure 4.11 Nascent RNA-seq strategy and initial data analysis**

A) Schematic representation of the Nascent RNA-seq experiment. B) Hierarchical clustering of the samples based on differentially expressed genes.

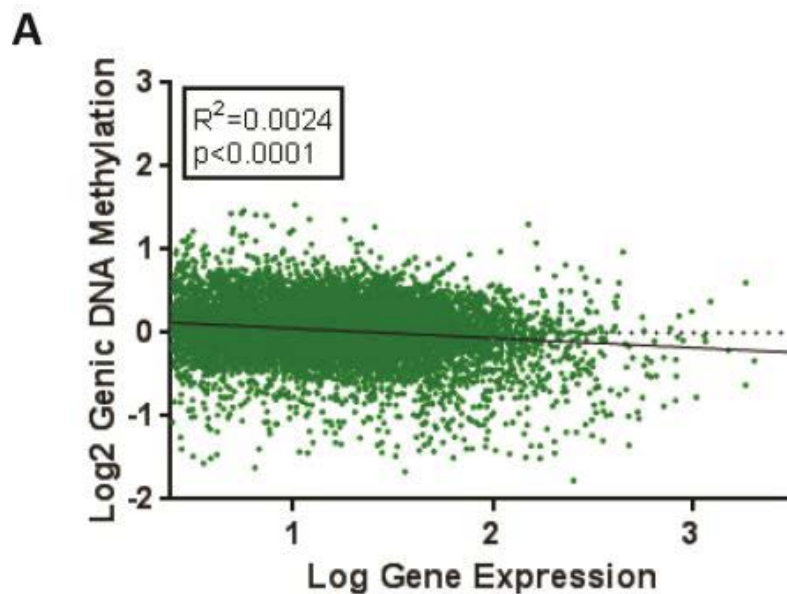
reads per sample (Table 4.3). However, hierarchical clustering of the data revealed close clustering of different treatments of the same biological replicates (Figure 4.11B). This is a strong indication that experimental variability exceeds the biological variation caused by deposition of 5hmC over the genome. This is not surprising, taking into account the technically complex Nascent RNA-seq experiment. Differential expression analysis was performed including all three biological replicates but no differentially expressed genes could be found. Therefore, based on the clustering results and differential expression analysis, replica 0 was eliminated, which skewed the analysis. This gave a list of significantly deregulated genes that could be used in further analyses (Table 4.4, Appendix Table S4).

Sample	Total Aligned Reads	Total Unique Reads
Rep 0 +DOX	84889916	35291318

Rep 0 -DOX	82649990	31615476
Rep 1 +DOX	104667343	51277176
Rep 1 -DOX	110430287	52875849
Rep 2 +DOX	89951027	37522514
Rep 2 -DOX	110876054	52401898

Table 4.3 Nascent RNA-seq sequencing metrics

Previous studies have reported various correlations between gene body methylation and gene expression in both cancer cells and normal tissues (Jjingo et al., 2012, Mellen et al., 2012). Therefore, we wanted to know how gene body methylation levels correlate with gene expression in untreated HEK293 cells. To compare gene body DNA methylation levels with expression, MeDIP values averaged over the transcribed regions were correlated with gene expression levels calculated from RNA-seq (Figure 4.12A). The data indicated that there was a weak but highly



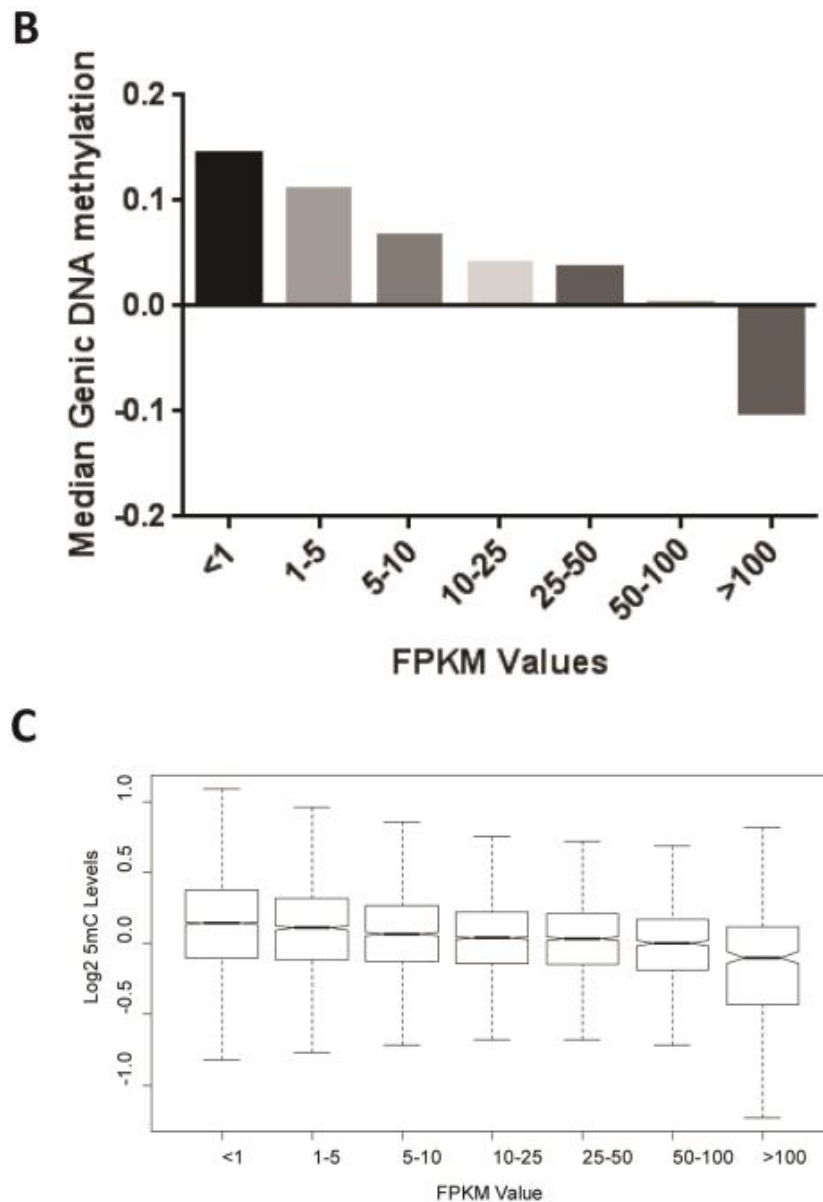


Figure 4.12 Correlation of gene body methylation and gene expression in HEK293 cells

A) Correlations between gene body DNA methylation and gene expression. Each gene is plotted as a single dot. P values represent the likelihood of the correlation being 0. B) Median DNA methylation levels over gene bodies correlated to gene expression based on the FPKM values. C) Correlation between genes binned based on their expression values and $\log_2$ DNA methylation levels over gene bodies. Each box is significantly different from the others.

significant negative correlation between gene body methylation and gene expression ($R^2 = 0.0024$, $p < 0.0001$). In order to explore whether the data exhibited a bell-shaped pattern as reported previously (Jjingo et al., 2012), all genes were binned into seven bins based on gene expression values (Figure 4.12B). Based on the median gene body methylation values, the correlation of genic 5mC and gene expression remained. The lowest expressed genes had a higher amount of DNA methylation compared to the most highly expressed genes. As can be seen in Figure 4.12C, variation of the methylation values in each bin was large. However, each bin was significantly different from the others with $p < 0.02$ (Welch two sample t-test). Interestingly, these correlation profiles resembled those of terminally differentiated cells (Mellen et al., 2012) more than those of cancer cells (Jjingo et al., 2012) reported previously. Despite previous efforts to link gene body DNA methylation and gene expression, evidence for it remains rather weak. Therefore, experimental approaches that allow the manipulation of DNA modifications over genic regions could be more conclusive.

To further investigate gene body DNA modification and its effects on gene expression, differential expression analysis was performed between untreated HEK293 cells and those that had been expressing TET-CD for 48 h. Eleven genes were found that changed significantly with $p < 0.05$ corrected for multiple t-testing (Table 4.4). The top hit was TET1, which was expected as the TET catalytic domain was being overexpressed. Interestingly, only one gene was observed to be significantly downregulated which indicated that 5hmC works mainly as an activating mark.

Gene	Fold Change	Log ₂ Fold Change	p value	p value (adjusted)
TET1	6.953	2.797	9.04 E-46	1.99 E-41
VPS9D1-AS1	4.719	2.238	6.256 E-20	6.89 E-16

GATA2	1.919	0.940	4.472 E-10	3.28 E-06
QDPR	2.123	1.086	3.732 E-08	0.00020
HAND1	2.859	1.515	2.294 E-07	0.00101
MGAT3	2.015	1.011	4.044 E-07	0.0014
HOXA-AS3	0.506	-0.980	8.093 E-07	0.0025
SOGA1	1.584	0.664	2.368 E-06	0.0065
PANX2	2.282	1.190	3.593 E-06	0.0088
FAM78A	1.782	0.834	7.056 E-06	0.0155
PPAPDC3	2.433	1.283	7.973 E-06	0.015
LZTS1	2.015	1.010	2.028 E-05	0.037

Table 4.4 Differentially expressed genes upon overexpression of the TET catalytic domain

Promoter DNA methylation is a well-documented mechanism of gene inactivation (Merlo et al., 1995). Therefore, when investigating genic DNA modifications, methylated promoters should be eliminated from the analysis. To do this, the promoters of significantly deregulated genes were assessed based on expression and DNA methylation status (Figure 4.13). Because promoter boundaries are hard to define, it was aimed to eliminate genes with methylated promoters based on the expression levels. It has been convincingly shown by numerous studies that DNA methylation over CGI promoter regions can completely shut down transcription (Merlo et al., 1995, Ehrlich, 2002). In addition, the analysis of significantly deregulated genes indicated that genes with methylated promoters can be efficiently eliminated by introducing a 2.5 fragments per kb per million reads (FPKM) expression value threshold. Following this correction, differential expression analysis

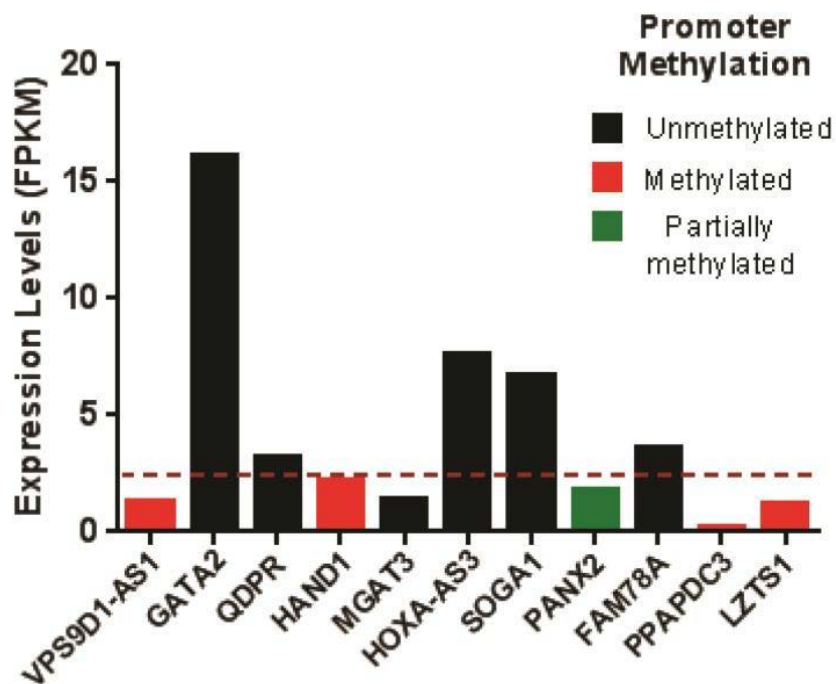


Figure 4.13 Upregulation of gene expression and promoter DNA methylation
Expression levels of genes significantly upregulated upon overexpression of the TET catalytic domain. Genes with black bars have unmethylated promoters, promoters of genes represented by red bars were methylated.

was repeated for all of the genes expressed at more than 2.5 FPKM. This successfully eliminated all of the genes with methylated CpG island promoters (Table 4.5). Interestingly, another gene (CEBPA) passed the significance threshold. Unfortunately, this type of analysis gave only five positive hits, which was useful for candidate gene analysis but insufficient for correlative analysis and making confident conclusions about global gene regulation. In order to have a larger dataset, we have decided to in-

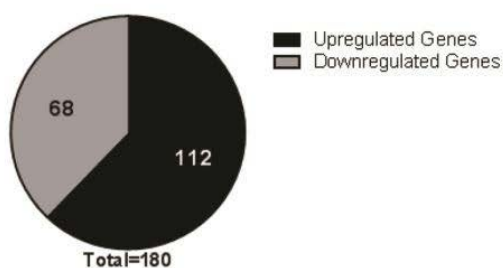
Gene	Fold Change	Log ₂ Fold Change	p value	p value (adjusted)
TET1	7.089	2.825	5.27E-40	5.78E-36

GATA2	1.954	0.966	1.52E-10	8.34E-07
SOGA1	1.613	0.690	2.89E-07	0.001056
QDPR	2.159	1.110	4.34E-07	0.00119
HOXA-AS3	0.515	-0.957	7.26E-06	0.0159
FAM78A	1.818	0.862	9.68E-06	0.0173
CEBPA	1.524	0.608	7.02E-04	0.040

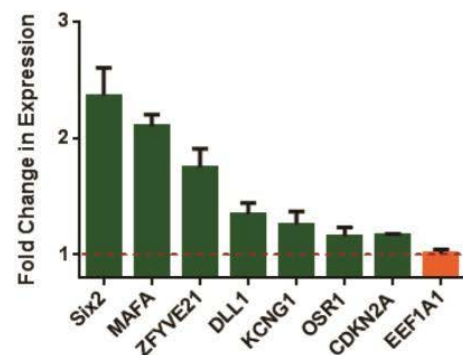
Table 4.5: Differentially expressed genes upon overexpression of the TET catalytic domain with a 2.5 FPKM cut-off.

crease false discovery rate (FDR) and obtained a new gene set that changed their expression levels with $p < 0.02$, not corrected for multiple hypothesis testing. This data set was likely to contain a larger fraction of false positives and true negatives, but the majority of the values would still be true positives. The new dataset contained 180 genes that were deregulated: 62% upregulated and 38% downregulated (Figure 4.14A). To confirm that the majority of the genes in this data set were indeed true positives, a panel of genes that were upregulated was randomly picked and their upregulation confirmed by RT-qPCR (Figure 4.14B). These data suggest that 5hmC

A



B



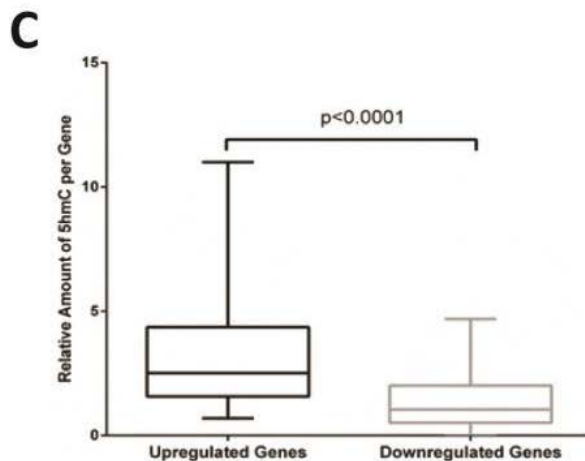


Figure 4.14 The metrics of significantly expressed genes

A) Pie chart representing the fraction of genes with expression going up or down ($p < 0.02$, not corrected for multiple hypothesis testing). B) Confirmation of upregulated genes by RT-qPCR. C) Comparison of 5hmC accumulation over gene bodies between upregulated and downregulated genes.

could act as transcriptional activator over genic regions. In addition, the downregulation of the number of genes could be caused by oxidation of 5hmC to the 5fC or 5CaC, or alternative secondary effects. To further confirm this, the gain of 5hmC deposited over the gene bodies of both groups was compared. Interestingly, there was significantly less (3-fold) 5hmC deposited over the downregulated genes, suggesting that they could be false positives or indirectly regulated (Figure 4.14C). The data indicate that 5hmC deposited over gene bodies may act as a genome-wide gene activator. To further investigate 5hmC-mediated upregulation of transcription, whether the extent of transcriptional upregulation was dependent on the amount of 5hmC deposited over genic regions was analysed. Therefore, the amount of 5hmC on the gene body was compared with the changes in gene expression of significantly upregulated genes (Figure 4.15A). As can be seen from Figure 4.15, genic 5hmC significantly correlated with a change in transcription. In other words, the more

5hmC was deposited over a gene body the more the gene was upregulated. Whether or not the amount of genic 5hmC correlated with the downregulation of genes was also assessed, however, no significant correlation between gene downregulation and deposition of 5hmC on these genes was observed (Figure 4.15B). Therefore, it was concluded that hydroxylation of 5mC over gene bodies can upregulate gene expression.

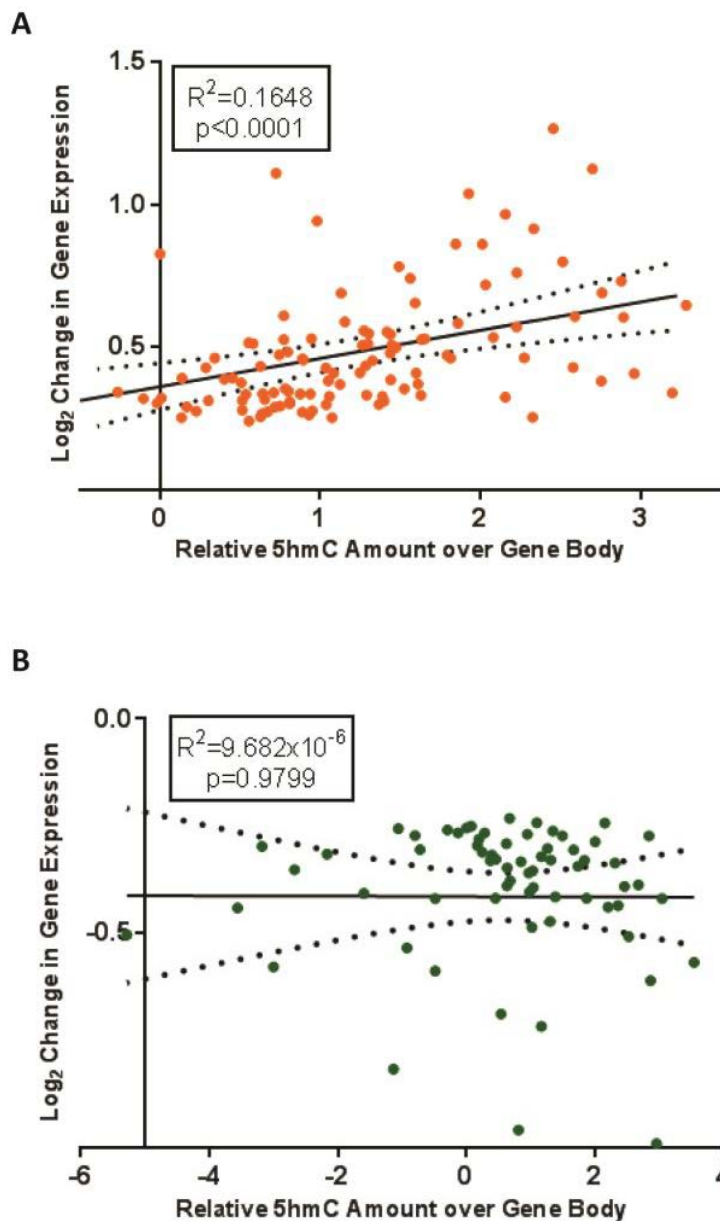


Figure 4.15 5hmC accumulation over gene bodies and deregulation of gene expression

A) Accumulation of 5hmC over gene bodies correlates with upregulation of gene expression (p value defines the probability of no correlation). Dotted lines represent 99% confidence intervals. B) Correlation of 5hmC accumulation over gene bodies with deregulation of unexpressed genes.

4.2.7. Genic DNA Modifications Mediated Transcriptional Regulation:

Mechanistic Insights from Genomic Data

In order to gain better insight into the mechanism of 5hmC-mediated genic upregulation, certain features of significantly upregulated genes were interrogated (Figure 4.16). Firstly, DNA methylation over gene bodies was examined. As shown in Figure 4.16A and B, the upregulated genes had significantly higher CpG densities and higher amounts of methylation over their genic regions compared to the whole genome. This suggests that methylation could act as a repressive mark over gene bodies. It has been previously proposed that MeCP2, which binds to methylated CpG dinucleotides, can specifically repress long genes (Gabel et al., 2015). Therefore, it was investigated whether gene body 5hmC-mediated activation is more prominent on long genes. However, no significant enrichment of long genes was observed in the dataset (Figure 4.16C). Finally, the absolute expression levels of the upregulated genes were examined (Figure 4.16D). Interestingly, the upregulated genes had lower expression values compared to the rest of the genes. This indicated that these genes are actively repressed in HEK293 cells. Taken together, these data suggest that gene body methylation actively represses transcription in a concentration-dependant manner and that 5hmC can alleviate this repression.

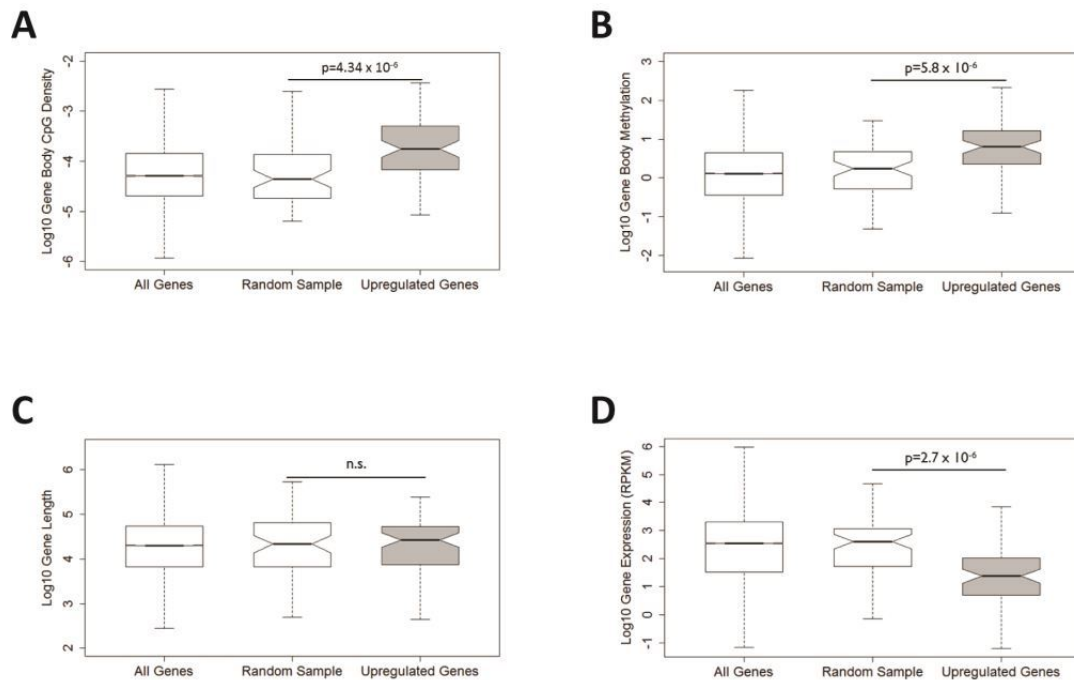
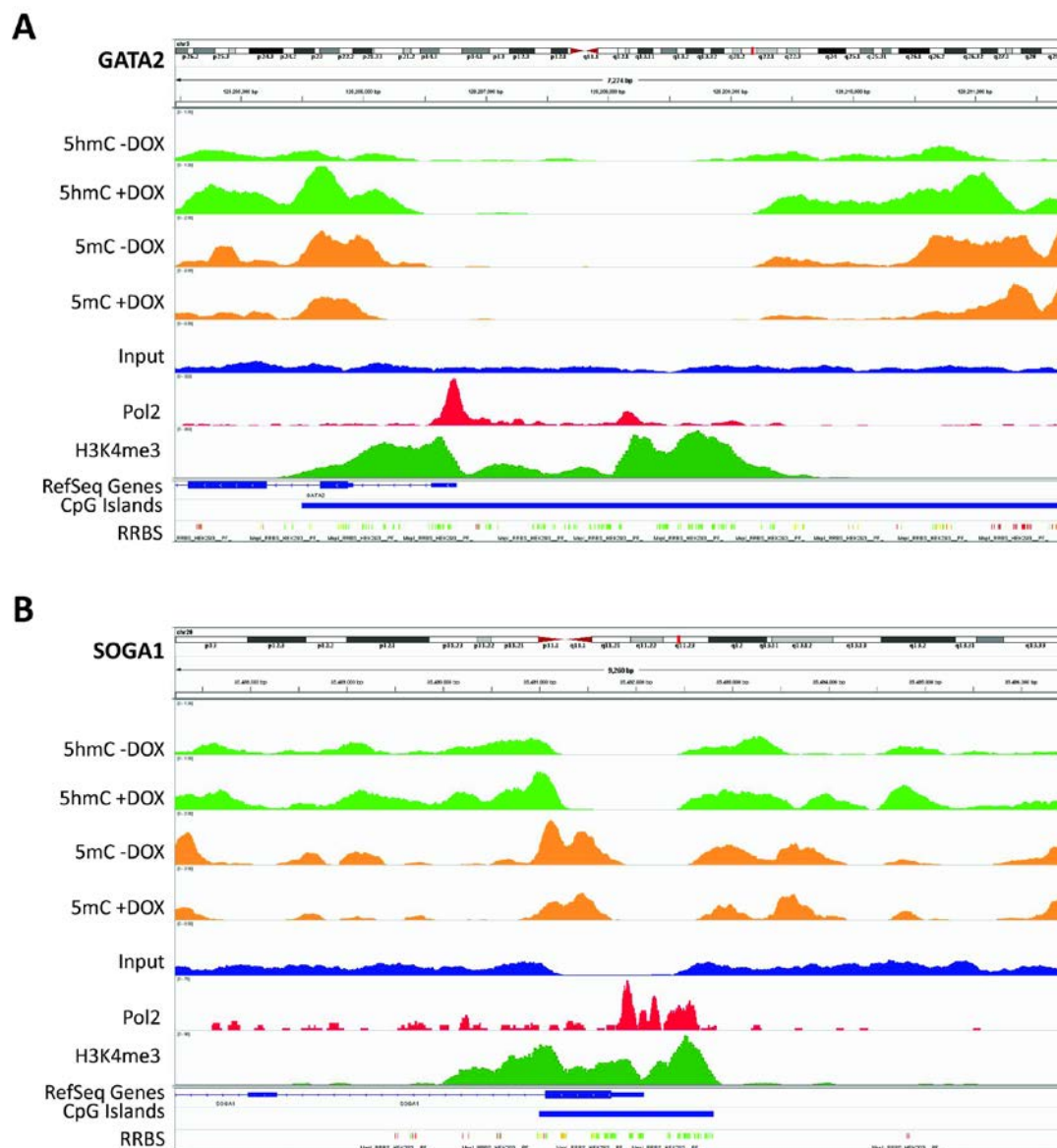


Figure 4.16 General features of the significantly upregulated genes set
 A) A comparison of CpG density over gene bodies between all genes, a random sample set and the set of upregulated genes. B) Gene body DNA methylation of all genes, a random sample and upregulated genes. C) Comparison of gene length between the upregulated gene set, a randomly sampled gene set of the same size and all the genes. D) Absolute expression levels of the upregulated genes compared to all genes and randomly sampled genes.

The next step was to better explore the DNA modification profiles of the significantly upregulated genes. By taking a closer look at single examples of upregulated genes it became apparent that genic regions, which contain large amounts of DNA methylation, accumulate 5hmC upon upregulation of the TET catalytic domain (Figure 4.17). Interestingly, upregulated genes had strong methylation signals over CpG island shores and even at the edges of CpG islands (Figure 4.17). This posed the possibility that expression can not only be affected by the amount of methylation over a genic region, e.g. via downregulation of the

polymerase elongation rate, but also that the proximity of DNA methylation to TSSs could play a role. Looking at the genomic snapshot for GATA2 it is obvious that the genic region around the TSS is unmodified. However, genic DNA methylation spreads into promoter CpG islands (Figure 4.17A). The same seems to be true for the other genes. Interestingly, DNA methylation spread towards the TSS for the SOGA1 gene is more pronounced from the genic region and there is no DNA methylation over the CpG island upstream of the TSS (Figure 4.17B). The same holds true for the CEBPA



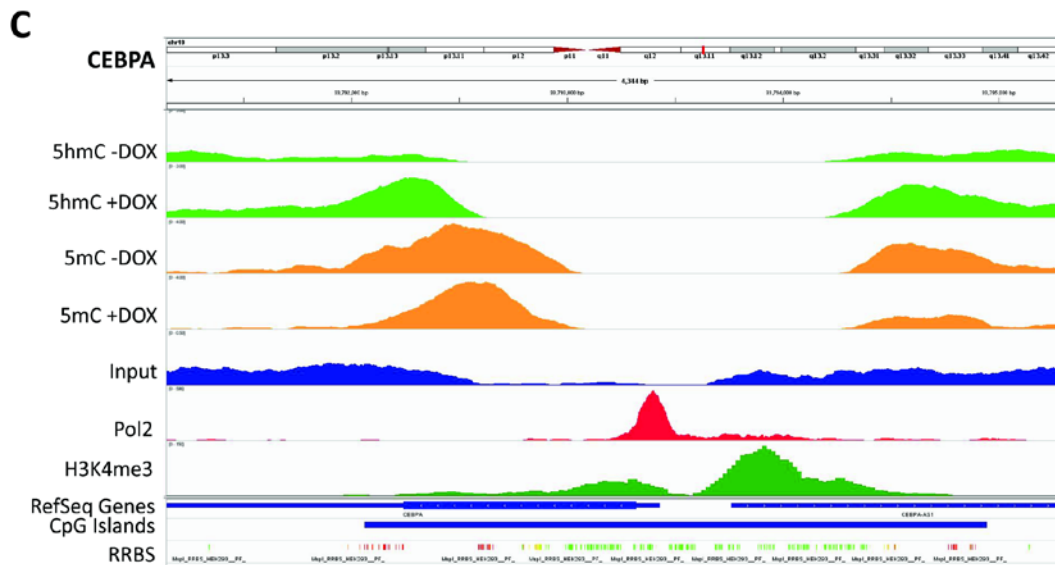
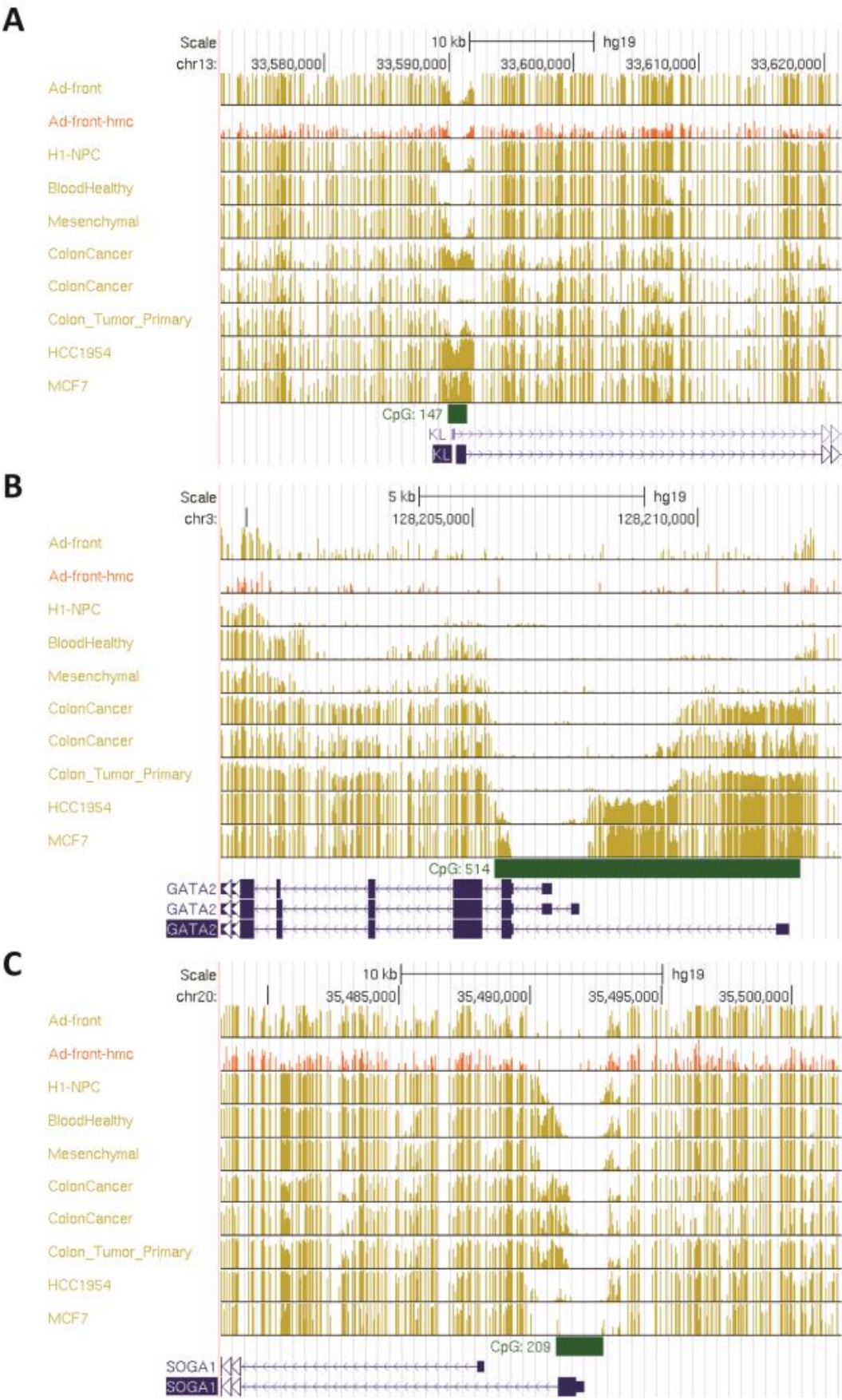


Figure 4.17 Promoter demethylation of significantly upregulated genes
 A) IGV snapshot of the GATA2 promoter. Tracks for 5hmC and 5mC are shown for the induced and uninduced states. H3K4me3, Pol2, and RRBS data has been taken from the ENCODE project for the HEK293 cell line. B) IGV tracks for the SOGA1 gene. C) IGV tracks for the CEBPA gene promoter.

gene (Figure 4.17C). However, it was noticed that the input for these genes was depleted over CpG island regions (Figure 4.17B,C). In general, this is a common problem of the Illumina sequencing platform (Benjamini and Speed, 2012). Therefore, we supported our data with the previously published RRBS dataset in the HEK293 cell line (GEO:GSM720355) (Figure 4.17 bottom panel). Most importantly, these genes are moderately expressed, suggesting that they are not inactivated by CpG island methylation.

To investigate whether the mechanism described above can be used by cancer cells to actively downregulate but not silence gene expression, the methylation maps generated by bisulfite sequencing in other studies of the promoters of these genes were examined (Figure 4.18). Figure 4.18A represents the *KL* gene, which has a CpG island promoter and is fully methylated in MCF7 and HCC1954 cells. Full methylation



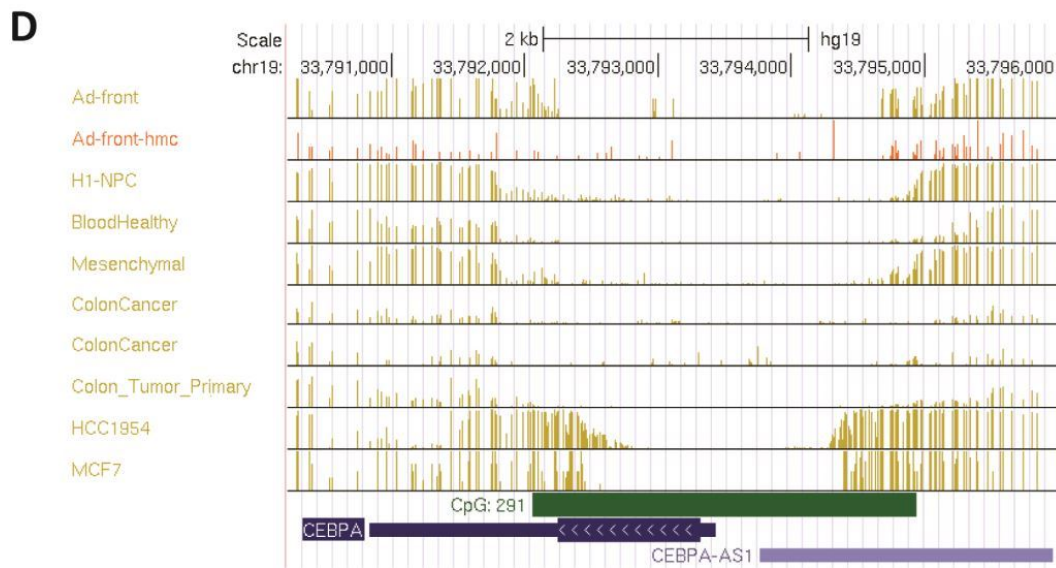


Figure 4.18 Promoter proximal DNA modifications in normal and cancer cells
 A) Single nucleotide resolution data for DNA methylation in frontal cortex (Ad-front) and other cell types for *KL* gene are shown in yellow, hydroxymethylation for frontal cortex (Ad-front-hmc) is shown in red. B) DNA modification data for *GATA2*. C) *SOGA1* gene. D) *CEBPA* gene.

of this promoter completely abolishes gene expression. However, in other tissues and clinical samples, differential methylation has been observed in the shores of CpG islands (Irizarry et al., 2009). Interestingly, in the brain, the DNA modifications in close proximity to the promoter are almost exclusively 5hmC (Figure 4.18A). The same is true for other genes such as *GATA2*, *SOGA1* and *CEBPA* (Figure 4.18A,B,C). In addition, single nucleotide resolution data revealed that 5hmC is not exclusively deposited over the gene bodies but also upstream of the promoter. This data, together with the gene expression analysis described here, indicates that DNA modifications can modulate promoter activity in *trans*. Importantly, by looking at DNA modification profiles of the brain and blood from healthy adults, differences in

promoter proximal DNA modifications emerge (Figure 4.18) indicating that this type of regulation may also occur in healthy tissues.

4.3. Discussion

The role of gene body hydroxymethylation in biological processes, and in particular in transcription, is not fully understood. In order to address this problem, a cell line was engineered that expressed the TET1 catalytic domain to deposit a substantial amount of 5hmC over the genome. It was demonstrated that DNA modification levels in HEK293 cells could be altered without a major impact on cell viability. In addition, upon withdrawal of DOX and downregulation of the TET catalytic domain, cells regained the initial amount of DNA modifications. This implies that there may be a mechanism of 'epigenetic memory' for global DNA methylation levels, but whether the same methylation landscape is re-established genome-wide regions remains to be seen.

Promoter CpG island methylation is a well-known mechanism of gene transcription repression (Merlo et al., 1995). However, genome-wide DNA modification mapping has revealed that dynamic changes in DNA modification are not only restricted to CGI promoter regions but also happen on gene bodies and distant regulatory elements (Lister et al., 2013). Here, the role of genic DNA modifications was investigated by converting around 25% of 5mC to 5hmC in the genome of the HEK293 cancer cell line. A similar amount of 5hmC has been observed in neurons (Kriaucionis and Heintz, 2009). In order to investigate how gene body DNA modifications affect transcription, 5mC and 5hmC were mapped genome-wide in induced and uninduced states. Together with differential expression analysis, this

study demonstrated that genic 5hmC is a *bona fide* activating mark and that it upregulates transcription in concentration-dependant manner. This observation extends our knowledge about 5hmC-mediated transcriptional regulation beyond the correlations reported previously (Mellen et al., 2012, Song et al., 2011a, Lister et al., 2013).

Interestingly, in this study, it was revealed that upregulated genes had highly methylated CpG island shores. Highly differential promoter proximal DNA methylation has been reported in colon cancer (Irizarry et al., 2009). Furthermore, the ability of promoter proximal DNA methylation to repress transcription has been shown in an artificial reporter system (Appanah et al., 2007). This is in agreement with the data described here, which suggests that promoter proximal DNA modifications could modulate transcription levels. However, the magnitude of the upregulation is small and few genes are significantly deregulated indicating that genic 5hmC might not affect transcription genome-wide. In addition, several important issues remain to be addressed. Firstly, 5–7% demethylation was observed upon the overexpression of the TET1 catalytic domain. This poses the possibility that demethylation creates transcription factor binding sites in promoter proximal regions. Secondly, the activation of gene expression can be upregulated by depositing 5hmC over distal regulatory elements, which often reside in genic regions. Thirdly, the effect on transcription could be entirely indirect. Most importantly, the identity of the molecular mechanism behind gene body DNA modification-mediated transcriptional regulation is, as yet, unknown. Attempts to address these issues and other questions are discussed in subsequent chapters.

Chapter 5

Gene Body Hydroxymethylation and Demethylation Can Reverse 5mC-Mediated Repression

6.1. Introduction

Gene body DNA modifications have been shown to exhibit modulatory effects on gene expression in several experimental systems (Karimi et al., 2011, Lorincz et al., 2004, Robertson et al., 2011, Allan et al., 2008, Hsieh, 1994, Hsieh, 1997) including this study, however, the mechanistic details of this phenomenon have not been elucidated. Furthermore, the activating role of 5hmC over gene bodies has only been speculated based on genome-wide correlations (Song et al., 2011a, Mellen et al., 2012). In the previous chapter, it was demonstrated that hydroxylation of 5-methylcytosine in genic regions can upregulate gene expression. Even though the data presented strongly suggests that 5hmC deposited over gene bodies can induce transcriptional activation it also has several drawbacks. Firstly, the definition of a gene body or transcribed region is the region between a transcription start site (TSS) and a transcription termination site (TTS) (Nanty et al., 2011), and complexity arises from the fact that gene bodies often overlap with other genomic elements such as enhancers or promoters, which makes the results hard to interpret. It is clear, however, that DNA modifications deposited over promoters and enhancers play a role in the regulation of the activity of these elements (Blattler et al., 2014). Secondly, gene expression rates can be modulated by transcription-coupled processes such as splicing. Importantly, exon splicing has been shown to be affected by genic DNA modifications (Khare et al., 2012). Although some of these problems can be addressed by careful examination of genome-wide expression data, different experimental approaches and techniques are required to further support the hypothesis above and make meaningful conclusions. In this chapter, a reductionist approach was taken and several reporter-based assays were developed to investigate the effect of gene body cytosine modifications on gene expression. Importantly, the advantage of reporter-

based assays is that they are very well defined. This allows direct and simple interrogation of the effects of gene body DNA modifications on gene expression.

6.2. Results

6.2.1. Engineering and Integration of a Differentially Methylated Reporter System into the Genome of the TRex HEK293 TET-CD Cell Line

To assess the effect of gene body DNA modifications on transcription, we aimed to design an artificial reporter system that would enable the interrogation of the unmethylated, methylated and hydroxymethylated state of the gene body. For this reason, we began with an expression plasmid devoid of CpG dinucleotides, which are the main sites for DNA methylation. The reporter system was constructed by integrating the firefly luciferase reporter, which has CpG island-like DNA sequence, into the CpGless plasmid (Figure 5.1). The firefly reporter gene was driven by the hybrid EF1/CMV promoter with all the CpG dinucleotides eliminated. In order to control for expression differences due to tandem integration events, a CpGless *Renilla* luciferase reporter gene was engineered and integrated, which was also to serve for expression data normalisation purposes. To avoid interference between different active elements in the system, each component was separated by Matrix Attached Regions (MARs) that act as insulators and ensure that each element in the system works as an independent unit. Methylation was established by treating the constructs with bacterial CpG DNA methyltransferase (M.SssI). The engineered reporter systems (either methylated or unmethylated) were integrated into the TRex HEK293 TET-CD cell line. In order to avoid possible experimental bias due to integration of the repor-

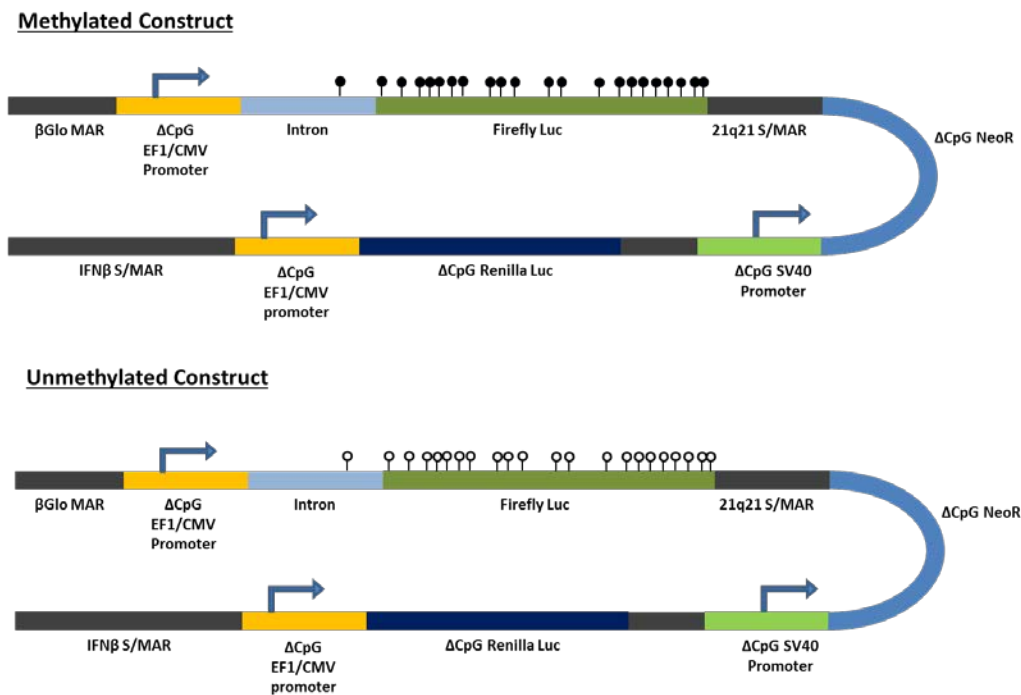


Figure 5.1 Schematic diagram of the integrated reporter construct

Methylated (filled lollipops) and unmethylated (hollow lollipops) CpG sites are shown. Firefly and *Renilla* luciferase genes are driven by the same promoter. The neomycin resistance gene was used for the selection of positive clones.

ter into specific genomic regions of active or repressed chromatin, single clones were not selected, but rather all clones were pooled that survived antibiotic selection. This gave a representation of integration events in various genomic regions. Seven passages after integration, the methylation status of the integrated reporters was investigated (Figure 5.2). The reporter methylation levels were assessed using either locus-specific bisulfite sequencing or by post-bisulfite digestion using specific restriction enzymes (COBRA assay) (Figure 5.2A and B). As can be seen from the data, even after seven passages the methylated reporter did not lose DNA methylation and the unmethylated reporter did not gain any. In agreement with this observation, the ability of exogenous, high CpG density DNA sequences integrated into cell lines to maintain methylated or unmethylated states has been reported previously (Lienert et

al., 2011). Knowing that reporters maintained their DNA modification status, luciferase reporter expression levels were assessed and compared between methylated and unmethylated reporter constructs (Figure 5.2C). Interestingly, reporters with methylated gene bodies have shown 7-fold lower expression levels compared to unmethylated reporters. This result indicates that DNA methylation deposited over a gene body can repress transcription of the gene. To exclude the possibility that the difference had been observed due to different integration between methylated and unmethylated reporters, methylated and unmethylated plasmids were transfected into TRex HEK293 cells and the same difference in expression was

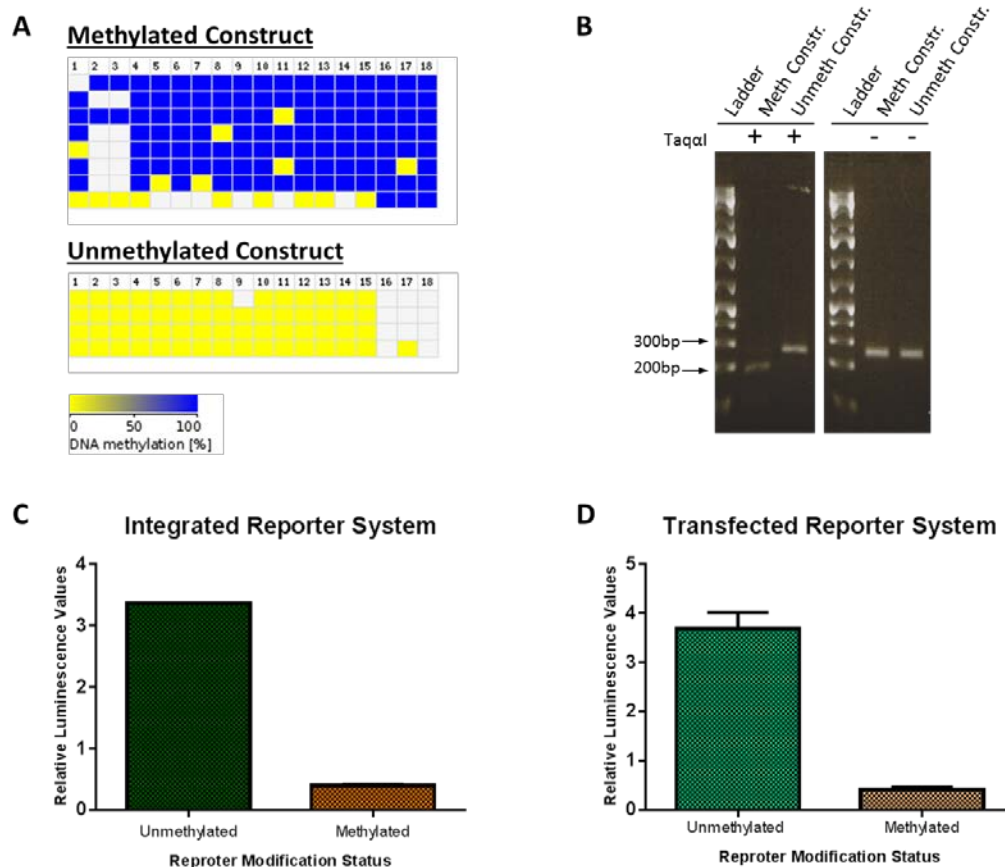


Figure 5.2 Modification status and expression levels of integrated reporter constructs

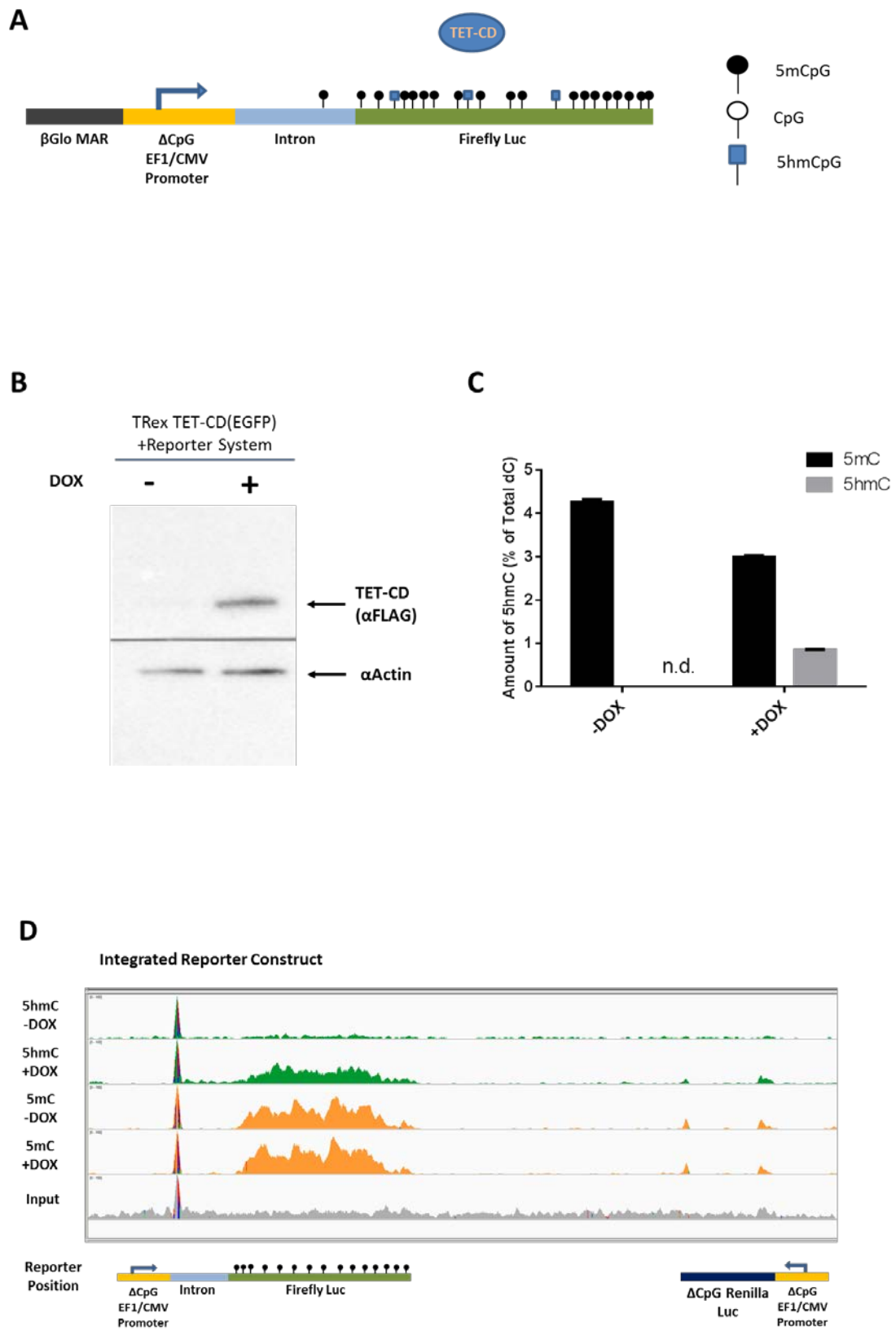
A) DNA methylation is maintained on the methylated reporter gene body for at least

seven passages after the end of selection (top panel) although some sporadic demethylation occurs in a small fraction of cells. The unmethylated reporter gene body does not acquire any methylation over the same period of time (bottom diagram). DNA methylation for each CpG is represented as yellow square for unmethylated and a blue square for methylated. B) Gene body methylation assessed by post-bisulfite digestion using TaqαI restriction enzyme. C) The expression levels of integrated reporters. Relative luminescence represents firefly luciferase luminescence normalised to *Renilla* luciferase luminescence values. D) Transfected differentially modified reporter expression levels. All error bars represent the standard error of the mean from at least 3 independent experiments.

observed as between the integrated methylated and unmethylated reporter (Figure 5.2D). In addition, very similar differences in reporter activity were evident indicating that similar mechanisms are responsible for the repression of transfected as well as integrated reporters. These results indicate that gene body methylation can repress expression of the gene.

6.2.2. Deposition of 5hmC Alleviates Genic 5mC-Mediated Repression

To investigate whether hydroxylation of 5mC deposited over gene bodies has any effect on gene expression, the TET1 catalytic domain was overexpressed in the cell lines containing the reporter systems. The aim was to specifically deposit 5hmC on the methylated reporter and compare changes in gene expression with the unmethylated reporter, where 5hmC cannot be deposited due to the lack of DNA methylation (Figure 5.3A). It was confirmed that TET-CD was expressed after induction with doxycycline (DOX) (Figure 5.3B) and that 5hmC was successfully deposited genome-wide using HPLC (Figure 5.3C). Using genome-wide sequencing data from hmSeal and MeDIP experiments, DNA modification signatures of the reporter construct were successfully reconstituted and it was shown that only the gene body of the firefly reporter was methylated. In addition, it was confirmed that



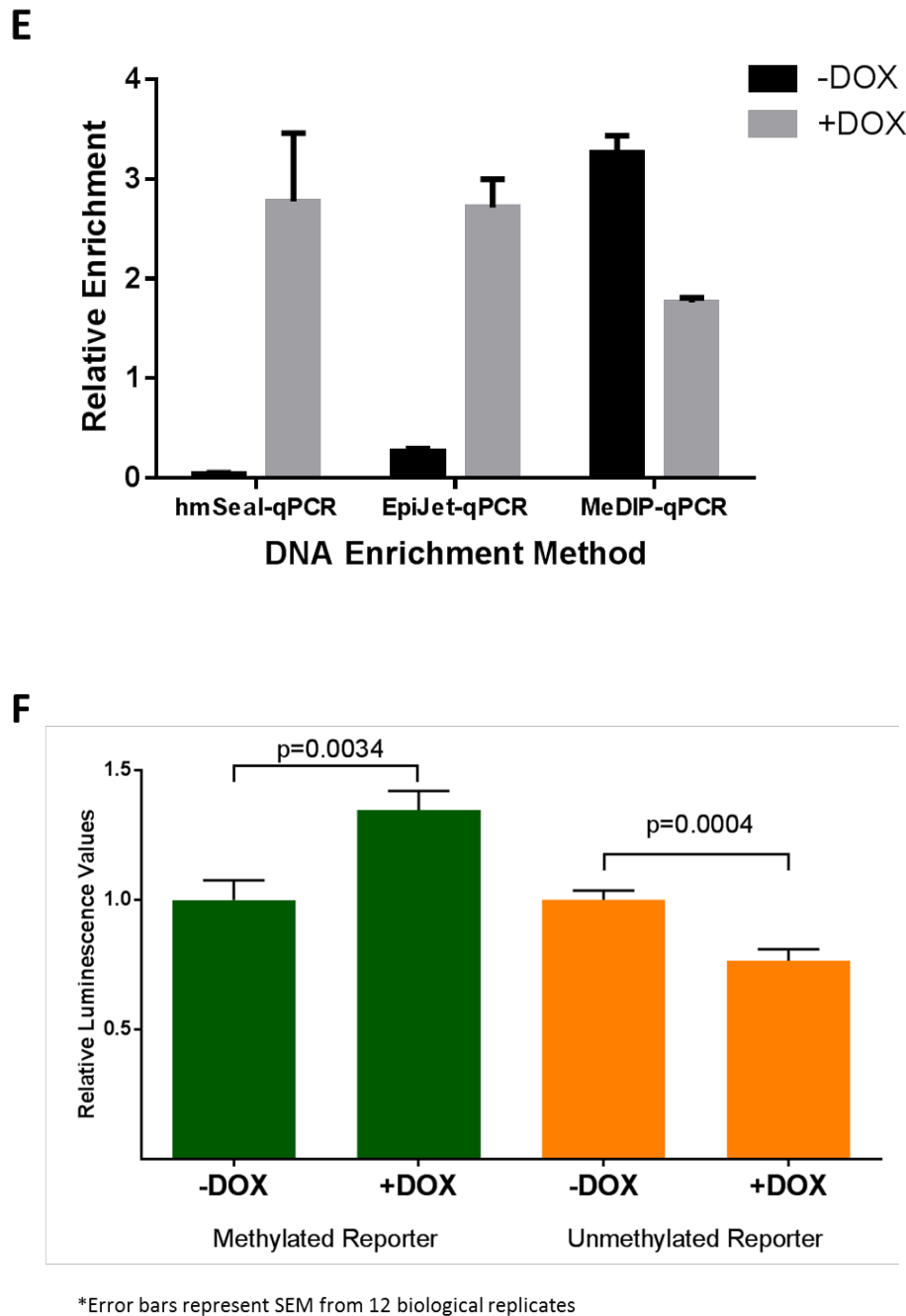


Figure 5.3 Hydroxylation of the methylated reporter gene body upregulates gene expression

A) Schematic representation of the hydroxylation of the methylated reporter. B) Immunoblot representing efficient overexpression of the TET1 catalytic domain upon DOX treatment. C) HPLC analysis indicates that TET1-CD can efficiently produce 5hmC over the genome upon DOX treatment. D) MeDIP-seq and hmSEAL-seq reads aligned to the reporter system indicate clear methylation and hydroxymethylation patterns. E) Deposition of 5hmC and a reduction of 5mC levels upon DOX treatment were confirmed by pull-down methods followed by quantitative PCR. Error bars

represent the standard error of the mean of three independent experiments. F) Significant upregulation of methylated, and downregulation of unmodified, reporter can be observed upon DOX treatment of cells for 48 h. Error bars represent the standard error of the mean of 12 independent assays.

the methylated firefly gene body was the only place that acquired 5hmC upon induction of the TET-CD (Figure 5.3D). The presence of 5hmC over the firefly luciferase gene body was also confirmed using hmSeal-qPCR and EpiJET-qPCR methods (Figure 5.3E). Finally, whether the deposition of 5hmC over the methylated reporter system has any effect on gene expression was assessed. Interestingly, luciferase assay indicated that deposition of 5hmC could upregulate the expression of methylated reporter by around 1.5-fold (Figure 5.3F). Surprisingly, the unmethylated reporter was significantly downregulated upon overexpression of TET-CD. This could be due to the fact that the unmethylated reporter acquires very low levels of sporadic DNA methylation that is very efficiently converted to 5fC or 5caC, which was shown to inhibit transcriptional elongation (Wang et al., 2015).

To further investigate the role of gene body modifications and their role in the modulation of gene expression, the artificial reporter system was further perturbed. Firstly, cells were treated with the widely used demethylating agent 5-azacytidine (5AzaC). 5AzaC works by inhibiting and depleting the pool of active DNA methyltransferases, which leads to subsequent DNA demethylation (Vesely and Cihak, 1978). Methylated reporter-containing cells were treated with 5AzaC for 24 h, following which significant alleviation of gene repression was observed (Figure 5.4). Interestingly, 5AzaC treatment significantly downregulated the expression levels from the unmethylated reporter. This might have been caused by the redistribution of transcriptional machinery or induction of a transcription block by cross-linking DNMT1 to DNA or RNA.

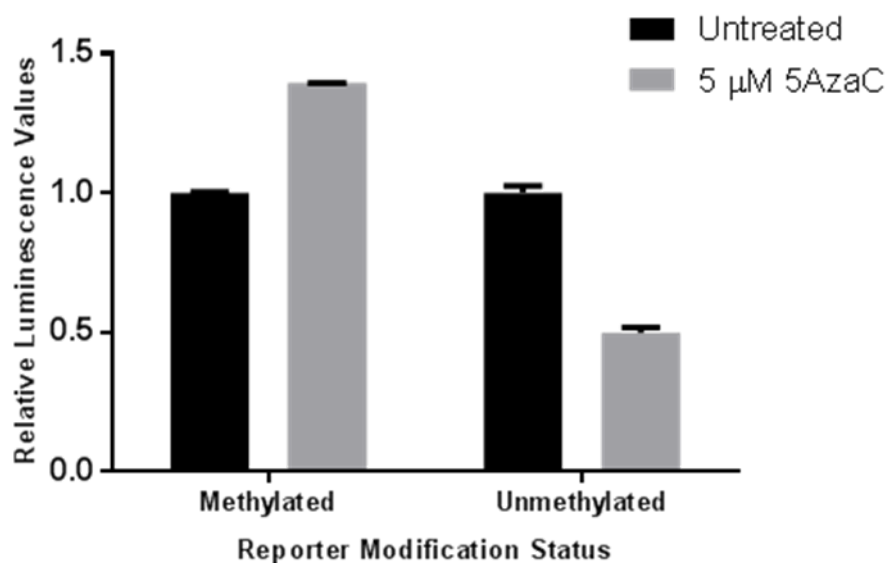


Figure 5.4 5-AzaC treatment can modulate the expression of differentially modified reporters

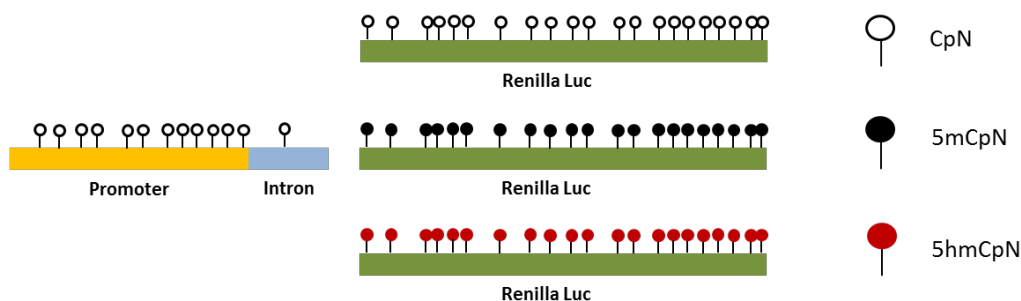
Cells containing methylated or unmethylated reporter were treated with 5AzaC and luminescence levels were quantified. Error bars represent the standard error of the mean of 4 independent experiments.

These data strongly suggest that the function of 5hmC is very similar to that of unmethylated cytosine. The integrated reporter experiments described here demonstrate that gene body hydroxymethylation and demethylation act as activators, and that gene body methylation acts as a repressive mark. Importantly, these integrated reporter assays allow differentiation between gene body modifications and modifications of gene regulatory elements that often reside in the gene bodies. In addition, these assays eliminate the possibility that gene expression differences are mediated by low levels of sporadic DNA methylation over the promoter region. However, because of TET-CD induced demethylation, discrimination between the effects of demethylation versus hydroxymethylation was not possible.

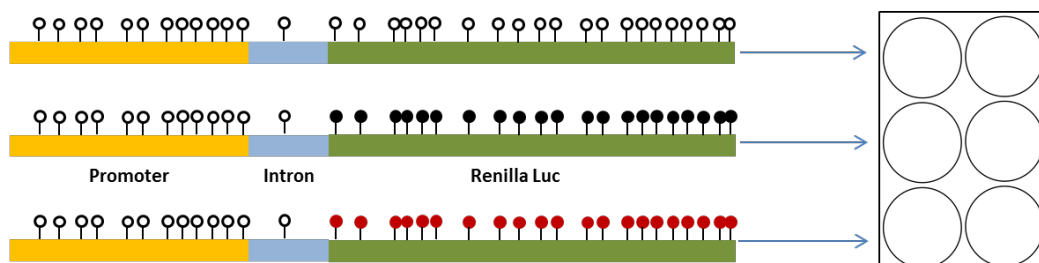
6.2.3. Transfected Reporter Assays: DNA Methylation but not Hydroxymethylation over Gene Bodies can Repress Transcription in a Concentration-Dependant Manner

One of the drawbacks of the integrated reporter experiment described above is that deposition of 5hmC is likely to cause DNA demethylation, as has been observed genome-wide. In order to eliminate this problem and investigate the role of 5hmC deposited over gene bodies, a transfected reporter assay was developed. The main idea of the assay was to randomly introduce a known amount of modified cytosine over the gene body of the reporter gene by PCR, then ligate it to an unmodified promoter and transfect it into cells (Figure 5.5). Transfection of equal amounts of the

1) PCR Amplification of Promoter and Modified Gene Bodies



2) Ligation and Transfection into the Selected Cell Type



3) Luciferase Assay after 24 hour Incubation

Figure 5.5 Schematic representation of the transfected reporter assay Reporter genes were synthesised by PCR. This leads to random incorporation of modified cytosines in the constructs. Constructs were ligated and the right product extracted from the agarose gel. Equal amounts of constructs were transfected into the cells and luciferase activity was measured 24 hours after transfection.

construct is crucial for this assay, therefore, samples were quantified and run on a gel to ensure equal concentrations and the high quality of each construct. To investigate the role of DNA modifications deposited over gene bodies, differentially modified reporters were transfected into HeLa cells (Figure 5.6). Interestingly, gene body 5mC repressed reporter activity in density-dependant manner. However, the reporter gene containing 5hmC over the gene body showed little or no repression (Figure 5.6A). The high quality and equal amounts of the constructs transfected into the cells are demonstrated in Figure 5.6B. As these artificial constructs are unlikely to contain mammalian replication origins, they were not replicated during cell division and were improbable to lose modifications by passive dilution. In order to make sure that there was no active dehydroxymethylation occurring in these cells, the transfected probes were recovered 24 h after transfection and subjected to bisulfite sequencing. As can be seen from the data, there was no significant difference in DNA modifications between untransfected and recovered constructs (Figure 5.6C). A time course experiment was performed to investigate whether gene body DNA modifications induce stable and consistent effects on gene expression (Figure 5.6D). The expression of the reporters was tested starting 6 h after transfection until 24 h, and stable maintenance of repression patterns was observed. This indicates that gene body-mediated transcriptional regulation is a very robust phenomenon. Finally, the effect of gene body-mediated repression was contrasted with the repression mediated by the promoter modifications (Figure 5.6E). As can be seen from the

graph, there was around 40-fold repression when the whole construct was fully

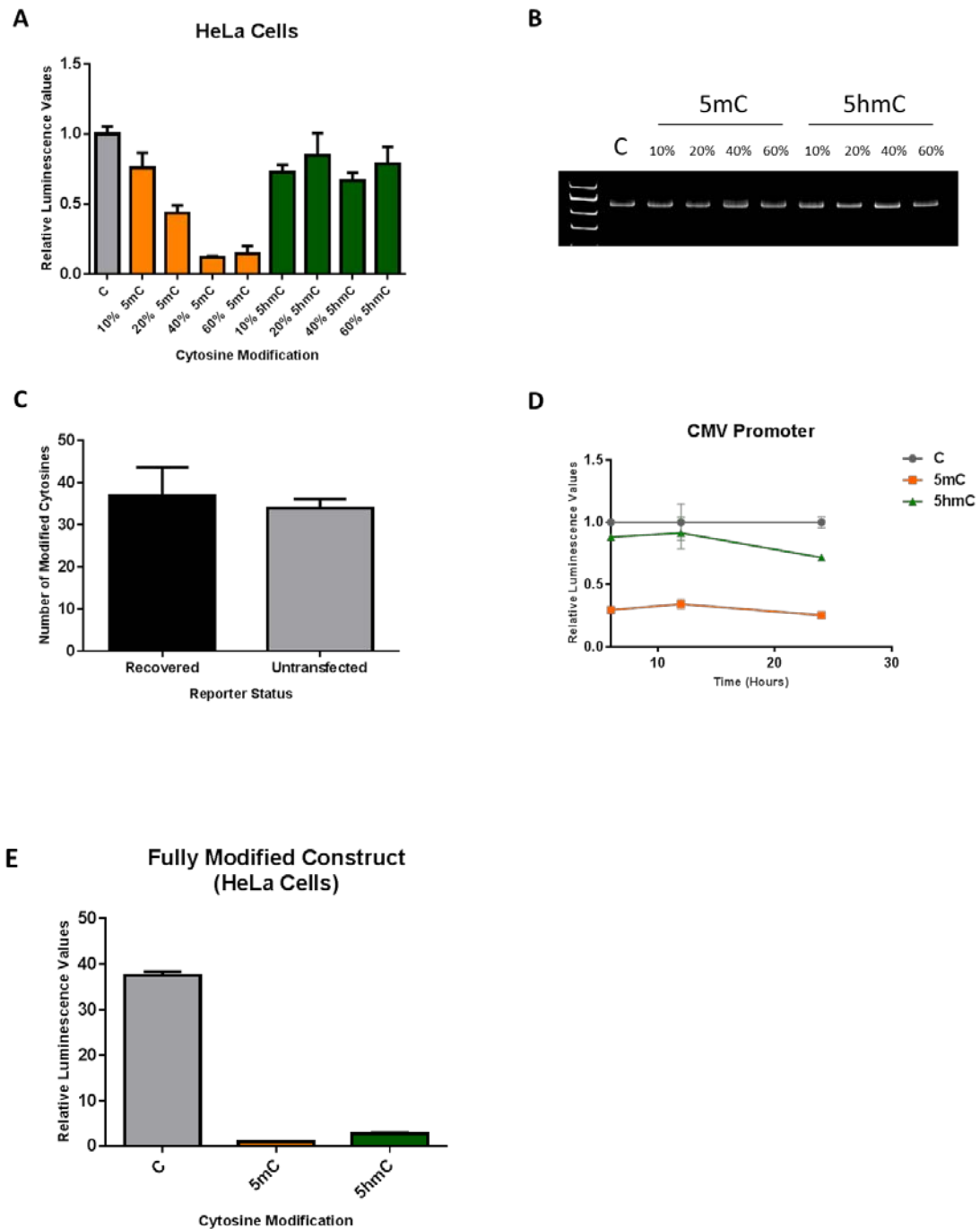
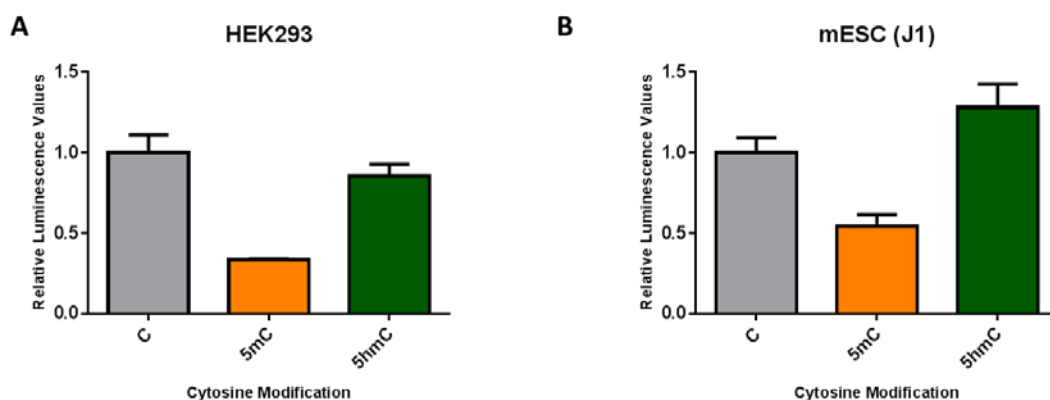


Figure 5.6 Gene body 5mC but not 5hmC can efficiently repress transcription

A) Gene body DNA methylation but not hydroxymethylation can repress transcription in a concentration-dependant manner in HeLa cells. B) Equal amounts and the high

quality of transfected probes were confirmed by running them on an acrylamide gel. C) Quantification of the DNA modifications in untransfected and recovered 5hmC constructs using bisulfite sequencing. D) Time course experiment demonstrating the stability of DNA modification-mediated gene modulation. E) DNA methylation over the promoter completely abolishes the expression of transfected reporter regardless of type or cytosine modification. All error bars represent the standard error of the mean from at least 3 independent experiments.

methylated or hydroxymethylated. This is likely to have arisen from promoter DNA modifications interfering with transcription factor binding, as reported previously (Robertson et al., 2011). This data indicates that hydroxymethylation, as well as demethylation, can alleviate gene body 5mC-mediated transcriptional repression. To further investigate the role of gene body DNA modifications, whether or not the effect could be observed in different cell lines was assessed (Figure 5.7). As can be seen from the data, the same trend was apparent in both HEK293 cells and mouse ESCs. The levels of promoter expression were higher for 5hmC and 5mC constructs in mouse ESCs compared to cancer cell lines. However, this is likely to have been caused by experimental variability.



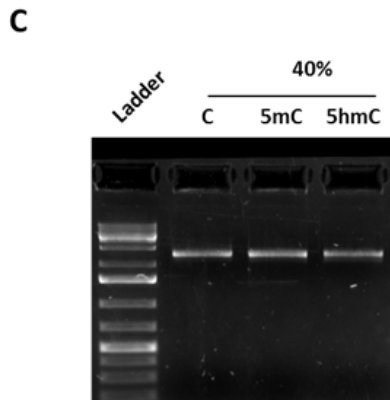


Figure 5.7 Gene body DNA modifications modulate expression in various cell types

A) Gene body DNA methylation, but not hydroxymethylation, represses transcription in HEK293 cells. B) Genic DNA modifications modulate transcription rates in mouse ESCs. C) Equal amounts and high qualities of the probes with 40% DNA modifications is demonstrated. Error bars represent the standard error of the mean of at least three independent experiments.

6.2.4. CMV, SV40 and p16, but not EEf1A1, Promoters are Sensitive to Gene Body Methylation

To investigate the role of gene body DNA methylation further, whether or not the effect could be observed for a number of promoters or was specific to the CMV promoter was investigated. To do this, the modified reporter was fused to different promoters and its activity assessed in HeLa cells (Figure 5.8). Interestingly, CDKN2A (p16) and SV40 promoters were repressed to a similar extent as the CMV promoter, indicating that this repression mechanism is likely to be universal for most promoters. To reinforce these findings, we attached the EEf1A1 promoter to a modi-

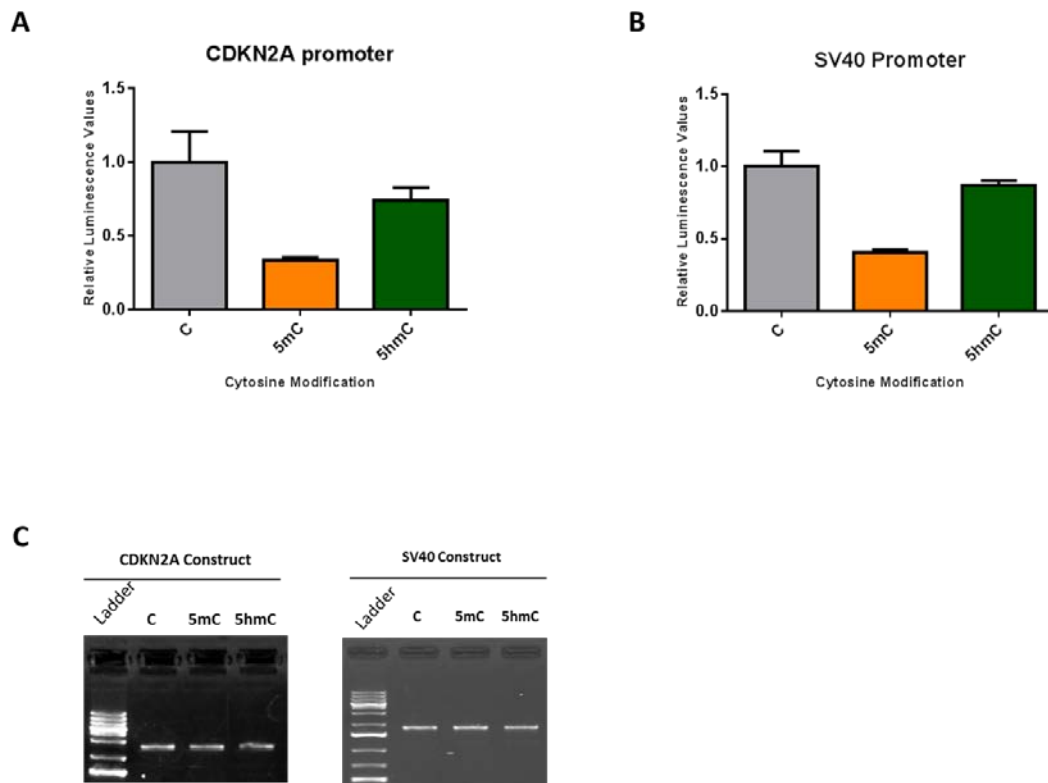


Figure 5.8 CDKN2A and SV40 promoters are modulated by gene body DNA modifications

A) CDKN2A is a mammalian promoter that is strongly repressed by gene body methylation but unaffected by genic 5hmC. B) SV40 promoter-driven expression of the luciferase reporter is downregulated by gene body DNA methylation but not 5-hydroxymethylation. C) The equal amounts and quality of the construct is demonstrated by running the 40% modification-containing constructs on an agarose gel. Error bars represent the standard error of four independent experiments.

fied *Renilla* reporter gene to enable the checking of more promoters. Surprisingly, gene body DNA methylation failed to repress the expression of the reporter fused to the *EEF1A1* promoter (Figure 5.9A). As this could have been an experimental artefact, the repression of the gene body with all the CpG dinucleotides methylated *in vitro* by M.SssI DNA methyltransferase was assessed, but this gave the same result (Figure 5.9B). Clearly, a mechanism exists that prevents the repression of *EEF1A1* promoter

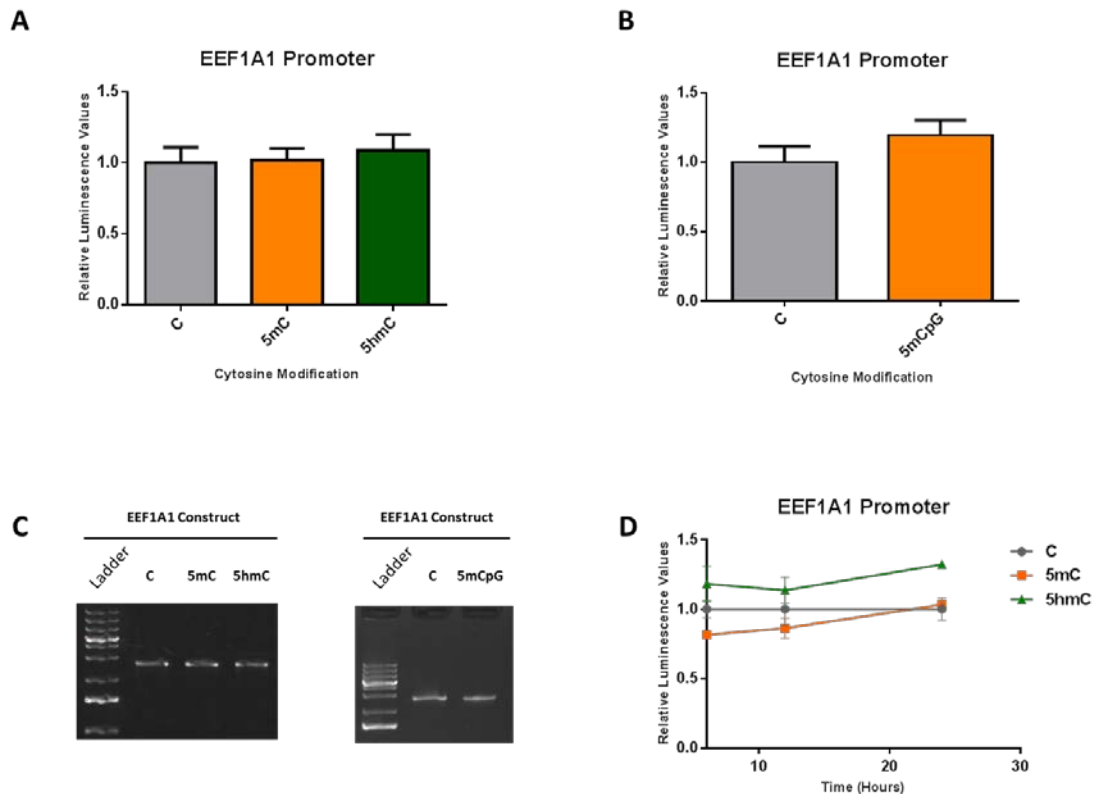


Figure 5.9 Gene body DNA modification effect on EEf1A1 promoter-driven transcription

A) 40% of modified cytosines in the genic region fused to EEf1A1 promoter do not have any effect on transcription. B) Transcription driven by EEf1A1 promoter was not affected when all CpGs of the *renilla* luciferase reporter were methylated. C) Agarose gel represents high quality and equal amounts of the transfected constructs. D) Time course experiment indicates that methylated reporter is slightly repressed when readings are taken 6 hours after transfection and then gradually reaches unmethylated levels at 24 hours after transfection. All error bars represent the standard error of the mean of three independent experiments.

activity by gene body DNA methylation. Several artificial promoters such as CMV are silenced in transgenic animals but EEf1A1 promoter is more resistant to such

silencing. To investigate this further, the activity of the modified EEF1A1 reporter was measured in a time course experiment (Figure 5.9 D). The results indicated that, even 6 hours after transfection, gene body DNA methylation showed very little repressive potential, which gradually diminishes with time. There are several reasons why the EEF1A1 promoter may behave in such a way. Firstly, it is possible that the methylated gene body is very efficiently demethylated or hydroxymethylated. This hypothesis is strongly supported by the time course experiment. Secondly, an insulator might exist that insulates the promoter from the repressive effect of the methylated gene body.

6.2.5. CTCF Does Not Play a Role in Desensitising the EEF1A1 Promoter to Gene Body Methylation

Gene body 5-hydroxymethylation and gene expression have been shown to be positively correlated in tissues and in cell lines (Mellen et al., 2012, Huang et al., 2014, Song et al., 2011a). Therefore, it was reasoned that the EEF1A1 promoter-driven methylated gene body could be efficiently hydroxymethylated by high levels of transcription. To test this hypothesis, the strengths of the different promoters that had been tested were compared (Figure 5.10). The data indicated that CMV was the strongest promoter and yet responded to gene body modifications as well as promoters with low expression levels. EEF1A1 was the second strongest promoter followed by SV40 and CDKN2A (Figure 5.10). This experiment suggests that the derepression of the EEF1A1 promoter-driven methylated construct is not transcription mediated. It was further reasoned that there might be an insulator protein that can protect the promoter from gene body-mediated repression, thus,

several candidates were selected from the literature. Sp1 and CTCF are both reported to be potent insulators and both bind to the EEF1A1 promoter (ENCODE Project).

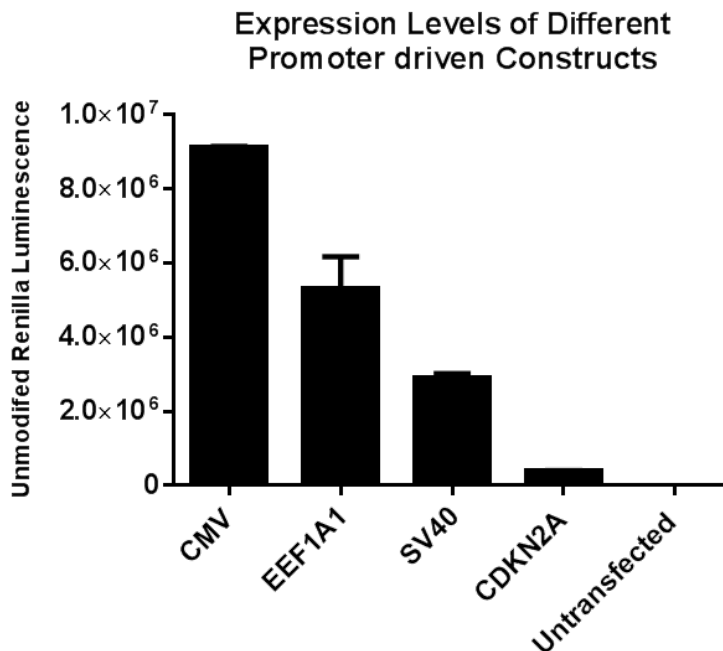


Figure 5.10 The promoter strengths of transfected reporter constructs

The strengths of different promoters used in the transfected reporter assays were compared. CMV promoter yielded the highest luminescence values while the CDKN2A promoter yielded the lowest. All error bars represent the standard error of the mean of three independent experiments.

After a thorough literature review, Sp1 was eliminated as a possible insulator because it is known to be the main driver of SV40 expression (Dynan and Tjian, 1983), which can be effectively modulated by gene body DNA modifications (Figure 5.8B). In order to check whether the activity of CTCF could drive the resistance of the EEF1A1 promoter to gene body DNA methylation, a strong CTCF consensus site was introduced to the CMV and CDKN2A promoters. The site could not be eliminated from the EEF1A1 promoter because the strong consensus site could not be mapped and promoter architecture could have been perturbed. As shown in Figure 5.11A, the strong consensus site was introduced just downstream of the CMV or CDKN2A

promoters. However, introduction of the CTCF consensus site did not change the expression pattern of the CMV or CDKN2A promoter-driven reporters (Figure 5.11C and D).

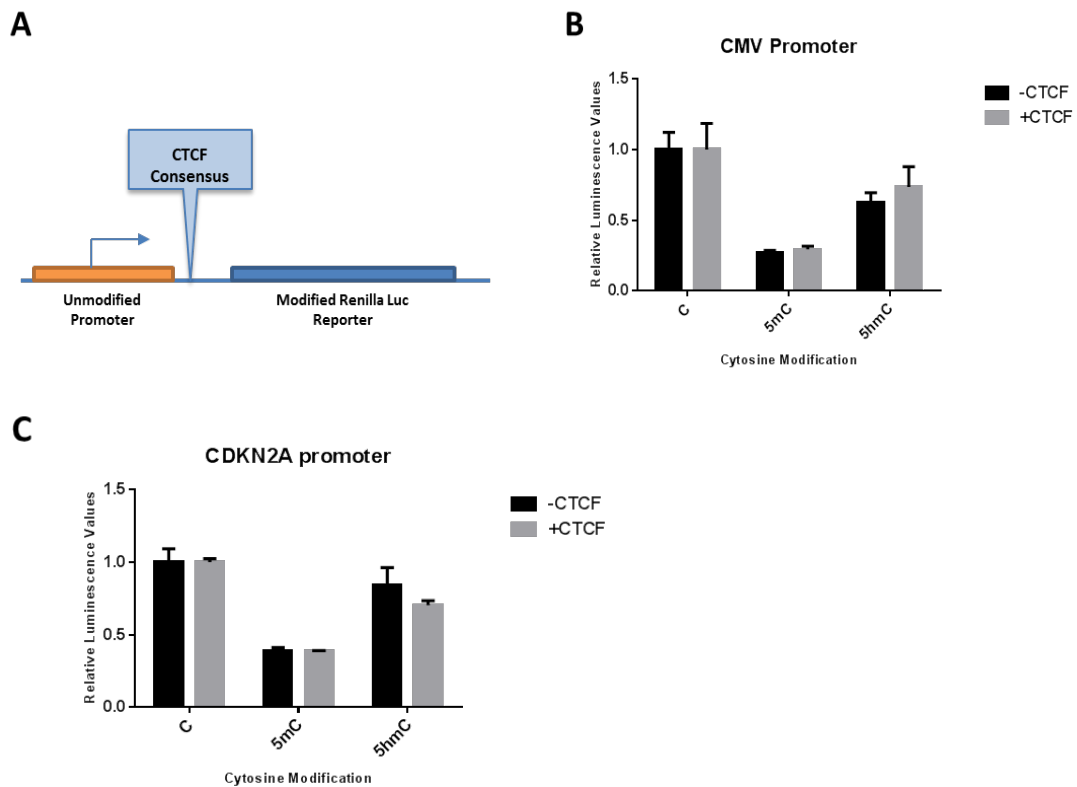


Figure 5.11 Introduction of a CTCF site does not alleviate gene body 5mC-mediated repression

A) Schematic representation of the transfected construct with a CTCF consensus site. B) Expression levels of the modified reporter gene fused to the CMV promoter with or without a CTCF consensus site. C) Expression levels driven by the CDKN2A promoter with or without CTCF consensus. All error bars represent the standard error of the mean of three independent experiments.

6.2.6. Promoter Proximal DNA Modifications Affect Transcription Initiation

Further investigation of the effects of gene body modifications on transcription was conducted using the transfected reporter system. Two different hypotheses were examined. Firstly, does genic 5mC modification interfere with transcriptional elongation and does a slower elongation rate of the polymerase result in lower levels compared to 5mC. Secondly, do gene body DNA modifications attract repressive chromatin and does this repressive activity spread into the promoter. To answer these questions, a modified region was ligated to the 5' end of the promoter and reporter construct (Figure 5.12A).

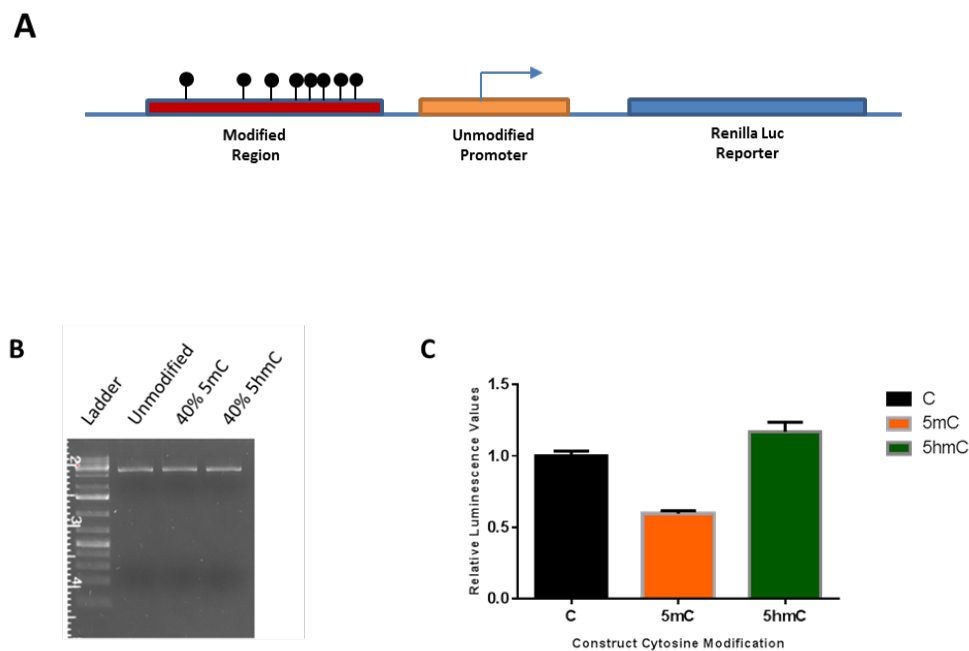


Figure 5.12 DNA methylation upstream of the promoter can modulate promoter activity

A) Schematic representation of the modified DNA region inserted upstream of the promoter. B) Equal amounts of modified DNA probes are represented on an agarose gel. C) DNA methylation but not hydroxymethylation placed upstream of the promoter can modulate CMV promoter activity in HeLa cells. All error bars represent the standard error of the mean from three independent experiments.

Upon transfection of the constructs into HeLa cells, it was noticed that a methylated region upstream of the promoter could repress reporter activity. However, the hydroxymethylated construct showed no repression (Figure 5.12C). This supports previously published results that regions containing DNA methylation can affect a promoter in *cis* (Appanah et al., 2007). Interestingly, 5hmC does not have the ability to repress the promoter at all. This indicates that hydroxylation of methylated DNA can repel repressive complexes that have a negative effect on transcription initiation. This data, however, does not invalidate the hypothesis that gene body DNA modifications have an effect on transcriptional elongation. Importantly, the levels of repression mediated by DNA methylation downstream of the promoter are compared with those upstream, we can see that downstream repression is more prominent (Figure 5.12C and 5.6A). In addition, slight repression by the hydroxymethylated promoter when it is placed downstream of the promoter can be observed. These observations indicate that gene body DNA modifications are likely to have an effect on transcriptional elongation.

6.3. Discussion

The importance of gene body methylation has been addressed in numerous studies (Appanah et al., 2007, Lorincz et al., 2004, Yang et al., 2014, Hsieh, 1994, Hsieh, 1997). In these studies, not only gene body DNA modifications have been linked to gene regulation in correlative genome-wide manner but also studies using more defined *in vitro* systems. Current models based on *in vitro* studies suggest that gene body DNA methylation acts as a repressive mark that affects transcriptional elongation and initiation if methylation is found in close proximity to the promoter (Appanah et al., 2007). Here, the role of 5hmC in the modulation of transcriptional activity has been investigated using various *in vitro* systems. Firstly, the methylated reporter system was integrated into the genome of HEK293 cells. In concordance with previously published data, the methylated reporter maintained its methylation status and showed around 7-fold repression compared to the unmodified reporter gene. Interestingly, upon deposition of 5hmC, a small but significant increase in the expression of methylated reporter was seen. The modified reporter has been separated from the promoter by 1.2 kb intron, the purpose of which is to prevent its repressive effect from spreading into the promoter. The effect of deposition of 5hmC is very similar to that of demethylation by 5-AzaC indicating that 5hmC has an effect similar to that of demethylation on transcriptional elongation. One of the drawbacks of the integrated reporter system is that it accumulates a relatively low amount of 5hmC, and this is likely to be insufficient to reverse the repressive effect of genic 5mC.

In order to address this problem, a switch was made to the transfection assays of the PCR modified reporter constructs. This system makes it possible to demonstrate that methylation over a gene body can repress transcription of the reporter in a concentration-dependant manner, however, 5hmC showed little or no repression.

These results were reproducible in different cell lines and with different promoters. Numerous studies have reported that DNA methylation over promoter regions can efficiently repress transcription. Here, it is demonstrated that the repression mediated by gene body methylation is rather mild compared to the promoter modification. Interestingly, 5hmC deposited over the promoter region can repress transcription nearly as efficiently as 5mC. This repression is likely to be mediated by blocking of the binding of transcription factors, as reported previously (Robertson et al., 2011). In the case of gene body DNA methylation-mediated repression, it could arise either from differential elongation rates of polymerase or the ability of the modified region to modulate promoter activity in *cis*, or both. Here, it is also shown that methylation of the DNA upstream of the promoter could efficiently repress initiation, however, 5hmC lacks such activity. This data implies that promoter proximal DNA modifications affect promoter activity rather than RNAPII elongation rate. Interestingly, 5hmC is found in the regions flanking active promoters and enhancers in neuronal tissues as well as in mouse embryonic cells (Lister et al., 2013, Yu et al., 2012). The observations described in this study pose the possibility that these regions are likely to be marked to prevent the spreading of repressive chromatin into these regions, as previously reported for DNA methylation (Appanah et al., 2007).

Interestingly, not all of the promoters responded to gene body DNA methylation-mediated repression. *EEF1A1* promoter is one of the strongest promoters in humans and the observation that it cannot be repressed leads to several hypotheses. Firstly, the gene body could be very quickly and efficiently demethylated, however, this demethylation activity would be independent of promoter strength. Secondly, an insulator could exist that efficiently prevents the spread of repressive chromatin into

the promoter. CTCF has been shown to have insulating activity by others (Herold et al., 2012), but the introduction of the CTCF site to the CMV or CDKN2A promoter did not reverse gene body methylation-mediated repressive activity in this system. The presence of an insulator would suggest that gene body DNA modifications do not alter the rate of transcriptional elongation. However, these questions remain to be addressed. The evidence presented so far strongly suggests that DNA methylation, but not hydroxymethylation, deposited over gene bodies can efficiently repress promoter activity and, possibly, the elongation rate of RNAPII. However, the mechanistic details of this mechanism remain to be determined.

Chapter 6

The Molecular Mechanism of Gene Body DNA Modification-Mediated Transcriptional Regulation

6.1. Introduction

Promoter DNA methylation-mediated effects on transcription have been thoroughly investigated and the major molecular mechanisms of repression established (Klose and Bird, 2006). There are two mechanisms by which promoter DNA methylation mediates transcriptional silencing. First, DNA methylation directly interferes with transcription factor binding to sites that are methylated. Second, methyl-CpG binding proteins (MBPs) bind to the methylated sequences and recruit factors such as HDACs, which establish closed and repressive chromatin (Aygün et al., 2013).

Recent studies have reported that gene bodies show highly dynamic DNA methylation patterns during brain development (Lister et al., 2013, Wen et al., 2014) and in other biological systems (Tsagaratou et al., 2014). Gene body DNA methylation has been proposed to be a transcriptional activator, based on the interrogation of transcription profiles after the treatment of cancer cells with the DNA demethylating agent 5-azacytidine (Yang et al., 2014). This downregulation of the genes after 5-azaC treatment could be caused by demethylation and activation of antisense transcripts, microRNAs or endogenous retroelements that would negatively affect transcription (Roulois et al., 2015). However, other studies using a reporter gene approach have concluded that gene body methylation silences transcription via gene body histone deacetylation and methylation (Appanah et al., 2007, Lorincz et al., 2004), although the exact mechanism of this regulation has not been reported. In addition, 5hmC has been shown to be enriched over gene bodies of highly expressed genes (Mellen et al., 2012, Huang et al., 2014). In this study, it has been shown that 5hmC deposited over gene bodies can upregulate gene expression by enhancing transcription initiation and

possibly transcriptional elongation. However, the exact mechanism of how this regulation is achieved remains unknown.

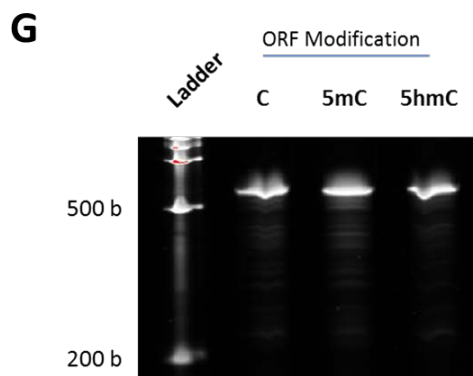
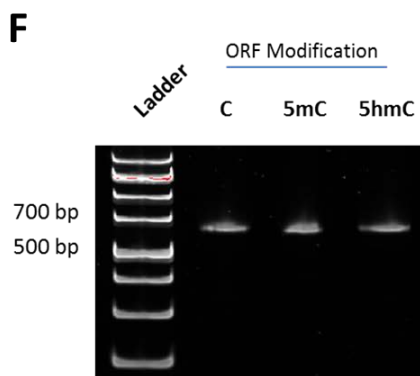
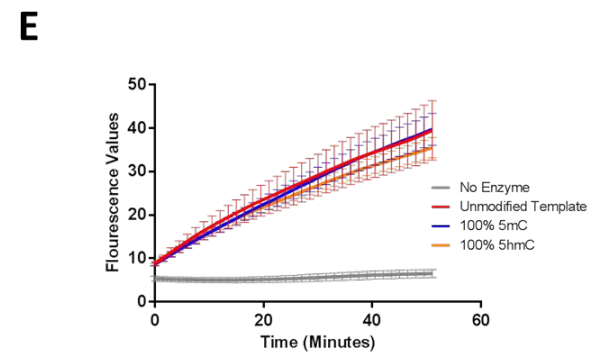
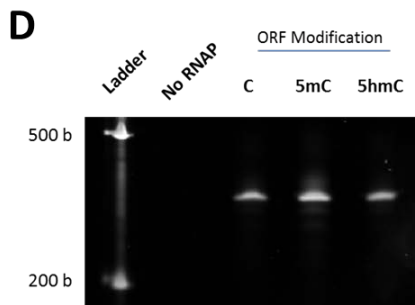
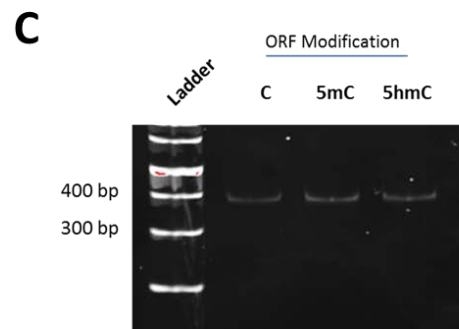
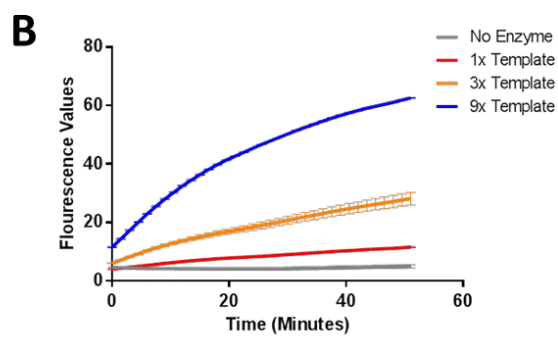
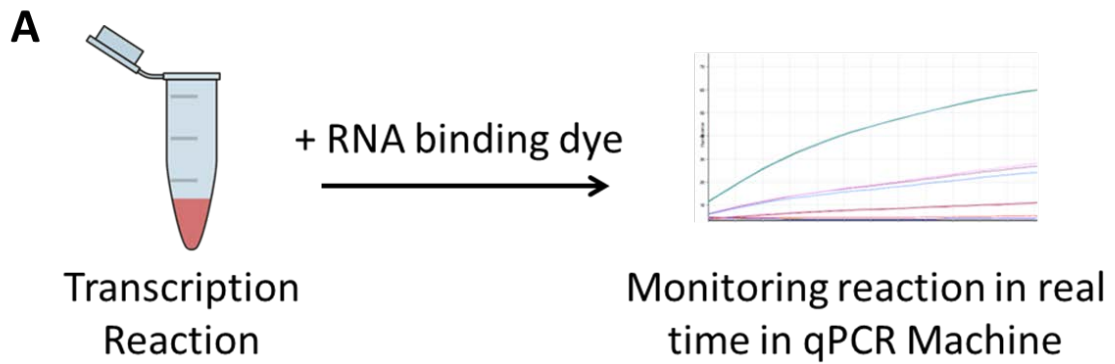
In the brain and ESCs, a large amount of 5hmC is deposited in promoter and enhancer proximal regions (Wen et al., 2014, Yu et al., 2012). Furthermore, methyl-CpG protein binding has been shown to be incompatible with hydroxylation of 5mC (Spruijt et al., 2013). Based on these reports, we reasoned that 5hmC could be displacing MBPs from promoter proximal regions and facilitating transcriptional initiation. In addition, a recent study reported that MeCP2, which binds to 5mC, represses the transcription of very long genes (Gabel et al., 2015), possibly by posing a block to transcriptional elongation. Alternatively, it has been previously shown that DNA methylation has the ability to stabilise the DNA duplex and increase the temperature required to melt it (Thalhammer et al., 2011). This stabilisation has been shown to be efficiently reversed by hydroxylation of 5mC. Since transcriptional machinery is required to melt double stranded DNA, it was reasoned that an increased melting temperature of methylated DNA could result in a slower RNA Polymerase II (RNAPII) elongation rate and subsequent downregulation of gene expression. In this chapter, testing of these hypotheses to determine the molecular mechanism of gene body DNA modification-mediated transcriptional regulation is discussed.

6.2. Results

6.2.1. T7 RNAP Elongation Rates through Modified Templates

To investigate the effect of modified DNA melting temperature differences on transcription, an *in vitro* T7 RNAP assay was developed. This strategy was chosen for several reasons. Firstly, this simple system allows the precise detection and quantification of the differences in transcription from different DNA templates. Secondly, all of the components in the assay are known and are only required for the transcription reaction. The lack of additional factors eliminates the possibility of any indirect effects on transcription. Finally, T7 RNAP uses the same elongation mechanism as multi-subunit RNAPs, where the energy to melt the DNA duplex is taken from triphosphate hydrolysis and hydrogen bond formation between NTPs and the DNA template in the active site (Steitz, 1998, Steitz, 1999, Cramer, 2002). However, the drawbacks of this assay are that the templates are not packaged into chromatin and that T7 RNAP is structurally different from the multi-subunit eukaryotic RNA polymerases.

In order to accurately and quantitatively measure the synthesis rate of RNA from differentially modified templates, the assay was designed to monitor the production of RNA in real time. This assay is based on monitoring the fluorescence of an RNA binding dye that binds to synthesised RNA using a qPCR instrument (Figure 6.1A). Different amounts of unmodified template were used to determine the sensitivity of the system (Figure 6.1B). Based on the calibration experiment, it was concluded that less than a two-fold difference in RNA production rate could be detected. Furthermore, it could be seen that the amount of RNAP was saturated in this reaction due to the large excess of polymerase compared to the template.



H

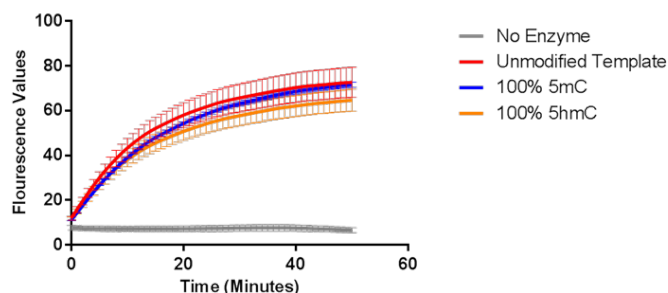


Figure 6.1 T7 RNAP-driven transcription through methylated or hydroxy-methylated templates

A) Schematic representation of transcription assay and monitoring of transcription using a qPCR instrument. B) Calibration experiment of *in vitro* transcription of different amounts of DNA template. C) Transcription template: 315 bp modified region ligated to unmodified T7 promoter sequence. D) RNA product produced in transcription reaction run on a denaturing acrylamide gel. E) RNA synthesis graph recorded by the qPCR instrument. F) Modified transcription template: 550 bp transcribed region ligated to unmodified T7 promoter. G) The final transcription product produced from a 550 bp fully modified template. H) The graph of RNA production over time depicted by detecting the fluorescence of RNA binding dye from the qPCR instrument. All error bars represent the standard error of the mean of 3 independent experiments.

In order to compare the transcription of differentially modified templates, DNA constructs were assembled that had the unmodified T7 promoter and a modified 315 bp transcribed region designed based on a physiological template (a fragment of *BDNF* promoter), or a *Zdhh2* cDNA based 550 bp template, which has been also used in the other *in vitro* experiments. Templates were quantified carefully and their integrity assessed by agarose gel analysis (Figure 6.1C,F). Contrary to our hypothesis, measuring the transcription rates showed no difference between different DNA modifications (Figure 6.1E,H). After the reaction, the RNA products were DNase treated and run on a denaturing acrylamide gel (Figure 6.1D,G). The results demonstrated that the same amounts of transcript were synthesised in all of the reactions. This indicates that T7 RNAP does not show any preference for different template DNA modifications.

6.2.2. *In vitro* Transcription in HeLa Nuclear Extracts Using Modified DNA Templates

T7 RNAP transcription *in vitro* does not show an ability to discriminate between differentially modified templates. However, it is possible that T7 RNAP has evolved in such a way that DNA melting temperature does not have any effect on the rate of transcription. In addition, the crowded environment in the cell nucleus is very different compared to the T7 RNAP reaction. In the nucleus, chromatin, transcription factors and many other components interact with the transcriptional machinery and DNA in various different ways. Therefore, the transcription of differentially modified DNA templates was investigated in an environment that better resembled the physiological state, via *in vitro* transcription in HeLa nuclear extracts. Transcription in nuclear extracts is a well-established method to measure transcriptional efficiency and transcription-coupled events such as splicing (Rooke and Underwood, 2001). The method is based on the extraction of HeLa nuclear contents, which contain active transcription machinery. Then DNA templates are supplied together with radioactively labelled NTPs. The product of transcription is visualised on a phosphor screen. Importantly, careful experimentation allows quantification of the RNA products (Hawley and Roeder, 1987). In order to assess the transcription efficiency of differentially modified constructs, DNA modifications were introduced by PCR and PCR products were ligated to the CMV promoter. A number of templates were synthesised ranging from artificial, such as the *Renilla* luciferase reporter, to semi-natural, such as the cDNA of endogenous transcripts. A 315 bp control template was also introduced into the reaction mix in order to normalise for unequal loading and other losses of RNA product. Interestingly, it was observed that the methylated *Renilla* template was transcribed less efficiently compared to the unmethylated or

hydroxymethylated template. However, it is important to note that the absolute differences were only around 30% (Figure 6.2). Based on these data, it

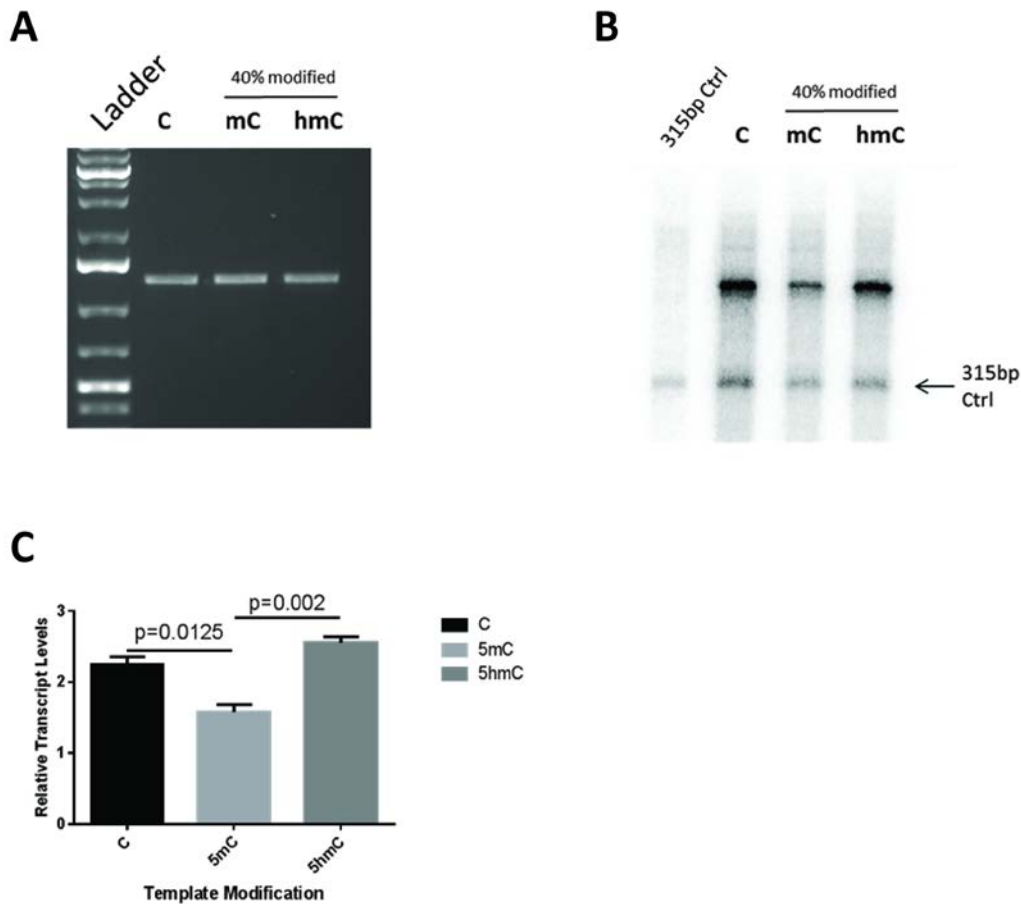


Figure 6.2 Differential DNA modifications of gene body modulate transcription efficiency in HeLa nuclear extracts

A) Agarose gel image representing the same amount of differentially modified DNA templates. B) Auto-radiogram of full-length RNA transcripts, 315 bp loading control indicated. C) Quantification of the full-length transcript produced in the transcription reaction. The signal of the transcript has been normalised to a 315 bp control template. Error bars represent the standard error of the mean of three independent experiments.

can be concluded that transcription efficiency in HeLa nuclear extracts is affected by gene body DNA modifications. However, it was unclear whether these differences arose from an effect on transcriptional elongation or initiation. In order to investigate

whether the effect was template independent, different DNA templates were incubated in HeLa nuclear extracts. The Cyclin D3-based template provided the same pattern of transcription efficiencies, with the 5mC template producing less end-point RNA compared to the unmodified or 5hmC templates (Figure 6.3 A). This indicated that the effect on transcription was specifically mediated by DNA modification in the transcribed region.

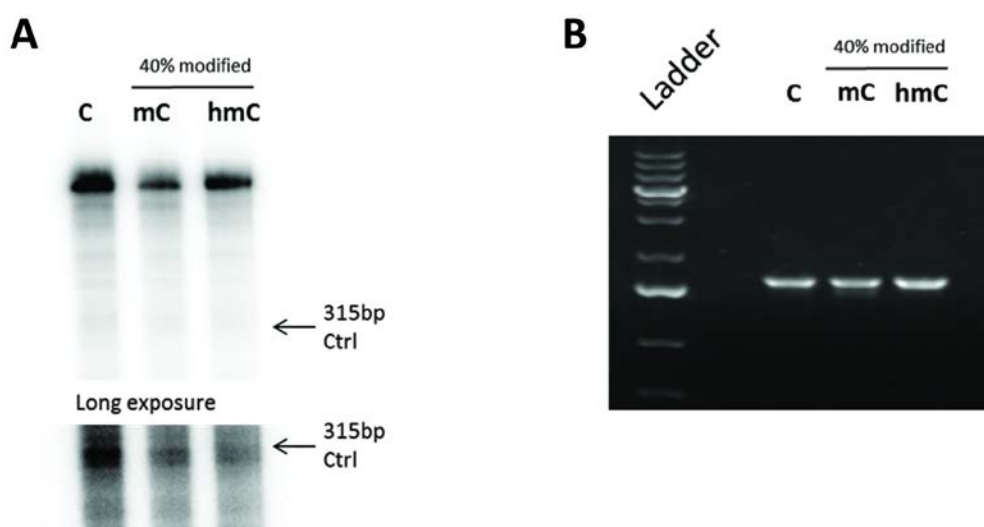


Figure 6.3 DNA modifications modulate the transcription of a Cyclin D3 cDNA-based template

A) Autoradiogram representing the transcription product of the fragment of Cyclin D3 cDNA. B) Agarose gel representing the DNA templates used in the transcription reaction.

Interestingly, when *in vitro* transcription was performed on the cDNA of the *Zdhhc2* gene, an additional signal was observed that indicated the production of a well-defined, shorter RNA of around 100 bp (Figure 6.4). The expression of the transcript was reduced in the constructs with 10% DNA modification and completely abolished when all cytosines were modified. This is in agreement with the transfected

reporter assays described in earlier chapters and data published previously (Robertson et al., 2011) that hydroxymethylation can suppress transcription initiation as efficiently as DNA methylation. Taken together, these data opens a

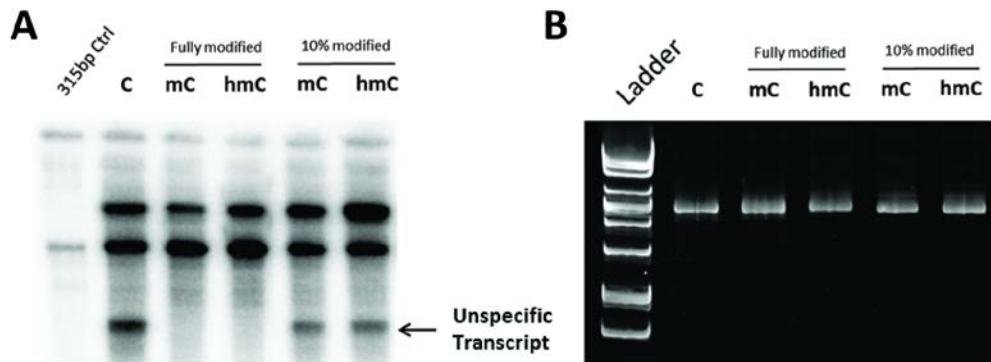


Figure 6.4 Genic 5mC and 5hmC inhibit unspecific transcription initiation in HeLa nuclear extracts

A) Autoradiogram of the RNA products of *in vitro* transcription from Zdhhc2 cDNA are shown. The unspecific transcript is marked by an arrow. The middle size transcript represents a spike in the 315 bp control template. B) An acrylamide gel showing the transcribed templates.

possibility that both 5hmC and 5mC deposited over transcribed region could prevent the initiation of transcription. In addition, genic 5mC has a mild repressive effect, while gene body 5hmC lacks such activity. However, the origin of this repressive activity remains ambiguous because the templates in the nuclear extracts were incubated for 60 minutes, which essentially gave an end-point measurement or the general efficiency of transcription, which could be affected by reduced elongation and/or initiation.

6.2.3. MeCP2 and Modulation of Gene Expression via Gene Body DNA Modifications

Histone acetylation is a well characterised mark associated with transcription initiation (Guenther et al., 2007) and has also been shown to facilitate transcriptional elongation (Lorincz et al., 2004). In addition, MBD family proteins have been demonstrated to repress transcription by recruiting of HDACs to methylated regions and establishing repressive chromatin (Klose and Bird, 2006). Therefore, it was hypothesised that genic DNA modifications could lead to establishing differential histone acetylation signatures and that this would result in altered construct expression levels. To test this, differentially modified reporter constructs (described in Chapter 4) were transfected into the cell lines in addition to Trichostatin A (TSA) treatment. TSA is a broad range HDAC inhibitor that has been shown to upregulate the expression of repressed genes (Yoshida et al., 1990). As expected, expression of modified reporter genes was upregulated to a greater extent than unmodified reporters (Figure 6.5). This indicates that DNA methylation, and to some extent DNA hydroxymethylation, downregulates gene expression via recruitment of HDACs.

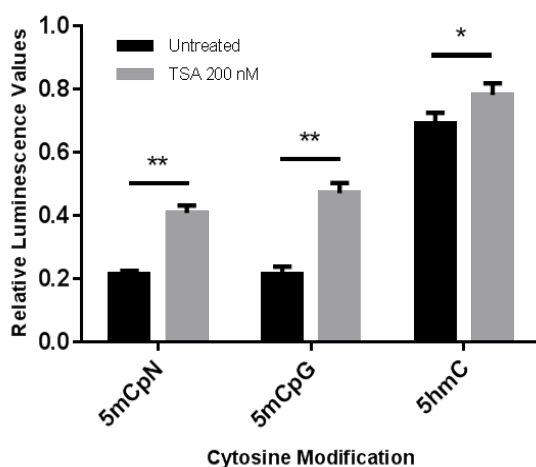


Figure 6.5 Gene body DNA modifications mediate transcriptional regulation via histone acetylation

TSA treatment of cell lines transfected with differentially modified reporters indicates that inhibition of HDACs can alleviate gene body methylation- and hydroxymethylation-mediated repression. The data were normalised to unmodified reporter expression values. Error bars represent the standard error of the mean of four independent experiments.

The hypothesis that DNA methylation-mediated repression is mediated by histone deacetylation through recruitment of MeCP2 has been previously proposed by others (Nan et al., 1998). In addition, MeCP2 has been reported to bind hydroxymethylated cytosine (Mellen et al., 2012). Considering that this binding could be the cause of the slight downregulation of the hydroxymethylated construct, MeCP2 binding to modified DNA templates was investigated in a more detail via an EMSA experiment, and the possible binding sites for MeCP2 were interrogated (Figure 6.6). Interestingly, it was observed that MeCP2 could not only bind to methylated DNA in a CpG context but also to 5mC and 5hmC in a CpA context. Because DNA modifications in the reporter constructs were introduced by PCR, they also contained 5mC and 5hmC in a CpA context. Therefore, it was reasoned that downregulation of expression from modified reporters might be mediated via binding of MeCP2 and that elimination of MeCP2 would alleviate genic DNA methylation-mediated repression.

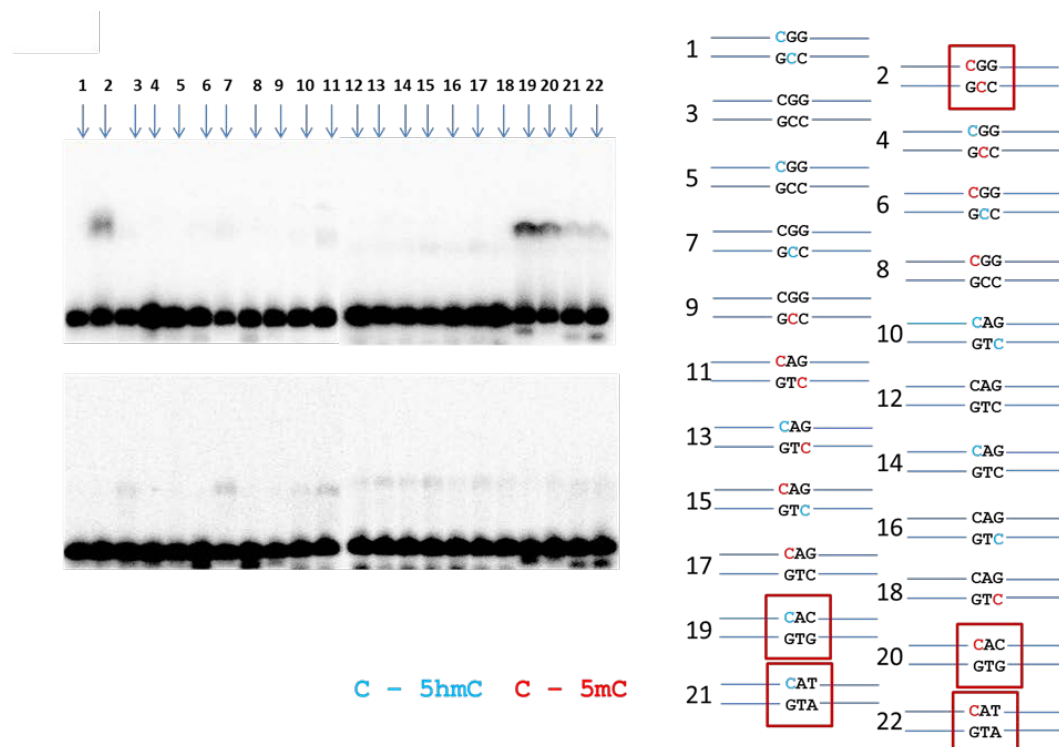


Figure 6.6 MeCP2 can bind to methylated and hydroxymethylated cytosine in a non-CpG context

Autoradiogram representing MeCP2 binding to radioactively labelled probes. The top gel has MeCP2 added and the bottom gel represents reactions with no MeCP2. Probe sequences and modifications are represented on the right panel. 5hmC in the probes is marked blue and 5mC is red.

In order to test the hypothesis that MeCP2 regulates transcription via bridging gene body DNA modifications to HDACs, MeCP2 was knocked out in HEK293 cells, which contain inducible TET catalytic domain, using CRISPR technology. After designing gRNAs to bind at the second and third exons of the gene (Figure 6.7A) a number of cell lines were successfully generated (Figure 6.7B). As shown in the western blot, clone 2B and clone 2C were a full knockout and heterozygous for *MeCP2*, respectively. However, the clone 2B lane seemed to be slightly underloaded (Figure 6.7B). In addition, clone 1A showed a very faint band for MeCP2, but this could be just a technical artefact. A previous study has suggested that MeCP2 may

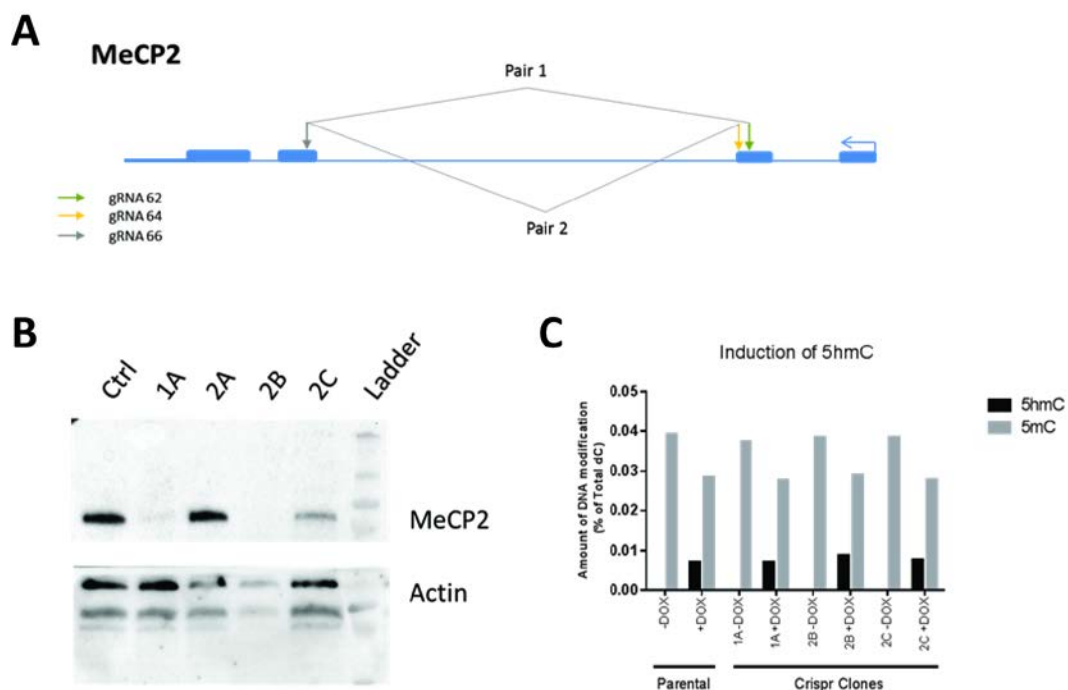


Figure 6.7 CRISPR-mediated knockout of MeCP2 protein in HEK293 T_{REX} TET-CD cells

A) Schematic diagram representing the MeCP2 gene and the binding sites for gRNAs. Pair 1 or Pair 2 of gRNAs was used to generate knockout cell lines. B) Immunoblot of the MeCP2 protein in the CRISPR clones. MeCP2 is not present in clones 1A and 2B. 2C is likely to have lost one copy of MeCP2. C) Quantification of genomic levels of DNA modifications before and after treatment with doxycycline which induces TET1 catalytic domain.

bind to methylated DNA and protect it from hydroxymethylation (Szulwach et al., 2011). Therefore, efforts were made to confirm that selected clones had maintained the ability to deposit 5hmC upon overexpression of TET-CD as efficiently as the parental cell lines. As indicated in Figure 6.7C, *MeCP2* deletion or downregulation did not change global DNA methylation levels nor affect the deposition of 5hmC by TET-CD.

Next, it was investigated whether MeCP2 plays a role in the repression of endogenous genes, which were identified as being upregulated upon induction of TET-CD (Chapter 4). To test this, the ability of parental and *MeCP2* KO cells to upregulate candidate genes upon deposition of 5hmC was compared. As mentioned previously, MeCP2 was convincingly demonstrated to repress long genes (Gabel et al., 2015). Therefore, three candidate genes were chosen that have different lengths: 10 kb, 20 kb and 100 kb for *GATA2*, *ZFYVE21* and *SOGA1*, respectively. All three genes were expressed and had highly methylated gene bodies (Chapter 4). As shown in Figure 6.8A *GATA2* gene expression levels were upregulated in *MeCP2* KO compared to parental cells. Although this indicated that MeCP2 was repressing the expression of *GATA2* to some extent, *GATA2* was still efficiently upregulated upon overexpression of TET-CD (Figure 6.8A). This data suggests that *GATA2* upregulation is either indirect or that MBD family proteins could be acting in a redundant fashion. Interestingly, other candidates identified to be upregulated in the genome-wide expression study showed marginal downregulation in *MeCP2* KO cells (Figure 6.8B,C). This suggests

that MeCP2 might act as an activator for these genes. In addition, *ZFYVE21* and *SOGA1* could still be efficiently upregulated upon DOX induction. Although it seems that MeCP2 is not a major player in genic DNA modification-mediated transcriptional regulation, these results are hard to interpret because knockout clones could have adapted transcriptional programmes that differ from parental clones or observed gene deregulation could be indirect. Therefore, the ability of these cells to upregulate the integrated reporter construct (described in Chapter 5) upon overexpression of TET-CD was tested (Figure 6.8D). Surprisingly, the results indicated that the reporter gene could be upregulated to the same extent as the reporter in the cell line that had MeCP2 present. This further supports the observations that MeCP2 is not the main modulator of gene expression via gene body DNA modifications in HEK293T cells.

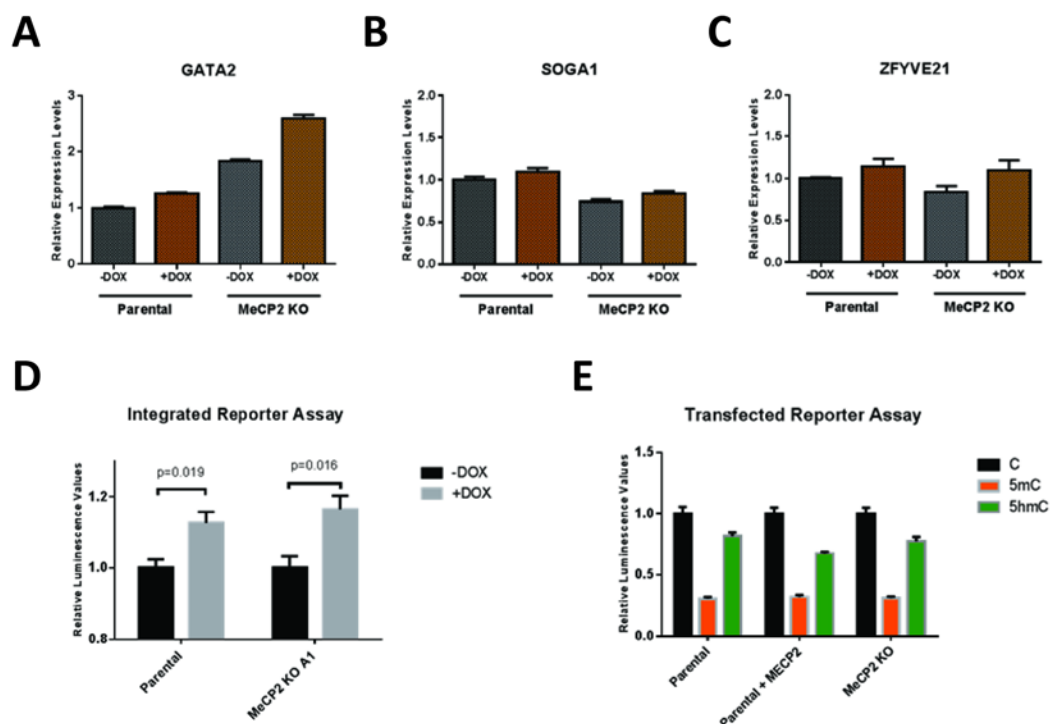


Figure 6.8 Changes in gene expression in *MeCP2* KO Cells

A, B, C) Expression levels of the genes in parental and *MeCP2* KO cells as detected by RT-qPCR. Both cell lines were also treated with DOX to induce the TET-CD. D) Expression levels of integrated reporter upon induction of the TET-CD in *MeCP2* KO and parental cell lines. E) Expression levels of transfected differentially modified reporters in parental and *MeCP2* KO cells and parental cells overexpressing MeCP2.

All error bars represent the standard error of the mean from at least three independent experiments.

Finally, the expression of transfected reporters was investigated in the *MeCP2* KO cell line. As shown in Figure 6.7E there was no difference in the expression levels of methylated reporter in parental, *MeCP2* KO or *MeCP2* overexpressing cell lines. Interestingly, in *MeCP2* overexpressing cells the expression levels of the hydroxymethylated reporter seemed to be reduced to some extent. Taken together, these results indicate that *MeCP2* is not the major player that modulates gene expression via gene body DNA modifications, or its role is redundant with other members in the MBD protein family.

6.2.4. MBD1 and Modulation of Gene Expression via Gene Body DNA Modifications

MBD1 is another member of the MBD family of proteins that has been shown to be involved in transcriptional repression (Fujita et al., 2003, Ichimura et al., 2005). Interestingly, MBD1 not only has a methyl-CpG binding domain but also possesses three CXXC domains, although not all of them have been demonstrated to bind DNA (Jorgensen et al., 2004). In order to check whether MBD1 plays a role in gene body-mediated repression, knockout of *MBD1* was attempted in the TRex HEK293 TET-CD cell line. As shown in Figure 6.9A, gRNAs were designed in the same fashion as for previous *MeCP2* knockout experiments. Three knockout clones were obtained, denoted clone 3B, 3C and 3D (Figure 6.9B), which were used in further analysis.

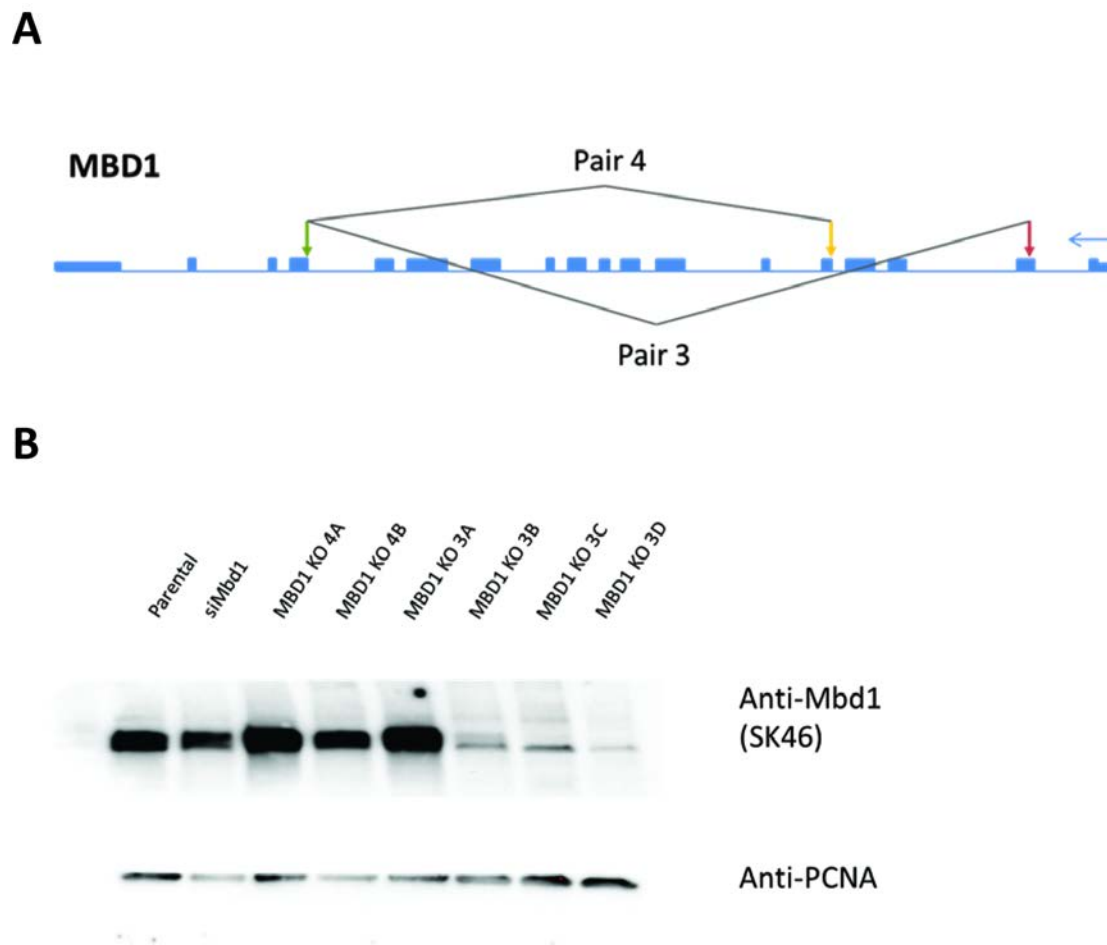


Figure 6.9 CRISPR-mediated knockout of MBD1 in HEK293 TRex TET-CD cells
 A) Schematic diagram representing the MBD1 gene and the binding sites for gRNAs. Pair 3 or Pair 4 of gRNAs was used to generate the knockout cell lines. B) Immunoblot of the MBD1 protein in the CRISPR clones. MBD1 is absent in clones 3B, 3C and 3D. Clone 4B is heterozygous for MBD1.

Firstly, the expression levels of the genes upregulated by induction of the TET-CD were compared between parental and knockout clones. Interestingly, both *GATA2* and *ZFYVE21* were upregulated in two different *MBD1* KO clones compared to the parental cells (Figure 6.10 A,B). This suggested that MBD1 plays a role in repressing these genes. However, the relative expression levels between the KO clones varied to some extent, indicating that selected clones had rather different gene expression

profiles. In addition to the upregulation of gene expression in *MBD1* KO clones, *GATA2* and *ZFYVE21* were still upregulated by expression of the TET-CD. Therefore, just by looking at the extent of the upregulation, it could be concluded that *MBD1* KO cells upregulated these genes less efficiently compared to parental or MeCP2 KO cells

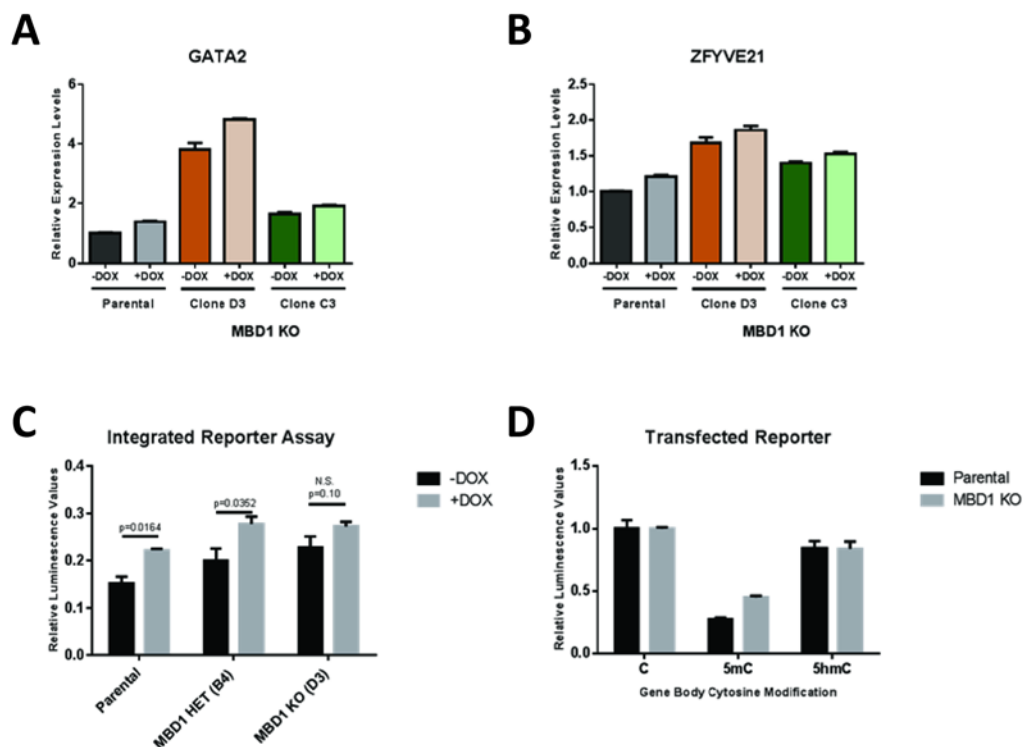


Figure 6.10 Changes in gene expression in MeCP2 KO Cells

A, B) Expression levels of *GATA2* and *ZFYVE21* in parental and two independent *MBD1* KO cell lines. The expression levels were determined by RT-qPCR. All cell lines were also treated with DOX to induce the TET-CD. C) Expression levels of integrated reporter upon induction of TET-CD in parental *MBD1* heterozygous and *MBD1* KO cell lines. E) Expression levels of transfected differentially modified reporters in parental and *MBD1* KO cells. All error bars represent the standard error of the mean of at least three independent experiments.

(Table 6.1). However, more clones would be needed to show that these observations are statistically significant. Taken together, these findings suggest that *MBD1* is likely to play a role in the regulation of gene expression via gene body DNA modifications.

Gene	Increase in expression upon 48 hour DOX treatment			
	Parental	MeCP2 KO	MBD1 KO C3	MBD1 KO D3
GATA2	38.9 %	42.1 %	17.1 %	25.7 %
ZFYVE21	21 %	30.5 %	9.1 %	10.7 %

Table 6.1 Changes in candidate gene expression upon DOX induction

To further support MBD1's involvement in gene body DNA modification-mediated transcriptional regulation the expression levels of the integrated reporter were tested. Methylated luciferase reporter activity was investigated in the parental, *MBD1* heterozygous and *MBD1* KO cell lines (Figure 6.9 C). Interestingly, a gradual upregulation of methylated reporter was observed in heterozygous and in KO cell lines. This suggests that MBD1 is likely to play a role in gene body DNA methylation-mediated repression. Notably, upon induction of the TET-CD, efficient upregulation of the reporter could be seen in parental cells, slightly less efficient induction in cells heterozygous for MBD1 and no significant induction in *MBD1* KO cell line. This data suggests that 5hmC-mediated upregulation is at least partially driven by the displacement of MBD1. To strengthen this hypothesis, differentially modified reporters were transfected into MBD1 KO cells (Figure 6.10D). Methylated reporter was significantly de-repressed in KO compared to parental cells, while no change was observed in the expression of the reporter with a hydroxymethylated gene body. These findings imply that MBD1 is a regulator of gene body DNA modification-mediated regulation of gene expression. However, it seems that it is not the sole player in this mechanism and more work is needed to completely understand genic DNA modification-mediated transcriptional regulation.

6.3. Discussion

The role of gene body DNA methylation in transcriptional regulation has been the topic of research for several decades and different studies have suggested that it could act as an activator (Yang et al., 2014) or repressor of gene expression (Hsieh, 1994, Hsieh, 1997, Lorincz et al., 2004). Although it is well established that DNA methylation over promoter regions effectively represses transcription, there is less consensus regarding the role of gene body DNA methylation. It does not seem unreasonable that the same repressive mechanisms that act on promoters could also act on gene bodies. However, because of the different biological functions of promoters compared to gene bodies, DNA modification-mediated regulation could also be rather different. Previous results (discussed in Chapter 4 and 5) indicate that the transition from a methylated to hydroxymethylated gene body upregulates gene expression but the mechanism of this regulation was not proposed. In this chapter, several potential mechanisms of gene body DNA modification-mediated transcriptional regulation were investigated.

It has been reported previously that DNA methylation can increase the stability of the DNA duplex compared to unmodified or hydroxymethylated cytosine (Thalhammer et al., 2011). Because DNA duplex melting is the core event in transcriptional elongation, it was reasoned that the increase in energy required to melt the DNA duplex would negatively affect transcription. To test this hypothesis, a sensitive assay of T7 RNAP transcription efficiency through modified DNA templates was designed. However, T7 RNAP showed no preference for any of the differentially modified templates used. This is highly surprising because all of the cytosines in the templates were modified, which should have significantly altered the DNA melting

temperatures. However, an explanation for this observation could lie within the polymerisation reaction itself. In order to melt DNA, RNA polymerases get energy from two major sources. Firstly, from formation of the duplex between DNA and the newly synthesised RNA strand, as the DNA-RNA duplex is more stable than the DNA-DNA duplex, which could compensate for the increased energy barrier mediated by DNA methylation. Secondly, and probably more likely, DNA melting and polymerase translocation is mainly mediated by the energy released upon hydrolysis of the phosphodiester bonds of incoming NTPs. The amount of energy released during hydrolysis is the main driver of polymerase movement and exceeds the energy required for melting of the DNA duplex nearly 10 fold (Thalhammer et al., 2011, Kamerlin and Warshel, 2009). Therefore, the mild increase in DNA melting temperature is unlikely to affect the transcription process. Furthermore, DNA melting temperature increases proportionally with the amount of DNA methylation (Thalhammer et al., 2011). However, densely methylated regions of DNA are not that common in the genome, although RNA Pol II encounters densely methylated regions known as intragenic CpG islands during the production of a transcript.

Although *in vitro* transcription by T7 RNAP can address the question of DNA melting, T7 RNAP is structurally very different from multi-subunit RNAPs such as mammalian RNAPII. Of interest was whether modified bases can form a structural block for transcription, as has been demonstrated for 5-carboxylcytosine in the DNA (Wang et al., 2015), or if transcription elongation rates can be modulated by transcription factors binding to genes. It has been shown previously that MeCP2 can reduce the efficiency of transcription of very long genes by a currently unknown mechanism (Gabel et al., 2015). This raises the possibility that transcription factor binding could pose a block for transcription elongation. To address this question, *in*

vitro transcription was performed in HeLa nuclear extracts. Interestingly, the methylated gene body construct was transcribed less efficiently compared to unmodified or hydroxymethylated templates. Because the end-point effect was being measured in this experiment, the effect on transcription could have been caused by gene body DNA modifications interfering with transcription initiation and/or elongation. In addition, proteins such as MBPs could possibly bind to and interfere with transcriptional initiation or elongation. One would expect that, if transcription factor binding is interfering with transcriptional elongation, the unmodified template would be the most repressed. However, it was not the case. In addition, an increased smear of RNA below the full length transcript was not observed, further indicating that an effect on transcriptional elongation is unlikely. In order to decouple the effect of initiation from transcriptional elongation, pulse-chase experiments in nuclear extracts could be performed (Zhu et al., 1997). However, the analysis of expression in HeLa nuclear extracts described here yielded the interesting observation that DNA methylation, as well as hydroxymethylation, can suppress the expression of unspecific transcripts in a DNA modification density-dependant manner. DNA methylation has previously been proposed to be the mechanism for the repression of retroelement expression (Karimi et al., 2011), which opens up the possibility that 5hmC can also efficiently repress spurious transcription.

In the mammalian cell, transcription is complex and has to be considered in the context of chromatin. It has been well documented that histone acetylation is associated with expressed genes and HDACs can efficiently repress gene expression (Guenther et al., 2007). Based on this evidence, efforts were made to determine whether gene body 5mC-mediated repression can be mediated by histone deacetylation. Cells were treated with the broad range HDAC inhibitor Trichostatin A

(TSA). It seems that TSA treatment can efficiently upregulate expression from methylated genes compared to those that are hydroxymethylated or unmethylated. Based on this data, it can be concluded that gene body DNA methylation-induced repression is at least partially mediated by histone deacetylation. Interestingly, HDAC enzymes were found to interact with MBD family proteins. MeCP2, MBD1 and MBD2 were all shown to interact with HDACs (Section 1.5.1). MBD family proteins have been shown to play a role in silencing methylated promoters via recruitment of histone modifying enzymes, therefore, it was reasoned that MBD family proteins could influence gene expression via histone deacetylation. To address this question, two MBP family members (MeCP2 and MBD1) were knocked out. These genes have been previously reported to act as repressors of transcription. The data described here suggests that MeCP2 does not seem to play a role in gene body-mediated repression in HEK293 cells, although MeCP2 may play a role when present at very high levels such as in neuronal cells (Skene et al., 2010) or affect the expression of only very long genes (Gabel et al., 2015). In order to address this, the regulation of the *SOGA1* gene was investigated, which is 100 kb long, and found to be slightly downregulated in the *MeCP2* KO cell line although the analysis of more clones is needed for confident conclusions to be drawn. *GATA2* was the only gene that showed upregulation in *MeCP2* KO cells. *GATA2* has extensive gene body and promoter proximal DNA methylation compared to other genes which could suggest that MeCP2 is a weak repressor. In addition, MeCP2 did not affect the expression of transfected reporter genes or an integrated reporter, which suggests that it does not play a major repressive role at least in this cell type. Alternatively, MeCP2 function could be redundant to other MBP family proteins. To test this hypothesis, one would need to knock out all of the family members together in combinatorial manner.

MBD1 is another member of the MBP family of proteins that, in addition to a methyl-CpG binding domain, has several CXXC domains that have been reported to bind unmethylated CpGs (Jorgensen et al., 2004). In order to check whether MBD1 plays a role in repressing gene expression, *MBD1* was knocked out using the CRISPR system. Interestingly, two independent clones of *MBD1* KOs showed consistently increased gene expression for the *GATA2* and *ZFYVE21* genes, indicating that MBD1 could be involved in the repression of these genes. In addition, the methylated reporter genes were de-repressed in a MBD1 concentration-dependant manner. Interestingly, deletion of MBD1 did not fully alleviate 5mC-mediated repression of reporters, nor did it completely abolish the induction of endogenous genes caused by expression the TET-CD. This suggests that other factors could be involved, possibly due to the redundancy of MBD family proteins or direct involvement of another family member such as MBD2. In addition, a number of other proteins have been shown to bind 5mC such as zinc finger proteins or Kaiso proteins (Spruijt et al., 2013). CHTOP protein has been reported to bind 5hmC and activate gene expression (Takai et al., 2014). CHTOP is highly expressed in HEK293 cells and could play a role in the regulation of gene body modification-mediated gene expression. Overall, this study has established a strategy to screen for regulators of genic DNA modification-mediated transcription. Elimination of MBP family proteins one-by-one and in combination, as well as other relevant binders, could help to elucidate the molecular mechanism of this regulation. In addition, a number of genome-wide techniques have been established as well as artificial reporter based assays, which can convincingly confirm the direct involvement of various proteins in gene body DNA modification-mediated transcriptional regulation.

Chapter 7

Discussion

6.4. Gene Body DNA Modifications: Potential Effects on Transcription

The discovery of 5hmC in mammalian genomes and, in particular, over gene bodies (Mellen et al., 2012, Song et al., 2011a) has made necessary a re-evaluation and thorough investigation of the biological role of genic DNA modifications. In this study, the role of genic DNA modifications in transcriptional regulation was of especial interest. But firstly, one has to consider the question of how a gene body itself affects transcription. In lower eukaryotes such as yeast, the length of a gene body is strictly dependent on the size of the protein that it encodes. However, in higher eukaryotes, the presence of introns causes large variation in gene length. It is reasonable to believe that the longer the gene, the longer it takes to transcribe it. Large number of long introns could result in lower mRNA production rates. Indeed, it has been demonstrated that gene length negatively correlates with gene expression in mammals (Grishkevich and Yanai, 2014). For example genes that encode histones (very highly abundant proteins) tend to be very short and do not contain introns. Nevertheless, genes can be transcribed by several polymerases at the same time. Therefore, one might argue that transcript production levels are controlled solely by the RNAPII firing rate. However, recent studies have proposed that DNA modifications deposited over genic regions can modulate the elongation rate of RNA polymerase, thus, affecting gene expression levels (Lorincz et al., 2004, Wang et al., 2015). However, it remains unknown how other DNA modifications, such as 5hmC over genic regions, interact with various regulatory elements and affect transcription. Therefore, the main aim of this study was to address this issue.

6.5. Genic 5hmC Acts as Transcriptional Activator

In this study, the TET1 catalytic domain (TET-CD) was used to deposit 5hmC over the genome. Importantly, it was demonstrated that TET-CD efficiently converts 5mC to 5hmC in genic regions. Interestingly, the highest enrichment was observed in the 5' UTRs of genes, indicating that a large amount of 5hmC was deposited in close proximity to promoters. In general, it was observed that 5hmC enriched regions were largely overlapping 5mC peaks indicating that TET1-CD preferentially oxidises substrate-rich regions. Finally, it was demonstrated that 5hmC deposited over gene bodies can efficiently upregulate gene expression (Chapter 4). Importantly, prior to the upregulation of genes by 5hmC deposition, the genes were already expressed and unmethylated in TSS proximal regions. The observation that genic 5mC act as a transcriptional repressor agrees with previous reports (Lorincz et al., 2004). However, the ability of 5mC oxidation to alleviate this repression has not been previously demonstrated.

Based on genomic data alone, it is difficult to conclude whether genic 5hmC facilitates intragenic enhancer activation, splicing, transcription elongation rates, or promoter stimulation in *trans*. However, the efficient upregulation of several short genes, which lack introns, suggests that intragenic enhancer demethylation or splicing events are unlikely. In addition, global transcription analysis does not allow effects of initiation and elongation to be distinguished. Therefore, measurements of RNAPII elongation rates would be necessary to confirm the importance of transcription elongation rate in this process. Altogether, these observations suggest that gene body DNA modifications could fine-tune gene expression.

In order to overcome the limitations of genome-wide experiments and gain more mechanistic insights into the regulatory role of gene body DNA modifications, several reporter-based transcription assays were established (Chapter 5). The reporters (genic regions) used in these assays were designed without enhancers and depleted of TF binding sites to eliminate the possibility of intragenic transcription initiation. Using this system, it was demonstrated that gene body DNA methylation can indeed repress transcription, as reported previously (Lorincz et al., 2004, Hsieh, 1994, Hsieh, 1997). In addition, repression by 5mC was mediated in a concentration-dependent manner, while 5hmC deposited over gene bodies showed little or no repression compared to the unmodified reporter. The same results were observed when the reporter was stably integrated into the genome. This set of data strongly supports the hypothesis that the presence of 5hmC over genic regions is not merely the consequence of transcription but can also alleviate gene body DNA methylation-mediated transcriptional repression. In addition, reporter based assays indicated that gene body DNA modifications can modulate promoter activity in *trans*. However, not all promoters can be regulated by gene body DNA modifications. The *EEF1A1* gene is one of the most highly expressed in the genome and could not be repressed by DNA methylation in the reporter-based assays, although it remains possible that the gene body fused to *EEF1A1* promoter was very rapidly demethylated. In support of this hypothesis is the fact that, in the vast majority of the cell lines studied, the genic region of the *EEF1A1* gene was demethylated. This poses the interesting possibility that this particular promoter is capable of recruiting strong demethylase activity, however, its ability to demethylate the reporter remains to be assessed.

6.6. Genic DNA Modifications Affect Transcription via MBD Family Proteins

Efforts were made to elucidate the molecular mechanism that leads to the gene body DNA modification-mediated repression (Chapter 6). The initial experiments, treating cells with TSA, showed that this repression is at least partially mediated by histone deacetylation, which is in agreement with previously published results (Lorincz et al., 2004). This led us to reason that MBD family members, which were demonstrated to recruit HDACs to the methylated DNA regions, could facilitate 5mC-mediated repression. To test this hypothesis, either *MeCP2* or *MBD1* were deleted. While *MeCP2* appeared to act as an activator on several candidate genes, *MBD1* deletion alleviated the repression of the genes repressed by gene body DNA methylation. It has been previously demonstrated that hydroxylation of 5mC abolishes binding of the *MBD1* MBD domain to DNA (Mellen et al., 2012). This observation poses the possibility that hydroxylation of methylated DNA can specifically exclude *MBD1* repressive activity from genic regions. Interestingly, *MeCP2* – a well-documented transcriptional repressor – did not show an ability to mediate gene body methylation-mediated repression in this system. This could be because of the lack of interaction partners. Alternatively, it is possible that *MeCP2* competes for methylated DNA sites with a more potent repressor. This may be *MBD1* which, when *MeCP2* is deleted, is able to bind to gene bodies without competition and repress them even further. However, this does not exclude the possibility that the *MeCP2* can mediate gene body DNA methylation-mediated repression in other cell types. In addition, *MBD1* deletion does not completely abolish gene body DNA methylation mediated-repression. This means that other repressors exist. The primary candidates would be other MBD family proteins such as *MBD2*. Hence, future

studies should knockout MBD proteins in a combinatorial manner to overcome the redundancy problem and establish their role in gene body DNA modification-mediated transcriptional regulation.

6.7. 5mC Hydroxylation Versus Demethylation

Overall, this study has observed that the effect on transcription of genic 5hmC is very similar to that of unmethylated cytosine. The obvious question, however, is why not demethylate? A demethylation strategy has been successfully established in plants (Penterman et al., 2007), where the presence of 5hmC has not been convincingly demonstrated. However, demethylation is multistep process that requires complex base excision machinery (Shen et al., 2013). In contrast, hydroxylation of 5mC is a simple one step mechanism (Hu et al., 2013). Furthermore, unmethylated residues have to be protected from subsequent re-methylation, while 5hmC cannot be re-methylated. Finally, it is thought that gene body DNA methylation evolved to silence spurious transcription (Karimi et al., 2011). In this study, it has been demonstrated that 5hmC can efficiently silence unspecific transcription initiation *in vitro* (Chapter 6). In addition, we and others (Robertson et al., 2011) have demonstrated that 5hmC deposited over promoter regions can efficiently silence transcription. Taken together, this data suggests that 5hmC, as opposed to unmodified DNA, could contribute to the silencing of spurious transcription initiation. However, whether hydroxymethylation can silence repetitive elements *in vivo* remains to be demonstrated.

7.5. The Model: Genic DNA Modifications Affect Transcription Efficiency

Based on the evidence gathered by this study, a model (Figure 7) is proposed whereby gene body DNA modifications fine-tune transcription through recruitment

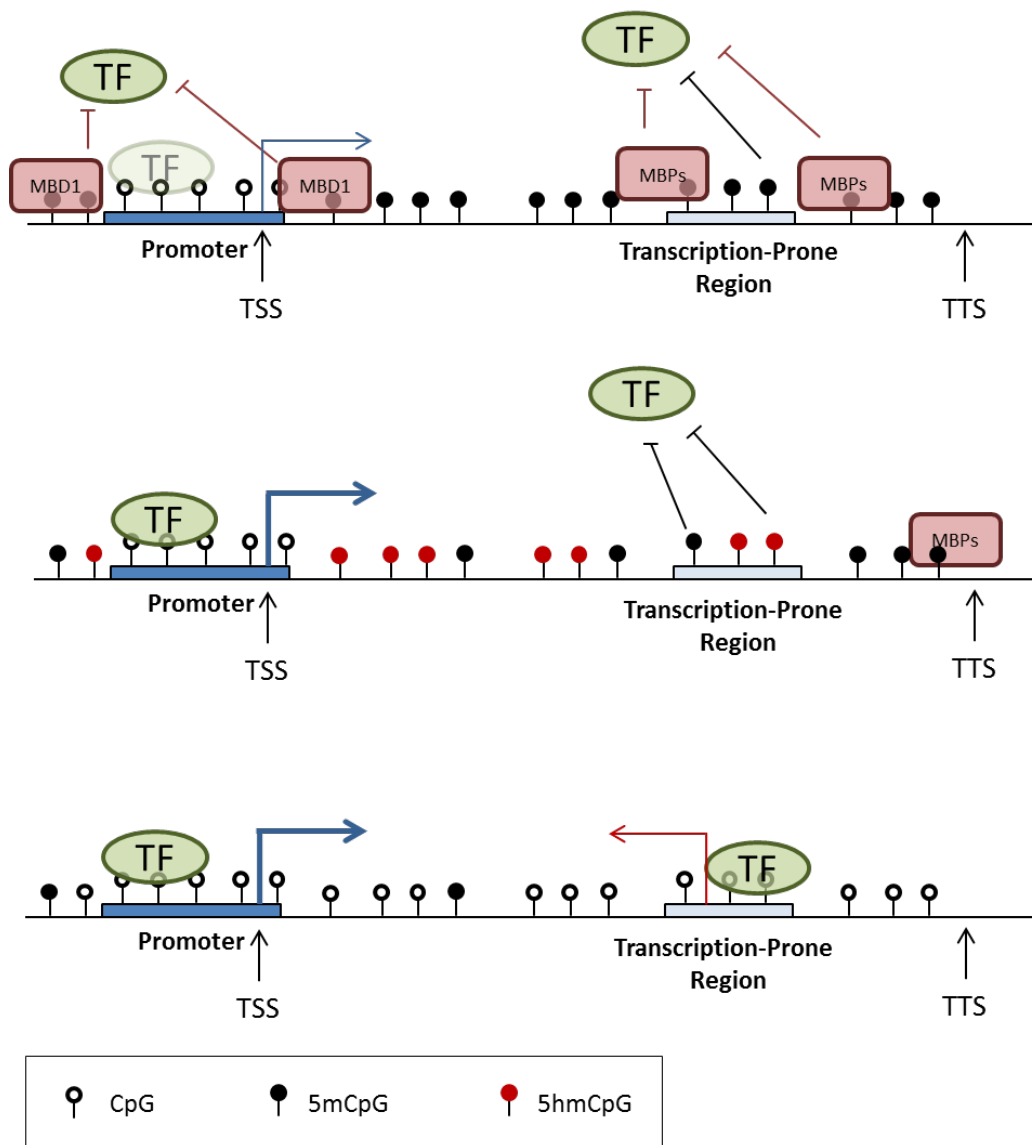


Figure 7 The model demonstrating the role of gene body 5hmC in transcriptional regulation

Gene body DNA methylation aids the repression of spurious transcription. However, this DNA methylation, especially that in close proximity to promoters, also reduces transcription *in cis* (top panel) via MBD1. 5hmC evolved to counteract 5mC mediated repressive activity in promoter proximal regions (middle panel). In addition,

hydroxymethylation of cytosine reduces the rate of spurious transcription initiation compared to unmodified cytosine (middle and bottom panels).

of chromatin modifying enzymes by creating or abolishing binding sites for MBD1 and possibly other proteins. This modulation depends on the proximity to the promoter and the density of DNA modifications. Genic 5hmC, in proximity to promoters and other regulatory elements, is suggested to counteract the repressive effects of DNA methylation. In addition, it reduces intragenic transcription initiation potential compared to the unmodified cytosine.

7.6. Future Perspectives

This study provides mechanistic evidence of how gene body DNA modifications affect or actively regulate gene expression (Figure 7.1). However, a large amount of mechanistic detail remains unknown. The polymerase bursting dynamics of differentially modified genes are of significant importance for our further understanding of the molecular mechanisms involved. In addition, direct measurement of polymerase elongation rates through differentially modified regions would provide evidence of whether or not DNA modifications can affect the speed of transcribing polymerases. Integration of this model into predictions of gene expression would also be of interest. A recent study has demonstrated that gene body DNA methylation is a better predictor of gene expression compared to promoter methylation (Lou et al., 2014). Provided that this mechanism is correct, incorporation of the methylation distance from the TSS parameter should make this type of model more accurate. In general, the insights from these studies open up the possibility of using genic DNA modifications for medical prognosis and diagnostics.

Gene body DNA methylation-mediated regulation could be important in cancer. In the process of clonal evolution, cells with the highest proliferative capacity are selected (Greaves and Maley, 2012). These clones adjust their global transcription to drive the genes that are important for proliferation while other genes are downregulated. Cancers tend to shut down several tumor suppressor genes by promoter methylation (Merlo et al., 1995), while the majority of the genes are inactivated by mutations (Hanahan and Weinberg, 2000). This study opens up the possibility that genes can be downregulated by deposition of higher densities of DNA methylation, and also methylation spreading towards promoters, without expression being completely repressed. Importantly, a well-documented loss of 5hmC in cancers might facilitate this process. This type of regulation may fine-tune gene expression at the single gene level without causing major perturbations in the levels of transcription factor networks. However, the importance of these mechanisms in the process of malignant transformation remains to be proven.

Finally, genome-wide studies indicate that 5hmC-enriched regions highly correlate with sites of active transcription (Wen et al., 2014). Despite there being no direct evidence to date that 5hmC is deposited co-transcriptionally, it is possible that a positive feedback loop exists between gene body hydroxymethylation and transcription. Increased transcription results in more 5hmC deposited over gene bodies and promoter proximal regions. As a consequence, the probability of a promoter being in a repressed state is reduced, which can be caused by proximal DNA methylation. This makes the gene more active or the induction response to be more rapid when needed. These processes might be of significant importance in memory formation and learning via making the genes more prone to respond to

various signals (Graff and Tsai, 2013). However, the role of genic DNA modifications in these processes remains to be proven.

Appendix

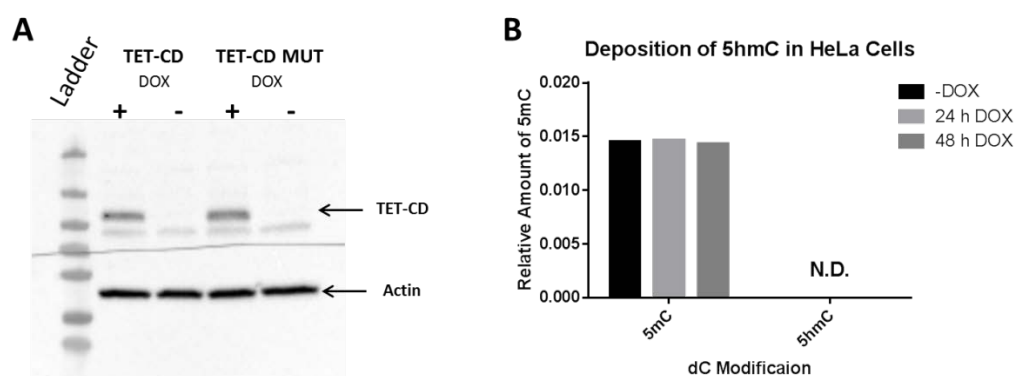


Figure S1 Induction of TET-CD and deposition of 5hmC in HeLa cells

A) Immunoblot showing induction of TET-CD 24 h after adding doxycycline to the media. B) HPLC analysis of HeLa genomic DNA 24 h and 48 h after induction of TET-CD. 5hmC levels were below the detection limits of the HPLC instrument.

Modification	Size (bp)	Sequence
C	264	TCAAGCTTGGTACCGAGCTCGGATCCACTAGTAACGGCCGCC AGTGTGCTGGAATTCGCCCTTGCGTGAATTTGCTAGGACTGG AAGTGAAAACATCTACAAAGCATGCAATGCCCTGGAACGGA ATTCTTCTAATAAAAGATGTATCATTTTAAATGCGCGGAAT TCTGATTCTGGTAATTCAAGGGCGAATTCTGCATATAATAAT ATATTTAAATTCATATAATAATATATTTAGAATGGCAGATA TCCATCACACTGGCG

5mC	287	TCAAGCTTGGTACCGAGCTCGGATCCACTAGTAACGGCCGCC AGTGTGCTGGAATTCGCCCTTGCGTGAATTTGCTAGGACTGG AAGTGAAAACATCTACAAAGCATGCAATGCCCTGGAACGGA ATTCTTCTAATAAAAAGATGTATCATTTTAAATGCGCGGAAT TCTGATTCTGGTAATTCAAGGGCGAATTCTGCAATTTAAATA TATTATTATATGCAATTTAAATATATTATTATATGCAATTT AAATATATTATTATATGCAGATATCCATCACACTGGCG
5hmC	310	TCAAGCTTGGTACCGAGCTCGGATCCACTAGTAACGGCCGCC AGTGTGCTGGAATTCGCCCTTGCGTGAATTTGCTAGGACTGG AAGTGAAAACATCTACAAAGCATGCAATGCCCTGGAACGGA ATTCTTCTAATAAAAAGATGTATCATTTTAAATGCGCGGAAT TCTGATTCTGGTAATTCAAGGGCGAATTCTGCATATAATAAT ATATTTAAATTGCATATAATAATATATTTAAATTGCATATA ATAATATATTTAAATTGCATATAATAATATATTTAAATTGC AGATATCCATCACACTGGCG

Table S1: Spiked in modified probes for enrichment quality control

Region	Length (bp)	CpG density (CGs/kb)	CpGs Observed/Expected
EGFP	738	84	1.34
Ren Luc	938	65	1.04
Ffly Luc	1816	81	1.3
EEF1 Prom	1189	80	1.28

p16 Prom	846	56	0.896
SV40 Prom	294	27	0.432
CMV Prom	803	49	0.777

Table S3: CpG densities in the constructs used in transfected reporter assays

ATGGCTTCCAAGGTGTATGACCCTGAGCAAAGGAAAAGGATGATCACTGGGCCTCAGTGG
TGGGCTAGGTGCAAGCAAATGAATGTGCTGGACTCCTTCATCAACTACTATGATAGTGAG
AAGCATGCAGAGAATGCTGTGATTTTTCTGCATGGTAATGCTGCCTCCAGCTACCTGTGGA
GGCATGTGGTGCCTCACATTGAGCCAGTGGCTAGATGCATCATCCCTGATCTGATTGGAAT
GGGTAAGAGTGGCAAGAGTGGGAATGGCTCATATAGACTCCTGGATCACTACAAGTACCT
CACAGCTTGGTTTGAGCTGCTGAACCTTCAAAGAAAATCATCTTTGTGGGCCATGACTGG
GGGGCTTGTCTGGCCTTTCCTACTCCTATGAGCACCAAGACAAGATCAAGGCCATTGTCC
ATGCTGAGAGTGTGGTGGATGTGATTGAGTCCTGGGATGAGTGGCCTGACATTGAGGAGG
ATATTGCCCTGATCAAGAGTGAAGAGGGGGAGAAAATGGTGCTTGAGAATAACTTCTTTG
TGGAGACCATGCTCCCAAGCAAGATCATGAGAAAATGGAGCCTGAGGAGTTTGCTGCCT
ACCTGGAGCCATTCAAGGAGAAGGGAGAGGTTAGAAGACCTACCCTCTCCTGGCCTAGAG
AGATCCCTCTGGTTAAGGGAGGCAAGCCTGATGTGGTCCAGATTGTCAGAACTACAATG
CCTACCTTAGAGCCAGTGATGATCTGCCTAAGATGTTTCATTGAGAGTGACCCTGGGTTCTT
TTCCAATGCTATTGTGGAGGGAGCTAAGAAGTTCCCTAACACAGAGTTTGTGAAGGTGAA
GGGCCTCCACTTCAGCCAGGAGGATGCTCCAGATGAAATGGGTAAGTACATCAAGAGCTTT
GTGGAGAGAGTGCTGAAGAATGAGCAGTAA

Figure S2 Sequence of CpGless *Renilla* Luciferase gene

Script to visualise genomic coverage

```

library(chipseq)
library(rtracklayer)
library(BSgenome.Hsapiens.UCSC.hg19)

seqname_1<-seqnames(Hsapiens)[1]
subject<-Hsapiens[[seqname_1]]
chr_length<-length(subject)

output_data_final<-GRanges(ranges=IRanges(start=1, end=chr_length),
seqnames=seqname_1)

seqlengths(output_data_final)<-chr_length
seqnames<-seqnames(Hsapiens)[2:25]

for (seqname in seqnames)
{
  subject<-Hsapiens[[seqname]]
  chr_length<-length(subject)
  output_data<-GRanges(ranges=IRanges(start=1, end=chr_length),
seqnames=seqname)

  seqlengths(output_data)<-chr_length
  output_data_final<-append(output_data_final, output_data)
}

input<-as(readAligned(".", "ej3_minus_dox_sorted_no_dupl.bam", type="BAM"),
"GRanges")
input_ranges_extended<-resize(input, width=250)
input_ranges_extended_chr<-subsetByOverlaps(input_ranges_extended,
output_data_final)
input_cov<-coverage(input_ranges_extended_chr)

export.bw(input_cov, "ej3_minus_dox_no_dupl_extend.bw")

```

Gene	FPKM Minus DOX	FPKM Plus DOX	log2FoldChange	p-value	p-adjusted	hmC Enrichment (Genic)
TET1	26.6015	184.9642702	2.797667107	9.05E-46	2.00E-41	1.664120873
GATA2	14.8829	28.56598141	0.940642426	4.47E-10	3.29E-06	4.451834315
QDPR	3.25427	6.910850946	1.08652941	3.73E-08	0.0002058	1.651572907
HAND1	2.99798	8.571692574	1.515589349	2.29E-07	0.001012	6.08952136
MGAT3	2.70396	5.449511393	1.011053073	4.04E-07	0.0014865	2.648767218
HOXA-AS3	6.50056	3.295449443	-0.980088769	8.09E-07	0.0025496	1.746662262
SOGA1	8.72667	13.82934639	0.664229825	2.37E-06	0.0065282	2.187290294
PANX2	1.71143	3.905732685	1.190390927	3.59E-06	0.0088063	6.724615633
FAM78A	4.8779	8.697174446	0.834286589	7.06E-06	0.0155626	3.590785365
PPAPDC3	0.636497	1.548961808	1.283075953	7.97E-06	0.0159844	3.133437173
LZTS1	1.47059	2.963369785	1.010843585	2.03E-05	0.0372768	2.781641333
SYNM	0.838309	1.671507587	0.99559588	4.83E-05	0.0819657	2.529027643
RPP25	2.22963	4.755699644	1.092853282	6.67E-05	0.1050845	6.484627497

CYB561	1.4369	3.225270884	1.166460673	7.92E-05	0.1164404	4.487227637
MAFA	2.46929	5.837659655	1.241293821	9.51E-05	0.1310449	5.476805665
H19	0.377952	2.774664814	2.876038573	0.0001139	0.1448529	6.973702381
ATL3	1.29349	2.186713869	0.757495556	0.0001182	0.1448529	1.692911154
CEBPA	29.9097	44.80044861	0.582899739	0.0001474	0.1710897	6.018396553
INHBB	2.49046	4.599713143	0.885131652	0.000248	0.2526381	5.030405277
HIST1H2BG	28.1797	50.14817943	0.831541074	0.000252	0.2526381	4.029569421
GCNT2	1.88164	2.890996517	0.619576236	0.0004621	0.4430659	1.647031686
C8orf82	6.19858	10.47409262	0.756815608	0.0005119	0.4525513	2.809345702
TERT	0.535049	1.017663928	0.927518282	0.000513	0.4525513	4.219711143
FASN	216.784	159.487593	-0.442814084	0.000538	0.4563491	8.216138261
USP2	3.22835	5.299195694	0.71497641	0.0006323	0.5135874	2.949506647
SPINT2	25.3119	36.55965237	0.530436547	0.0006521	0.5135874	2.669240528
HES4	68.2679	44.93617278	-0.603330119	0.0007111	0.5407901	0.125043505
DUX2	0	0	0.60600328	0.0009161	0.5460027	4.256121843

PTGER2	1.83215	3.459549981	0.917046761	0.0009813	0.5694832	1.971763207
SLC10A4	1.27984	2.785166242	1.12179998	0.0010636	0.5768298	3.572659829
MYOM2	3.21553	4.614288275	0.521051589	0.0010722	0.5768298	2.709558931
LINC00649	0.540142	0.940875161	0.800664581	0.0010724	0.5768298	1.302170105
TCAM1P	0.0276595	0.111875364	2.016045391	0.0011827	0.6209987	1.062883355
ALDH1A3	0.439875	0.996796741	1.180205742	0.0014234	0.7300196	2.078580525
GREB1	0.758603	1.152523052	0.603378625	0.0015266	0.7633749	2.347195008
WNT3A	2.00854	3.395758852	0.757586817	0.0015577	0.7633749	1.998560029
CPT1A	1.60984	2.382330373	0.565456187	0.0015982	0.7661781	2.780851
ADM	37.1878	25.75012392	-0.530250027	0.001669	0.783093	0.025741256
GRM5	0.00636215	0.009042005	0.507128315	0.0019659	0.8758845	0.166743939
FGF18	1.33941	0.803281491	-0.737620105	0.0019829	0.8758845	2.631823543
SH2B3	11.2493	15.78600816	0.488811169	0.0019859	0.8758845	2.456874754
NKX2-2	6.46467	10.53573416	0.704642216	0.0021404	0.9255499	7.350210708
FBXO44	4.82205	7.465121752	0.63051918	0.0022044	0.9348679	3.012983979

NACAD	0.582335	1.179012045	1.017657221	0.0022946	0.9547529	3.517518659
NEURL1B	0.833638	1.385132628	0.732531176	0.0023874	0.9749777	2.24591766
FUT1	2.21813	3.699128688	0.73784157	0.0024466	0.9810133	3.676312175
LHX1	4.33707	6.473775406	0.577886587	0.0026941	1	7.417728008
IL17RD	2.35789	3.376398207	0.517988648	0.0027403	1	1.551385561
ALX4	2.3337	3.352317764	0.522539795	0.0028654	1	2.46515706
KCNG3	9.40915	12.67403337	0.429739414	0.002874	1	1.859678327
KANK1	7.49848	9.904225541	0.401445988	0.0030394	1	1.219475479
LRRC24	1.22426	2.099644619	0.778235181	0.0032541	1	2.70917743
SNORA23	425.145	237.2120716	-0.841777541	0.0033506	1	0.454740564
COL4A2	5.81545	7.722301117	0.409139979	0.0033926	1	2.430404764
FGD5	0.0947851	0.163166251	0.783610488	0.003398	1	1.67789263
CTSH	0.730926	1.426815066	0.965001096	0.0034086	1	2.425963467
ITPRIPL2	9.10255	14.73919946	0.695315503	0.0037783	1	4.084343153
SP6	0.491194	1.148154275	1.224951663	0.0038101	1	3.532378783

TMBIM1	12.7152	18.04696878	0.505202384	0.00384	1	1.924111366
FAM114A1	1.04539	1.547614668	0.566005043	0.0038599	1	1.330477284
EBLN2	8.12066	4.954026128	-0.712995508	0.0039154	1	1.446948533
DMRTA2	3.3671	5.627420551	0.740967215	0.003922	1	4.675870906
LRIT3	2.17666	1.475122984	-0.561280832	0.0039424	1	0.526186559
SCARNA3	55.6297	16.46117639	-1.756787884	0.0040641	1	0
ADAMTS4	0.150606	0.443592879	1.558456959	0.0042038	1	2.230183559
LHFPL4	6.91686	5.191957982	-0.413838548	0.004205	1	1.545893196
SIX2	4.30575	6.820634087	0.663641311	0.004211	1	6.75896414
FAM84B	6.14148	8.565626968	0.479972485	0.004219	1	2.724377063
AFAP1-AS1	0.0321111	0.159314245	2.310731277	0.0042206	1	3.166201562
ZNF280D	3.00785	4.220702676	0.488750581	0.004347	1	0.643988012
ADRB2	1.30594	2.936551956	1.169034548	0.0046141	1	3.711750322
ERC1	22.5594	31.63783656	0.487922252	0.0046689	1	1.468295325
SCUBE1	0.479578	0.702652389	0.551045668	0.0049189	1	3.191829023

SEMA3F	1.82568	2.872228761	0.653736743	0.005067	1	2.804416638
ERVMER34-1	3.97733	5.952281628	0.58164252	0.0051337	1	1.708643335
STAB1	0.0257515	0.072112901	1.485600916	0.0051805	1	3.59926339
DHX38	18.4757	13.84545821	-0.416216226	0.0052851	1	2.050425653
PBX4	0.503539	0.691981938	0.458630859	0.0057133	1	2.316570926
SOX2-OT	0.103822	0.135533607	0.384538447	0.0058362	1	0.690105532
GDF15	2.92215	5.864149178	1.004891568	0.0065066	1	3.797631013
GJB2	7.42004	10.91426518	0.556716132	0.0066682	1	3.624390851
PNMA6C	0.00883662	0.015437255	0.80484968	0.006683	1	7.286199765
PNMA6A	0	0	0.80484968	0.006683	1	7.274977597
C9orf172	5.50073	3.648928698	-0.592150131	0.0067695	1	11.57660714
COL7A1	0.999979	0.620486096	-0.688498915	0.0068563	1	3.252116156
FAM213A	8.07043	11.31129874	0.487047138	0.0068856	1	1.495553131
NOP56	281.23	220.3675877	-0.351838457	0.0069759	1	1.372344639
ZNF707	11.31	8.32515359	-0.442050136	0.0070593	1	3.639128633

SNX33	3.55862	5.256139733	0.562685744	0.007226	1	2.223521625
CARD9	5.99777	4.51606736	-0.409359195	0.0074283	1	6.357920001
TRIM47	4.23127	7.242882014	0.775473126	0.0075118	1	5.700334707
SERPINF1	1.74252	2.849934445	0.70975352	0.0075144	1	3.392831392
SH3PXD2B	5.85959	7.748761799	0.403166074	0.0077003	1	2.048987793
NKD2	4.33531	3.147656589	-0.461857008	0.0079484	1	4.585505223
SKI	140.253	109.0478618	-0.363070149	0.0079541	1	4.944841444
AS3MT	2.21969	3.13028265	0.495934727	0.008095	1	1.726337809
EVL	11.267	8.823224883	-0.352725467	0.0082756	1	2.473553733
SNHG6	43.3674	32.17794915	-0.430538573	0.0082892	1	0.331097766
G RTP1	4.55788	6.169768545	0.436853427	0.0085219	1	3.51575835

Table S4: Top 100 deregulated genes without applying 2.5FPKM cut-off. Upregulated genes are marked in green, downregulated are marked in red.

Bibliography

- ALLAN, A. M., LIANG, X. M., LUO, Y. P., PAK, C. H., LI, X. K., SZULWACH, K. E., CHEN, D. H., JIN, P. & ZHAO, X. Y. 2008. The loss of methyl-CpG binding protein 1 leads to autism-like behavioral deficits. *Human Molecular Genetics*, 17, 2047-2057.
- AMIR, R. E., VAN DEN VEYVER, I. B., WAN, M., TRAN, C. Q., FRANCKE, U. & ZOGHBI, H. Y. 1999. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature Genetics*, 23, 185-188.
- APPANAH, R., DICKERSON, D. R., GOYAL, P., GROUDINE, M. & LORINCZ, M. C. 2007. An unmethylated 3' promoter-proximal region is required for efficient transcription initiation. *Plos Genetics*, 3, 241-253.
- ARBER, W. 1974. DNA modification and restriction. *Prog Nucleic Acid Res Mol Biol*, 14, 1-37.
- AST, G. 2004. How did alternative splicing evolve? *Nature Reviews Genetics*, 5, 773-782.
- AVERY, O. T., MACLEOD, C. M. & MCCARTY, M. 1944. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med*, 79, 137-58.
- AYGUN, O., MEHTA, S. & GREWAL, S. I. 2013. HDAC-mediated suppression of histone turnover promotes epigenetic stability of heterochromatin. *Nat Struct Mol Biol*, 20, 547-54.
- BACHMAN, M., URIBE-LEWIS, S., YANG, X. P., WILLIAMS, M., MURRELL, A. & BALASUBRAMANIAN, S. 2014. 5-Hydroxymethylcytosine is a predominantly stable DNA modification. *Nature Chemistry*, 6, 1049-1055.
- BAKER, S. A., CHEN, L., WILKINS, A. D., YU, P., LICHTARGE, O. & ZOGHBI, H. Y. 2013. An AT-Hook Domain in MeCP2 Determines the Clinical Course of Rett Syndrome and Related Disorders. *Cell*, 152, 984-996.
- BALL, M. P., LI, J. B., GAO, Y., LEE, J. H., LEPROUST, E. M., PARK, I. H., XIE, B., DALEY, G. Q. & CHURCH, G. M. 2009. Targeted and genome-scale

- strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnology*, 27, 361-368.
- BASHTRYKOV, P., JANKEVICIUS, G., JURKOWSKA, R. Z., RAGOZIN, S. & JELTSCH, A. 2014. The UHRF1 Protein Stimulates the Activity and Specificity of the Maintenance DNA Methyltransferase DNMT1 by an Allosteric Mechanism. *Journal of Biological Chemistry*, 289, 4106-4115.
- BECKER, P. B., RUPPERT, S. & SCHUTZ, G. 1987. Genomic Footprinting Reveals Cell Type Specific DNA-Binding of Ubiquitous Factors. *Cell*, 51, 435-443.
- BEDNARIK, D. P., DUCKETT, C., KIM, S. U., PEREZ, V. I., GRIFFIS, K., GUENTHNER, P. C. & FOLKS, T. M. 1991. DNA CpG Methylation Inhibits Binding of Nf-Kappa-B Proteins to the Hiv-1 Long Terminal Repeat Cognate DNA Motifs. *New Biologist*, 3, 969-976.
- BENJAMINI, Y. & SPEED, T. P. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res*, 40, e72.
- BERGER, S. L., KOUZARIDES, T., SHIEKHATTAR, R. & SHILATIFARD, A. 2009. An operational definition of epigenetics. *Genes Dev*, 23, 781-3.
- BESTOR, T., LAUDANO, A., MATTALIANO, R. & 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. *J Mol Biol*, 203, 971-83.
- BESTOR, T. H. 2000. The DNA methyltransferases of mammals. *Hum Mol Genet*, 9, 2395-402.
- BHATTACHARYA, S. K., RAMCHANDANI, S., CERVONI, N. & SZYF, M. 1999. A mammalian protein with specific demethylase activity for mCpG DNA. *Nature*, 397, 579-83.
- BINISZKIEWICZ, D., GRIBNAU, J., RAMSAHOYE, B., GAUDET, F., EGGAN, K., HUMPHERYS, D., MASTRANGELO, M. A., JUN, Z., WALTER, J. & JAENISCH, R. 2002. Dnmt1 overexpression causes genomic hypermethylation, loss of imprinting, and embryonic lethality. *Mol Cell Biol*, 22, 2124-35.
- BIRD, A., TAGGART, M., FROMMER, M., MILLER, O. J. & 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 Res*, 8, 1499-504.

- BLASCHKE, K., EBATA, K. T., KARIMI, M. M., ZEPEDA-MARTINEZ, J. A., GOYAL, P., MAHAPATRA, S., TAM, A., LAIRD, D. J., HIRST, M., RAO, A., LORINCZ, M. C. & RAMALHO-SANTOS, M. 2013. Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. *Nature*, 500, 222-6.
- BLATTLER, A., YAO, L. J., WITT, H., GUO, Y., NICOLET, C. M., BERMAN, B. P. & FARNHAM, P. J. 2014. Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. *Genome Biology*, 15.
- BOOTH, M. J., BRANCO, M. R., FICZ, G., OXLEY, D., KRUEGER, F., REIK, W. & BALASUBRAMANIAN, S. 2012. Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. *Science*, 336, 934-937.
- BOSTICK, M., KIM, J. K., ESTEVE, P. O., CLARK, A., PRADHAN, S. & JACOBSEN, S. E. 2007. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science*, 317, 1760-1764.
- BRODY, Y., NEUFELD, N., BIEBERSTEIN, N., CAUSSE, S. Z., BOHNLEIN, E. M., NEUGEBAUER, K. M., DARZACQ, X. & SHAV-TAL, Y. 2011. The In Vivo Kinetics of RNA Polymerase II Elongation during Co-Transcriptional Splicing. *Plos Biology*, 9.
- BRZYZEK, G. & SWIEZEWSKI, S. 2015. Mutual interdependence of splicing and transcription elongation. *Transcription*, 6, 37-9.
- CAREY, B. W., FINLEY, L. W. S., CROSS, J. R., ALLIS, C. D. & THOMPSON, C. B. 2015. Intracellular alpha-ketoglutarate maintains the pluripotency of embryonic stem cells. *Nature*, 518.
- CARNINCI, P., SANDELIN, A., LENHARD, B., KATAYAMA, S., SHIMOKAWA, K., PONJAVIC, J., SEMPLE, C. A. M., TAYLOR, M. S., ENGSTROM, P. G., FRITH, M. C., FORREST, A. R. R., ALKEMA, W. B., TAN, S. L., PLESSY, C., KODZIUS, R., RAVASI, T., KASUKAWA, T., FUKUDA, S., KANAMORI-KATAYAMA, M., KITAZUME, Y., KAWAJI, H., KAI, C., NAKAMURA, M., KONNO, H., NAKANO, K., MOTTAGUI-TABAR, S., ARNER, P., CHESI, A., GUSTINCICH, S., PERSICHETTI, F., SUZUKI, H., GRIMMOND, S. M., WELLS, C. A., ORLANDO, V., WAHLESTEDT, C., LIU, E. T., HARBERS, M., KAWAI, J., BAJIC, V. B., HUME, D. A. & HAYASHIZAKI, Y. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, 38, 626-635.
- CHEDIN, F., LIEBER, M. R. & HSIEH, C. L. 2002. The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. *Proc Natl Acad Sci U S A*, 99, 16916-21.

- CHEN, K. & RAJEWSKY, N. 2007. The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics*, 8, 93-103.
- CHEN, L., CHEN, K. F., LAVERY, L. A., BAKER, S. A., SHAW, C. A., LI, W. & ZOGHBI, H. Y. 2015. MeCP2 binds to non-CG methylated DNA as neurons mature, influencing transcription and the timing of onset for Rett syndrome. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 5509-5514.
- CHEN, L., MACMILLAN, A. M., CHANG, W., EZAZ-NIKPAY, K., LANE, W. S. & VERDINE, G. L. 1991. Direct identification of the active-site nucleophile in a DNA (cytosine-5)-methyltransferase. *Biochemistry*, 30, 11018-25.
- CHEN, Q., CHEN, Y. B., BIAN, C. J., FUJIKI, R. & YU, X. C. 2013. TET2 promotes histone O-GlcNAcylation during gene transcription. *Nature*, 493, 561-+.
- CHENG, B. & PRICE, D. H. 2007. Properties of RNA polymerase II elongation complexes before and after the P-TEFb-mediated transition into productive elongation. *Journal of Biological Chemistry*, 282, 21901-21912.
- CHIAPPINELLI, K. B., STRISSEL, P. L., DESRICHARD, A., LI, H. L., HENKE, C., AKMAN, B., HEIN, A., ROTE, N. S., COPE, L. M., SNYDER, A., MAKAROV, V., BUHU, S., SLAMON, D. J., WOLCHOK, J. D., PARDOLL, D. M., BECKMANN, M. W., ZAHNOW, C. A., MERGOUB, T., CHAN, T. A., BAYLIN, S. B. & STRICK, R. 2015. Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell*, 162, 974-986.
- CHUANG, L. S. H., IAN, H. I., KOH, T. W., NG, H. H., XU, G. L. & LI, B. F. L. 1997. Human DNA (cytosine-5) methyltransferase PCNA complex as a target for p21(WAF1). *Science*, 277, 1996-2000.
- CLEMONSON, C. M., MCNEIL, J. A., WILLARD, H. F. & LAWRENCE, J. B. 1996. XIST RNA paints the inactive X chromosome at interphase: Evidence for a novel RNA involved in nuclear chromosome structure. *Journal of Cell Biology*, 132, 259-275.
- COMB, M. & GOODMAN, H. M. 1990. CpG Methylation Inhibits Proenkephalin Gene-Expression and Binding of the Transcription Factor Ap-2. *Nucleic Acids Research*, 18, 3975-3982.
- COOPER, D. N., TAGGART, M. H. & BIRD, A. P. 1983. Unmethylated domains in vertebrate DNA. *Nucleic Acids Res*, 11, 647-58.
- CORDAUX, R. & BATZER, M. A. 2009. The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, 10, 691-703.

- CORE, L. J., WATERFALL, J. J. & LIS, J. T. 2008. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science*, 322, 1845-1848.
- CORTAZAR, D., KUNZ, C., SELFRIDGE, J., LETTIERI, T., SAITO, Y., MACDOUGALL, E., WIRZ, A., SCHUERMANN, D., JACOBS, A. L., SIEGRIST, F., STEINACHER, R., JIRICNY, J., BIRD, A. & SCHAR, P. 2011. Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. *Nature*, 470, 419-U210.
- CORTELLINO, S., XU, J. F., SANNAL, M., MOORE, R., CARETTI, E., CIGLIANO, A., LE COZ, M., DEVARAJAN, K., WESSELS, A., SOPRANO, D., ABRAMOWITZ, L. K., BARTOLOMEI, M. S., RAMBOW, F., BASSI, M. R., BRUNO, T., FANCIULLI, M., RENNER, C., KLEIN-SZANTO, A. J., MATSUMOTO, Y., KOBI, D., DAVIDSON, I., ALBERTI, C., LARUE, L. & BELLACOSA, A. 2011. Thymine DNA Glycosylase Is Essential for Active DNA Demethylation by Linked Deamination-Base Excision Repair. *Cell*, 146, 67-79.
- COSTANZI, C. & PEHRSON, J. R. 1998. Histone macroH2A1 is concentrated in the inactive X chromosome of female mammals. *Nature*, 393, 599-601.
- CRAMER, P. 2002. Common structural features of nucleic acid polymerases. *Bioessays*, 24, 724-729.
- CROSS, S. H., MEEHAN, R. R., NAN, X. & BIRD, A. 1997. A component of the transcriptional repressor MeCP1 shares a motif with DNA methyltransferase and HRX proteins. *Nat Genet*, 16, 256-9.
- DAHM, R. 2005. Friedrich Miescher and the discovery of DNA. *Dev Biol*, 278, 274-88.
- DASKALOS, A., NIKOLAIDIS, G., XINARIANOS, G., SAVVARI, P., CASSIDY, A., ZAKOPOULOU, R., KOTSINAS, A., GORGOULIS, V., FIELD, J. K. & LILOGLOU, T. 2009. Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. *Int J Cancer*, 124, 81-7.
- DAWLATY, M. M., BREILING, A., LE, T., RADDATZ, G., BARRASA, M. I., CHENG, A. W., GAO, Q., POWELL, B. E., LI, Z., XU, M. J., FAULL, K. F., LYKO, F. & JAENISCH, R. 2013. Combined Deficiency of Tet1 and Tet2 Causes Epigenetic Abnormalities but Is Compatible with Postnatal Development. *Developmental Cell*, 24, 310-323.
- DAWLATY, M. M., GANZ, K., POWELL, B. E., HU, Y. C., MARKOULAKI, S., CHENG, A. W., GAO, Q., KIM, J., CHOI, S. W., PAGE, D. C. & JAENISCH, R. 2011. Tet1 Is Dispensable for Maintaining Pluripotency and Its Loss Is Compatible with Embryonic and Postnatal Development. *Cell Stem Cell*, 9, 166-175.

- DE WERT, G. & MUMMERY, C. 2003. Human embryonic stem cells: research, ethics and policy. *Human Reproduction*, 18, 672-682.
- DEATON, A. M. & BIRD, A. 2011. CpG islands and the regulation of transcription. *Genes & Development*, 25, 1010-1022.
- DEPLUS, R., DELATTE, B., SCHWINN, M. K., DEFRANCE, M., MENDEZ, J., MURPHY, N., DAWSON, M. A., VOLKMAR, M., PUTMANS, P., CALONNE, E., SHIH, A. H., LEVINE, R. L., BERNARD, O., MERCHER, T., SOLARY, E., URH, M., DANIELS, D. L. & FUKS, F. 2013. TET2 and TET3 regulate GlcNAcylation and H3K4 methylation through OGT and SET1/COMPASS. *Embo Journal*, 32, 645-655.
- DHAYALAN, A., RAJAVELU, A., RATHER, P., TAMAS, R., JURKOWSKA, R. Z., RAGOZIN, S. & JELTSCH, A. 2010. The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *J Biol Chem*, 285, 26114-20.
- DIGNAM, J. D., LEOVITZ, R. M. & ROEDER, R. G. 1983. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res*, 11, 1475-89.
- DYNAN, W. S. & TJIAN, R. 1983. The Promoter-Specific Transcription Factor-Sp1 Binds to Upstream Sequences in the Sv40 Early Promoter. *Cell*, 35, 79-87.
- EDEN, A., GAUDET, F., WAGHMARE, A. & JAENISCH, R. 2003. Chromosomal instability and tumors promoted by DNA hypomethylation. *Science*, 300, 455-455.
- EHRICH, M., NELSON, M. R., STANSSENS, P., ZABEAU, M., LILOGLOU, T., XINARIANOS, G., CANTOR, C. R., FIELD, J. K. & VAN DEN BOOM, D. 2005. Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 15785-15790.
- EHRlich, M. 2002. DNA methylation in cancer: too much, but also too little. *Oncogene*, 21, 5400-5413.
- EHRlich, M. 2009. DNA hypomethylation in cancer cells. *Epigenomics*, 1, 239-259.
- EHRlich, M. & WANG, R. Y. 1981. 5-Methylcytosine in eukaryotic DNA. *Science*, 212, 1350-7.
- FARCAS, A. M., BLACKLEDGE, N. P., SUDBERY, I., LONG, H. K., MCGOURAN, J. F., ROSE, N. R., LEE, S., SIMS, D., CERASE, A., SHEAHAN, T. W., KOSEKI, H.,

- BROCKDORFF, N., PONTING, C. P., KESSLER, B. M. & KLOSE, R. J. 2012. KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. *Elife*, 1.
- FENG, J., ZHOU, Y., CAMPBELL, S. L., LE, T., LI, E., SWEATT, J. D., SILVA, A. J. & FAN, G. P. 2010a. Dnmt1 and Dnmt3a maintain DNA methylation and regulate synaptic function in adult forebrain neurons. *Nature Neuroscience*, 13, 423-U37.
- FENG, J. X., LIU, T., QIN, B., ZHANG, Y. & LIU, X. S. 2012. Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, 7, 1728-1740.
- FENG, S. H., COKUS, S. J., ZHANG, X. Y., CHEN, P. Y., BOSTICK, M., GOLL, M. G., HETZEL, J., JAIN, J., STRAUSS, S. H., HALPERN, M. E., UKOMADU, C., SADLER, K. C., PRADHAN, S., PELLEGRINI, M. & JACOBSEN, S. E. 2010b. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 8689-8694.
- FICZ, G. & GRIBBEN, J. G. 2014. Loss of 5-hydroxymethylcytosine in cancer: cause or consequence? *Genomics*, 104, 352-7.
- FICZ, G., HORE, T. A., SANTOS, F., LEE, H. J., DEAN, W., ARAND, J., KRUEGER, F., OXLEY, D., PAUL, Y. L., WALTER, J., COOK, S. J., ANDREWS, S., BRANCO, M. R. & REIK, W. 2013. FGF Signaling Inhibition in ESCs Drives Rapid Genome-wide Demethylation to the Epigenetic Ground State of Pluripotency. *Cell Stem Cell*, 13, 351-359.
- 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., TALLMAN, M. S., SUN, Z. X., WOLNIAK, K., PEETERS, J. K., LIU, W., CHOE, S. E., FANTIN, V. R., PAIETTA, E., LOWENBERG, B., LICHT, J. D., GODLEY, L. A., DELWEL, R., VALK, P. J. M., THOMPSON, C. B., LEVINE, R. L. & MELNICK, A. 2010. Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation. *Cancer Cell*, 18, 553-567.
- FORREST, A. R. R., KAWAJI, H., REHLI, M., BAILLIE, J. K., DE HOON, M. J. L., HABERLE, V., LASSMANN, T., KULAKOVSKIY, I. V., LIZIO, M., ITOH, M., ANDERSSON, R., MUNGALL, C. J., MEEHAN, T. F., SCHMEIER, S., BERTIN, N., JORGENSEN, M., DIMONT, E., ARNER, E., SCHMIDL, C., SCHAEFER, U., MEDVEDEVA, Y. A., PLESSY, C., VITEZIC, M., SEVERIN, J., SEMPLE, C. A., ISHIZU, Y., YOUNG, R. S., FRANCESCOTTO, M., ALAM, I., ALBANESE, D., ALTSCHULER, G. M., ARAKAWA, T., ARCHER, J. A. C., ARNER, P., BABINA, M., RENNIE, S., BALWIERZ, P. J., BECKHOUSE, A. G., PRADHAN-BHATT, S., BLAKE, J. A., BLUMENTHAL, A., BODEGA, B., BONETTI, A., BRIGGS, J., BROMBACHER, F., BURROUGHS, A. M., CALIFANO, A.,

- CANNISTRACI, C. V., CARBAJO, D., CHEN, Y., CHIERICI, M., CIANI, Y., CLEVERS, H. C., DALLA, E., DAVIS, C. A., DETMAR, M., DIEHL, A. D., DOHI, T., DRABLOS, F., EDGE, A. S. B., EDINGER, M., EKWALL, K., ENDOH, M., ENOMOTO, H., FAGIOLINI, M., FAIRBAIRN, L., FANG, H., FARACH-CARSON, M. C., FAULKNER, G. J., FAVOROV, A. V., FISHER, M. E., FRITH, M. C., FUJITA, R., FUKUDA, S., FURLANELLO, C., FURUNO, M., FURUSAWA, J., GEIJTENBEEK, T. B., GIBSON, A. P., GINGERAS, T., GOLDOWITZ, D., GOUGH, J., GUHL, S., GULER, R., GUSTINCICH, S., HA, T. J., HAMAGUCHI, M., HARA, M., HARBERS, M., HARSHBARGER, J., HASEGAWA, A., HASEGAWA, Y., HASHIMOTO, T., HERLYN, M., HITCHENS, K. J., SUI, S. J. H., HOFMANN, O. M., HOOF, I., HORI, F., HUMINIECKI, L., et al. 2014. A promoter-level mammalian expression atlas. *Nature*, 507, 462-+.
- FRAUER, C., HOFFMANN, T., BULTMANN, S., CASA, V., CARDOSO, M. C., ANTES, I. & LEONHARDT, H. 2011. Recognition of 5-Hydroxymethylcytosine by the Uhrf1 SRA Domain. *Plos One*, 6.
- FROMMER, M., MCDONALD, L. E., MILLAR, D. S., COLLIS, C. M., WATT, F., GRIGG, G. W., MOLLOY, P. L. & 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 of the United States of America*, 89, 1827-1831.
- FUJITA, N., WATANABE, S., ICHIMURA, T., TSURUZOE, S., SHINKAI, Y., TACHIBANA, M., CHIBA, T. & NAKAO, M. 2003. Methyl-CpG binding domain 1 (MBD1) interacts with the Suv39h1-HP1 heterochromatic complex for DNA methylation-based transcriptional repression. *Journal of Biological Chemistry*, 278, 24132-24138.
- FUKS, F., BURGERS, W. A., BREHM, A., HUGHES-DAVIES, L. & KOUZARIDES, T. 2000. DNA methyltransferase Dnmt1 associates with histone deacetylase activity. *Nat Genet*, 24, 88-91.
- GABEL, H. W., KINDE, B., STROUD, H., GILBERT, C. S., HARMIN, D. A., KASTAN, N. R., HEMBERG, M., EBERT, D. H. & GREENBERG, M. E. 2015. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature*, 522, 89-93.
- GASTON, K. & FRIED, M. 1995. CpG Methylation Has Differential-Effects on the Binding of Yy1 and Ets Proteins to the Bidirectional Promoter of the Surf-1 and Surf-2 Genes. *Nucleic Acids Research*, 23, 901-909.
- GOLL, M. G. & BESTOR, T. H. 2005. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem*, 74, 481-514.
- GOLL, M. G., KIRPEKAR, F., MAGGERT, K. A., YODER, J. A., HSIEH, C. L., ZHANG, X., GOLIC, K. G., JACOBSEN, S. E. & BESTOR, T. H. 2006. Methylation of

- trNAAsp by the DNA methyltransferase homolog Dnmt2. *Science*, 311, 395-8.
- GOMMERS-AMPT, J. H., VAN LEEUWEN, F., DE BEER, A. L., VLIEGENTHART, J. F., DIZDAROGLU, M., KOWALAK, J. A., CRAIN, P. F. & BORST, P. 1993. beta-D-glucosyl-hydroxymethyluracil: a novel modified base present in the DNA of the parasitic protozoan *T. brucei*. *Cell*, 75, 1129-36.
- GONG, Z., MORALES-RUIZ, T., ARIZA, R. R., ROLDAN-ARJONA, T., DAVID, L. & ZHU, J. K. 2002. ROS1, a repressor of transcriptional gene silencing in *Arabidopsis*, encodes a DNA glycosylase/lyase. *Cell*, 111, 803-14.
- GONZALEZ-VALLINAS, J., PAGES, A., SINGH, B. & EYRAS, E. 2015. A semi-supervised approach uncovers thousands of intragenic enhancers differentially activated in human cells. *Bmc Genomics*, 16.
- GOSSEN, M. & BUJARD, H. 1992. Tight Control of Gene-Expression in Mammalian-Cells by Tetracycline-Responsive Promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 89, 5547-5551.
- GRAFF, J. & TSAI, L. H. 2013. Histone acetylation: molecular mnemonics on the chromatin. *Nature Reviews Neuroscience*, 14, 97-111.
- GREAVES, M. & MALEY, C. C. 2012. Clonal evolution in cancer. *Nature*, 481, 306-13.
- GREGORY, T. R. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biological Reviews*, 76, 65-101.
- GRISHKEVICH, V. & YANAI, I. 2014. Gene length and expression level shape genomic novelties. *Genome Research*, 24, 1497-1503.
- GU, T. P., GUO, F., YANG, H., WU, H. P., XU, G. F., LIU, W., XIE, Z. G., SHI, L. Y., HE, X. Y., JIN, S. G., IQBAL, K., SHI, Y. J. G., DENG, Z. X., SZABO, P. E., PFEIFER, G. P., LI, J. S. & XU, G. L. 2011. The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature*, 477, 606-U136.
- GUENTHER, M. G., LEVINE, S. S., BOYER, L. A., JAENISCH, R. & YOUNG, R. A. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130, 77-88.
- GUO, H., ZHU, P., YAN, L., LI, R., HU, B., LIAN, Y., YAN, J., REN, X., LIN, S., LI, J., JIN, X., SHI, X., LIU, P., WANG, X., WANG, W., WEI, Y., LI, X., GUO, F., WU, X., FAN, X., YONG, J., WEN, L., XIE, S. X., TANG, F. & QIAO, J. 2014. The DNA methylation landscape of human early embryos. *Nature*, 511, 606-10.

- GUO, J. U., SU, Y. J., ZHONG, C., MING, G. L. & SONG, H. J. 2011. Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain. *Cell*, 145, 423-434.
- HANAHAN, D. & WEINBERG, R. A. 2000. The hallmarks of cancer. *Cell*, 100, 57-70.
- HANSEN, R. S., WIJMENGA, C., LUO, P., STANEK, A. M., CANFIELD, T. K., WEEMAES, C. M. R. & GARTLER, S. M. 1999. The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome. *Proceedings of the National Academy of Sciences of the United States of America*, 96, 14412-14417.
- HARGREAVES, D. C., HORNG, T. & MEDZHITOV, R. 2009. Control of Inducible Gene Expression by Signal-Dependent Transcriptional Elongation. *Cell*, 138, 129-145.
- HATA, K. & SAKAKI, Y. 1997. Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene*, 189, 227-234.
- HAWLEY, D. K. & ROEDER, R. G. 1987. Functional Steps in Transcription Initiation and Reinitiation from the Major Late Promoter in a Hela Nuclear Extract. *Journal of Biological Chemistry*, 262, 3452-3461.
- HE, Y. F., LI, B. Z., LI, Z., LIU, P., WANG, Y., TANG, Q. Y., DING, J. P., JIA, Y. Y., CHEN, Z. C., LI, L., SUN, Y., LI, X. X., DAI, Q., SONG, C. X., ZHANG, K. L., HE, C. & XU, G. L. 2011. Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science*, 333, 1303-1307.
- HENDRICH, B. & BIRD, A. 1998. Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol Cell Biol*, 18, 6538-47.
- HENDRICH, B., GUY, J., RAMSAHOYE, B., WILSON, V. A. & 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. & BIRD, A. 1999. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature*, 401, 301-304.
- HERMANN, A., GOYAL, R. & JELTSCH, A. 2004. The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. *J Biol Chem*, 279, 48350-9.
- HEROLD, M., BARTKUHN, M. & RENKAWITZ, R. 2012. CTCF: insights into insulator function during development. *Development*, 139, 1045-1057.

- HO, K. L., MCNAE, L. W., SCHMIEDEBERG, L., KLOSE, R. J., BIRD, A. P. & WALKINSHAW, M. D. 2008. MeCP2 binding to DNA depends upon hydration at methyl-CpG. *Molecular Cell*, 29, 525-531.
- HODGES, C., BINTU, L., LUBKOWSKA, L., KASHLEV, M. & BUSTAMANTE, C. 2009. Nucleosomal Fluctuations Govern the Transcription Dynamics of RNA Polymerase II. *Science*, 325, 626-628.
- HODGES, E., MOLARO, A., DOS SANTOS, C. O., THEKKAT, P., SONG, Q., UREN, P. J., PARK, J., BUTLER, J., RAFII, S., MCCOMBIE, W. R., SMITH, A. D. & HANNON, G. J. 2011. Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hematopoietic Compartment. *Molecular Cell*, 44, 17-28.
- HOLMAN, D. J. 2000. Principles of population genetics, third edition. *American Journal of Physical Anthropology*, 112, 433-435.
- HON, G. C., HAWKINS, R. D., CABALLERO, O. L., LO, C., LISTER, R., PELIZZOLA, M., VALSESIA, A., YE, Z., KUAN, S., EDSALL, L. E., CAMARGO, A. A., STEVENSON, B. J., ECKER, J. R., BAFNA, V., STRAUSBERG, R. L., SIMPSON, A. J. & REN, B. 2012. Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Research*, 22, 246-258.
- HOTCHKISS, R. D. 1948. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *J Biol Chem*, 175, 315-32.
- HSIEH, C. L. 1994. Dependence of Transcriptional Repression on CpG Methylation Density. *Molecular and Cellular Biology*, 14, 5487-5494.
- HSIEH, C. L. 1997. Stability of patch methylation and its impact in regions of transcriptional initiation and elongation. *Molecular and Cellular Biology*, 17, 5897-5904.
- HUANG, Y., CHAVEZ, L., CHANG, X., WANG, X., PASTOR, W. A., KANG, J., ZEPEDA-MARTINEZ, J. A., PAPE, U. J., JACOBSEN, S. E., PETERS, B. & RAO, A. 2014. Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 1361-1366.
- HUGHES, L. A., MELOTTE, V., DE SCHRIJVER, J., DE MAAT, M., SMIT, V. T., BOVEE, J. V., FRENCH, P. J., VAN DEN BRANDT, P. A., SCHOUTEN, L. J., DE MEYER, T., VAN CRIEKINGE, W., AHUJA, N., HERMAN, J. G., WEIJENBERG, M. P. & VAN ENGELAND, M. 2013. The CpG island methylator phenotype: what's in a name? *Cancer Res*, 73, 5858-68.
- ICHIMURA, T., WATANABE, S., SAKAMOTO, Y., AOTO, T., FUJITA, N. & NAKAO, M. 2005. Transcriptional repression and heterochromatin formation by

- MBD1 and MCAF/AM family proteins. *Journal of Biological Chemistry*, 280, 13928-13935.
- ILLINGWORTH, R. S. & BIRD, A. P. 2009. CpG islands - 'A rough guide'. *Febs Letters*, 583, 1713-1720.
- ILLINGWORTH, R. S., GRUENEWALD-SCHNEIDER, U., WEBB, S., KERR, A. R. W., JAMES, K. D., TURNER, D. J., SMITH, C., HARRISON, D. J., ANDREWS, R. & BIRD, A. P. 2010. Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome. *Plos Genetics*, 6.
- IRIZARRY, R. A., LADD-ACOSTA, C., WEN, B., WU, Z. J., MONTANO, C., ONYANGO, P., CUI, H. M., GABO, K., RONGIONE, M., WEBSTER, M., JI, H., POTASH, J. B., SABUNCYAN, S. & FEINBERG, A. P. 2009. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature Genetics*, 41, 178-186.
- ITO, S., SHEN, L., DAI, Q., WU, S. C., COLLINS, L. B., SWENBERG, J. A., HE, C. & ZHANG, Y. 2011. Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science*, 333, 1300-1303.
- IYER, L. M., TAHILIANI, M., RAO, A. & 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.
- JANULAITIS, A., KLIMASAUSKAS, S., PETRUSYTE, M. & BUTKUS, V. 1983. Cytosine Modification in DNA by Bcni Methylase Yields N-4-Methylcytosine. *Febs Letters*, 161, 131-134.
- JELTSCH, A. 2008. Reading and writing DNA methylation. *Nature Structural & Molecular Biology*, 15, 1003-1004.
- JIN, B., YAO, B., LI, J. L., FIELDS, C. R., DELMAS, A. L., LIU, C. & ROBERTSON, K. D. 2009. DNMT1 and DNMT3B modulate distinct polycomb-mediated histone modifications in colon cancer. *Cancer Res*, 69, 7412-21.
- JJINGO, D., CONLEY, A. B., YI, S. V., LUNYAK, V. V. & JORDAN, I. K. 2012. On the presence and role of human gene-body DNA methylation. *Oncotarget*, 3, 462-474.
- JONES, P. A. 1999. The DNA methylation paradox. *Trends in Genetics*, 15, 34-37.
- JORGENSEN, H. F., BEN-PORATH, I. & BIRD, A. P. 2004. Mbd1 is recruited to both methylated and nonmethylated CpGs via distinct DNA binding domains. *Molecular and Cellular Biology*, 24, 3387-3395.

- KAER, K. & SPEEK, M. 2013. Retroelements in human disease. *Gene*, 518, 231-241.
- KAMERLIN, S. C. L. & WARSHEL, A. 2009. On the Energetics of ATP Hydrolysis in Solution. *Journal of Physical Chemistry B*, 113, 15692-15698.
- KANIUSAITE, M., ASTROMSKAS, E., ALZBUTAS, G., BRUZAITE, R. & LAGUNAVIČIUS, A. 2015. Fast and convenient 5-hydroxymethylcytosine enrichment workflow for next-generation sequencing. *Nat Meth*, 12.
- KARIMI, M. M., GOYAL, P., MAKSAKOVA, I. A., BILENKY, M., LEUNG, D., TANG, J. X., SHINKAI, Y., MAGER, D. L., JONES, S., HIRST, M. & LORINCZ, M. C. 2011. DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell Stem Cell*, 8, 676-87.
- KEREN, H., LEV-MAOR, G. & AST, G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics*, 11, 345-355.
- KHARE, T., PAI, S., KONCEVICIUS, K., PAL, M., KRIUKIENE, E., LIUTKEVICIUTE, Z., IRIMIA, M., JIA, P. X., PTAK, C., XIA, M. H., TICE, R., TOCHIGI, M., MORERA, S., NAZARIANS, A., BELSHAM, D., WONG, A. H. C., BLENCOWE, B. J., WANG, S. C., KAPRANOV, P., KUSTRA, R., LABRIE, V., KLIMASAUSKAS, S. & PETRONIS, A. 2012. 5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary. *Nature Structural & Molecular Biology*, 19, 1037-U94.
- KIM, T. K., HEMBERG, M., GRAY, J. M., COSTA, A. M., BEAR, D. M., WU, J., HARMIN, D. A., LAPTEWICZ, M., BARBARA-HALEY, K., KUERSTEN, S., MARKENSCOFF-PAPADIMITRIOU, E., KUHLE, D., BITO, H., WORLEY, P. F., KREIMAN, G. & GREENBERG, M. E. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465, 182-U65.
- KLOSE, R. J. & BIRD, A. P. 2006. Genomic DNA methylation: the mark and its mediators. *Trends in Biochemical Sciences*, 31, 89-97.
- KO, M., AN, J., BANDUKWALA, H. S., CHAVEZ, L., AIJO, T., PASTOR, W. A., SEGAL, M. F., LI, H., KOH, K. P., LAHDESMÄKI, H., HOGAN, P. G., ARAVIND, L. & RAO, A. 2013. Modulation of TET2 expression and 5-methylcytosine oxidation by the CXXC domain protein IDAX. *Nature*, 497, 122-6.
- 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., LIU, X. S., ARAVIND, L., AGARWAL, S., MACIEJEWSKI, J. P. & RAO, A. 2010. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature*, 468, 839-43.

- KOVESDI, I., REICHEL, R. & NEVINS, J. R. 1987. Role of an Adenovirus E2 Promoter Binding-Factor in E1a-Mediated Coordinate Gene-Control. *Proceedings of the National Academy of Sciences of the United States of America*, 84, 2180-2184.
- KRIAUCIONIS, S. & HEINTZ, N. 2009. The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science*, 324, 929-930.
- KUDO, Y., TATEISHI, K., YAMAMOTO, K., YAMAMOTO, S., ASAOKA, Y., IJICHI, H., NAGAE, G., YOSHIDA, H., ABURATANI, H. & KOIKE, K. 2012. Loss of 5-hydroxymethylcytosine is accompanied with malignant cellular transformation. *Cancer Science*, 103, 670-676.
- KULIS, M. & ESTELLER, M. 2010. DNA Methylation and Cancer. *Epigenetics and Cancer, Pt A*, 70, 27-56.
- KULIS, M., HEATH, S., BIBIKOVA, M., QUEIROS, A. C., NAVARRO, A., CLOT, G., MARTINEZ-TRILLOS, A., CASTELLANO, G., BRUN-HEATH, I., PINYOL, M., BARBERAN-SOLER, S., PAPASAIKAS, P., JARES, P., BEA, S., RICO, D., ECKER, S., RUBIO, M., ROYO, R., HO, V., KLOTZLE, B., HERNANDEZ, L., CONDE, L., LOPEZ-GUERRA, M., COLOMER, D., VILLAMOR, N., AYMERICH, M., ROZMAN, M., BAYES, M., GUT, M., GELPI, J. L., OROZCO, M., FAN, J. B., QUESADA, V., PUENTE, X. S., PISANO, D. G., VALENCIA, A., LOPEZ-GUILLERMO, A., GUT, I., LOPEZ-OTIN, C., CAMPO, E. & MARTIN-SUBERO, J. I. 2012. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nature Genetics*, 44, 1236-1242.
- LAIRD, P. W. 2010. Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11, 191-203.
- LAW, J. A. & JACOBSEN, S. E. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet*, 11, 204-20.
- LAZCANO, A., GUERRERO, R., MARGULIS, L. & ORO, J. 1988. The evolutionary transition from RNA to DNA in early cells. *J Mol Evol*, 27, 283-90.
- LE GUEZENNEC, X., VERMEULEN, M., BRINKMAN, A. B., HOEIJMAKERS, W. A. M., COHEN, A., LASONDER, E. & STUNNENBERG, H. G. 2006. MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties. *Molecular and Cellular Biology*, 26, 843-851.
- LEE, D. H., SINGH, P., TSAI, S. Y., OATES, N., SPALLA, A., SPALLA, C., BROWN, L., RIVAS, G., LARSON, G., RAUCH, T. A., PFEIFER, G. P. & SZABO, P. E. 2010. CTCF-Dependent Chromatin Bias Constitutes Transient Epigenetic

- Memory of the Mother at the H19-Igf2 Imprinting Control Region in Prospermatogonia. *Plos Genetics*, 6.
- LEE, J. H. & SKALNIK, D. G. 2005. CpG-binding protein (CXXC finger protein 1) is a component of the mammalian set1 histone H3-Lys(4) methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *Journal of Biological Chemistry*, 280, 41725-41731.
- LEUNG, D., JUNG, I., RAJAGOPAL, N., SCHMITT, A., SELVARAJ, S., LEE, A. Y., YEN, C. A., LIN, S., LIN, Y., QIU, Y., XIE, W., YUE, F., HARIHARAN, M., RAY, P., KUAN, S., EDSALL, L., YANG, H., CHI, N. C., ZHANG, M. Q., ECKER, J. R. & REN, B. 2015. Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, 518, 350-4.
- LEY, T. J. & GRANT, G. A. P. P. 2011. DNMT3A Mutations in Acute Myeloid Leukemia. *Blood*, 118, 1815-1815.
- LI, B., CAREY, M. & WORKMAN, J. L. 2007. The role of chromatin during transcription. *Cell*, 128, 707-719.
- LI, E., BESTOR, T. H. & JAENISCH, R. 1992. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, 69, 915-26.
- LI, Y., WANG, H. Y., MUFFAT, J., CHENG, A. W., ORLANDO, D. A., LOVEN, J., KWOK, S. M., FELDMAN, D. A., BATEUP, H. S., GAO, Q., HOCKEMEYER, D., MITALPOVA, M., LEWIS, C. A., VANDER HEIDEN, M. G., SUR, M., YOUNG, R. A. & JAENISCH, R. 2013. Global Transcriptional and Translational Repression in Human-Embryonic-Stem-Cell-Derived Rett Syndrome Neurons. *Cell Stem Cell*, 13, 446-458.
- LI, Z., CAI, X. Q., CAI, C. L., WANG, J. P., ZHANG, W. Y., PETERSEN, B. E., YANG, F. C. & XU, M. J. 2011. 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. F., CEOL, C., WU, F. Z., LARSON, A., DRESSER, K., XU, W. Q., TAN, L., HU, Y. G., ZHAN, Q., LEE, C. W., HU, D., LIAN, B. Q., KLEFFEL, S., YANG, Y. J., NEISWENDER, J., KHORASANI, A. J., FANG, R., LEZCANO, C., DUNCAN, L. M., SCOLYER, R. A., THOMPSON, J. F., KAKAVAND, H., HOUVRAS, Y., ZON, L. I., MIHM, M. C., KAISER, U. B., SCHATTON, T., WODA, B. A., MURPHY, G. F. & SHI, Y. J. G. 2012. Loss of 5-Hydroxymethylcytosine Is an Epigenetic Hallmark of Melanoma. *Cell*, 150, 1135-1146.
- LIENERT, F., WIRBELAUER, C., SOM, I., DEAN, A., MOHN, F. & SCHUBELER, D. 2011. Identification of genetic elements that autonomously determine DNA methylation states. *Nature Genetics*, 43, 1091-U78.

- LIENHARD, M., GRIMM, C., MORKEL, M., HERWIG, R. & CHAVEZ, L. 2014. MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. *Bioinformatics*, 30, 284-286.
- 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., YU, M., TONTI-FILIPPINI, J., HEYN, H., HU, S. J., WU, J. C., RAO, A., ESTELLER, M., HE, C., HAGHIGHI, F. G., SEJNOWSKI, T. J., BEHRENS, M. M. & ECKER, J. R. 2013. Global Epigenomic Reconfiguration During Mammalian Brain Development. *Science*, 341, 629-+.
- 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., EDSALL, L., ANTOSIEWICZ-BOURGET, J., STEWART, R., RUOTTI, V., MILLAR, A. H., THOMSON, J. A., REN, B. & ECKER, J. R. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462, 315-322.
- LONG, H. K., BLACKLEDGE, N. P. & KLOSE, R. J. 2013. ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. *Biochemical Society Transactions*, 41, 727-740.
- LORINCZ, M. C., DICKERSON, D. R., SCHMITT, M. & GROUDINE, M. 2004. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nature Structural & Molecular Biology*, 11, 1068-1075.
- LOU, S. K., LEE, H. M., QIN, H., LI, J. W., GAO, Z. B., LIU, X., CHAN, L. L., LAM, V. K. L., SO, W. Y., WANG, Y., LOK, S., WANG, J., MA, R. C. W., TSUI, S. K. W., CHAN, J. C. N., CHAN, T. F. & YIP, K. Y. 2014. Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome Biology*, 15.
- LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15.
- LUCO, R. F., ALLO, M., SCHOR, I. E., KORNBLIHTT, A. R. & MISTELI, T. 2011. Epigenetics in Alternative Pre-mRNA Splicing. *Cell*, 144, 16-26.
- LUND, K., COLE, J. J., VANDERKRAATS, N. D., MCBRYAN, T., PCHELINTSEV, N. A., CLARK, W., COPLAND, M., EDWARDS, J. R. & ADAMS, P. D. 2014. DNMT inhibitors reverse a specific signature of aberrant promoter DNA methylation and associated gene silencing in AML. *Genome Biology*, 15.

- LUNDIN, D., GRIBALDO, S., TORRENTS, E., SJOBERG, B. M. & POOLE, A. M. 2010. Ribonucleotide reduction - horizontal transfer of a required function spans all three domains. *Bmc Evolutionary Biology*, 10.
- LUO, T., CUI, S. J., BIAN, C. J. & YU, X. C. 2013. Uhrf2 is important for DNA damage response in vascular smooth muscle cells. *Biochemical and Biophysical Research Communications*, 441, 65-70.
- LYON, M. F. 1961. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature*, 190, 372-3.
- LYST, M. J., EKIERT, R., EBERT, D. H., MERUSI, C., NOWAK, J., SELFRIDGE, J., GUY, J., KASTAN, N. R., ROBINSON, N. D., ALVES, F. D., RAPPSILBER, J., GREENBERG, M. E. & BIRD, A. 2013. Rett syndrome mutations abolish the interaction of MeCP2 with the NCoR/SMRT co-repressor. *Nature Neuroscience*, 16, 898-U268.
- MAITI, A. & DROHAT, A. C. 2011. Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine POTENTIAL IMPLICATIONS FOR ACTIVE DEMETHYLATION OF CpG SITES. *Journal of Biological Chemistry*, 286, 35334-35338.
- MAOR, G. L., YEARIM, A. & AST, G. 2015. The alternative role of DNA methylation in splicing regulation. *Trends in Genetics*, 31, 274-280.
- MAUNAKEA, A. K., CHEPELEV, I., CUI, K. R. & ZHAO, K. J. 2013. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Research*, 23, 1256-1269.
- MAYER, W., NIVELEAU, A., WALTER, J., FUNDELE, R. & HAAF, T. 2000. Embryogenesis - Demethylation of the zygotic paternal genome. *Nature*, 403, 501-502.
- MEEHAN, R. R., LEWIS, J. D. & BIRD, A. P. 1992. Characterization of MeCP2, a vertebrate DNA binding protein with affinity for methylated DNA. *Nucleic Acids Res*, 20, 5085-92.
- MELLEN, M., AYATA, P., DEWELL, S., KRIAUCIONIS, S. & HEINTZ, N. 2012. MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell*, 151, 1417-30.
- MERLO, A., HERMAN, J. G., MAO, L., LEE, D. J., GABRIELSON, E., BURGER, P. C., BAYLIN, S. B. & SIDRANSKY, D. 1995. 5' CpG island methylation is associated with transcriptional silencing of the tumour suppressor p16/CDKN2/MTS1 in human cancers. *Nat Med*, 1, 686-92.
- MIKKELSEN, T. S., KU, M., JAFFE, D. B., ISSAC, B., LIEBERMAN, E., GIANNOUKOS, G., ALVAREZ, P., BROCKMAN, W., KIM, T. K., KOCH, R.

- P., LEE, W., MENDENHALL, E., O'DONOVAN, A., PRESSER, A., RUSS, C., XIE, X., MEISSNER, A., WERNIG, M., JAENISCH, R., NUSBAUM, C., LANDER, E. S. & BERNSTEIN, B. E. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448, 553-60.
- MILLER, C. A. & SWEATT, J. D. 2007. Covalent modification of DNA regulates memory formation. *Neuron*, 53, 857-869.
- MOHANDAS, T., SPARKES, R. S. & SHAPIRO, L. J. 1981. Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science*, 211, 393-6.
- MULLER, T., GESSI, M., WAHA, A., ISSELSTEIN, L. J., LUXEN, D., FREIHOFF, D., FREIHOFF, J., BECKER, A., SIMON, M., PIETSCH, T. & WAHA, A. 2012. Nuclear exclusion of TET1 is associated with loss of 5-hydroxymethylcytosine in IDH1 wildtype gliomas. *Klinische Padiatrie*, 224, 222-223.
- NAN, X. S., NG, H. H., JOHNSON, C. A., LAHERTY, C. D., TURNER, B. M., EISENMAN, R. N. & BIRD, A. 1998. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*, 393, 386-389.
- NANTY, L., CARBAJOSA, G., HEAP, G. A., RATNIEKS, F., VAN HEEL, D. A., DOWN, T. A. & RAKYAN, V. K. 2011. Comparative methylomics reveals gene-body H3K36me3 in *Drosophila* predicts DNA methylation and CpG landscapes in other invertebrates. *Genome Research*, 21, 1841-1850.
- NEIDLE, S. 2008. Principles of Nucleic Acid Structure. *Principles of Nucleic Acid Structure*, 1-289.
- NG, H. H., ZHANG, Y., HENDRICH, B., JOHNSON, C. A., TURNER, B. M., ERDJUMENT-BROMAGE, H., TEMPST, P., REINBERG, D. & BIRD, A. 1999. MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nature Genetics*, 23, 58-61.
- NG, R. K., DEAN, W., DAWSON, C., LUCIFERO, D., MADEJA, Z., REIK, W. & HEMBERGER, M. 2008. Epigenetic restriction of embryonic cell lineage fate by methylation of Elf5. *Nature Cell Biology*, 10, 1280-U68.
- 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., VERHAAK, R. G. W., HOADLEY, K. A., HAYES, D. N., PEROU, C. M., SCHMIDT, H. K., DING, L., WILSON, R. K., VAN DEN BERG, D., SHEN, H., BENGTTSSON, H., NEUVIAL, P., COPE, L. M., BUCKLEY, J., HERMAN, J. G., BAYLIN, S. B., LAIRD, P. W., ALDAPE, K. & NETWORK, C. G. A. R. 2010.

- Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. *Cancer Cell*, 17, 510-522.
- OHKI, I., SHIMOTAKE, N., FUJITA, N., JEE, J., IKEGAMI, T., NAKAO, M. & SHIRAKAWA, M. 2001. Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. *Cell*, 105, 487-97.
- OKANO, M., BELL, D. W., HABER, D. A. & LI, E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99, 247-57.
- ONO, R., TAKI, T., TAKETANI, T., TANIWAKI, M., KOBAYASHI, H. & HAYASHI, Y. 2002. LCX, leukemia-associated protein with a CXXC domain, is fused to MLL in acute myeloid leukemia with trilineage dysplasia having t(10;11)(q22;q23). *Cancer Research*, 62, 4075-4080.
- OOI, S. K. & BESTOR, T. H. 2008. The colorful history of active DNA demethylation. *Cell*, 133, 1145-8.
- PANNING, B. & JAENISCH, R. 1996. DNA hypomethylation can activate Xist expression and silence X-linked genes. *Genes & Development*, 10, 1991-2002.
- PASTOR, W. A., PAPE, U. J., HUANG, Y., HENDERSON, H. R., LISTER, R., KO, M., MCLOUGHLIN, E. M., BRUDNO, Y., MAHAPATRA, S., KAPRANOV, P., TAHILIANI, M., DALEY, G. Q., LIU, X. S., ECKER, J. R., MILOS, P. M., AGARWAL, S. & RAO, A. 2011. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature*, 473, 394-397.
- PATIL, V. S. & KAI, T. 2010. Repression of Retroelements in Drosophila Germline via piRNA Pathway by the Tudor Domain Protein Tejas. *Current Biology*, 20, 724-730.
- PATTON, J. T. & SPENCER, E. 2000. Genome replication and packaging of segmented double-stranded RNA viruses. *Virology*, 277, 217-225.
- PENTERMAN, J., ZILBERMAN, D., HUH, J. H., BALLINGER, T., HENIKOFF, S. & FISCHER, R. L. 2007. DNA demethylation in the Arabidopsis genome. *Proc Natl Acad Sci U S A*, 104, 6752-7.
- PFAFFENEDER, T., HACKNER, B., TRUSS, M., MUNZEL, M., MULLER, M., DEIML, C. A., HAGEMEIERS, C. & CARELL, T. 2011. The Discovery of 5-Formylcytosine in Embryonic Stem Cell DNA. *Angewandte Chemie-International Edition*, 50, 7008-7012.
- PRENDERGAST, G. C. & ZIFF, E. B. 1991. Methylation-Sensitive Sequence-Specific DNA-Binding by the C-Myc Basic Region. *Science*, 251, 186-189.

- RADDATZ, G., GUZZARDO, P. M., OLOVA, N., FANTAPPIE, M. R., RAMPP, M., SCHAEFER, M., REIK, W., HANNON, G. J. & LYKO, F. 2013. Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proc Natl Acad Sci U S A*, 110, 8627-31.
- RAN, F. A., HSU, P. D., WRIGHT, J., AGARWALA, V., SCOTT, D. A. & ZHANG, F. 2013. Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, 8, 2281-2308.
- REIK, W., COLLICK, A., NORRIS, M. L., BARTON, S. C. & SURANI, M. A. 1987. Genomic imprinting determines methylation of parental alleles in transgenic mice. *Nature*, 328, 248-51.
- ROBERTSON, J., ROBERTSON, A. B. & KLUNGLAND, A. 2011. The presence of 5-hydroxymethylcytosine at the gene promoter and not in the gene body negatively regulates gene expression. *Biochem Biophys Res Commun*, 411, 40-3.
- ROHDE, C., ZHANG, Y. Y., REINHARDT, R. & JELTSCH, A. 2010. BISMA - Fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences. *Bmc Bioinformatics*, 11.
- ROLOFF, T. C., ROPERS, H. H. & NUBER, U. A. 2003. Comparative study of methyl-CpG-binding domain proteins. *BMC Genomics*, 4, 1.
- ROMERO, I. G., RUVINSKY, I. & GILAD, Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. *Nat Rev Genet*, 13, 505-16.
- ROOKE, N. & UNDERWOOD, J. 2001. In vitro RNA splicing in mammalian cell extracts. *Curr Protoc Cell Biol*, Chapter 11, Unit 11 17.
- ROULOIS, D., LOO YAU, H., SINGHANIA, R., WANG, Y., DANESH, A., SHEN, S. Y., HAN, H., LIANG, G., JONES, P. A., PUGH, T. J., O'BRIEN, C. & DE CARVALHO, D. D. 2015. DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. *Cell*, 162, 961-73.
- SAKSOUK, N., AVVAKUMOV, N., CHAMPAGNE, K. S., HUNG, T., DOYON, Y., CAYROU, C., PAQUET, E., ULLAH, M., LANDRY, A. J., COTE, V., YANG, X. J., GOZANI, O., KUTATELADZE, T. G. & COTE, J. 2009. HBO1 HAT Complexes Target Chromatin throughout Gene Coding Regions via Multiple PHD Finger Interactions with Histone H3 Tail. *Molecular Cell*, 33, 257-265.
- SARRAF, S. A. & STANCHEVA, I. 2004. Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly. *Molecular Cell*, 15, 595-605.

- SCHMITTGEN, T. D. & LIVAK, K. J. 2008. Analyzing real-time PCR data by the comparative C(T) method. *Nat Protoc*, 3, 1101-8.
- SCHUBELER, D. 2015. Function and information content of DNA methylation. *Nature*, 517, 321-6.
- SCHUETTENGROBER, B., CHOURROUT, D., VERVOORT, M., LEBLANC, B. & CAVALLI, G. 2007. Genome regulation by polycomb and trithorax proteins. *Cell*, 128, 735-745.
- SHARP, A. J., STATHAKI, E., MIGLIAVACCA, E., BRAHMACHARY, M., MONTGOMERY, S. B., DUPRE, Y. & ANTONARAKIS, S. E. 2011. DNA methylation profiles of human active and inactive X chromosomes. *Genome Research*, 21, 1592-1600.
- SHEN, E. S. & WHITLOCK, J. P., JR. 1989. The potential role of DNA methylation in the response to 2,3,7,8-tetrachlorodibenzo-p-dioxin. *J Biol Chem*, 264, 17754-8.
- SHEN, L., WU, H., DIEP, D., YAMAGUCHI, S., D'ALESSIO, A. C., FUNG, H. L., ZHANG, K. & ZHANG, Y. 2013. Genome-wide Analysis Reveals TET- and TDG-Dependent 5-Methylcytosine Oxidation Dynamics. *Cell*, 153, 692-706.
- SHEN, L. & ZHANG, Y. 2012. Enzymatic Analysis of Tet Proteins: Key Enzymes in the Metabolism of DNA Methylation. *Nucleosomes, Histones & Chromatin, Pt A*, 512, 93-105.
- SHUKLA, S., KAVAK, E., GREGORY, M., IMASHIMIZU, M., SHUTINOSKI, B., KASHLEV, M., OBERDOERFFER, P., SANDBERG, R. & OBERDOERFFER, S. 2011. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, 479, 74-U99.
- SKENE, P. J., ILLINGWORTH, R. S., WEBB, S., KERR, A. R., JAMES, K. D., TURNER, D. J., ANDREWS, R. & BIRD, A. P. 2010. Neuronal MeCP2 is expressed at near histone-octamer levels and globally alters the chromatin state. *Mol Cell*, 37, 457-68.
- SMITH, Z. D., CHAN, M. M., HUMM, K. C., KARNIK, R., MEKHOUBAD, S., REGEV, A., EGGAN, K. & MEISSNER, A. 2014. DNA methylation dynamics of the human preimplantation embryo. *Nature*, 511, 611-5.
- SMRT, R. D., EAVES-EGENES, J., BARKHO, B. Z., SANTISTEVAN, N. J., ZHAO, C., AIMONE, J. B., GAGE, F. H. & ZHAO, X. 2007. Mecp2 deficiency leads to delayed maturation and altered gene expression in hippocampal neurons. *Neurobiology of Disease*, 27, 77-89.

- SONG, C. X., SZULWACH, K. E., FU, Y., DAI, Q., YI, C. Q., LI, X. K., LI, Y. J., CHEN, C. H., ZHANG, W., JIAN, X., WANG, J., ZHANG, L., LOONEY, T. J., ZHANG, B. C., GODLEY, L. A., HICKS, L. M., LAHN, B. T., JIN, P. & HE, C. 2011a. Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nature Biotechnology*, 29, 68-72.
- SONG, J. K., RECHKOBLIT, O., BESTOR, T. H. & PATEL, D. J. 2011b. Structure of DNMT1-DNA Complex Reveals a Role for Autoinhibition in Maintenance DNA Methylation. *Science*, 331, 1036-1040.
- SPIRIN, A. S. 2010. Ancient RNA World. *Paleontological Journal*, 44, 737-746.
- SPRUIJT, C. G., GNERLICH, F., SMITS, A. H., PFAFFENEDER, T., JANSEN, P. W. T. C., BAUER, C., MUNZEL, M., WAGNER, M., MULLER, M., KHAN, F., EBERL, H. C., MENSINGA, A., BRINKMAN, A. B., LEPHIKOV, K., MULLER, U., WALTER, J., BOELEN, R., VAN INGEN, H., LEONHARDT, H., CARELL, T. & VERMEULEN, M. 2013. Dynamic Readers for 5-(Hydroxy)Methylcytosine and Its Oxidized Derivatives. *Cell*, 152, 1146-1159.
- STADLER, M. B., MURR, R., BURGER, L., IVANEK, R., LIENERT, F., SCHOLER, A., WIRBELAUER, C., OAKELEY, E. J., GAIDATZIS, D., TIWARI, V. K. & SCHUBELER, D. 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480, 490-495.
- STEITZ, T. A. 1998. Structural biology - A mechanism for all polymerases. *Nature*, 391, 231-232.
- STEITZ, T. A. 1999. DNA polymerases: Structural diversity and common mechanisms. *Journal of Biological Chemistry*, 274, 17395-17398.
- STENGLEIN, M. D., BURNS, M. B., LI, M., LENGUEL, J. & HARRIS, R. S. 2010. APOBEC3 proteins mediate the clearance of foreign DNA from human cells. *Nature Structural & Molecular Biology*, 17, 222-U13.
- STIRZAKER, C., TABERLAY, P. C., STATHAM, A. L. & CLARK, S. J. 2014. Mining cancer methylomes: prospects and challenges. *Trends in Genetics*, 30, 75-84.
- STONE, A., ZOTENKO, E., LOCKE, W. J., KORBIE, D., MILLAR, E. K. A., PIDSLEY, R., STIRZAKER, C., GRAHAM, P., TRAU, M., MUSGROVE, E. A., NICHOLSON, R. I., GEE, J. M. W. & CLARK, S. J. 2015. DNA methylation of oestrogen-regulated enhancers defines endocrine sensitivity in breast cancer. *Nature Communications*, 6.
- SWISHER, J. F., RAND, E., CEDAR, H. & MARIE PYLE, A. 1998. Analysis of putative RNase sensitivity and protease insensitivity of demethylation activity in extracts from rat myoblasts. *Nucleic Acids Res*, 26, 5573-80.

- SZULWACH, K. E., LI, X. K., LI, Y. J., SONG, C. X., WU, H., DAI, Q., IRIER, H., UPADHYAY, A. K., GEARING, M., LEVEY, A. I., VASANTHAKUMAR, A., GODLEY, L. A., CHANG, Q., CHENG, X. D., HE, C. & JIN, P. 2011. 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nature Neuroscience*, 14, 1607-U150.
- TAHILIANI, M., KOH, K. P., SHEN, Y., PASTOR, W. A., BANDUKWALA, H., BRUDNO, Y., AGARWAL, S., IYER, L. M., LIU, D. R., ARAVIND, L. & RAO, A. 2009a. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, 324, 930-5.
- TAHILIANI, M., KOH, K. P., SHEN, Y. H., PASTOR, W. A., BANDUKWALA, H., BRUDNO, Y., AGARWAL, S., IYER, L. M., LIU, D. R., ARAVIND, L. & RAO, A. 2009b. 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., MORISHITA, Y., KOZUKA-HATA, H., OYAMA, M., TODO, T., INO, Y., MUKASA, A., SAITO, N., TOYOSHIMA, C., SHIRAHIGE, K. & AKIYAMA, T. 2014. 5-Hydroxymethylcytosine Plays a Critical Role in Glioblastomagenesis by Recruiting the CHTOP-Methylosome Complex. *Cell Reports*, 9, 48-60.
- TAN, L., XIONG, L. J., XU, W. Q., WU, F. Z., HUANG, N., XU, Y. F., KONG, L. C., ZHENG, L. J., SCHWARTZ, L., SHI, Y. & SHI, Y. G. 2013. Genome-wide comparison of DNA hydroxymethylation in mouse embryonic stem cells and neural progenitor cells by a new comparative hMeDIP-seq method. *Nucleic Acids Research*, 41.
- TATE, P. H. & BIRD, A. P. 1993. Effects of DNA methylation on DNA-binding proteins and gene expression. *Curr Opin Genet Dev*, 3, 226-31.
- THALHAMMER, A., HANSEN, A. S., EL-SAGHEER, A. H., BROWN, T. & SCHOFIELD, C. J. 2011. Hydroxylation of methylated CpG dinucleotides reverses stabilisation of DNA duplexes by cytosine 5-methylation. *Chemical Communications*, 47, 5325-5327.
- TOYOTA, M., AHUJA, N., OHE-TOYOTA, M., HERMAN, J. G., BAYLIN, S. B. & ISSA, J. P. J. 1999. CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 96, 8681-8686.
- TRAPNELL, C., ROBERTS, A., GOFF, L., PERTEA, G., KIM, D., KELLEY, D. R., PIMENTEL, H., SALZBERG, S. L., RINN, J. L. & PACHTER, L. 2014. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks (vol 7, pg 562, 2012). *Nature Protocols*, 9, 2513-2513.

- TSAGARATOU, A., AIJO, T., LIO, C. W., YUE, X., HUANG, Y., JACOBSEN, S. E., LAHDESMAKI, H. & RAO, A. 2014. Dissecting the dynamic changes of 5-hydroxymethylcytosine in T-cell development and differentiation. *Proc Natl Acad Sci U S A*, 111, E3306-15.
- TSUMURA, A., HAYAKAWA, T., KUMAKI, Y., TAKEBAYASHI, S., SAKAUE, M., MATSUOKA, C., SHIMOTOHNO, K., ISHIKAWA, F., LI, E., UEDA, H. R., NAKAYAMA, J. & OKANO, M. 2006. Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. *Genes to Cells*, 11, 805-814.
- TURCAN, S., ROHLE, D., GOENKA, A., WALSH, L. A., FANG, F., YILMAZ, E., CAMPOS, C., FABIUS, A. W. M., LU, C., WARD, P. S., THOMPSON, C. B., KAUFMAN, A., GURYANOVA, O., LEVINE, R., HEGUY, A., VIALE, A., MORRIS, L. G. T., HUSE, J. T., MELLINGHOFF, I. K. & CHAN, T. A. 2012. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature*, 483, 479-U137.
- VELLA, P., SCELFO, A., JAMMULA, S., CHIACCHIERA, F., WILLIAMS, K., CUOMO, A., ROBERTO, A., CHRISTENSEN, J., BONALDI, T., HELIN, K. & PASINI, D. 2013. Tet Proteins Connect the O-Linked N-acetylglucosamine Transferase Ogt to Chromatin in Embryonic Stem Cells. *Molecular Cell*, 49, 645-656.
- VESELY, J. & CIHAK, A. 1978. 5-Azacytidine - Mechanism of Action and Biological Effects in Mammalian-Cells. *Pharmacology & Therapeutics Part a-Chemotherapy Toxicology and Metabolic Inhibitors*, 2, 813-840.
- VIRE, E., BRENNER, C., DEPLUS, R., BLANCHON, L., FRAGA, M., DIDELOT, C., MOREY, L., VAN EYNDE, A., BERNARD, D., VANDERWINDEN, J. M., BOLLEN, M., ESTELLER, M., DI CROCE, L., DE LAUNOIT, Y. & FUKS, F. 2006. The Polycomb group protein EZH2 directly controls DNA methylation. *Nature*, 439, 871-874.
- VOO, K., CARLONE, D., JACOBSEN, B., FLODIN, A. & SKALNIK, D. 1999. Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. *Blood*, 94, 469a-469a.
- WADA, T., TAKAGI, T., YAMAGUCHI, Y., FERDOUS, A., IMAI, T., HIROSE, S., SUGIMOTO, S., YANO, K., HARTZOG, G. A., WINSTON, F., BURATOWSKI, S. & HANDA, H. 1998. DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes & Development*, 12, 343-356.

- WALSH, C. P., CHAILLET, J. R. & BESTOR, T. H. 1998. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genetics*, 20, 116-117.
- WAN, J., OLIVER, V. F., ZHU, H., ZACK, D. J., QIAN, J. & MERBS, S. L. 2013. Integrative analysis of tissue-specific methylation and alternative splicing identifies conserved transcription factor binding motifs. *Nucleic Acids Research*, 41, 8503-8514.
- WANG, H., XING, J., GROVER, D., HEDGES, D. J., HAN, K., WALKER, J. A. & BATZER, M. A. 2005. SVA elements: a hominid-specific retroposon family. *J Mol Biol*, 354, 994-1007.
- WANG, L. F., ZHOU, Y., XU, L., XIAO, R., LU, X. Y., CHEN, L., CHONG, J., LI, H. R., HE, C., FU, X. D. & WANG, D. 2015. Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex. *Nature*, 523, 621-+.
- WANG, Z., ZANG, C., ROSENFELD, J. A., SCHONES, D. E., BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T. Y., PENG, W., ZHANG, M. Q. & ZHAO, K. 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet*, 40, 897-903.
- WATSON, J. D. & CRICK, F. H. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171, 737-8.
- WEBER, M., DAVIES, J. J., WITTIG, D., OAKELEY, E. J., HAASE, M., LAM, W. L. & SCHUBELER, D. 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics*, 37, 853-862.
- WEBER, M., HELLMANN, I., STADLER, M. B., RAMOS, L., PAABO, S., REBHAN, M. & SCHUBELER, D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics*, 39, 457-466.
- WEN, L., LI, X. L., YAN, L. Y., TAN, Y. X., LI, R., ZHAO, Y. Y., WANG, Y., XIE, J. C., ZHANG, Y., SONG, C. X., YU, M., LIU, X. M., ZHU, P., LI, X. Y., HOU, Y., GUO, H. S., WU, X. L., HE, C., LI, R. Q., TANG, F. C. & QIAO, J. 2014. Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biology*, 15.
- WILLIAMS, K., CHRISTENSEN, J., PEDERSEN, M. T., JOHANSEN, J. V., CLOOS, P. A. C., RAPPSILBER, J. & HELIN, K. 2011. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature*, 473, 343-U472.

- WYSOCKA, J., SWIGUT, T., XIAO, H., MILNE, T. A., KWON, S. Y., LANDRY, J., KAUER, M., TACKETT, A. J., CHAIT, B. T., BADENHORST, P., WU, C. & ALLIS, C. D. 2006. A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*, 442, 86-90.
- XING, J., HEDGES, D. J., HAN, K. D., WANG, H., CORDAUX, R. & BATZER, M. A. 2004. Alu element mutation spectra: Molecular clocks and the effect of DNA methylation. *Journal of Molecular Biology*, 344, 675-682.
- XING, J., ZHANG, Y., HAN, K., SALEM, A. H., SEN, S. K., HUFF, C. D., ZHOU, Q., KIRKNESS, E. F., LEVY, S., BATZER, M. A. & JORDE, L. B. 2009. Mobile elements create structural variation: analysis of a complete human genome. *Genome Res*, 19, 1516-26.
- XU, C., BIAN, C. B., LAM, R., DONG, A. P. & MIN, J. R. 2011a. The structural basis for selective binding of non-methylated CpG islands by the CFP1 CXXC domain. *Nature Communications*, 2.
- XU, Y., XU, C., KATO, A., TEMPEL, W., ABREU, J. G., BIAN, C., HU, Y., HU, D., ZHAO, B., CEROVINA, T., DIAO, J., WU, F., HE, H. H., CUI, Q., CLARK, E., MA, C., BARBARA, A., VEENSTRA, G. J., XU, G., KAISER, U. B., LIU, X. S., SUGRUE, S. P., HE, X., MIN, J., KATO, Y. & SHI, Y. G. 2012. Tet3 CXXC domain and dioxygenase activity cooperatively regulate key genes for *Xenopus* eye and neural development. *Cell*, 151, 1200-13.
- XU, Y. F., WU, F. Z., TAN, L., KONG, L. C., XIONG, L. J., DENG, J., BARBERA, A. J., ZHENG, L. J., ZHANG, H. K., HUANG, S., MIN, J. R., NICHOLSON, T., CHEN, T. P., XU, G. L., SHI, Y., ZHANG, K. & SHI, Y. G. 2011b. Genome-wide Regulation of 5hmC, 5mC, and Gene Expression by Tet1 Hydroxylase in Mouse Embryonic Stem Cells. *Molecular Cell*, 42, 451-464.
- YAMAGUCHI, Y., TAKAGI, T., WADA, T., YANO, K., FURUYA, A., SUGIMOTO, S., HASEGAWA, J. & HANDA, H. 1999. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell*, 97, 41-51.
- YANG, X. J., HAN, H., DE CARVALHO, D. D., LAY, F. D., JONES, P. A. & LIANG, G. N. 2014. Gene Body Methylation Can Alter Gene Expression and Is a Therapeutic Target in Cancer. *Cancer Cell*, 26, 577-590.
- YEARIM, A., GELFMAN, S., SHAYEVITCH, R., MELCER, S., GLAICH, O., MALLM, J. P., NISSIM-RAFINIA, M., COHEN, A. H. S., RIPPE, K., MESHORER, E. & AST, G. 2015. HP1 Is Involved in Regulating the Global Impact of DNA Methylation on Alternative Splicing. *Cell Reports*, 10, 1122-1134.
- YOKOMORI, N., MOORE, R. & NEGISHI, M. 1995. Sexually dimorphic DNA demethylation in the promoter of the Slp (sex-limited protein) gene in mouse liver. *Proc Natl Acad Sci U S A*, 92, 1302-6.

- YOON, H. G., CHAN, D. W., REYNOLDS, A. B., QIN, J. & WONG, J. M. 2003. N-CoR mediates DNA methylation-dependent repression through a methyl CpG binding protein Kaiso. *Molecular Cell*, 12, 723-734.
- YOSHIDA, M., KIJIMA, M., AKITA, M. & BEPPU, T. 1990. Potent and specific inhibition of mammalian histone deacetylase both in vivo and in vitro by trichostatin A. *J Biol Chem*, 265, 17174-9.
- YU, M., HON, G. C., SZULWACH, K. E., SONG, C. X., ZHANG, L., KIM, A., LI, X. K., DAI, Q., SHEN, Y., PARK, B., MIN, J. H., JIN, P., REN, B. & HE, C. 2012. Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome. *Cell*, 149, 1368-1380.
- ZHANG, L., LU, X. Y., LU, J. Y., LIANG, H. H., DAI, Q., XU, G. L., LUO, C., JIANG, H. L. & HE, C. 2012. Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. *Nature Chemical Biology*, 8, 328-330.
- ZHANG, Y., NG, H. H., ERDJUMENT-BROMAGE, H., TEMPST, P., BIRD, A. & 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.
- ZHAO, Y. Q., JORDAN, I. K. & LUNYAK, V. V. 2013. Epigenetics Components of Aging in the Central Nervous System. *Neurotherapeutics*, 10, 647-663.
- ZHU, Y. R., PEERY, T., PENG, T. M., RAMANATHAN, Y., MARSHALL, N., MARSHALL, T., AMENDT, B., MATHEWS, M. B. & PRICE, D. H. 1997. Transcription elongation factor P-TEFb is required for HIV-1 Tat transactivation in vitro. *Genes & Development*, 11, 2622-2632.
- ZILLER, M. J., GU, H. C., MULLER, F., DONAGHEY, J., TSAI, L. T. Y., KOHLBACHER, O., DE JAGER, P. L., ROSEN, E. D., BENNETT, D. A., BERNSTEIN, B. E., GNIRKE, A. & MEISSNER, A. 2013. Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 500, 477-481.