Regulation of DNA methylation by DNA glycosylases MBD4 and TDG

by

Abdulkhaleg IBRAHIM

Thesis submitted to the University of Strasbourg for the degree of

DOCTOR OF PHILOSOPHY

Discipline: Life and Health Sciences

Specialization: Molecular and Cellular Biology

Public PhD defence MAY 19th, 2015

Thesis Jury

Thesis Director:..... **Dr Christian BRONNER**

Rapporteur:..... **Pr Gilles SALBERT**

Rapporteur:..... **Dr Pierre-Antoine DEFOSSEZ**

Examiner:..... **Pr Philippe BOUCHER**

Guest:..... **Dr Ali HAMICHE**

Acknowledgements

I am deeply grateful to my supervisor, **Dr Christian Bronner,** for his guidance with kindness and patience in helping me during my thesis. I appreciated very much the opportunity he gave me to discover the epigenetic field and the time he spent to have highly interesting scientific discussion with me.

I would like to express my sincere gratitude to our team leader **Dr Ali Hamiche** for giving me the opportunity to join his valuable laboratory. He helped me how to think scientifically, and provided me many helpful suggestions, important expert advice and constant encouragement over the years of my thesis. I deeply appreciate what you have done for me Ali.

My special thanks are going to **Dr Pierre-Antoine Defossez, Pr Gilles Salbert** and **Pr Philippe Boucher** for their acceptance to be members in my thesis jury and for their valuable time spent to critically read and comment on my thesis work.

I am especially grateful to **Dr Christophe Papin** for his enthusiasm, guidance and stimulating discussions throughout all the period we worked together. He worked with me to figure out reasons whenever I had problems in experiments and provided kindly recommended solutions.

In no particular order, I thank my colleagues; **Khalid**, **Catherine, Philippe**, **Isabelle**, **Hatem**, **Chrysa**, **Maria** and **Liam**. Thank you all, for useful instruction, interesting scientific discussions, and hearty laughs.

I am grateful for my financial support from the National Authority for Scientific Research and the Biotechnology Reseach Center **(BTRC),** Libya.

Thanks for the **Affaires Culturelles Bureau** in the libyan embassy in Paris for their efforts to facilitate my staying in France during my study and for their kind collaboration.

Finally, I wish to dedicate this thesis to my mother, my father and all my family.

Summary

In mammals, methylation is an epigenetic mark targeting the 5th carbon of cytosine mainly in a CpG context producing a methylcytosine (5mC). The majority (70 to 80%) of CpGs are methylated but non-methylated in CpG-rich regions, referred as CpG islands (CGI). Repeated sequences cover almost half the genome in mammals and are highly methylated in healthy cells. In cancer cells, they can be specifically targeted by a demethylation process. However, the interest of repeated sequences has long been ignored in favor of the more functionally important regulatory regions such as CpG islands.

DNA methylation is not stable regarding that 5mC is highly sensitive to a spontaneous or enzymatic deamination leading to thymine and thus forming G/T mismatch. 5mC can also be successively oxidized to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5FC) and 5-carboxylcytosine (5caC) by TET protein family (TET1, 2 and 3). These modifications (deamination and oxidation) of 5mC are considered as participating in active demethylation processes. But how the methylation state of these CGIs at silenced promoters is preserved, remains, however unclear. In mammals, the thymine in G/T mismatch (deamination product of 5mC) is recognized and cleaved by TDG and MBD4 glycosylase proteins. TDG is able also to specifically recognize and excise the 5FC and 5caC modifications. Therefore, TDG and MBD4 could play an essential role in the active demethylation process.

MBD4 is especially intriguing since it is the sole mammalian protein having a methylated DNA binding domain MBD (Methyl Binding Domain) associated with a glycosylase domain. *In vitro,* MBD4 possesses a monofunctional glycosylase activity that specifically cleaves the N-glycosidic bond of thymidine mismatched to a guanine (G/T), leading to the formation of an abasic site.

This thesis was to clarify the function of TDG and MBD4 in the dynamics of 5mC. We started by identifying proteins associated with MBD4 *in vivo* trying to understand its mechanism of action. To this end, we purified MBD4 complex from HeLa cells. Analysis by mass spectrometry of this complex shows that MBD4 is associated with PMS2, MLH1, MSH2 and MSH6 proteins, four proteins involved in DNA mismatch repair (MMR, mismatch repair). The *in vitro* enzymatic tests show that MBD4/MMR complex has a bifunctional glycosylase/ lyase activity specific for G/T and directed by methylation. Biochemical analysis of point mutants of MBD4 reveals that the integrity of MBD and glycosylase domains is required for this function.

Our data suggest an activator and/or enzymatic role of MMR proteins in the bifunctional activity of MBD4 complex. To study this hypothesis, we purified recombinant MBD4, PMS2/MLH1 dimer and MSH2/MSH6 dimer proteins, and verified that MBD4 protein physically interacts with the MMR proteins. *In vitro,* low nuclease activity was detected with the MBD4 protein suggesting that the enzymatic activity of the native complex MBD4/MMR is catalyzed by MBD4 and regulated by MMR proteins.

Considering that the MLH1 protein is the most abundant subunit of MBD4 complex in our analysis by mass spectrometry, we wondered whether MLH1 could act as activator of MBD4 within the MBD4/MMR complex. Consistent with this hypothesis, we reconstituted a MBD4/MLH1 complex (Fig. 8A). Interestingly, while the nuclease activity of MBD4 is low and is not sensitive to methylation, the complex MBD4/MLH1 shows intense nuclease activity strongly induced by the methylation. The lack of enzymatic activity detected with the purified complex from a catalytic mutant MBD4 (MBD4 D554A/MLH1), demonstrates that the cleavage reaction is catalyzed by MBD4 and is dependent on its binding to MLH1.

We then sought to verify the role of TDG in the genome-wide DNA methylation dynamic. To this goal, we targeted this enzyme by shRNA technique in MEF cells and characterized the distribution of modified cytosine (5mC, 5hmC, 5FC and 5caC). Strikingly, our results show an enrichment of modified cytosines specifically at repeat sequences level. While 5mC target all the repeated sequences, with a preference for SINEs, oxidized forms (5hmC, and 5caC 5FC) are found preferentially in microsatellite (simple repeats). The distribution of 5mC, 5hmC, 5FC and 5caC at SINEs level in MEFs reveals a dynamic of these changes, regulated by TDG, only at the SINEs B1m and B2m family.

The genomic distribution analysis of SINEs B1m and B2m showed that the most preserved mouse specific SINEs are those the most close to a TSS (Transcription Start Site), suggesting an important function in the regulation of gene expression.

The dynamic changes was also observed at LINEs and LTRs. Similar to the obtained results at SINEs, the TDG-dependent regulation is observed only at the evolutionary newer retro-elements (mouse specific lineage), the L1Md family for LINEs and IAP family for the LTRs. We concluded that TDG regulates the dynamics of methylation/demethylation at repeated sequences during differentiation. Conversely, MBD4 targets promoters repressed by methylation and protects them against deamination mechanisms.

Table of contents

List of figures

List of abbreviations

ADD	ATRX-DNMT3-DNMT3L-type zinc finger domain
AID	Activation-induced Cytidine Deaminase
AP	Apurinic/Apyrimidinic
APE1	APurinic/apyrimidinic Endonuclease 1
APOBEC	Apolipoprotein B mRNA Editing enzyme, Catalytic polypeptide
ATRX	Alpha Thalassemia/Mental Retardation Syndrome X-Linked
BAH	Bromo-Adjacent Homology domain
BDNF	Brain-Derived Neurotrophic Factor
BER	Base Excision Repair
C	Cytosine
CFP1	CXXC Finger Protein 1
CGI	CpG Island
CMT3	ChromoMethylase 3
CpG	Cytosine-phosphate-Guanine
CREB	C-AMP Response Element-Binding protein
CtBP	C-terminal Binding Protein
CXXC	Cysteine rich region
CYP27B1	CYtochrome P450 27B1
DIP-seq	DNA-ImmunoPrecipitation-sequencing
DNMT	DNA methyltransferase
DRM2	Domains Rearranged Methyltransferase 2
DSBH	Double-Stranded B-Helix
ElP3	Elongator complex Protein 3
ERα	Oestrogen Receptor-α
ERVs	Endogenous RetroViruses
ESC	Embryonic Stem Cells
EXO1	Exonuclease 1
FaPy	2,6-diamino-4-hydroxy-5-N-methylformamidopyrimidine

G	Guanine
Gadd45	Growth arrest and DNA-damage-inducible protein 45
GFP	Green Fluorescent Protein
HA	HemAgglutinin
H3K27ac	Histone H3 Lys27 acetylation
H3K36me2	Histone H3 di-methylated on lysine 36
H3K4me3	Histone H3 tri-methylated on lysine 4
H3K9ac	Histone H3 Lys27 acetylation
HAT	Histone AcetylTransferase
HDAC2	Histone DeACetylase 2
HEK293	Human Embryonic Kidney 293 cells
HP1	Heterochromatin Protein1
IAP	Intracisternal A Particle
ING4	INhibitor of Growth protein 4
JBP1	J-Binding Protein1
KDM2A	Lysine-specific DeMethylase 2A
LINEs	Long INterspersed Elements
LTRs	Long Terminal Repeats
MBD	Methyl-CpG Binding Domain
MBPs	Methyl-CpG-Binding Proteins
MeCP2	Methyl-CpG-binding Protein 2
MED	Methyl-CpG-binding Endonuclease
MLH1	MutL homolog 1
MMR	Misatch Repair
MSH2	MutS protein homolog 2
MTA2	Metastasis-associated protein
Mtase	Methyltransferase domain.
NCoR	Nuclear receptor Co-Repressor
NER	Nucleotide Excision Repair
NLS	Nuclear localization signal
NuRD	Nucleosome Remodelling and histone Deacetylation

NURF	NUcleosome Remodeling Factor
ORF	Open Reading Frame
PARP1	Poly-ADP-Ribose Polymerase 1
PHD	Plant Homeo Domain
PGC	Primordial Germ Cells
Pol III	RNA polymerase III
PWWP	Proline-tryptophan-tryptophan-Proline
REs	Repetitive DNA Elements
RMSK	UCSC Repeat-Masker
ROS1	Repressor Of Silencing 1
RRBS	Reduced Representation Bisulfite Sequencing
SAM	S-Adenosyl Methionine
SETDB1	SET Domain, Bifurcated 1
shRNA	short hairpin RNA
SINEs	Short INterspersed Elements
siRNA	Small interfering RNA
SRA	Set and Ring Associated domain
SUMOs	Small Ubiquitin Like Modifiers
SWI/SNF	SWItch/Sucrose NonFermentable
T	Thymine
TDG	Thymine DNA Glycosylases
TET	Ten-Eleven Translocation
TFF1	TreFoil Factor 1
TFIID	Transcription Factor IID
TFIIIC	Transcription Factor IIIC
TFs	Transcription Factors
Tg	Thymine glycol
THF	TetraHydroFuran
TpG	Thymine-phosphate-Guanine
TRD	Transcriptional Repression Domain
TSA	TrichoStatin A

TSGs	Tumor Suppressor Genes
TSS	Transcription Start Site
U	Uracil
UHRF1	Ubiquitin-like containing plant Homeo-domain and RING Finger domain 1
UTR	UnTranslated Region
ZBTB38	Zinc finger and BTB domain containing 38
ZF-CXXC	Zinc Finger domain
5caC	5-carboxylCytosine
5fC	5-formylCytosine
5-FU	5-Fluorouracil
5hmU	5-hydroxymethylUracil
5hmC	5-hydroxymethylCytosine
5-hm	5-hydroxymethyl
5mC	5-methyl-Cytosine
8-oxoG	8-oxo-7,8-dihydroguanine
3-meA	3-methyl-Adenine
7-meG	7-methyl-Guanine

CHAPTER 1
INTRODUCTION

1.1 Overview of Epigenetics

All cells of a multicellular mammalian organism, except germinal cells, contain the same DNA in terms of nucleotide sequence. Considering that DNA is the layer of heredity and cell identity, how can arise cell diversity and differentiation from the same DNA sequence? The question is important and is one of the major challenges for the scientific community. Epigenetics is the research field that tries to answer this question by deciphering a tremendous number of cellular mechanisms of gene regulation embedded in the chromatin but not related to changes in DNA sequences.

The term “Epigenetic” comes from the Greek prefix “Epi” which means “above” or “over” and the term “Genetic” which is related to the study of genes. Therefore, “Epigenetic” literally means the study of gene expression variation that is not due to alterations of DNA sequences. In other words, it refers to external modifications to DNA that turn *genes* "on" or "off.

Since it was originally introduced by Conrad Waddington in 1942, epigenetics showed a large progression in its understanding. Waddington used the term epigenetic to describe heritable changes in a cellular phenotype not related to alterations in the DNA sequence. He hypothesized that environmental stimulus could be converted into an internal genetic factor. Since then, different definitions of the term epigenetics have been proposed. In 2001, Rakyan has redefined the epigenetic as the mechanism of physically marking the DNA or its associated proteins to maintain stably gene expression which allows genotypically identical cells to be phenotypically distinct (Rakyan et al. 2001). Other definition is that *“epigenetic”* is the stably inherited phenotype resulting from changes in a chromosome without alterations in the DNA sequence (Berger et al. 2009). These tissue-specific patterns are essential for normal cellular function, and control of different stages of development by affecting gene expression (Eckhardt et al. 2006, Jones P. L. et al. 1998, Razin and Riggs 1980). Disturbance in these modifications or failure to maintain them can induce irregular gene transcription which lead to various forms of disease (Robertson 2005).

Three categories of signals have been proposed to be accompanied in the establishment of a stably heritable epigenetic state: an *‘‘Epigenator’’* which is a signal acquired from the environment and triggers an intracellular pathway; the second signal is

the *''Epigenetic Initiator''* and this responds to the *Epigenator* and defines the precise location of the modification on chromatin. The Initiator could be a DNA-binding protein or a noncoding RNA. The third type of signal is the *''Epigenetic Maintainer''* to sustain the chromatin environment in the first and subsequent cell generations. *Epigenetic Maintainers* involve many different epigenetic marks. These various maintainer marks act together as an epigenetic machinery complex (Illingworth and Bird 2009).

At the molecular level, an epigenetic mark is by definition a chemical signal outside DNA or above DNA that regulates gene expression and that is transmitted to its descent. There are many different epigenetic marks, including DNA methylation, histone post-translational modifications, histone variants, micro-RNAs (Berger et al. 2009, Bronner et al. 2010) and we might imagine that more new epigenetic marks can be uncovered in a near future.

As my thesis work is dealing with DNA methylation, my manuscript will mainly focus on this epigenetic mark. However, we have to keep in mind that DNA methylation is intemely connected to other epigenetic marks such as for instance methylation of lysine 9 of histone H3, mediated by UHRF1 (Liu X. et al. 2013).

1.2 DNA methylation

DNA methylation occurs on the 5th carbon of cytosine in eukaryotes leading to form 5'-methylcytosine (5mC), whereas in prokaryotes, it happens on various bases giving rise to 5-methylcytosine, N6-methyladenine and N4-methylcytosine, and contributes to host restriction systems and protects the cells from foreign genetic material such as viral DNA and destruction by proper restriction enzymes. DNA methylation is a stable epigenetic mark that plays a key role in several biological functions including gene expression regulation, transposon silencing, X chromosome inactivation, imprinting, cell differentiation, development and other diverse processes (Bird A. 2002, Cedar and Bergman 2012). DNA methylation is a post-replicative mechanism that refers to transfer of a methyl group from the methyl donor S-adenosyl methionine (SAM) to cytosine (Figure 1).

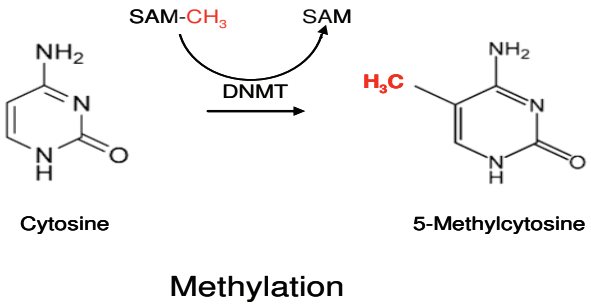


Figure 1 – Cytosine methylation on the 5-position in a reaction catalysed by DNMT enzymes using S-adenosyl methionine as a substrate.

The methylated cytosine was first reported in 1948 by using paper chromatography (Hotchkiss 1948). Later, it was suggested that this modification of DNA can serve as a heritable epigenetic modification that conveys the epigenetic memory to progeny cells (Holliday and Pugh 1975, Riggs 1975). The fraction of methylated cytosine in humans is 0.76-1% depending on tissues, corresponding to 4-5% of all cytosine bases (Ehrlich et al. 1982). The human sperm DNA is highly methylated as that of eggs. However, after fertilization the paternal becomes rapidly demethylated, whereas that of the maternal undergoes slower demethylation (Figure 5). Paternal and maternal genomes become remethylated at the right place and at the right moment, through as a yet mysterious mechanism, which might also affect non-CpG regions.

DNA methylation is established and maintained by a family of DNA methyltransferases (DNMTs), which will be now reviewed.

1.3 DNA methyltransferases

The DNA methyltransferases (DNMTs) family fits into two general classes based on their preferred DNA substrate, and the way to achieve this process. First, DNMT3A and B establish *de novo* methylation patterns at previously unmethylated cytosine after embryos implantation during early stages of mammalian development, that leads to a wide spread of global DNA methylation, and second, the established global methylation patterns are faithfully maintained through the action of DNA methyltransferase1 (DNMT1) enzyme. This enzyme is able, with the help of its obligate functional partner UHRF1 (Ubiquitin-like containing plant homeo-domain and RING finger domain 1), to recognize the hemi-methylated DNA resulted from DNA replication. The hemi-methylated double strand DNA is specifically recognized by the SRA (Set and Ring

Associated) domain of UHRF1. The SRA-DNA interaction may serve as an anchor to keep UHRF1 at a hemi-methylated CpG site, where it recruits DNMT1 for DNA methylation maintenance, *i.e.,* methylation of the newly synthesized DNA strand (Arita et al. 2008, Avvakumov et al. 2008, Bostick et al. 2007, Hashimoto et al. 2008, Sharif et al. 2007). An attractive model has been proposed, which suggests that UHRF1 is sliding along the DNA in order to detect the hemimethylated CpG (m1/2CpG) and once it had, a "message" would be transmitted to DNMT1 through a yet unidentified mechanism (Bronner et al. 2010) and then UHRF1 would slide to the next m1/2CpG. This will ensure the duplication of symmetric DNA methylation patterns.

It is evident that this model cannot be applied for asymmetric DNA methylation inheritance. Accordingly, it has been proposed that DNMT3A is involved in DNA methylation maintenance in non-CpG regions (Ramsahoye et al. 2000).

A fourth DNA methyltransferase is called DNMT2. This DNMT family member shows a weak DNA methyl-transferase activity *in vitro* (Hermann et al. 2003). But it seems not to have an essential role in setting DNA methylation patterns because targeted deletion of the DNMT2 gene in embryonic stem cells causes no detectable effect on global DNA methylation (Okano et al. 1998). An other DNMT-related protein, called DNMT3L, has not DNA methyltransferase activity, but associates physically with DNMT3A and DNMT3B and modulates their catalytic activities (Okano et al. 1999). Most of DNMT proteins share common C-termini that are responsible for the methyltransferase activity, while the N-terminal part of DNMT3A and DNMT3B consists of a cysteine-rich domain and a tryptophan-rich region (PWWP) domain (Figure 3). This tryptophan-rich domain is required for directing enzymes to the major satellite repeats at pericentric heterochromatin (Chen T. et al. 2004).

DNA methyltransferases are essential for viability as mice with disrupted alleles of DNA-methyltransferase1 (Dnmt1) or double-null for *Dnmt3a* and *Dnmt3b* die early in embryogenesis and show inappropriate expression of a large number of genes (Li E. et al. 1992, Okano et al. 1999).

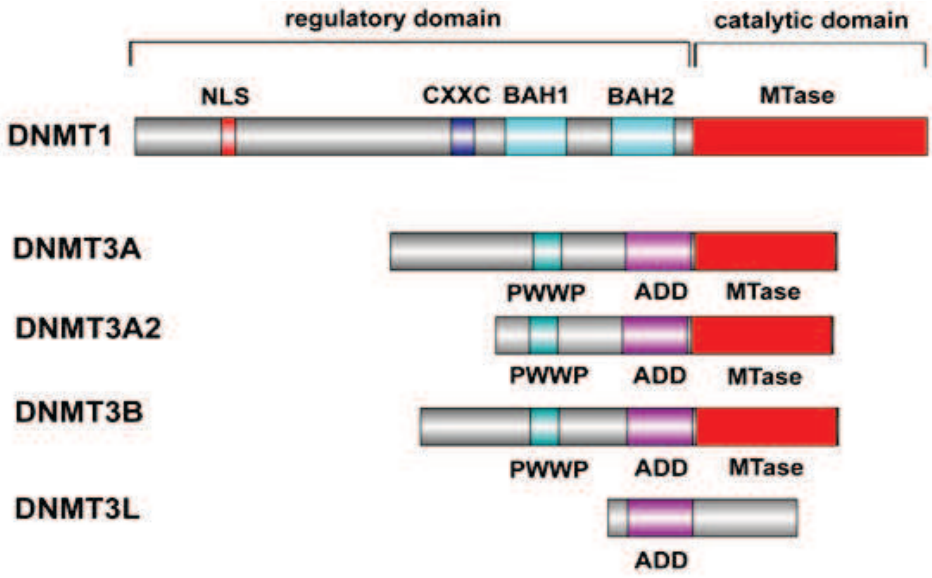


Figure 2; Structure of DNA methyltransferases. The N-terminal contains motifs of interaction with proteins or DNA. The C-terminal contains the conserved methyltransferase domains. NLS stands for nuclear localization signal; CXXC for cysteine rich region; BAH for bromo-adjacent homology domain; PWWP for proline-tryptophan-tryptophan-proline domain; ADD for ATRX-DNMT3-DNMT3L-type zinc finger domain; and Mtase stands for methyltransferase domain.

1.4 Genomic distribution of DNA methylation

In adult mammals, cytosine methylation occurs almost exclusively at CpG dinucleotides (a cytosine and a guanine separated by a phosphate). An interesting study performed on two cell lines, i.e., IMR90 (fetal lung fibroblasts) and H1 (human embryonic stem cells) showed that in fetal cells 99,98% of cytosine methylation is located within CpG dinucleotides whereas in embryonic stem cells cytosine methylation was found in different context: in CpG dinucleotides (75,5%), in CHG (17,3%) and in CHH (7,2%), where H accounts for A, T or C (Lister et al. 2009). These results have been confirmed later by several studies supporting that cytosine methylation can be found outside of CpG, providing that cells arise from embryonic stem cells (Laurent et al. 2010, Lister et al. 2009, Ziller et al. 2011). A comparative study among different ESC lines showed that the highly methylated non-CpG sites were conserved at TACAG (Chen P. Y. et al. 2011).

Altogether, these studies support the notion that gametic genomes, which are heavily methylated, undergo demethylation followed by remethylation through CpG dinucleotides but also through CHH and CHG. The requirement through these later is not

clear but may involve DNA methyltransferases other than DNMT1. In *Arabidopsis thaliana,* DNA methylation frequently occurs at CpG, CHG and CHH, which is maintained by CMT3 and DRM2, respectively (Law and Jacobsen 2010). It appears that in mammals DNMT3A and DNMT3B are involved in the maintainance of non-CpG methylation (Arand et al. 2012). However, so far the mechanisms of non-CpG methylation maintenance, establishment and functions are still elusive.

1.4.1 CpG islands methylation

As seen before, the methylation happens mostly on cytosine within CpG dinucleotides. The majority of CpG sites (approximately 70-80%) throughout the genome are methylated (Bird A. 2002, Ehrlich et al. 1982). These methylated CpGs are not randomly distributed across the genome, they are instead mainly distributed within repetitive elements such as SINEs (Short INterspersed Elements), LINEs (Long Interspersed Elements) or LTRs (Long Terminal Repeats) and coding regions of functional genes, while unmethylated CpGs are often found clustered in CpG rich DNA regions called CpG islands (CGI). CGIs are defined as DNA regions of around 500 bp in length with a G-C content greater than 55% (Takai and Jones 2002). Most of CGIs co-localize with promoters in mammalian genome, and approximately 70% of mammalian gene promoters contain CGIs. These CGIs remain unmethylated in the majority of promoters, especially those located at the promoters of housekeeping genes and tumor suppressor genes (TSGs) (Illingworth and Bird 2009). However, in pathological situations, such as cancer, promoters can be abnormally methylated, particularly those of TSGs (Alhosin et al. 2011).

Repetitive DNA elements, also known as retroelements (REs), have been reported to exhibit various deleterious effects on genome structure and functioning (Figure 2). Given that the most of the methylated cytosines in mammalian genome reside in repetitive elements, it has been proposed that DNA methylation evolved primarily to suppress the activity of transposable elements and to protect the host cell (Yoder et al. 1997). Supporting to this, hypomethylation of REs was demonstrated to be associated with genomic instability in cancer (Daskalos et al. 2009).

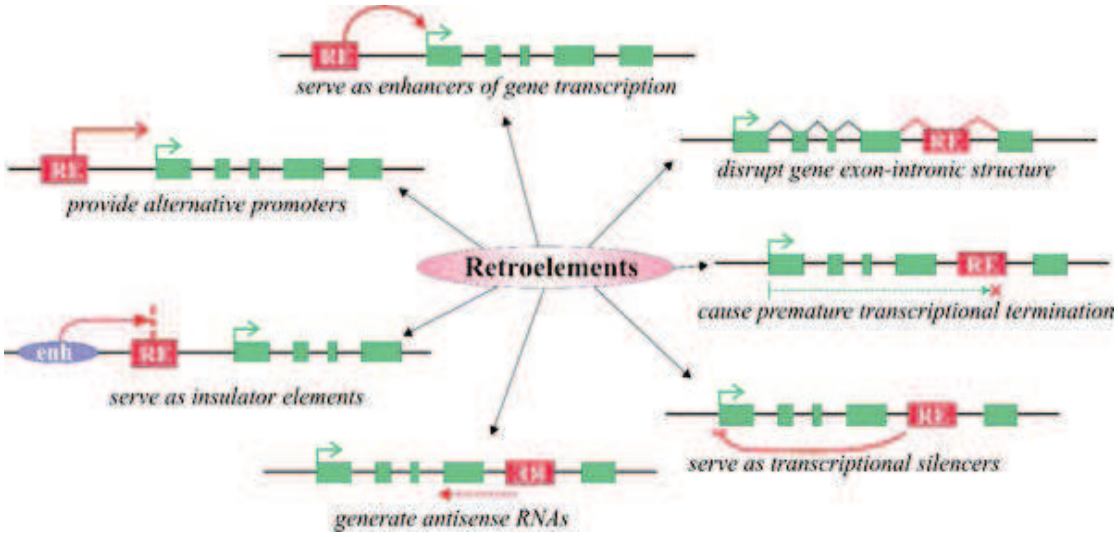


Figure 3; Different mechanisms of RE influence on gene transcription. Red boxes retroelements, green boxes gene exons, green arrow gene transcriptional start site, purple oval enhancer element. (Gogvadze and Buzdin 2009).

Currently, more than 175 families of SINEs have been described in a wide variety of eurkaryotes, whereas in mammals around 70 families are supposed to exist (Vassetzky and Kramerov 2013). SINEs are DNA sequences of typically 100 to 500 base pairs in length (Singer 1982). In contrast, LINEs are 5000 to 8000 base pairs length (Treangen and Salzberg 2011). They can occupy more than 10% of the genome under the existence of more than 1 million of copies (Ichiyanagi T. et al. 2013). SINEs are nonautonomous retrotransposons transcribed by the cellular RNA polymerase III (pol III) from an internal promoter, and their reverse transcription depends on the reverse transcriptase of partner LINEs (Gogvadze and Buzdin 2009). In contrast, LINEs can be autonomous, *i.e.,* they encode the proteins necessary for their proliferation and transposition. They are transcribed by the cellular RNA polymerase II (Vassetzky and Kramerov 2013). Regions enriched in SINEs and LINEs are found in different parts of the genome, *i.e.,* in gene-rich and gene-poor respectively (Lander et al. 2001).

Human Alu, and mouse B1 & B2 are among the most represented retrotransposons in mammalian genomes. These elements can amplify their copy numbers by transposition via transcription. The human Alu of approximately 290 bp in length carries up to 25 CpG sites both inside and outside of its Pol III promoter. These CpG sites are highly methylated in somatic tissues. This methylation results in inhibition of Pol III binding and therefore in prevention of Alu and tRNA transcription (Besser et al. 1990, Englander et al. 1993, Liu W. M. and Schmid 1993). The underlying mechanism is a hindering of the

TFIIIC (a co-factor of PolIII) to the A- and B- boxes, and thus, Pol III cannot load to the promoter.

The mouse B1 (145bp) contains up to 8 CpG sites in its sequence and is highly methylated in somatic cells but weakly in germ cells and in preimplantation embryos. The level of methylation in B1 is negatively correlated with RNA abundance (Ichiyanagi K. et al. 2011). Considering that SINEs are closely located to transcription start sites (TSS) of PolII genes, it has been suggested that methylation of mouse B1 SINEs can harbor gene expression regulatory mechanism (Ichiyanagi et al. 2013).

1.4.2 Non-CpG methylation

Non-CpG methylation, *i.e.,* methylation that occurs at cytosine in CHH or CHG trinucleotides, is known to be enriched in germ cells and ES cells (Lister et al. 2009, Tomizawa et al. 2011). Non-CpG methylation, are mainly found in SINEs, LINEs and LTRs. These retrotransposable elements constitute about half of mammalian genomes (Deininger and Batzer 2002) and have long been considered as selfish or junk DNA (Orgel and Crick 1980). Recent findings, however, suggest that some of these elements may exert gene regulation as well as chromatin structure organization functions.

Non-CpG methylation, in contrast to CpG methylation, is asymmetric, hence it is of interest to study the methylation patterns of DNA on both strands. This was recently done in H1 cells, (embryonic stem cells) by Guo et al., (2014), who observed that in introns, non-CpG sites are more heavily methylated on the antisense strand than those on the sense strand. Such skew of non-CpG methylation was not observed in exons but was more pronounced at intron boundaries (Guo et al. 2014).

1.5 CpG islands and Establishment of a permissive chromatin

Non-methylated CGIs are recognized by a family of zing finger (ZF-CXXC) proteins which contributes to the establishment of a permissive chromatin. The CXXC finger protein 1 (CFP1) binds to non-methylated CGIs and recruits the SET1 methyltransferase activity complex to these regions (Lee and Skalnik 2005). SET1

complex catalyzes the histone H3 lysine 4 tri-methylation (H3K4me3) (Thomson et al. 2010). H3K4me3 serves as a platform recruitment of PHD domain proteins (Plant Homeo Domain) such as TFIID, ING4 containing histone acetyltransferase (HAT) and NURF, which are known to be involved in the initiation of transcription (Blackledge and Klose 2011). Similar to CFP1, the lysine-specific demethylase 2A (KDM2A) specifically binds non-methylated CGI, but acts in a different way. KDM2A catalyzes the demethylation of H3K36me2 (Blackledge et al. 2010), a modification recognized by the histone deacetylase HDAC activity complex, known to exert gene transcription inhibitory activity. Thus, ZF-CXXC domain proteins cooperate at a single CGI to form a unique chromatin architecture (without H3K36me2 and enriched with H3K4me3), which provide ideal conditions to promote the initiation of transcription.

Non-methylated CGIs are also recognized by the TET1 protein, which unlike CFP1 and KDM2A, is involved in the recruitment of polycomb repressor complex family PRC1 and PRC2 (Wu H. et al. 2011). TET1 recruits PRC2, which catalyzes the histone H3 lysine 27 tri-methylation (H3K27me3). This modification is then recognized by PRC1, which in turn inhibits transcription elongation by promoting histone H2A ubiquitination and chromatin compaction (Eskeland et al. 2010, Stock et al. 2007). This mechanism of transcription repression is independent of CGI methylation.

The analysis of methylation distribution across the whole genome (or methylome) in somatic cells showed that, a fraction of CGI becomes methylated in a tissue-specific manner during developpement (Mohn et al. 2008, Weber et al. 2005). The majority of the CGI-promoters, that acquires methylation during cell differentiation, is already suppressed by the polycomb proteins within ESCs (Mohn et al. 2008), demonstrating that DNA methylation is not an initiator event in transcriptional repression, but rather acts as a heritable epigenetic marker which maintains repression in time

1.6 DNA methylation and regulation of transcription

DNA methylation is now well-known to play a role in regulating gene expression and chromatin structure beside its role in imprinting, X-chromosome inactivation and a possible role in silencing of repetitive DNA elements (Li E. 2002).

DNA methylation has long been found to be associated with a repressed chromatin

state (Bird A. P. and Wolffe 1999). In 2011, by associating RNA-sequencing approach with Whole-Genome-Bisulfite-Sequencing (WGBS), Bell and colleagues showed a genome-wide significant negative correlation between DNA methylation and gene expression level, and that the hypomethylated regions colocalize with the CGI-containing TSS of highly expressed genes (Bell et al. 2011). The effect of DNA methylation on gene control depends on the distribution of methylation (position of 5mC) on the transcriptional unit. For example, methylation at TSS blocks transcriptional initiation, whereas the methylation in gene body could stimulate transcription elongation, and may also have an impact on splicing (Jones P. A. 2012). At repeat regions such as centromeres, methylation has been suggested to suppress the expression of transposable elements, while at the same time allowing transcription of the host gene to run through them, and thus to have a role in genome stability (Li E. 2002, Yoder et al. 1997). Enhancers have low CpG content, however, methylation at these regions has been suggested to regulate their activity (Lister et al. 2009), whereas oxidation of 5mC to 5hmC leads to enhancer activation (Serandour et al. 2012).

There are two known mechanisms by which DNA methylation affects gene expression:

First, modification of cytosine bases can interrupt the association of some DNA binding transcription factors (TFs) with their cognate DNA recognition sequences. Most of transcription factors in mammals have GC-rich binding sites, and binding of several of these factors is impeded or abolished by methylation of CpG (Bird A. P. and Wolffe 1999, Watt and Molloy 1988). Consistent with these observations, it has been found that CGIs that remain unmethylated in normal and in malignant cells contain specific sequence motifs that are identical to the consensus sequence for general TFs (Gebhard et al. 2010). The absence of methylation, at these transcription factor binding sites, is important for facilitating the binding of these factors to their target genes (Bell et al. 2011). Some examples validating this model include E2F1, CREB and c-myc (Campanero et al. 2000, Iguchi-Ariga and Schaffner 1989).

Second, proteins that recognize and bind methylated cytosine at CpG (methyl-CpG-binding proteins MBPs) recruit transcriptional co-repressor molecules to silence the transcription of genes (Nan et al. 1993). These recruited co-repressors modify the surrounding chromatin to be in a repressive state providing a link between DNA methylation and chromatin remodelling modification during the regulation of gene-expression (Cedar and Bergman 2009, Thomson et al. 2010, Zhang Y. et al. 1999). It has

also been shown that DNA methylation levels correlate negatively with the presence of histone marks that target active genes such as H3K27ac, H3K4me3 and H3K9ac which are positively correlated with transcription levels (Heintzman et al. 2009, Lister et al. 2009).

1.7 Methyl-CpG binding proteins as intermediates in transcriptional repression

So far, three families of proteins are known to bind methylated DNA, methyl-CpG binding domain (MBD) protein family, Kaiso and Kaiso-like proteins and SET and Ring finger Associated (SRA) protein family.

At present, it is established that methyl-CpG binding proteins interact with histone deacetylases and histone methylase activities, that modify chromatin leading to the prevention of transcription initiation (Bird A. P. and Wolffe 1999). Each of these MBPs has a role in repressing the transcription in a DNA methylation-dependent manner. These MBPs will now be discussed.

1.7.1 Methyl-CpG binding domain proteins (MBDs)

The MBD family consists of the methyl-CpG-binding protein 2 (MeCP2), the methyl-CpG-binding-domain proteins MBD1, MBD2, MBD3 and MBD4 (Buck-Koehntop and Defossez 2013, Defossez and Stancheva 2011). These family members have been shown to play an intermediate role in gene repression.

1.7.1.a MeCP2

MeCP2 was the first discovered methyl-CpG binding protein member by Adrian Bird in 1992 (Lewis et al. 1992). This protein contain two fonctional domains, a N-terminal methyl-CpG binding domain (MBD) and a C-terminal transcriptional repression domain (TRD) (Nan et al. 1993). MeCP2 was co-purified with Sin3A/HDAC2 complex in mammalian cells, which interaction is essential for MeCP2-mediated transcription

repression (Jones P. L. et al. 1998). Nan and colleagues showed by co-immunoprecipitation that the interaction of MeCP2 with Sin3A/HDAC2 complex is occurring via the TRD domain, and they went to confirm the transcriptional repression fonction of this interaction by transfection, using a Gal4-targeted TRD. Furthermore, they could show that the HDAC inhibitor, trichostatin A (TSA) can partially modulate the induced repression of a reporter gene that contain GAL4 DNA-binding sites (Nan et al. 1998). MeCP2 interacts also with ATRX (Alpha Thalassemia/Mental Retardation Syndrome X-Linked) and Brahma (Brm1), which belong to SWI/SNF chromatin remodelling complex, this interaction results in transcription repression of the associated genes (Kokura et al. 2001). In spite of its role as a transcriptional repressor, MeCP2 was found to associate with the transcription activator CREB1 at the promoters of active genes. Moreover lack or overexpressed MeCP2 in the hypothalamus of mice leads to changes either positively or negatively of expression levels of thousands of genes, and most of these genes seemed to be activated by MeCP2 (Chahrour et al. 2008).

1.7.1.b **MBD1**

MBD1 protein acts as a histone deacetylation-independent transcriptional repressor (Ng et al. 2000). In contrast, MBD1 seems to act on histone H3 lysine 9 (H3K9) mrthylation as it is found to interact with two (H3K9) methylase activities, SETDB1 and SUV39H, as well as the heterochromatin protein HP1 (Sarraf and Stancheva 2004). It has been described that H3K9me2/3 recruits HP1α/β, which then recruits the H3K9 methyltransferase SUV39H1, which in turn methylates more heavily H3K9. This propagation of the H3K9me2/3 mark and HP1, itself, then serve to bind additional proteins leading to heterochromatin formation conducting to gene silencing (Lachner et al. 2001). In addition the C-terminus of MBD1 binds AM/MCAF the co-factor of SETDB1 (SET domain, bifurcated 1). This binding leads to stimulate SETDB1 activity in order to allow more efficient di- and trimethylation of H3K9 (Fujita et al. 2003, Wang et al. 2003).

1.7.1.c **MBD2/MBD3**

MBD2 and MBD3 have almost similar structure and outside the MBD domain they have 77 % identity to each other. But in spite of this similarity, they do not have the same affinity toward the methylated DNA. In contrast to MBD2, MBD3 cannot specifically

bind methylated DNA because it has a phenylalanine instead of the conserved tyrosine at position 34 (Fraga et al. 2003). These two proteins co-purified with the protein complex NuRD (nucleosome remodelling and histone deacetylation), which contains chromatin remodelling factors such as ATPase Mi-2, HDAC1 and HDAC2 histone deacetylases (Zhang Y. et al. 1999).

1.7.1.d **MBD4**

MBD4 has a special structure within the MBD protein family. Beside its MBD domain, it exhibits a DNA glycosylase domain. The MBD of this protein has a high affenity to bind to symmetrically methylated DNA. It has also been shown, by GFP fusion MBD4, that this protein localised in a methylation-dependent manner to the peri-centromeric heterochromatin (Hendrich and Bird 1998). MBD4 has been described as an HDAC-dependant transcriptional repressor as it is able to recruit HDAC complexes to the hypermethylated promoters of the $p16^{INK4a}$ and *hMLH1* genes, and the knockdown of *Mbd4* leads to an upregulation of these genes (Kondo et al. 2005).

Unlike to the other MBDs, MBD4 can bind to T:G mismatches (deamination product of methyl-Cytosine). Via its glycosylase domain, MBD4 can efficiently remove thymine from T:G mismatch to restore 5mC at the CpG sites (Hendrich et al. 1999). The ability of MBD4 to excise the T (deaminated m5C) from TpG to be replaced by a new C, which will be subsequently remethylated, could be a key mechanism in maintaining and/or repairing the methylation state of these regions. By this action, MBD4 participates to maintain the CGIs in a methylated state at the promoters of the methylation-dependent repressed genes. This role could reflect another indirect way by which MBD4 ensures the role as a methyl-CpG-dependent transcritional repressor.

1.7.2 **Kaiso and Kaiso-like proteins**

Similar to MBD family, this group of proteins specifically bind to methylated CpGs. Kaiso and Kaiso-like proteins have no structure similarity with MBDs and consist of an N-terminal POZ domain and C-terminal zinc finger domain. They use the zinc finger domain to bind the methylated CpGs. Via the zinc finger domain, Kaiso binds also to DNA regions that lack CpG dinucleotides with a mild affinity (Daniel et al. 2002).

Kaiso family proteins repress transcription in HDAC-dependent pathway. Indeed, Kaiso was shown to co-purify with NCoR (nuclear receptor co-repressor), from HeLa

cell nuclear extracts, containing histone deacetylase HDAC3. Furthermore, it has been reported that the association of Kaiso with NCoR is required for silencing of methylated *MTA2* (Metastasis-associated protein) promoter (Yoon et al. 2003). Moreover, it has been shown that the depletion of Kaiso protein leads to derepression of methylated genes in Xenopus embryos (Ruzov et al. 2004). Another member of Kaiso-like proteins, ZBTB38 is able to interact with histone deacetylases1, 3 (HDAC1, HDAC3) as well as the co-repressor CtBP (C-terminal binding protein) (Sasai et al. 2005).

1.7.3 SRA domain proteins

The SRA (SET- and RING-associated domains) family of proteins bind to methylated DNA via SRA domain. The founder member of this family is UHRF1, also known as ICBP90 in human and Np95 in mouse (Bronner et al. 2007). In human another UHRF1 paralog UHRF2 is defined to have similar domains to UHRF1. UHRF1 and 2 can bind to hemi-methylated DNA. UHRF1 has higher affinity towards to hemi-methylated than binding to fully-methylated DNA (Avvakumov et al. 2008). This protein interacts with and recruit DNMT1 to methylate the newly synthesized DNA strand during the replication (Bronner et al. 2010). UHRF1 recruits G9a (Histone H3 lysine 9 methyltransferase) and thanks to this function is considered as a transcriptional repressor (Kim J. K. et al. 2009a).

1.8 DNA demethylation

Although it is considered that 5mC is a stable epigenetic mark and strongly maintained, DNA methylation patterns are dynamic during development and during pathological situations such as cancer. Changes in the methylation patterns happens both at the global and local levels. We will now review the proposed different pathways of DNA demethylation, which can result from either passively at the new DNA strand after replication or actively, by a replication-independent process.

1.8.1 **Passive DNA demethylation**

During DNA replication, the symmetrical CpG methylation pattern is restored by the maintenance proteins UHRF1 through the methylation of the unmodified cytosine in the nascent DNA strand. Therefore, the passive loss or dilution of 5mC can be achieved in the absence of functional DNMT1/UHRF1 through successive cycles of DNA replication (Figure 4). One example of such a mechanism of demethylation is the global erasure of 5mC in the maternal genome during mouse preimplantation development. The first S phase in the newly formed zygote is initiated 8 to10 hours after fertilization, this give rise to the first opportunity for passive demethylation to act as a demethylation mechanism (Rougier et al. 1998). Supporting to this mechanism, it has been shown that in this stage of development, the abundance of DNMT1 is largely reduced by the active exclusion of DNMT1 protein from the nucleus to the subcortical region (Carlson et al. 1992, Howell et al. 2001), where it resides throughout most of the period of preimplantation, thus supporting a step-wise loss of DNA methylation from the maternal genome (Ratnam et al. 2002). Furthermore, alteration in UHRF1/DNMT1 tandem, via UHRF1 over-expression or through disrupted interaction between UHRF1 and DNMT1, induces global genome-wide demethylation, which is a hallmark of cancer cells (Mudbhary et al. 2014, Pacaud et al. 2014).

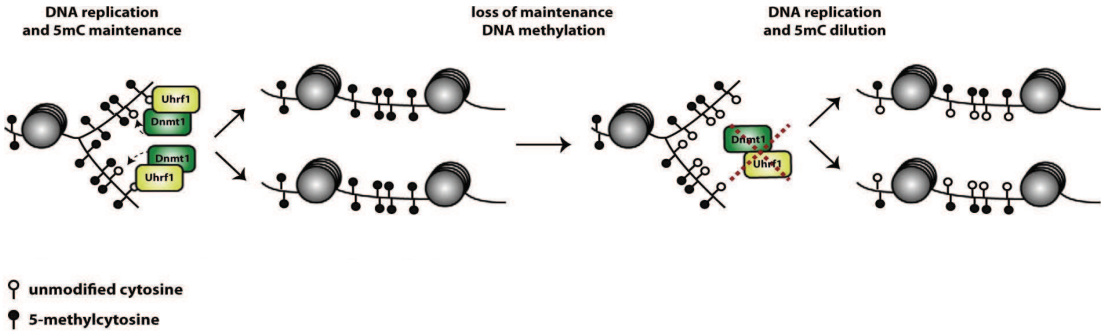


Figure 4; Passive DNA demethylation. Lack of methylation maintenance during DNA replication leads to 5mC dilution, eventually leading to fully unmethylated DNA. (Hill et al., 2014, Genomics).

1.8.2 **Active DNA demethylation**

Active DNA demethylation refers to an enzymatic process that results in the changing or removal of the methyl group from 5mC. After fertilization, the paternal and

maternal genomes show a global loss of the methylation pattern in mammalian fertilized oocytes (Mayer et al. 2000, Oswald et al. 2000). The excessive wave of demethylation observed on the paternal genetic material starts 4-8 hours after fertilization (Mayer et al. 2000) (Figure 5) gave rise to the concept of active DNA demethylation after it has long been controversial. This global fast reduction of DNA methylation levels takes place before the first round of DNA replication begins (Oswald et al. 2000). Thus, it is unlikely to be achieved by the passive dilution pathway. Moreover, when zygotes were treated with aphidicolin, a DNA replication inhibitor, paternal genome demethylation was still detected (Kishigami et al. 2006). Active DNA demethylation has also been reported in somatic cells, which happens at specific genomic loci in response to certain signals. One example, within 20 minutes of stimulation, activated T lymphocytes undergo active demethylation at the interleukin-2 promoter-enhancer region in the absence of DNA replication (Bruniquel and Schwartz 2003). The locus-specific demethylation has also been observed at the promoter of brain-derived neurotrophic factor (BDNF), which is, when methylated, recognized and bound by the MeCP2. Following the depolarization with KCl, BDNF is upregulated, coinciding with the release of MeCP2 and demethylation of the promoter (Martinowich et al. 2003).

The local active DNA demethylation has also been reported to take place during nuclear hormone-regulated gene activation. For example, the pS2 (also known as TFF1) promoter exhibits periodic methylation and demethylation that coincides with cyclical binding of oestrogen receptor-α (ERα) and expression of pS2 (Metivier et al. 2008).

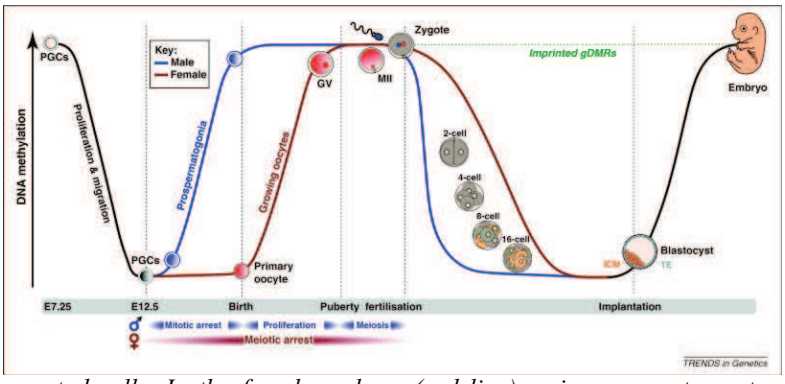


Figure 5. DNA methylation changes during developmental epigenetic reprogramming. Following sex-determination, new DNA-methylation landscapes are established in germ-cell precursors in an asymmetrical fashion in male and female embryos. In the male embryo (blue line), de novo methylation takes place before meiosis in mitotically arrested cells. In the female embryo (red line), primary oocytes enter meiosis and arrest in prophase-I (diplotene stage); DNA methylation is established after birth during the follicular/oocyte growth phase. Following fertilisation, a new wave of DNA demethylation takes place that is distinct on the parental genomes. In the zygote, DNA methylation of the paternal genome is rapidly erased by an active mechanism (blue line). Demethylation of the maternal genome is slower (red line) and is dependent on DNA replication (passive demethylation). Concomitant with blastocyst implantation and cell-lineage determination, new methylation landscapes become established, associated with cellular differentiation. (Smallwood and Kelsey, Trends in Genetics, 2012).

1.8.3 Mechanisms of active DNA demethylation

Regarding the importance of DNA methylation in diverse biological processes, beside the observations of active DNA demethylation in embryonic development and somatic cells, extensive efforts have been made to identify a DNA demethylase, and/or to understand the mechanisms by which this process is achieved.

So far, various mechanisms have been proposed as possible pathways by which active DNA demethylation can occur, including enzymatic removal of the methyl group of 5mC, radical SAM mechanism, base excision repair (BER) through direct excision of 5mC, deamination of 5mC to T, followed by BER of the T:G mismatch, nucleotide excision repair (NER) and the oxidation of 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) (Franchini et al. 2012, Li C. J. 2013, Wu H. and Zhang 2014, Wu S. C. and Zhang 2010). We will now review hereafter the different putative mechanisms underlying "DNA demethylation".

1.8.3.1 Enzymatic removal of the methyl group of 5mC

Initially, the search for DNA demethylation mechanisms focused on the identification of an enzymatic activity that directly removes the methyl group from 5-methylcytosine, which would be, if occurs, the most straightforward way to achieve DNA demethylation. However, this direct pathway needs a thermodynamically unfavorable reaction to break the carbon–carbon bond, which makes this pathway unlikely to occur in living cells. Nevertheless, a study proposed earlier, that the direct removal of methyl group could be catalyzed by the methyl-CpG-binding domain protein 2 (MBD2) leading to the release of methanol (Bhattacharya et al. 1999). Since then, no other laboratory could reproduce this enzymatic activity of MBD2, leaving this mechanism largely controversial. Moreover, other studies have shown that MBD2 can stably bind methylated DNA (Hendrich and Bird 1998, Ng et al. 1999) which give rise to the question of how MBD2 can bind to 5mC if it can efficiently remove the methyl group. Moreover, further study has shown that MBD2 null mice are viable and also exhibited normal methylation patterns (Hendrich et al. 2001). Importantly, the paternal pronucleus of MBD2-null zygotes still exhibits normal demethylation activity (Santos et al. 2002). All these findings have raised serious doubts on the capacity of MBD2 to serve as a DNA demethylase and is now completely given up.

1.8.3.2 Radical SAM mechanism

Efforts to define a protein(s) responsible for the observed paternal genome demethylation in zygotes led to propose a possible role of elongator complex protein 3 (ElP3) in DNA demethylation of embryonic development, and that Fe–S radical SAM domain of this protein is required for the demethylation process (Okada et al. 2010). ElP3 is a member of the core elongator complex (ElP1–ElP3), which associates with an other subcomplex (ElP4–ElP6) to form the holo-elongator complex (Hawkes et al. 2002). Knockdown of the ElP1 and ElP4 components impaired paternal genome demethylation, suggesting that it is likely that the entire elongator complex may be involved in the demethylation process (Okada et al. 2010).

Although the proposed role of SAM domain could provide a clue for an enzymatic mechanism of ElP3, other studies suggest that the Cys-rich domain of ElP3 is required for the integrity of the elongator complex (Greenwood et al. 2009) raising the possibility that the Fe–S radical SAM motif may has a structural rather than an enzymatic role. In spite of this proposed role, direct biochemical evidence of the demethylation activity of ELP or the elongator complex and genetic evidence using ElP3-null oocytes are still lacking.

1.8.3.3 Nucleotide Excision Repair (NER) to erase 5mC

Theoretically, the erasure of 5mC is possible by the excision repair of short genomic regions that contain methylated cytosine nucleotides. Via the NER pathway, cells can repair bulky DNA lesions formed by exposure to radiation or chemicals. Once the bulky DNA lesion is recognized, specific enzymatic activities introduce dual incisions flanking the damage region resulting in a single-stranded gap (usually 24–32 nucleotides). The resulted gap is then filled in by DNA repair polymerases and ligases. Supporting to this way, it has been reported that the Gadd45 (growth arrest and DNA-damage-inducible protein 45), stimulates active DNA demethylation via NER (Barreto et al. 2007). However, the role of this protein family in DNA demethylation remains uncertain as Gadd45a- and Gadd45b null mice are quite normal and have no global alteration in DNA methylation levels (Engel et al. 2009, Ma et al. 2009).

1.8.3.4 Direct excision of 5mC followed by Base excision repair (BER)

One hypothesis, proposed for some time, is that DNA demethylation can be achieved through the base excision repair (BER) pathway. This mechanism of repair involves a DNA glycosylase that cleaves the N-glycosidic bond between the 5-mC base and the deoxyribose to remove the target base resulting in an abasic (apurinic and apyrimidinic: AP) site. The DNA backbone is subsequently nicked by an AP lyase activity to generate a 5′ phosphomonoester and a 3′ sugar phosphate residue. An AP endonuclease then removes the 3′ sugar group leaving a single nucleotide gap that is ultimately filled in by DNA repair polymerases and ligases (Sancar et al. 2004). These steps lead to remove 5mC without previous modification of this base. This way of active demethylation has been reported in flowering plants (*e.g., Arabidopsis thaliana*), where the reaction is mediated by the Demeter (DME)/repressor of silencing 1 (ROS1) family of DNA glycosylases and base excision repair (BER) machinery (Gehring et al. 2006, Gong et al. 2002). These findings led to think that such a mechanism could also occur in mammalian cells. However, mammalian orthologs of DME/ROS1 enzymes have not been identified. Therefore, mammalian cells may achieve active DNA demethylation using different mechanisms rather than directly removing of 5mC.

Evidences demonstrating that mammalian cells first tend to modify the 5mC followed by the repair of this modified base. Indeed, two pathways has been reported to modify 5mC in mammalian cells, first; hydrolytic deamination of 5mC to thymine (T) leading to form a G:T mismatch, and second; oxidation of 5mC to 5hmC, 5fC and 5caC by the Tet family (Figure 6). Here we review these two pathways;

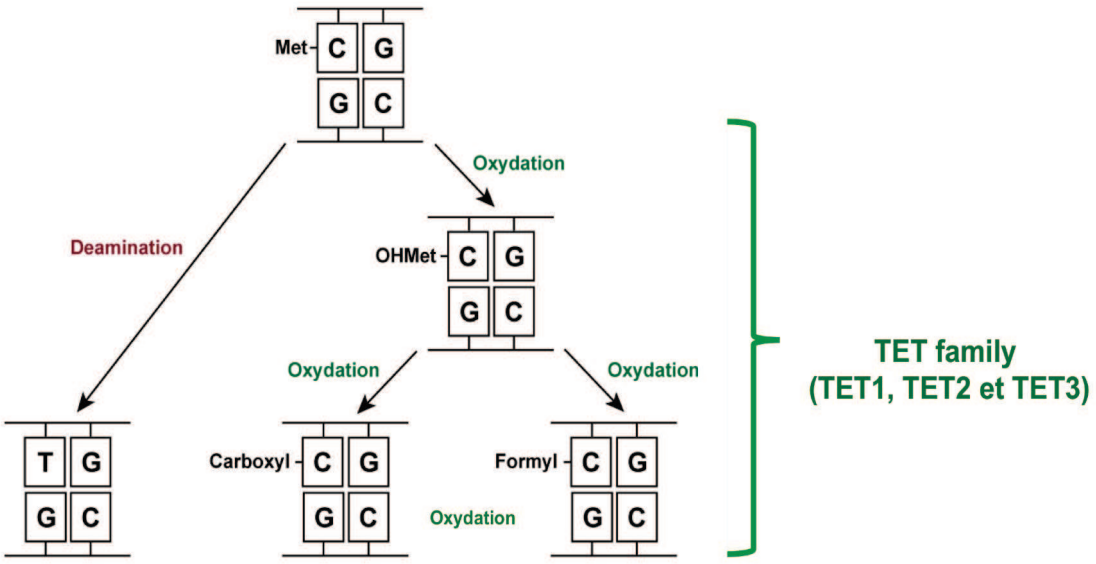


Figure 6. 5meC modifying pathways. Deamination of 5mC to thymine (T) leading to form a G:T mismatch (Left part), and oxidation of 5mC by the Tet family leads to 5hmC, 5fC and 5caC (Right part).

1.8.3.5 Hydrolytic deamination of 5mC followed by BER

The 5mC undergoes spontaneous hydrolytic deamination, which subsequently generates thymine (Gehring et al. 2006). Unmethylated cytosine could also be deaminated but with a lower rate resulting in the formation of uracil. Both deaminase activities towards C or mC are considered mutagenic, because if left unrepaired, the resulting G:U and G:T mispairs will give rise to a C to T transitions upon replication.

It has been suggested that some members of the vertebrate-specific Activation-induced Cytidine Deaminase (AID)/Apolipoprotein B mRNA Editing enzyme, Catalytic polypeptide (APOBEC) family such as AID (Morgan H. D. et al. 2004, Popp et al. 2010), APOBEC1, APOBEC2 (Guo et al. 2011, Morgan H. D. et al. 2004), could play a key role in active DNA demethylation. AID and APOBEC proteins are zinc-dependent cytidine deaminases acting on single-stranded polynucleotides and deaminate cytosines in different contexts (Chelico et al. 2006). Among the AID/APOBEC family, Aid and APOBEC1 are expressed in mammalian oocytes and embryos, which points to a possible role in global DNA demethylation occurs in these stages (Morgan H. D. et al. 2004). Moreover, it has been reported that locus-specific and global demethylation in zebrafish embryos is mediated by the AID/Apobec family of deaminases and the DNA glycosylase MBD4 (Rai et al. 2008). Overexpression of AID and MBD4 together in zebrafish

embryos, causes demethylation of the bulk genome and injected methylated DNA fragments. Importantly, overexpression of the glycosylase alone, does not lead to such changes demonstrating a deaminase activity of AID towards the 5mC (Rai et al. 2008). Moreover, a bisulphite sequencing study indicated that DNA methylation levels of male and female PGCs derived from AID-null embryos increased about 4% (from 18% to 22%) and 13% (from 7% to 20%), respectively, when compared to their wildtype counterpart (Popp et al. 2010), suggesting that AID may contribute to primordial germ cells (PGC) demethylation. However, DNA methylation levels in AID-null PGCs (~20%) are still relatively low compared with ES or somatic cells (70–80%) and significant demethylation still occurs in the absence of AID, indicating that other factors could be responsible for PGC demethylation.

In addition to AID and APOBEC, other studies showed that DNMTs are implicated in 5mC deamination, even though they are commonly known for their ability to catalyse DNA methylation. The starting evidence, indicating their involvement in the deamination process, initially came from studies in bacteria where the methyltransferases M. EcoRII (Wyszynski et al. 1994) and M. HpaII (Zingg et al. 1996) were shown to possess deaminase activities. Consistent with studies on bacteria, the mammalian counterparts, DNMT3A and DNMT3B, have been shown to possess deaminase activity toward 5mC *in vitro*, but this proposed deaminase activity of DNMT1 can only occur under conditions where SAM concentrations are very low or absent (Metivier et al. 2008). However, given that SAM is present at relatively high levels in all cell types, the physiological relevance of this reaction remains uncertain.

In support of 5mC deamination/BER mechanism, inhibitors of poly-ADP-ribose polymerase 1 (PARP1), a member of the BER pathway proteins, and of apurinic/apyrimidinic endonuclease 1 (APE1), an essential downstream enzyme required to generate an abasic site, were found to alter paternal-specific DNA methylation (Hajkova et al. 2010).

Altogether, these observations lead to accept that DNA demethylation can be achieved by deamination of 5mC followed by BER to replace the mismatched T with unmethylated C.

1.8.3.6 Oxidative modification of 5mC

The searching for enzymatic activity modifying 5mC through oxidation was prompted by the discovery of the biosynthesis of ‘‘base J’’ (b-D-glucosyl-hydroxymethyluracil), a modified base present in the genome of the parasite Trypanosome brucei. Biosynthesis of Base J is achieved by a two-step process. First, thymine (T) is oxidized to form 5-hydroxymethyluracil (5hmU) by J-binding protein (JBP) 1 and 2. These two proteins are members of the Fe (II)/ α ketoglutarate (α-KG)-dependent dioxygenase family (Loenarz and Schofield 2011). Then, a glycosyltransferase completes the synthesis of base J by adding glucose groups to 5hmU (Borst and Sabatini 2008). It has also been shown that further oxidation of 5hmU leads to form 5-formyluracil and 5-carboxyluracil. This later base then undergoes a decarboxylation process by isoorotate decarboxylase completing a cycle from T to U, with a putative similar mechanistic to 5mC demethylation (Wu S. C. and Zhang 2010).

Efforts to find mammalian homologues with similarity to the dioxygenase domains of JBP proteins led to the identification of the ten-eleven translocation family of proteins (Tahiliani et al. 2009). TET1 was first identified in Purkinje cells of the brain, and later in mouse embryonic stem cells (Kriaucionis and Heintz 2009, Tahiliani et al. 2009). Since then, two other TET family members, TET2 and TET3, harbouring the dioxygenase motif, have been identified (Ito et al. 2010).

All TET proteins contain a C-terminal catalytic domain that includes a Cys-rich insert and a large double-stranded b-helix (DSBH) domain. TET1 and TET3 proteins also contain a N-terminal CXXC domain (Kohli and Zhang 2013) (Figure 7 a). Concerning the oxidation activity toward 5mC, overexpression of TET1 in cultured cells results in reduction of genomic 5mC level, and recombinant TET1 proteins can oxidize 5mC *in vitro* resulting in 5hmC. Similar to that achieved by JBP1/2, this reaction, requires Fe(II) and alpha-ketoglutarate binding to complete the oxidation of 5mC to 5hmC (Tahiliani et al. 2009) (Figure 7 b). The same activity was later demonstrated with all TET (TET1-3) proteins in mouse (Ito et al. 2010).

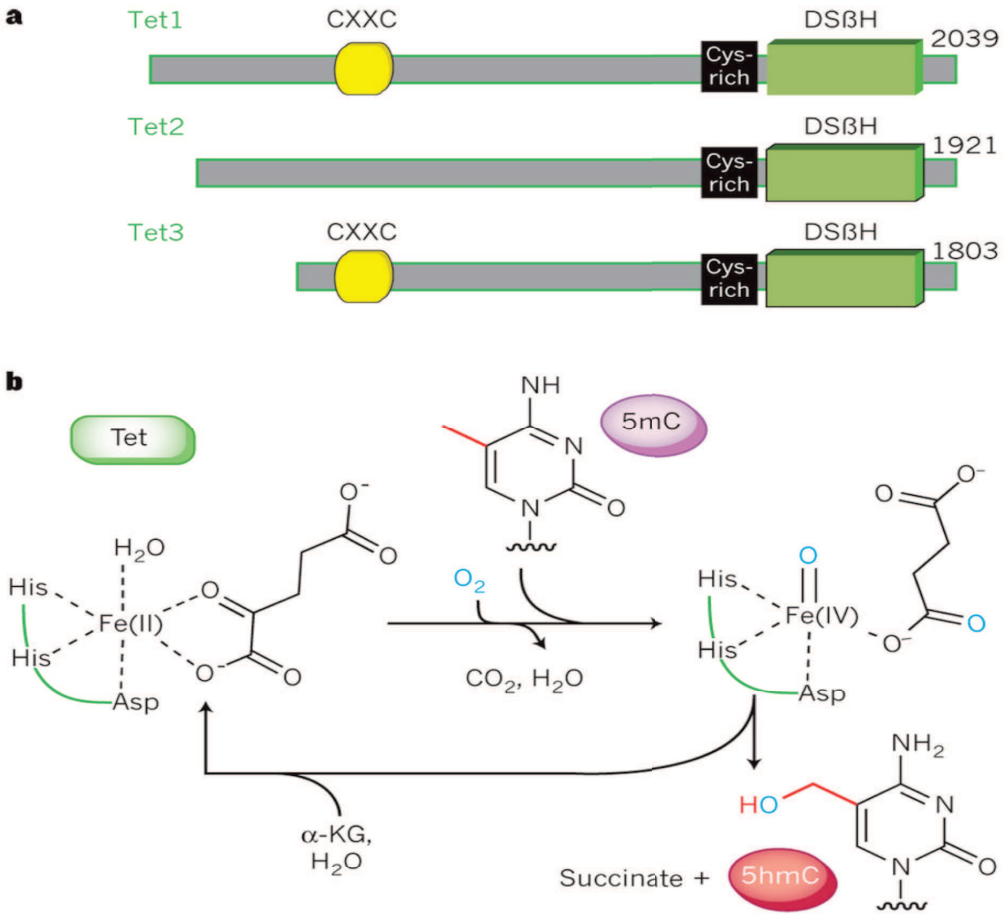


*Figure 7; Schematic and Catalytic reaction of Tet enzymes. **a**, Schematic of mouse Tet enzymes, showing the double-stranded β-helix (DSβH) fold core oxygenase domain, a preceding cysteine(Cys)-rich domain and a CXXC domain in Tet1 and Tet3. **b**, Catalytic mechanism for generation of 5hmC by Tet enzymes. An active site Fe(II) (left) is bound by conserved His-His-Asp residues in Tet and coordinates water and α-ketoglutarate (α-KG). A two-electron oxidation of α-KG by molecular oxygen yields CO2 and enzyme-bound succinate, and results in a high-valent Fe(IV)-oxo intermediate (Costello et al.). The intermediate reacts with 5mC to yield 5hmC, with a net oxidative transfer of the single oxygen atom to the substrate, resulting in regeneration of the Fe(II) species. (Kohli and Zhang., 2013).*

Ito and colleagues have later reported that TET proteins are capable of further oxidizing 5hmC to 5fC and 5caC (Ito et al. 2011) (Figure 8). Indeed, the important observation that the loss of paternal DNA methylation coincides with a highly increase in 5hmC levels (Inoue and Zhang 2011, Iqbal et al. 2011, Wossidlo et al. 2011) and 5fC/5caC (Inoue and Zhang 2011) strengthens the proposed oxidative demethylation mechanism as a main pathway in the paternal genome demethylation after fertilization. Within the TET family, TET3 is highly enriched (30-fold) in the zygote comparing to TET1 and TET2, which makes it as the most likely candidate responsible for the

oxidation of 5mC in the paternal genome during this development period (Iqbal et al. 2011). This role was later confirmed by small interfering RNA (siRNA) approach against TET3. Targeting TET3 lead to abrogate 5hmC and to an important increase of 5mC levels (Wossidlo et al. 2011). The oxidative demethylation pathway could further include a passive dilution of the oxidation products of 5mC (5hmC, 5fC or 5caC) (in a replication-dependent manner), or another enzymatic activity to remove the final oxidation products (Shen et al. 2013).

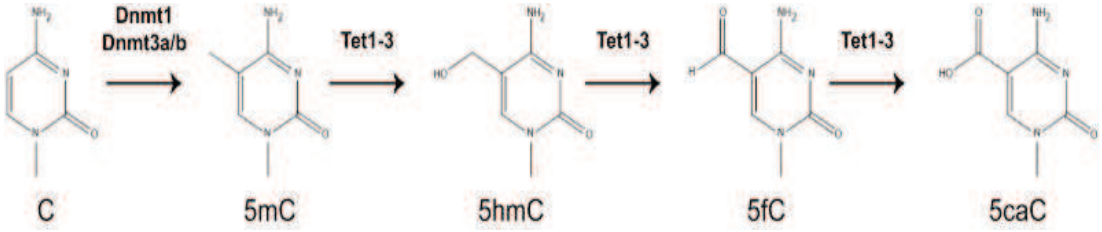


Figure 8; TET-induced 5meC oxidation. TET proteins are able to oxidize 5mC to 5hmC and further to 5fC and 5caC.

1.8.3.6.a **Passive dilution of oxidized 5mC**

Similar to 5mC, the oxidized products (5hmC, 5fC and 5caC) could be diluted by a replication-dependent manner (Figure 9a). The dilution of the oxidized 5mC may be effective even in the presence of a functional methylation maintenance machinery (DNMT1/UHRF1) because the affinity of this machinery towards the oxidized bases is less than towards 5mC. It has been established that DNMT1 is significantly less efficient (10-60 fold) in methylating hemihydroxi-methylated (hCG:GC) than hemimethylated (mCG:GC) sites *in vitro* (Hashimoto et al. 2012). The decreased efficiency of DNMT1 to methylate the newly strand could also occur, but it is not yet proven, in the presence of hemiformylcytosine (fCG:GC) or hemicarboxylcytosine (caCG:GC).

In summary, TET proteins may initiate a two-step demethylation process in proliferating cells that involves initial active modification of 5mC through oxidation and subsequent replication-dependent (passive) dilution of 5hmC and potentially 5fC/ 5caC (Wu H. and Zhang 2014).

1.8.3.6.b Active removal of oxidized methyl group

Beside the passive dilution of the oxidized bases, it has also been proposed that active demethylation could be achieved by two different enzymatic steps, starting with the oxidation of 5mC, then followed by one of two possible mechanisms. First, removal of the oxidized methyl group, through either dehydroxymethylation of 5hmC by DNMT3A/3B under oxidizing conditions but in the absence of SAM (Chen C. C. et al. 2012), or by decarboxylation of 5caC by a putative decarboxylase (Figure 9b) (Schiesser et al. 2012) that may directly convert the oxidized 5mC base (5caC) to unmodified cytosine. Second, excision of oxidized 5mC bases by a specific glycosylase including removal of 5hmU (deamination product of 5hmC) (Guo et al. 2011) or 5fC/5caC (He et al. 2011), which lead to abasic sites that are further repaired by BER to restore unmethylated cytosine. The first proposed way (dehydroxymethylation of 5hmC) remains unclear regarding that the levels of SAM are relatively abundant in all cell types.

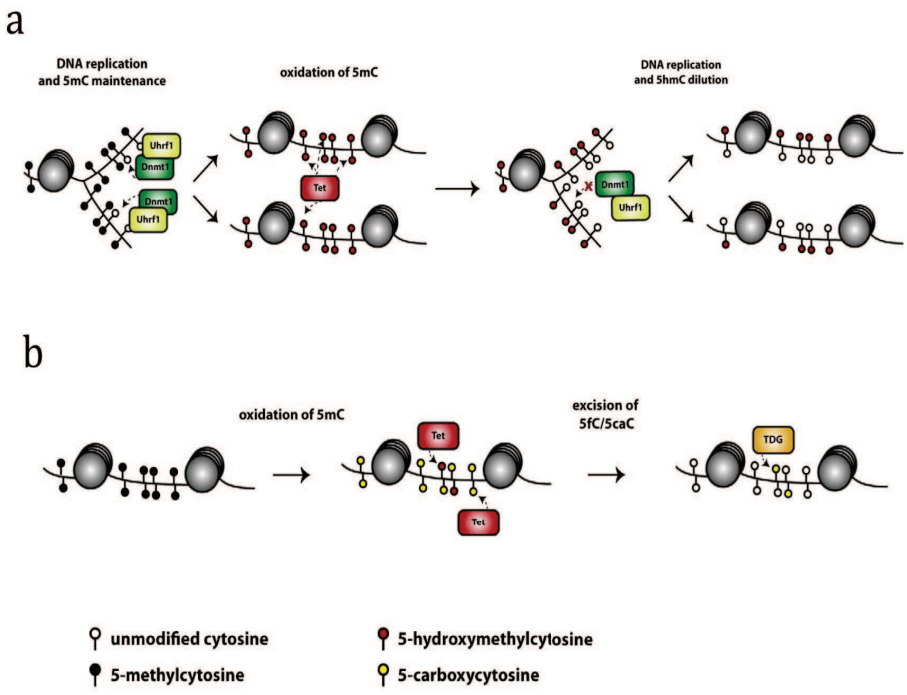


Figure 9; TET-induced DNA demethylation. (a), Conversion of 5mC to 5hmC impacts on the maintenance activity of Dnmt1 (in vitro) resulting in dilution of 5hmC through DNA replication, eventually leading to fully unmethylated DNA. (b) 5hmC can be further converted by Tet enzymes to 5fC and 5caC that are targeted for excision by TDG glycosylase followed by BER repair. (Hill et al., 2014, Genomics).

1.9 DNA glycosylases

After base deamination, a glycosylase activity is needed as an initiator of the BER pathway. When cytosine deaminases act on unmodified cytosine, the resulting uracil represents a foreign base in DNA independently of whether it is present in a matched or mismatched condition. This base is recognised and repaired by uracil DNA glycosylases (Olsen et al. 1989). However, the resulting thymine (T) from deamination of 5mC, is a normal base in genomic DNA. Thus, mismatch repair proteins should discriminate between thymines in a mismatch with guanosines (G:T) from the correctly paired thymines with adenosines (A:T). Physical interaction of deaminases with glycosylases could be critical for recognition of mismatches generated by deamination of 5mC, and therefore participating in the demethylation process.

So far, two glycosylases in mammals, Thymine DNA glycosylases (TDG), and MBD4, have been reported to selectively recognize G:T mismatches. This selectivity comes from their ability to interact not only with the T but also with the opposing base (Maiti et al. 2008, Manvilla et al. 2012, Yoon et al. 2003). DNA glycosylases are catalytically subdivided into monofunctional and bifunctional enzymes. Monofunctional glycosylases perform base excision only, using an activated water molecule for nucleophilic attack on the N-glycosidic bond, while bifunctional glycosylases use an amino group of a lysine side chain for the same purpose, and subsequently cleave the DNA backbone 3′ to the lesion (Figure 10) (Jacobs and Schar 2012). Regarding this subdivision, MBD4 and TDG have been classified as monofunctional DNA glycosylases that do not possess lyase activities.

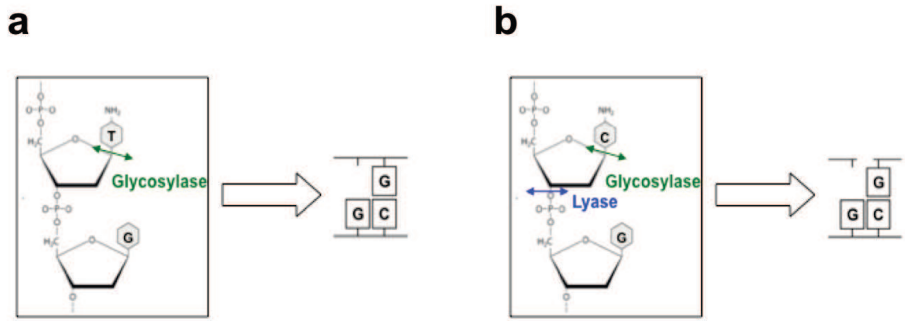


Figure 10; Mono- and Bi-functional glycosylases. a, Monofunctional glycosylase hydrolyses N-glycosidic bond between base and ribose creating an abasic site. b, Bifunctional glycosylase beside hydrolysing the N-glycosidic bond, has a lyase activity to break DNA backbone at the 3' side of the abasic site creating a nick.

1.9.1 Glycosylase activity of MBD4

MBD4, also known as Methyl-CpG-binding Endonuclease (MED1), is a 66 kDa protein that belongs to the methyl-CpG-binding domain protein family (MBDs). The glycosylase activity of MBD4 was initially described by Hendrich and colleagues who identified this protein in a database screening of MBD sequences (Hendrich et al. 1999). Later, MBD4 was reported to act as a G:T and G:U mismatch-specific thymine and uracil glycosylase (Petronzelli et al. 2000b). This protein consists of an N-terminal MBD domain and a C-terminal DNA glycosylase domain. The C-terminal region of MBD4 shows homology to the glycosylase/endonuclease domains of bacterial repair proteins (Bellacosa et al. 1999). Such homology led to suggest two possible functions of MBD4, *i.e.,* DNA glycosylase and endonuclease activities. However, all the further *in vitro* experiments failed to provide evidence that MBD4 has endonuclease activity (Drummond and Bellacosa 2001).

As the CpG dinucleotide is largely represented in a methylated form (Bird A. P. 1980), the spontaneous hydrolytic deamination of methylated cytosine causes mCpG-TpG transitions, whereas non-methylated CpG mutates to UpG. Obviously, MBD4 is able to excise both mutated nucleotides in order to be repaired (Hendrich et al. 1999). The presence of MBD and glycosylase domains increases the possibility of MBD4 involvement in DNA repair associated with methylated CpG sites (Scharer and Jiricny 2001). Indeed, MBD4-null mice showed a two to three times higher number of mCpG-TpG transitions (Millar et al. 2002). These observations demonstrate that the glycosylase activity of MBD4 contributes to G:T processing to reduce the mCpG-TpG mutation in living cells, and that its role can not be fully compensated by the presence of other glycosylases. Moreover, a demethylation of the bulk genome and injected methylated DNA fragments has been observed by overexpression of MBD4 with AID in zebrafish embryos. Interestingly, this demethylation effect could not be observed with an overexpressed catalytically inactive MBD4 (Rai et al. 2008).

At the locus specific level, MBD4 was reported to control CpG methylation in the context of parathyroid (PTH) hormone-induced gene activation. This was shown for the CYP27B1 promoter, which undergoes active demethylation upon hormone stimulation (Kim M. S. et al. 2009b). Both, promoter activation and cytosine demethylation, coincided with and depended on the physical association of MBD4 with downstream BER factors.

1.9.2 MBD4/Substrates interaction

As mentioned above, the specificity of MBD4 towards its substrate is provided by its ability to recognize and connect the opposite base in the complementary DNA strand. Recent crystal structure of MBD4-glycosylase domain with DNA containing abasic site (G:THF [tetrahydrofuran] mismatch, THF was used as an analog for the abasic site) was solved at 2.76 Å resolution (Manvilla et al. 2012). This crystal structure showed that the glycosylase domain interacts with DNA in the minor groove region and bends the target DNA site by 57°, thereby, flipping the abasic nucleotide (THF) completely out of the DNA duplex into the MBD4 active site (Figure 11). The complementary DNA strand containing the guanine base was nestled into a recognition pocket. In general, the major molecular interactions happens via the guanidinium side chain of Arg468, which penetrates the DNA minor groove, where it plugs the void in the DNA helix that is created by nucleotide flipping and forms the hydrogen bonds with the two "pinched" phosphates. These phosphates contacts help in the maintaining of the nucleotide in a flipped state.

Other crystal structures study on MBD domain of MBD4 bound to a DNA fragment containing the 5mCG/5mCG site or its deamination product, 5mCG/TG showed that the MBD domain can also bind to the T opposite to mC, which could facilitate a synergetic action with the glycosylase domain towards the target substrates (Otani et al. 2013).

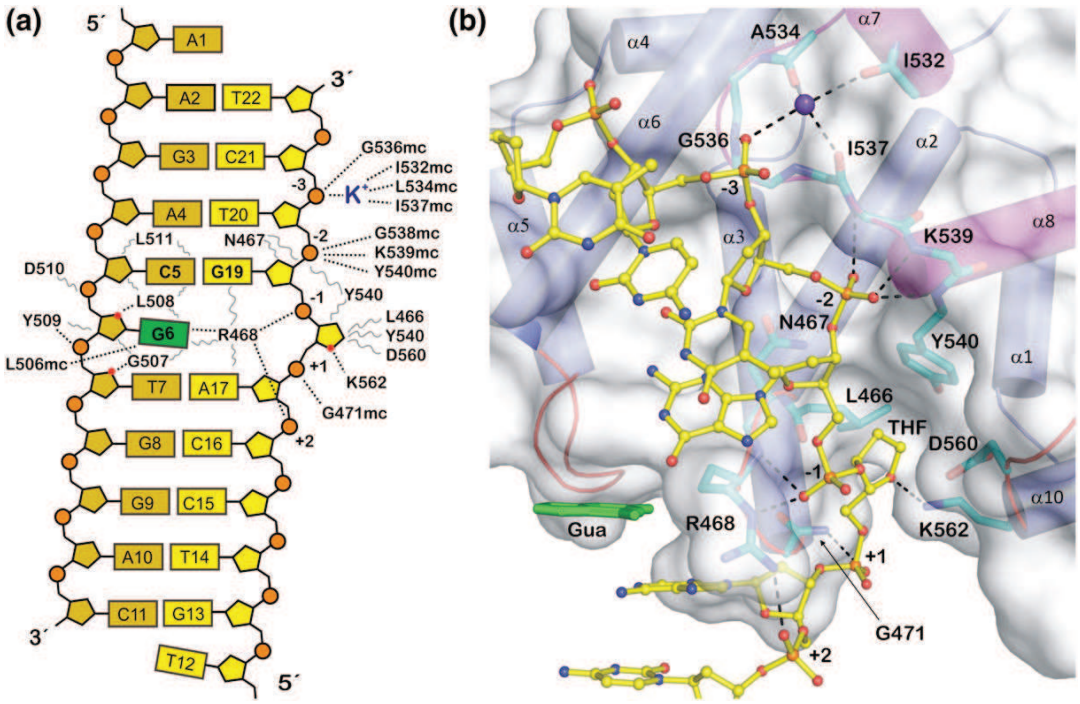


Figure 11. Interactions between MBD4 glycosylase domain and DNA sbstrate. (a) An interaction map summarizes the electrostatic and van der Waals interactions observed in the crystal structure (broken and curvy lines, respectively). The catalytic loops colored in red and include α2–α3 (N467–S470), α5–α6 (K504–L508), and α9–α10 (V556–D560). Nucleotides in the target DNA strand are yellow, including the flipped THF abasic analog. Complementary DNA is light orange, and the mismatched guanine is in green. (b) Interactions between MBD4 glycosylase and the lesion-containing DNA strand. MBD4 residues that contact the DNA are colored cyan, and the K+ ion is in purple. The complementary strand is omitted, except for the mismatched guanine (green). (Manvilla et al. 2012)

1.9.3 Glycosylase activity of TDG;

In the search for a specialized activity able to remove mismatched thymines, Jiricny Brown and Jiricny (1988) were able (by using transfection experiments with G:T mismatched SV40 DNA) to identify the first G:T directed repair activity in African green monkey kidney cells that efficiently replace the T with C (Brown and Jiricny 1988).

The discovery of the human TDG came as a result of subsequent purification of a G:T binding and processing enzyme from nuclear extracts of HeLa cells and the molecular cloning of the respective cDNA (Neddermann and Jiricny 1993). It was the first mismatch-specific DNA glycosylase able to hydrolyze thymine and uracil from G:T and G:U mispairs *in vitro* (Neddermann et al. 1996). Since then, orthologs of TDG has been defined in bacteria, yeast, insects, frogs, and vertebrates (Gallinari and Jiricny 1996, Hardeland et al. 2003). Sequence analysis of vertebrate orthologs shows that TDG

possesses a highly conserved central region containing the active site, and more divergent amino and carboxy-terminal regions (Cortazar et al. 2007).

Interestingly, beside its role in processing G:T/U mismatch repair, TDG has recently been shown of being able to excise the oxidation products of 5hmC (5fC and 5caC) (Maiti and Drohat 2011, Zhang L. et al. 2012), which further implicates the thymine glycosylases in active demethylation. Supporting to this pathway, overexpression of TET and TDG in HEK293 cells, rapidly depleted 5fC and 5caC (He et al. 2011). Moreover, other study reported that TDG-deficiency in mouse ESCs caused a 5- to 10- fold increase in 5fC and 5caC levels (Shen et al. 2013). The molecular basis for excision of 5fC/5caC by TDG could be explained by the effect of 5fC and 5caC in destabilizing the base-sugar binding (N-glycosidic bond) resulting in a cytosine with a weakened N-glycosidic bond. Destabilizing N-glycosidic bond has been described as a recruiting event of TDG to its substrates (Bennett et al. 2006).

1.9.4 TDG/Substrate interaction

It has been shown that hTDG has an activity of around 18,000 fold higher on T/G than that on T/A (Morgan M. T. et al. 2007). This fact leads to think that the substrate selectivity of TDG depends on its ability to specifically recognize and interact with the nucleotide in complementary position to avoid the potential attack of the normal DNA bases. The crystallographic analysis provided insight to the mechanistic bases of G:T/U specificity of TDG. Maiti et al solved a crystal structure of hTDG catalytic domain (hTDGcat, residues 111–308) binds to a 22-bp DNA containing a tetrahydrofuran nucleotide (THF), a chemically stable mimic of the natural AP product, this structure showed that TDG forms with the substrate a 2:1 complex: one subunit at the abasic site (product complex) and the other at an undamaged site (nonspecific complex), and that the catalytic pocket establish a specific hydrogen-bonding interactions with the guanine base opposite to the AP-site (in the complementary strand) (Figure 12) (Maiti et al. 2008). Following the base release, contacts with G base and the non-specific DNA contacts mediated by the N-terminus cooperate to prevent free dissociation of TDG from the AP-site (Cortazar et al. 2007). A structure analysis of the TDG core, along with kinetic studies of both core and full-length proteins, has shown that the amino-terminus is critical for the TDG glycosylase function. Experimental evidence supports a mechanistic model in which the N-terminal domain of TDG forms a flexible “clamp” that holds the

glycosylase onto the DNA (Steinacher and Schar 2005). In this state, TDG may slide along the DNA in search of a G mismatched substrate. Without the N-terminus region, TDG binds less stably to DNA and loses the ability to excise thymine from G:T mispairs while retaining G:U processing activity (Steinacher and Schar 2005).

Other independent crystallographic analysis performed on hTDG catalytic domain bound with 5caC-containing dsDNA. This study showed that the active site of TDG specifically recognizes and binds with the backbone of the 5caC-containing strand via electrostatic interaction and bends the dsDNA backbone by ~45° towards the active site, then it inserts the side chain of Arginine 275 through the dsDNA minor groove to push the 5caC base out of the DNA groove. Besides to the arginine 275, the catalytic domain of TDG recognizes by other small pocket the the 5-carboxyl moiety of 5caC that strengths the binding. (Figure 13) (Zhang L. et al. 2012).

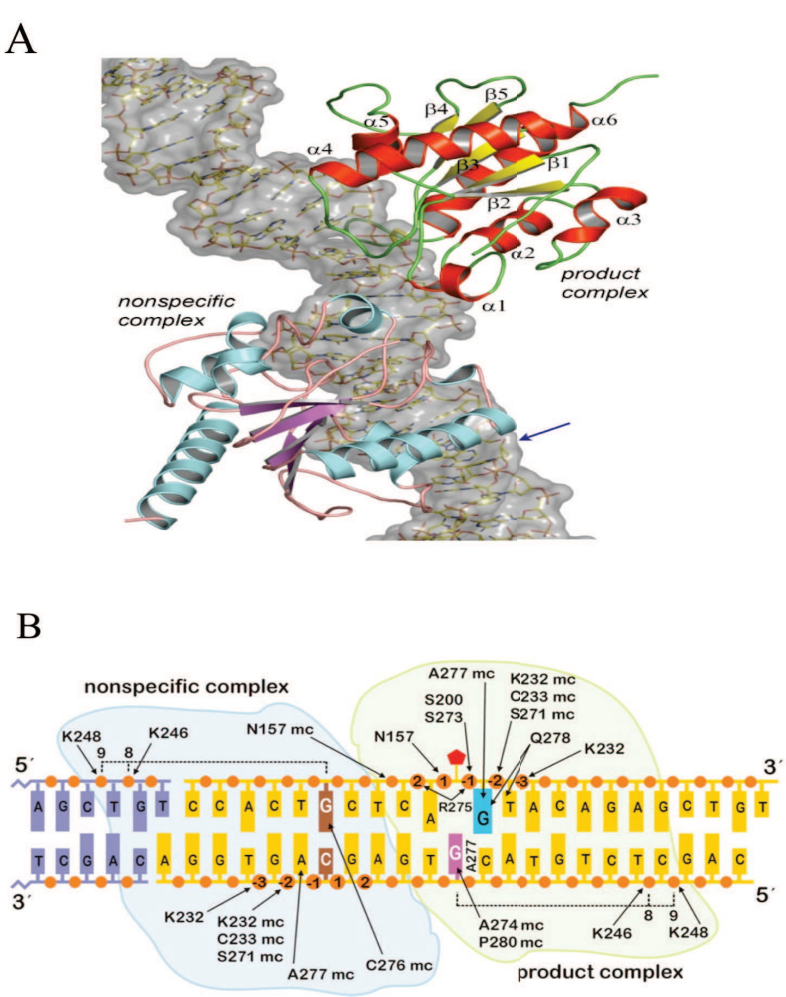


Figure 12; Overview of TDG/ substrate structure. (A) The hTDG catalytic domain binds a 22-bp DNA containing a tetrahydrofuran (THF) in a 2:1 complex: nucleotide one subunit at the abasic site (product complex) and the other at an undamaged site (nonspecific complex). DNA shown includes a full 22-bp duplex and part of the adjacent duplex joined by 3′ A/T overhangs (blue arrow). Overall, the two subunits are highly similar. (B) Schematic overview of the enzyme-DNA interactions and the dimer interface. The 22-bp DNA is yellow with phosphates shown as orange circles. The adjoining DNA fragment (purple) shows contacts with K246 and K248 from the NS subunit. The arrows represent hydrogen bonds involving side-chain or main-chain (mc) atoms of the enzyme. In the product complex, the flipped (THF) is a red pentagon, the "opposing G" is magenta, and the "3′-G" is cyan. A277 intercalates the complementary strand, disrupting base-stacking interactions between the opposing G and its 5′ neighbor. Contacts involving N157, S273, and A274 for hTDGcat. (Maiti A et al. 2008).

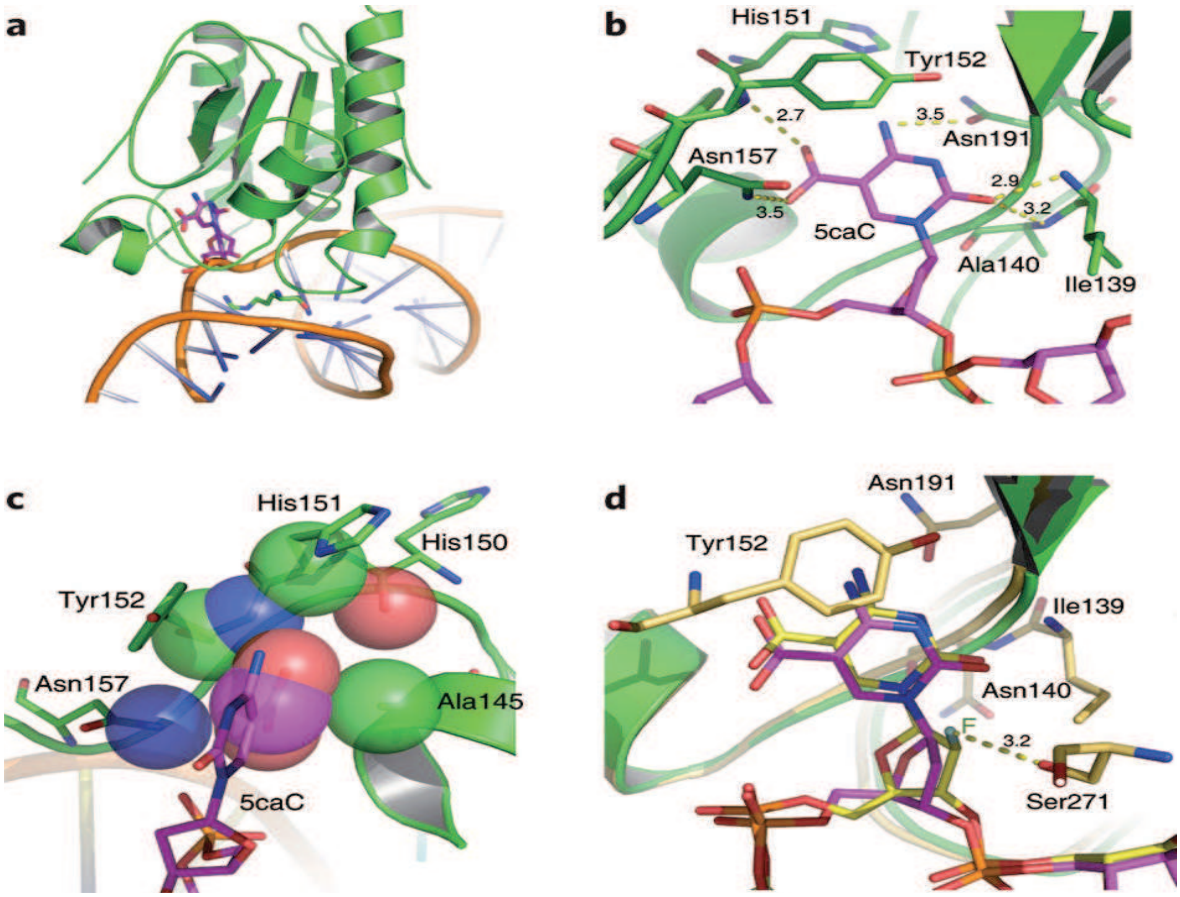


Figure 13; (a) Overview of the structure of hTDGcat (N140A) bound with 5caC-containing dsDNA. The 5caC (colored in magenta) penetrates into the active site pocket and holded by wedge residue Arg275 (shown as sticks and colored in green). (b) A network of hydrogen bonds in the active site of hTDG specifically recognizes 5caC. Residues involved in the interactions are labeled and shown as sticks, and hydrogen bonds are shown as yellow dashes. (c) The interactions involved in the 5-carboxyl–binding pocket. The atoms involved are presented as transparent spheres with nitrogen in blue, oxygen in red and carbon in green. (d) Superposition of 5caC and β-F-5caC in the active site pocket of the hTDGcat (N140A)–A·5caC and hTDGcat–G·β-F-5caC structures. Residues involved in the interactions in the β-F-5caC structure are shown as sticks and colored in yellow and orange. The fluorine atom is labeled in dark green. The hydrogen bond interaction between β-F-5caC and Ser271 is shown in yellow dashes. (Zhang et al., 2012).

1.9.5 TDG/AP site dissociation

The strong fixation of TDG on the substrate DNA, after the base release, predicts the need for a release factor that stimulates the displacement of TDG so that BER can further proceed. In this context, the carboxy-terminal of TDG plays a role in its turnover from the abasic site. It has been shown that the TDG is modified by the Small Ubiquitin-Like Modifiers (SUMOs) SUMO-1 and SUMO-3, and the covalent attachment of SUMO to lysine residue (K330) in the carboxy-terminal of human TDG reduces the affinity for

the AP-site, *in vitro,* and allowing the turnover of TDG from its substrate (Hardeland et al. 2002). This SUMO modification might be triggered by the presence of downstream acting BER factors. Interestingly, the N-termial is also needed in C-terminal-induced TDG turnover. The SUMOylation of C-terminal in TDG truncated of its N-terminal domain did not favor the detach of TDG from the AP-site, suggesting that the C-terminal SUMOylation may influence the structure of N-terminal region in the full-length TDG (Steinacher and Schar 2005). Another possible candidate has been proposed to induce or participate in the release of TDG from the AP-site, which is the AP-endonuclease 1 (APE1), a BER enzyme acting downstream of TDG. This appears likely to occur since the AP-sites are chemically unstable and lack base coding potential. In this case the binding of the glycosylase might serve to protect cells against their cytotoxic and mutagenic effects, so that the release of the AP-site is coordinated with the recruitment and assembly of the downstream acting BER factors. Indeed, experiments with purified human proteins showed that APE1 is able to stimulate the turnover of TDG from a G:T substrate (Waters and Swann 1998). However, the fact that any other AP-site interacting protein tested has a similar stimulatory impact on TDG turnover implicates that these are passive rather than active and specific effects.

In light of all these findings, we can conclude that two main models, for a role of TDG in active DNA demethylation, are likely to occur, both involving an initial modification of 5mC to another moiety followed by TDG glycosylase activity. First, TDG could functionally cooperate with a DNA deaminase, propably AID or APOBEC, which generates G-T, G-U, or G-5hmU mispairs from G-5mC, G-C, and G-5hmC respectively. TDG can then excise each of these mispairs to be repaired via BER mechanisim (Cortazar et al. 2007). In support of this model, Cortellino and colleagues showed that TDG forms a complex with AID and Gadd45a, a BER protein, and this complex is implicated in DNA demethylation (Cortellino et al. 2011). Additionally, global erasure of DNA methylation during reprogramming of mouse primordial germ cells requires AID (Popp et al. 2010). Moreover, at the loci specific level, Metivier et al showed that TDG plays an important role in the cyclical DNA methylation of the transcriptionally active estrogen responsive *pS2/TFF1* gene. This process takes place when TDG is recruited to the promoter along with DNMT3A and B, as well as the BER proteins APE, DNA ligase, and DNA polymerase β (Metivier et al. 2008).

Second, TET proteins oxidize 5mC to 5hmC and further to 5fC and 5caC. Both 5fC

and 5caC are TDG substrates, and subsequently, BER generates unmodified cytosine. In support of this second model, He and colleagues (2011) found that TET-2 catalyzes the formation of 5caC in human cells and that modified base accumulates in the absence of TDG (He et al. 2011). Indeed, among all DNA glycosylases, only TDG can efficiently excise 5fC/5caC (Maiti and Drohat 2011). Importantly, unlike other DNA glycosylases, TDG is indispensable for embryonic development, since TDG knockout in mice led to embryonic lethality at E12.5 (Cortazar et al. 2011, Cortellino et al. 2011). Of note, embryos carrying mutation in the glycosylase domain (N151A) led to the same developmental defects observed in *Tdg*-null embryos, demonstrating an important role of the glycosylase activity of TDG (Cortellino et al. 2011). These findings suggest that TDG plays an important unique role of in embryonic development.

The ability of TDG to cooperate with and excise both 5mC deamination and oxidation products give rise to the needs for further efforts to understand the substrate selectivity and affinity of TDG towards its substrates. It is possible that, under certain circumstances, or by the control of other partner protein(s), TDG is guided to a substrate, where its glycosylase activities is needed.

1.10 Aims of this study

DNA methylation has long been viewed as a stable epigenetic mark, but it has become apparent that in some circumstances DNA methylation pattern can rapidly change by mechanisms involving active DNA demethylation processes. Two pathways have been proposed to manage DNA demethylation. In a first hypothesis, a cytosine deaminase catalyzes the deamination of a 5-methylcytosine (5mC) into a thymine, leading to the formation of a G/T mismatch. In a second mechanism, 5mC undergoes several steps of oxidation leading to the formation of 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). In both cases, the DNA demethylation process is accomplished by DNA glycosylases, which cleave the N-glycosidic bond between the base and the sugar, leading to the formation of an abasic site, which is a signal to initiate base excision repair (BER). In mammalian cells, G/T mismatch can be specifically recognized by the glycosylases MBD4 and TDG. TDG has also the ability to bind and excise 5fC and 5caC. Each of these glycosylases (TDG and MBD4) shows specific biochemical features suggesting specific function in DNA methylation dynamics. TDG shows a wide substrate diversity by acting on the 5mC deamination and oxidation products, whereas MBD4 has an unique architecture having a methyl-CpG-binding domain associated with the glycosylase domain, which could target this protein only to highly methylated region.

In order to clarify the function of MBD4 and TDG in DNA methylation dynamics, it is essential to identify proteins interacting specifically with MBD4 and TDG *in vivo*. Toward this goal, we aimed to purify the MBD4 and TDG complexes from HeLa cell lines and from mouse embryonic fibroblasts, as well. This would allow us to identify several functional partners specifically interacting with MBD4 and TDG *in vitro* and *in vivo*. Biochemical studies of MBD4 and TDG associated or not to their functional partners, will allow us to highlight the specific enzymatic activities and to suggest specific function of MBD4 and TDG in DNA methylation dynamics *in vivo*. The putative roles will be validated by analyzing the impact of down-regulating of TDG, MBD4 or the different identified partners on 5mC, 5hmC, 5fC and 5caC genomic patterns. We sought to carry out these studies at genome-wide using high-throughput sequencing techniques (DIP-seq, RRBS, RNA-seq and ChIP-seq).

CHAPTER 2
RESULTS

The methyl-directed nuclease activity of MBD4-MLH1 complex is required to protect silenced promoters from demethylation.

Christophe Papin[1], Abdulkhaleg Ibrahim[1], Khalid Ouararhni[1], Arnaud Obri[1], Christian Bronner[1], Mikhail Grigoriev[2], Alfonso Bellacosa[3], Stefan Dimitrov[4] and Ali Hamiche[1,§]

[1]Institut de Génétique et Biologie Moléculaire et Cellulaire (IGBMC), UdS, CNRS, INSERM, Equipe labélisée Ligue contre le Cancer, 1 rue Laurent Fries, B.P. 10142, 67404 Illkirch Cedex, France.[2] LBME, UMR 5099, CNRS, F-31000 Toulouse; France [3]Cancer Biology Program, Cancer Epigenetics Program, Fox Chase Cancer Center, Philadelphia PA 19111. [4]INSERM/UJF, Institut Albert Bonniot, U823, Site Santé-BP 170, 38042 Grenoble Cedex 9, France.

[§] To whom correspondence should be addressed. e-mail: hamiche@igbmc.fr

Runing title: MBD4 is required for DNA methylation maintenance

Abstract

Methylation of CpG islands is a hallmark of silenced gene promoters. How is the promoter methylation preserved, remains, however, elusive. Here we show that the methyl binding domain protein MBD4 is required for the preservation of promoter methylation. MBD4 forms *in vivo* a complex with the mismatch repair proteins (MMR), which exhibits high bi-functional glycosylase/AP-lyase endonuclease specific activity towards methylated DNA substrates containing a G/T mismatch. Experiments using recombinant proteins reveal that the association of MBD4 with the MMR protein MLH1 is required for this activity. Genome-wide transcriptome and methylome analyses *in vivo* shows that the absence of MBD4 results in alterations in both promoter methylation and transcriptional activity of a large number of genes. The described data identify MBD4 as an enzyme specifically designed to repair deaminated 5-methylcytosines in methylated-CpG islands and allows to understand how MBD4 functions in normal and pathological conditions.

Significance

MBD4, is a methyl-CpG binding protein, whose exact function is unknown. Here, we show that MBD4 is associated *in vivo* with several proteins, including the MMR proteins MLH1 and PMS2. The MBD4 complex exhibits both bifunctional glycosylase/AP-lyase activity and marked preference for methylated DNA. MBD4 seems to be specifically designed to repair the product of cytosine deamination (G/T mismatches) in a methylated-CGIs context. Genome-wide transcriptome and methylome analyses reveal that the absence of MBD4 correlates with both the derepression of a high number of MBD4-target genes and demethylation of their

promoters. These data, taken as a whole, illustrate that MBD4 is required for both preserving the methylation status of its target genes and maintaining them in a repressive state.

\body

Introduction

DNA methylation is an essential epigenetic modification in vertebrates and occurs predominantly within the CpG dinucleotides. 5-methylcytosine (5mC) is highly sensitive to spontaneous deamination which leads to the formation of thymine and thus to point mutations (1, 2). The genome of vertebrates are consequently largely CpG-deficient. The globally methylated and CpG-poor genomic landscape is, however, punctuated by CpG-rich regions referred as CpG islands (CGIs) (3, 4). Most of CGIs colocalize with promoters (5) and are protected from methylation, creating a transcriptionally permissive chromatin (6). However, there are well-known examples of CGIs that become methylated during development, in a tissue-specific manner, leading to stable silencing of the associated promoters (7-9).

The CGIs methylation inhibitory effect on transcription is mediated by two main mechanisms. First, the methylation of CGIs can directly affect both the recognition and the binding of transcription factors to promoter DNA. The second mechanism is indirect and involves the recruitment of transcription repressive complexes to methylated promoters through the specifically bound methyl-CpG binding proteins (MBPs) to methylated CGIs (10, 11).

Three different families of MBPs are identified: the MBD (Methyl Binding Domain) family, the zing finger family and the SRA family (12, 13). The structure of several members (alone or in complex with methylated DNA) of these families were solved by either solution NMR spectroscopy or by X-ray crystallography (14-20). The available data show that the MBPs recognize and bind to methylated DNA in a very specific manner. Intriguingly, the discrimination between methylated and non-methylated DNA is achieved via distinct MBPs protein folds (21).

The MBD family is composed of seven members, and four of them (MeCP2, MBD1, MBD2 and MBD4) are shown to preferentially bind to methylated DNA through their conserved MBD (22). Among both the MBD family and the other two families of MBPs, MBD4 is the only protein, which exhibits enzymatic activity. Indeed, MBD4, in addition to its methyl-binding domain, has a glycosylase domain and possesses thymine and uracil glycosylase activity (23-25). Note that MBD4 was cloned in the past by a two-hybrid approach using MLH1, a mismatch repair (MMR) protein, as a “bait” (23). The DNA mismatch repair system depends, in addition to MLH1, on several other factors, including the proteins MSH2, PMS2 and MSH6 (26). Reduced levels of MLH1, PMS2, MSH2 and MSH6 were detected in MBD4-deficient cells, suggesting that MBD4 might be involved in both the integrity and stability of the MMR complex (27).

The available data suggest that MBD4 is implicated in Base Excision Repair (BER) of G/T mismatches resulting from deamination of 5mC at CpG dinucleotide. This should allow to avoid mutations and maintain genome stability. In agreement with this, *MBD4*$^{-/-}$ mice exhibit a marked increase in C to T mutations at CpG sites and higher occurrence of these mutations has been demonstrated in colon tumors in crosses of *MBD4*$^{-/-}$ mice with *Apc*^{Min} mice (28, 29). In addition, between 26 and 43%

of human gastric, colorectal, endometrial and pancreatic tumors exhibiting microsatellite instability have also mutations in MBD4 (30-33).

Here, we have studied the role of MBD4 *in vivo* and have deciphered its function in a series of *in vitro* experiments. We show that MBD4 is associated *in vivo* with several proteins, including the MMR proteins MLH1 and PMS2. The MBD4 complex exhibits both bifunctional glycosylase/AP (apurinic or apyrimidinic site) - 3'-phosphomonoester lyase activity and marked preference for methylated DNA. Genome-wide transcriptome and methylome analyses reveal that the absence of MBD4 correlates with both the derepression of a high number of MBD4-target genes and demethylation of their promoters. Our data identify MBD4 as a key factor responsible for the preservation of both methylation patterns of promoters and their transcriptional repressive states. Therefore, MBD4 seems to be specifically designed to repair the product of cytosine deamination (G/T mismatches) in a methylated-CGIs context.

Results

MBD4 is associated *in vivo* with core MMR proteins

To gain insight into the role that MBD4 may play *in vivo*, we sought to study its enzymatic properties. The properties of purified recombinant MBD4 have previously been analyzed and the reported data, suggests that the recombinant protein exhibits G/T and G/U mismatch specific monofunctional glycosylase activity (23, 24, 34, 35). The native MBD4 complex could, however, have features distinct from those of the MBD4 protein alone. To test this, we purified the epitope-tagged MBD4 complex (e-MBD4.com) from HeLa cells stably expressing hemagglutinin (HA) and FLAG epitope

tagged MBD4 (Figure 1a). MBD4 together with the DNA helicases TIP49A/B and the MMR proteins (MLH1 and PMS2) were identified as major components of e-MBD4.com by both mass spectrometry (Figure 1b) and Western blotting (Figure 1a, lower panel). Fractionation of the e-MBD4 complex on a glycerol gradient confirmed that MLH1 and PMS2 are stable components of this complex (Figure 1c). This interaction was further validated by immunoprecipitating either the endogenous MBD4 (Figure 1d) or MLH1 (Figure 1e) complexes from non-tagged HeLa cell nuclear extracts using specific antibodies. We next analyzed the composition of the MBD4 complex in mouse embryonic fibroblast (MEF) cell lines stably expressing e-MBD4. Western blotting demonstrates that the MEF e-MBD4 complex exhibits the same composition as the e-MBD4 complex isolated from HeLa cells and thus, it should mechanistically function in the same way (Figure 1f). We conclude that MBD4 forms *in vivo* a complex with the PMS2/MLH1 heterodimer.

The MBD4 complex shows methyl-directed G/T mismatch specific endonuclease activity

Some common glycosylases are known to exhibit an endonuclease mismatch activity (36, 37). To test if this is the case for the e-MBD4.com, we carried out nuclease assays on substrate DNA containing different types of mismatches (Supplemental Figures 1a,b). The data clearly show that: (i) the e-MBD4 complex is able to cleave the G/T (or G/U) mismatch, but not the cytosine, the methyl-, the hydroxymethyl-, the formyl- or the carboxyl-cytosine substrates (Supplemental Figure 1c), and (ii) the cleavage is achieved at the abasic site on the “T”-containing strand of a G/T mismatch substrate (Supplemental Figure 1a). Importantly, no NaOH treatment of the

reaction products was needed for generation of the cleavage products (Supplemental Figure 1a). Therefore, the e-MBD4.com exhibits G/T mismatch specific endonuclease activity.

To analyze whether the e-MBD4.com endonuclease activity was dependent on the methylation status of the substrate, we carried out similar experiments, but with fully methylated (on both strands) G/T mismatch containing substrates by using identical amounts of either highly purified MBD4 alone (Figure 2a) or MBD4 in the context of the e-MBD4.com. The purified MBD4 protein was able to induce some weak, non-methylation dependent cleavage of the substrate (Figure 2b, upper panel and Figure 2c). The e-MBD4.com shows ~ 3-fold higher activity for unmethylated DNA relative to that of the MBD4 protein (Figures 2b,c). The e-MBD4.com endonuclease activity was, however, strongly methylation dependent and ~ 8-10 fold higher e-MBD4.com induced cleavage (compared to this for the MBD4 protein) for fully methylated substrates was measured (Figures 2b,c). These data indicate that the association of MBD4 with its partners modulates its enzymatic properties and, as a result, MBD4 acquires a much higher G/T specific endonuclease activity, which is dependent on the methylation of the DNA substrate.

We next addressed the role of the methyl binding domain of MBD4 by generating stable HeLa cell line expressing R97G mutated e-MBD4 (the substitution of R97 with G results in a dead methyl binding domain (14)). Both protein gel (Figure 2d) and Western blotting (Figure 2e) showed that the composition of the purified methyl binding dead e-MBD4 complex (e-MBD4.com R97G) is identical to the native e-MBD4.com. Nuclease assays revealed that the R97G mutant e-MBD4.com does not discriminate between unmethylated, hemi-methylated and fully methylated mismatch-containing substrates. In all cases, cleavage in absence of carrier DNA is

very low (~10%) and, with increasing concentrations of carrier DNA, this activity progressively decreases (Figures 2f,g, right panels). In contrast, the native complex (e-MBD4.com WT) discriminates clearly between methylated and unmethylated substrates and its cleavage efficiency, compared to the mutant e-MBD4.com R97G, is much higher at the respective carrier DNA concentrations (Figures 2f,g, left panels). In particular, dramatic differences in cleavage efficiencies for both complexes are measured for the fully methylated substrates (Figures 2f,g). These results demonstrate that the e-MBD4.com endonuclease activity strongly depends on the degree of methylation of the DNA substrate and that the methyl-binding domain of MBD4 targets the MBD4-MMR complex on methylated DNA.

MBD4 is an endonuclease with a bifunctional glycosylase/AP lyase activity

We next asked whether the MBD4 glycosylase domain is important for the observed G/T specific endonuclease activity. To this end we substituted amino acid residue D554 with alanine (A) and created a glycosylase dead mutant e-MBD4 protein (38). Then, we established stable HeLa cell lines expressing the e-MBD4 D554A mutant. Both protein gel analysis and Western blotting show that the glycosylase dead mutant e-MBD4 D554A complex purified from the stable HeLa cell lines has a protein composition identical to the WT e-MBD4 complex (Figures 3a,b). The mutant e-MBD4.com D554A exhibits, however, no endonuclease activity (Figure 3c). We conclude that the activity of the MBD4 glycosylase domain is required for its endonuclease activity.

Past studies have demonstrated that MBD4 possesses G/T specific glycosylase activity (23, 24, 35). Several glycosylases show, however, also AP lyase

activity, i.e., the excision of the base is followed by AP lyase-mediated cleavage of the phosphate backbone at the abasic site (36, 37). If the AP lyase product is generated through β elimination, it retains the abasic residue at its 3' end (Figure 3e and (39, 40)). This product migrates slower on sequencing PAGE than the respective product obtained from the Gilbert&Maxam sequencing reaction. Upon treatment of such AP-lyase generated products with NaOH, the phospho-diester bond is cleaved (δ elimination), the abasic residue is released and the mobility of the resulting product becomes identical to that generated by the Maxam&Gilbert sequencing reaction.

Our data suggest that MBD4 possesses an AP-lyase type endonuclease activity. To test this, we have used the above described procedure (Figures 3d,e). Treatment with NaOH of either MBD4 or e-MBD4.com reaction products resulted in clear increase of their migration rate in PAGE under denaturing conditions. The migration position of the NaOH treated products was identical to this of the products of the Maxam&Gilbert sequencing reaction obtained upon cleavage at the respective thymine (Figure 3d). These observations reveal that MBD4 possesses an AP-lyase endonuclease activity, which operates through β elimination (Figure 3e).

The methyl-directed nuclease activity of MBD4 is dependent on its physical interaction with MLH1.

Having demonstrated the AP-lyase endonuclease activity of MBD4, we next sought to analyze the role of the MMR proteins in the regulation of this newly identified activity. We have expressed either separately or co-expressed together the recombinant MBD4, MLH1 and PMS2 proteins in the baculovirus system. We were able to reconstitute the MBD4/MLH1 and the MBD4/MLH1/PMS2 complexes but not

the MBD4/PMS2 complex, suggesting a direct physical interaction between MBD4 and MLH1 (Figure 4a). Reconstitution of the *in vitro* nuclease assay using individual recombinant proteins shows that MBD4, but not the MMR proteins, has some weak endonuclease activity (Figure 4b), a result in agreement with the data presented in Figure 2b and excluding a potential enzymatic function of MMR proteins in e-MBD4.com endonuclease activity. Remarkably, the presence of MLH1 results in dramatic increase of the endonuclease activity of MBD4 with marked preference for fully methylated substrates (Figure 4b,c). Note that the enzymatic dead MBD4 D554A mutant protein alone or complexed with MLH1 did not exhibit endonuclease activity (Figure 4c). Similarly, none of the purified recombinant proteins exhibited endonuclease activity towards DNA substrate containing abasic sites (cleavable by the recombinant apurinic/apyrimidinic endonuclease APE1), ruling out the presence of a non-specifically associated contaminating AP-lyase activity (Figure 4b, right panel and Figure 4c, right panel). All together, these data reveal that the recombinant MLH1 alone and not some biochemically undetectable contaminant is responsible for stimulating the MBD4 endonuclease activity. The nuclease time-course assay (Figures 4d,e) shows that the purified MBD4/MLH1 complex cuts much more efficiently the G/T mismatch substrate compared to MBD4 alone. The most drastic differences in the cleavage efficiency are measured for fully methylated substrates, where the cleavage kinetics (as assessed by the initial slope of the nuclease assay time course curve) of the MBD4/MLH1 is at least 10 fold higher than for MBD4 alone (Figures 4d,e). Therefore, by using recombinant components we have been able to reconstitute *in vitro* the enzymatic properties of the isolated native MBD4 complex and demonstrate the crucial role that MLH1 plays in the MBD4 complex endonuclease activity.

Genome-wide hypomethylation of CGIs in the absence of MBD4

Having demonstrated that the MBD4 complex has DNA methylation dependent glycosylase and AP/lyase endonuclease activity stimulated by MLH1, we next asked how the cell uses these specific properties of the MBD4 complex. To this end we used primary MEFs WT (*MBD4*$^{+/+}$) and KO (*MBD4*$^{-/-}$) for *Mbd4* (Figure 5). We hypothesized that the absence of MBD4 would generate alterations in gene promoter methylation patterns, which would in turn affect their transcriptional status. The DNA methylation was analyzed by using reduced representation bisulfite sequencing (RRBS) technic (41). RRBS provides single-nucleotide resolution and quantitative DNA methylation measurements for the majority of CGI-containing promoters and other relevant genomic regions (42). Our RRBS data covered more than 1.1 millions (51%) of CpGs within CGIs in both WT and KO cells (Figure 5a and Supplemental Figure 2a), which corresponds to a coverage depth of ~75% for the majority of CGIs (Figure 5b). This analysis identified 45,784 and 43,923 5mCGs within CGIs in WT and KO cells respectively (Figure 5c and Supplemental Tables 1,2), which correspond to 4% of the corresponding CGs. While the majority of these 5mCGs were 30–40% methylated in WT cells, the methylation level of a large part of them decreased to less than 10% in KO cells (Figure 5d). We reasonably concluded that MBD4 protects methylated CGs from demethylation. Interestingly, methylation loss was preferentially observed in CGIs exhibiting both higher CG density (Figure 5e and supplemental Figure 2b) and low methylation level (Figure 5f and supplemental Figure 2c).

We next investigated how this alteration in the DNA methylation affects gene expression. Genome-wide transcriptome analysis of *MBD4*$^{-/-}$ cells identifies in total 142 genes (with $P < 0.01$) having strong transcriptional de-regulation compared to the control *MBD4*$^{+/+}$ cells (76 of these genes were found up-regulated and 66 were found down-regulated, Supplemental Figures 2d,e and Supplemental Table 4). We hypothesized that the hypomethylated CGI promoters are up-regulated. To test this we selected genes containing the most significantly demethylated proximal region (methylation level KO/WT > 2, number of CG > 10, $P < 0.01$, distance to nearest TSS < 5 kb ; Supplemental Table 3) and correlated their methylation level (Figure 5g) to their transcriptional states (Figure 5h). Accordingly, the vast majority of the hypomethylated CGI promoters were transcriptionnally up-regulated in absence of MBD4 (Figure 5h). Clonal bisulfite sequencing further confirmed the hypomethylation phenotype at proximal regions of Zic5, Tox and Lrtm2 genes (Figure 5i). These data, taken as a whole, illustrate that MBD4 is required for both preserving the methylation status of its target genes and maintaining them in a repressive state (Figure 5j).

Discussion

The described *in vitro* data demonstrate that MBD4 is an unusual glycosylase having two domains essential for its functions. In addition to its glycosylase activity, the MBD4 catalytic domain exhibits an AP lyase activity. These two activities are required for both the removal of the thymine base and cleavage of the DNA phosphate backbone. In cells, MBD4 forms a complex with the MMR proteins MLH1 and PMS2 as well as with other proteins. This MBD4-MMR protein complex possesses a much higher cleavage efficiency than MBD4 alone. Experiments with

highly purified recombinant proteins show the MMR protein MLH1 is required for this effect. Similar “boosting” function for MLH1 has already been observed for EXO1 and PMS1 proteins, two nucleases implicated in MMR pathway in eukaryotes. Indeed, the physical interaction between MLH1 and EXO1 is required for the endonuclease function of EXO1 in MMR pathway (43). A recent structural study has also revealed that the highly conserved C terminus of MLH1 forms part of the PMS1 endonuclease site (44). All together, these data define MLH1 as a nuclease effector protein. In addition, the MBD4 complex has a clear preference for methylated G/T mismatch containing substrates, which is determined by its methyl-binding domain (our data). Therefore, MBD4 appears to be specifically designed to repair G/T mismatches in the vicinity of methylated CpGs.

The absence of MBD4 in primary MEFs correlates with a marked methylation loss affecting preferentially CGIs exhibiting high CG density and low methylation level. Genome-wide transcriptome analysis revealed that the absence of MBD4 also correlates with the marked increase of transcription of 76 genes with a good correlation between methylation level and transcriptional states. Three genes strongly up-regulated in *MBD4*$^{-/-}$ cells, namely Zic5, Tox and Lrtm2, were further analyzed by clonal bisulfite sequencing and confirmed to be hypomethylated at their proximal regions. Therefore, MBD4 is directly involved in the preservation of the methylation status of the promoter of these genes through direct binding to the methylated CGIs. This is in contrast to the genes associated with other MBD proteins (MBD1, MBD2 or MeCP2), where the siRNA depletion of these proteins resulted only in derepression of the respective genes and not in demethylation of their promoters (45-50). This makes the function of MBD4 unique within the MBD class of proteins

and stresses the role of its glycosylase domain in preserving the methylation level of the target gene promoters.

We propose the following simplistic model for the function of MBD4 (Figure 5j). MBD4 is bound through its MBD to the methylated CGI-containing promoters of its target genes. In this way, MBD4, either by steric hindrance or/and by recruiting repressive complexes, keeps the promoter silenced. As a result of spontaneous deamination, the 5mC is mutated to T and thus, a G/T mismatch is formed. Since MBD4 is present at high concentration at the methylated promoter, it easily excises the T via its glycosylase/AP lyase activity and the BER machinery further repairs the "gap". Subsequent methylation of the repaired CpG allows the binding of another MBD4 molecule to the methylated dinucleotide though its MBD, and thus, both the methylated and the repressive states of the promoter are preserved. If MBD4 is absent, the T-G promoter mismatch cannot be repaired efficiently and the methylation of the promoter is lost as observed in *MBD4*$^{-/-}$ cells. In addition, in the absence of MBD4, C to T transitions at CpG sites will be generated which would lead to genome instability, as determined in *MBD4*$^{-/-}$ mice (28, 29).

Acknowledgments

We thank Irwin Davidson for critical reading of the manuscript. This work was supported by institutional funds from CNRS, INSERM, Université de Strasbourg (UDS), Université de Grenoble Alpes and by grants from, INCA (INCa_4496), INCA (INCa_4454), ANR (VariZome, contract n° ANR-12-BSV8-0018-01; Nucleoplat, contract n° NT09_476241), the Association pour la Recherche sur le Cancer, La Fondation pour la Recherche Medicale, La Ligue Nationale contre le Cancer Equipe labellisée (A.H. and S.D.), the European Community's Grant agreement number 289611 (''HEM_ID'') to S.D. C.P. acknowledges the Fondation pour la Recherche Médicale for financial support (A.H.). A.O. acknowledges the Association pour la

Recherche sur le Cancer for Financial support.

References

1. Coulondre C, Miller JH, Farabaugh PJ, & Gilbert W (1978) Molecular basis of base substitution hotspots in Escherichia coli. *Nature* 274(5673):775-780.

2. Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res* 8(7):1499-1504.

3. Bird A, Taggart M, Frommer M, Miller OJ, & Macleod D (1985) A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* 40(1):91-99.

4. Cooper DN, Taggart MH, & Bird AP (1983) Unmethylated domains in vertebrate DNA. *Nucleic Acids Res* 11(3):647-658.

5. Saxonov S, Berg P, & Brutlag DL (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* 103(5):1412-1417.

6. Ramirez-Carrozzi VR*, et al.* (2009) A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. *Cell* 138(1):114-128.

7. Stein R, Razin A, & Cedar H (1982) In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. *Proc Natl Acad Sci U S A* 79(11):3418-3422.

8. Mohn F & Schubeler D (2009) Genetics and epigenetics: stability and plasticity during cellular differentiation. *Trends Genet* 25(3):129-136.

9. Payer B & Lee JT (2008) X chromosome dosage compensation: how mammals keep the balance. *Annu Rev Genet* 42:733-772.

10. Klose RJ & Bird AP (2006) Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci* 31(2):89-97.

11. Bogdanovic O & Veenstra GJ (2009) DNA methylation and methyl-CpG binding proteins: developmental requirements and function. *Chromosoma* 118(5):549-565.

12. Parry L & Clarke AR (2011) The Roles of the Methyl-CpG Binding Proteins in Cancer. *Genes Cancer* 2(6):618-630.

13. Defossez PA & Stancheva I (2011) Biological functions of methyl-CpG-binding proteins. *Prog Mol Biol Transl Sci* 101:377-398.

14. Ohki I*, et al.* (2001) Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. *Cell* 105(4):487-497.

15. Scarsdale JN, Webb HD, Ginder GD, & Williams DC, Jr. (2011) Solution structure and dynamic analysis of chicken MBD2 methyl binding domain bound to a target-methylated DNA sequence. *Nucleic Acids Res* 39(15):6741-6752.

16. Ho KL*, et al.* (2008) MeCP2 binding to DNA depends upon hydration at methyl-CpG. *Mol Cell* 29(4):525-531.

17. Arita K, Ariyoshi M, Tochio H, Nakamura Y, & Shirakawa M (2008) Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature* 455(7214):818-821.

18. Hashimoto H*, et al.* (2008) The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. *Nature* 455(7214):826-829.

19. Liu Y, Toh H, Sasaki H, Zhang X, & Cheng X (2012) An atomic model of Zfp57 recognition of CpG methylation within a specific DNA sequence. *Genes Dev* 26(21):2374-2379.

20. Avvakumov GV*, et al.* (2008) Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. *Nature* 455(7214):822-825.

21. Buck-Koehntop BA & Defossez PA (2013) On how mammalian transcription factors recognize methylated DNA. *Epigenetics* 8(2):131-137.

22. Hendrich B & Bird A (1998) Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Mol Cell Biol* 18(11):6538-6547.

23. Bellacosa A*, et al.* (1999) MED1, a novel human methyl-CpG-binding endonuclease, interacts with DNA mismatch repair protein MLH1. (Translated from eng) *Proc Natl Acad Sci U S A* 96(7):3969-3974 (in eng).

24. Hendrich B, Hardeland U, Ng HH, Jiricny J, & Bird A (1999) The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. (Translated from eng) *Nature* 401(6750):301-304 (in eng).

25. Turner DP*, et al.* (2006) The DNA N-glycosylase MED1 exhibits preference for halogenated pyrimidines and is involved in the cytotoxicity of 5-iododeoxyuridine. (Translated from eng) *Cancer research* 66(15):7686-7693 (in eng).

26. Kunz C, Saito Y, & Schar P (2009) DNA Repair in mammalian cells: Mismatched repair: variations on a theme. (Translated from eng) *Cell Mol Life Sci* 66(6):1021-1038 (in eng).

27. Cortellino S*, et al.* (2003) The base excision repair enzyme MED1 mediates DNA damage response to antitumor drugs and is associated with mismatch repair system integrity. *Proc Natl Acad Sci U S A* 100(25):15071-15076.

28. Millar CB*, et al.* (2002) Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. *Science* 297(5580):403-405.

29. Wong E*, et al.* (2002) Mbd4 inactivation increases Cright-arrowT transition mutations and promotes gastrointestinal tumor formation. *Proc Natl Acad Sci U S A* 99(23):14937-14942.

30. Riccio A*, et al.* (1999) The DNA repair gene MBD4 (MED1) is mutated in human carcinomas with microsatellite instability. (Translated from eng) *Nat Genet* 23(3):266-268 (in eng).

31. Miquel C*, et al.* (2007) Frequent alteration of DNA damage signalling and repair pathways in human colorectal cancers with microsatellite instability. *Oncogene* 26(40):5919-5926.

32. Bader S*, et al.* (1999) Somatic frameshift mutations in the MBD4 gene of sporadic colon cancers with mismatch repair deficiency. *Oncogene* 18(56):8044-8047.

33. Yamada T*, et al.* (2002) Frameshift mutations in the MBD4/MED1 gene in primary gastric cancer with high-frequency microsatellite instability. *Cancer Lett* 181(1):115-120.

34. Kim MS*, et al.* (2009) DNA demethylation in hormone-induced transcriptional derepression. (Translated from eng) *Nature* 461(7266):1007-1012 (in eng).

35. Petronzelli F*, et al.* (2000) Biphasic kinetics of the human DNA repair protein MED1 (MBD4), a mismatch-specific DNA N-glycosylase. *J Biol Chem* 275(42):32422-32429.

36. van der Kemp PA, Charbonnier JB, Audebert M, & Boiteux S (2004) Catalytic and DNA-binding properties of the human Ogg1 DNA N-glycosylase/AP lyase: biochemical exploration of H270, Q315 and F319, three amino acids of the 8-oxoguanine-binding pocket. (Translated from eng) *Nucleic Acids Res* 32(2):570-578 (in eng).

37. van der Kemp PA, Thomas D, Barbey R, de Oliveira R, & Boiteux S (1996) Cloning and expression in Escherichia coli of the OGG1 gene of Saccharomyces cerevisiae, which codes for a DNA glycosylase that excises 7,8-dihydro-8-oxoguanine and 2,6-diamino-4-hydroxy-5-N-methylformamidopyrimidine. (Translated from eng) *Proc Natl Acad Sci U S A* 93(11):5197-5202 (in eng).

38. Wu P*, et al.* (2003) Mismatch repair in methylated DNA. Structure and activity of the mismatch-specific thymine glycosylase domain of methyl-CpG-binding protein MBD4. (Translated from eng) *The Journal of biological chemistry* 278(7):5285-5291 (in eng).

39. Gehring M, Reik W, & Henikoff S (2009) DNA demethylation by DNA repair. (Translated from eng) *Trends in genetics : TIG* 25(2):82-90 (in eng).

40. McCullough AK, Dodson ML, & Lloyd RS (1999) Initiation of base excision repair: glycosylase mechanisms and structures. *Annu Rev Biochem* 68:255-285.

41. Gu H*, et al.* (2011) Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. (Translated from eng) *Nat Protoc* 6(4):468-481 (in eng).

42. Meissner A*, et al.* (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. (Translated from eng) *Nucleic Acids Res* 33(18):5868-5877 (in eng).

43. Dherin C*, et al.* (2009) Characterization of a highly conserved binding site of Mlh1 required for exonuclease I-dependent mismatch repair. *Mol Cell Biol* 29(3):907-918.

44. Gueneau E*, et al.* (2013) Structure of the MutLalpha C-terminal domain reveals how Mlh1 contributes to Pms1 endonuclease site. *Nat Struct Mol Biol* 20(4):461-468.

45. Jones PL*, et al.* (1998) Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat Genet* 19(2):187-191.

46. Nan X*, et al.* (1998) Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* 393(6683):386-389.

47. Fujita N*, et al.* (2000) Mechanism of transcriptional regulation by methyl-CpG binding protein MBD1. *Mol Cell Biol* 20(14):5107-5118.

48. Sarraf SA & Stancheva I (2004) Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly. *Mol Cell* 15(4):595-605.

49. Boeke J, Ammerpohl O, Kegel S, Moehren U, & Renkawitz R (2000) The minimal repression domain of MBD2b overlaps with the methyl-CpG-binding domain and binds directly to Sin3A. *J Biol Chem* 275(45):34963-34967.

50. Ng HH*, et al.* (1999) MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nat Genet* 23(1):58-61.

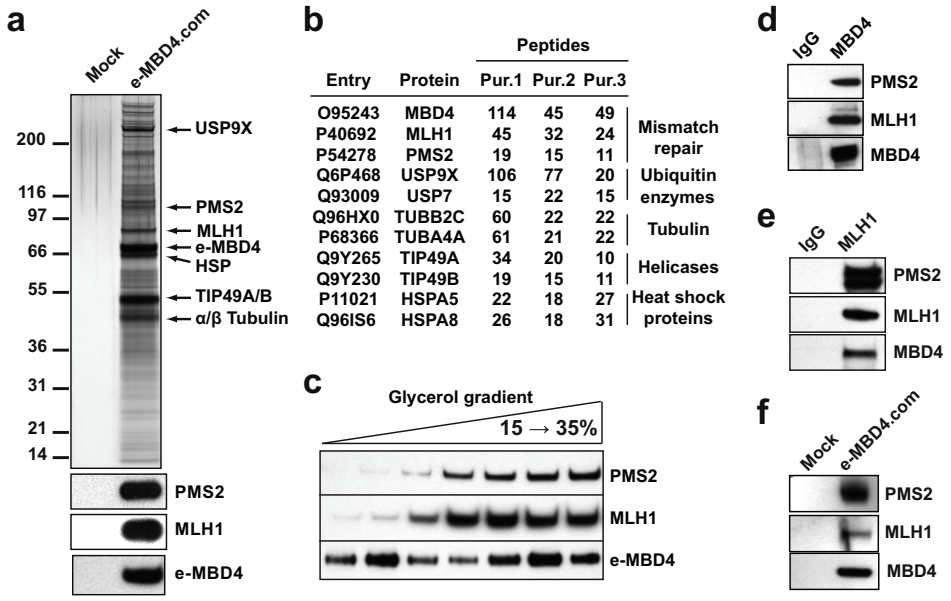


Figure 1. MBD4 interacts with mismatch repair proteins *in vivo*. (**a**) MBD4complex (e-MBD4.com) was purified by double immunoaffinity from HeLa cell linestably expressing MBD4 fused with N-terminal Flag- and HA-epitope tags (e-MBD4)and was run on a SDS PAGE. Silver staining of the SDS gel (top panel) andimmunoblotting detection (bottom panel) of the proteins associated with e-MBD4 areshown. (**b**) The major polypeptides detected by mass spectrometry analyses of threeindependent e-MBD4.com purifications. (**c**) Glycerol gradient fractionation of the e-MBD4 complex and western blot analysis of the different glycerol fractions separatedon SDS PAGE by using antibodies against the designated proteins. (**d**) EndogenousMBD4 specifically co-precipated with MLH1 and PMS2. (**e**) Endogenous MLH1 co-precipated with MBD4 and PMS2. (**f**) Western blot analysis of e-MBD4.com purifiedby double immunoaffinity from MEFs.

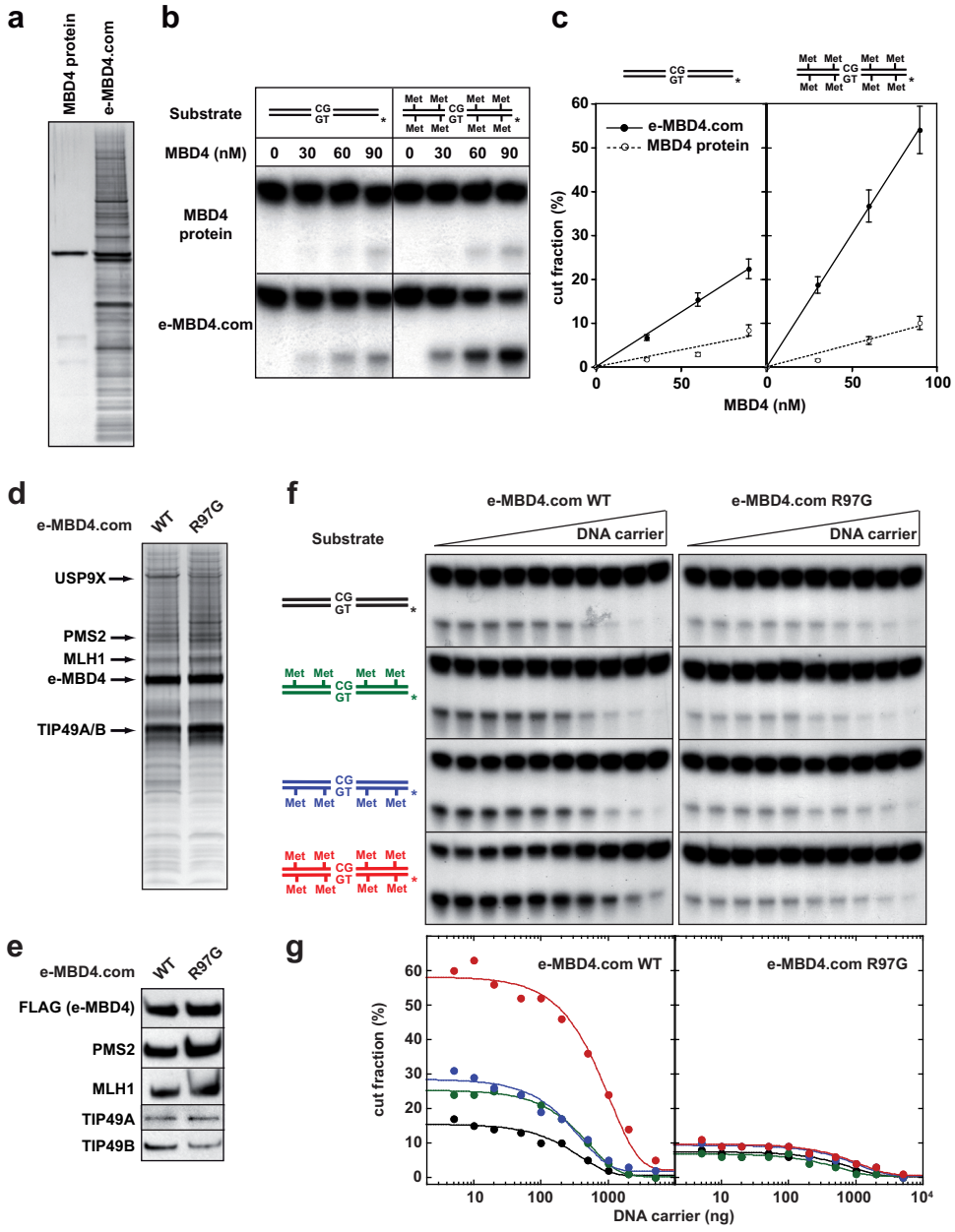


Figure 2. MBD4 is a methyl-directed mismatch endonuclease. (**a**) SDS PAGE silver staining of purified e-MBD4 and e-MBD4-com. (**b**) Unmethylated or full methylated G/T mismatch containing substrates were incubated with increasing amounts of recombinant protein MBD4 or with e-MBD4.com and analyzed as described Supplemental Figure 1a. The reaction products were not treated with NaOH. (**c**) Quantification of the data presented in (b). The means of three independent experiments are shown. **(d-g)** The methyl binding domain of MBD4 is required for its G/T mismatch and 5mCG-dependent endonuclease activity. R97G single point mutation was introduced in the coding sequence of e-MBD4 to generate a dead methyl binding MBD4 mutant (e-MBD4 R97G) and HeLa cell lines stablyexpressing e-

MBD4 R97G were generated. SDS PAGE gel silver staining (d) and immunobloting analysis (e) of e-MBD4.com (WT) and e-MBD4.com R97G complexes. (**f**) Nucleases assays for e-MBD4.com (WT) and e-MBD4.com R97G. Full-methylated (red), hemi-methylated (blue and green) or unmethylated (black) G/T mismatch containing substrates were incubated in the presence of either e- MBD4.com (WT) or e-MBD4.com R97G and increasing amounts of carrier DNA. The experiments were carried out as described Supplemental Figure 1a. The reaction products were not treated with NaOH. (**g**) Quantification of the results presented in (f). Red, green and blue, and black curves correspond to full-, hemi- and un-methylated G/T containing substrates, respectively.

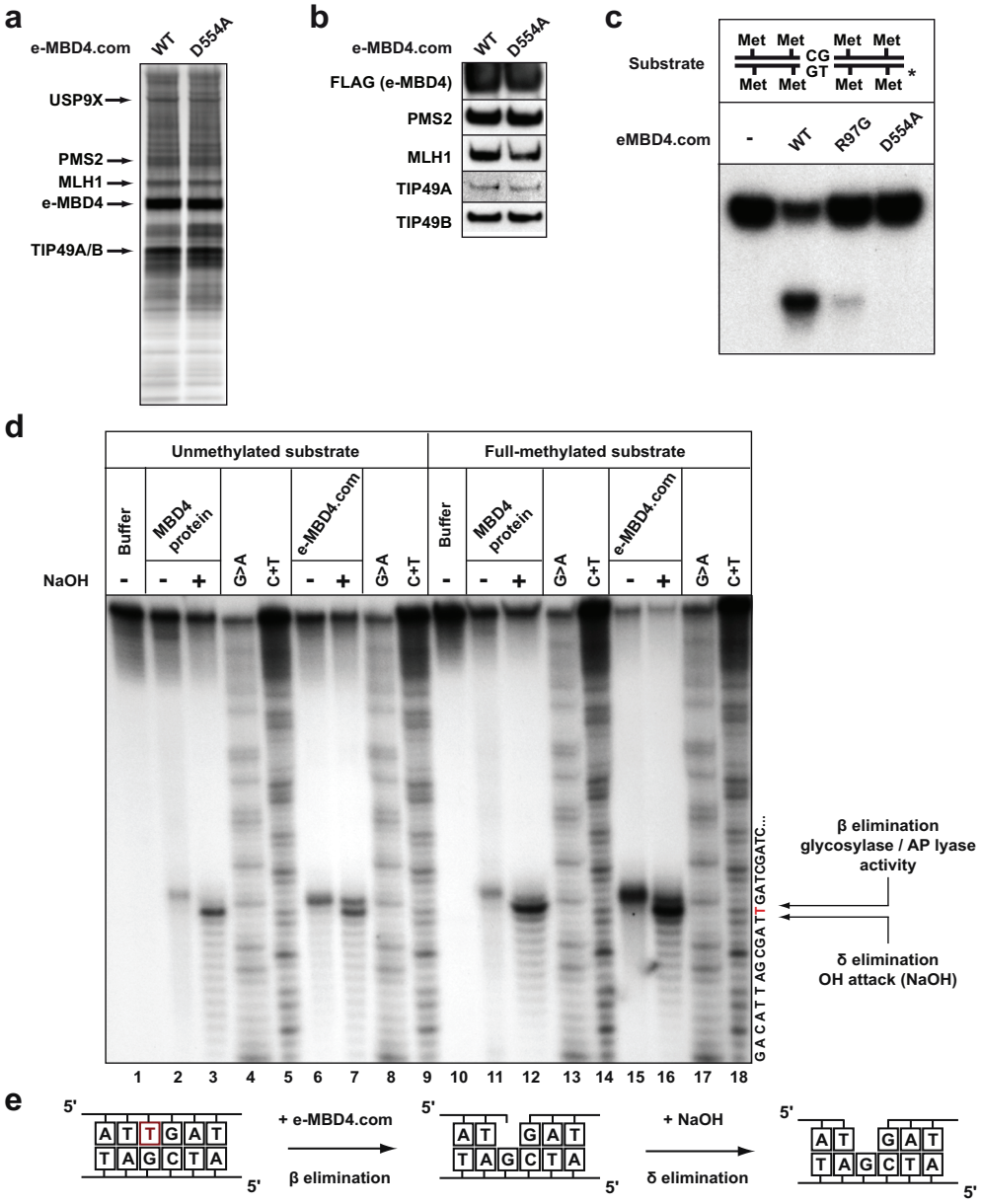


Figure 3. MBD4 is a bifunctional DNA glycosylase/AP lyase enzyme. (**a-c**) The glycosylase domain of MBD4 is required for its endonuclease activity. HeLa cell lines stably expressing the e-MBD4 glycosylase dead mutant (e-MBD4 D554A) were used to purify by double immunoaffinity the e-MBD4 D554A complex. Silver staining of the SDS PAGE gel (a) and immunoblotting (b) for e-MBD4.com (WT) and the mutant e- MBD4 D554A.com. (**c**) Nuclease assay for e-MBD4.com (WT) and both mutant e- MBD4.com R97G and e-MBD4 D554A.com. The experiments were carried out as described Supplemental Figure 1a. The reaction products were not treated with NaOH (**d**) Unmethylated and full-methylated G/T mismatch substrates were incubated with MBD4 protein or with e-MBD4 complex as described Supplemental Figure 1a. Both reaction products as well as the products of

Maxam&Gilbert sequencing assay were separated on PAGE under denaturing conditions. The « T » in red indicates the migration of the respective cleavage « T » product obtained by the Maxam&Gilbert sequencing reaction. (**e**) Schematics of the bi-functional glycosylase/AP-lyase activity of MBD4.

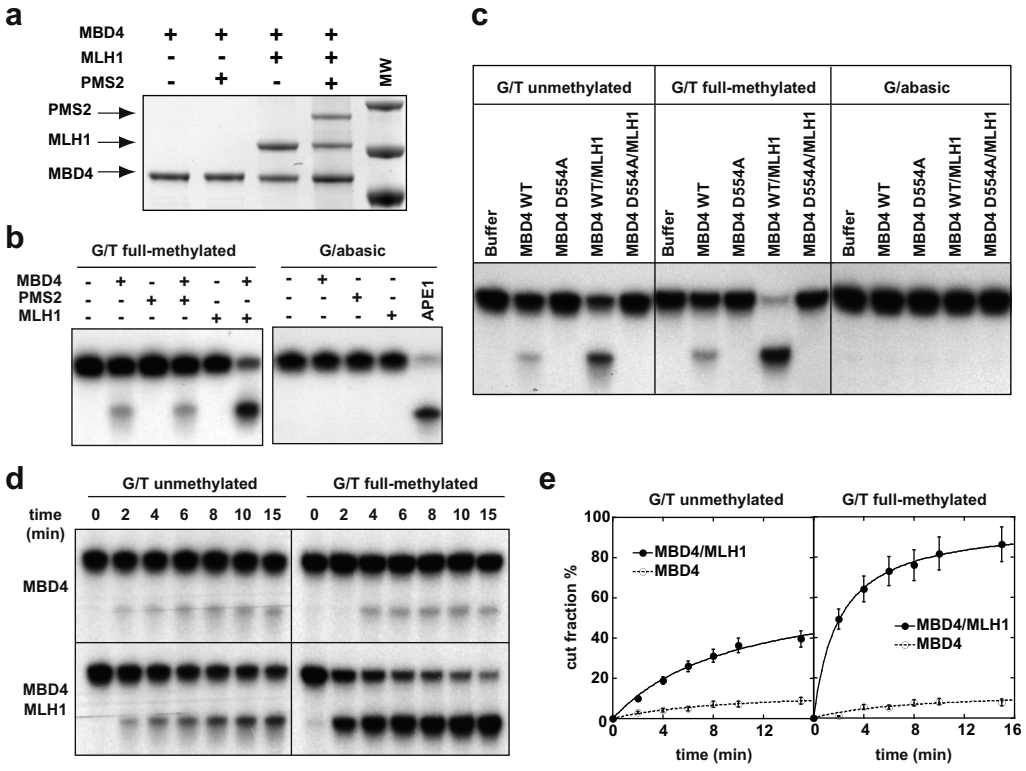


Figure 4. The methyl-directed nuclease activity of MBD4 is dependent to its physical interaction with MLH1. (**a**) SDS PAGE coomassie staining of the purified MBD4 protein, MBD4/MLH1 and MBD4/MLH1/PMS2 complexes co-expressed in the baculovirus system. (**b**) Nuclease assays. The indicated combinations of recombinant proteins were mixed with methylated G/T or abasic-containing substrates and the cleavage reaction was carried out and analyzed as described Supplemental Figure 1a. The reaction products were not treated with NaOH. (**c**) Nuclease assays using either the recombinant MBD4 protein or the purified MBD4/MLH1 complex on unmethylated substrates (left panel), fully methylated substrates (middle panel), or abasic site-containing substrates (right panel). Note that any activity was detected with substrates containing an abasic-site excluding the presence of nuclease contaminant. (**d**) Unmethylated or full-methylated substrates were incubated with identical amount of either MBD4 alone (upper panel) or MBD4 in complex with MLH1 (lower panel) for the indicated times and analyzed as described in Supplemental Figure 1a. The reaction products were not treated with NaOH. (**e**) Quantification of the data presented in (d). The means of three independent experiments are shown.

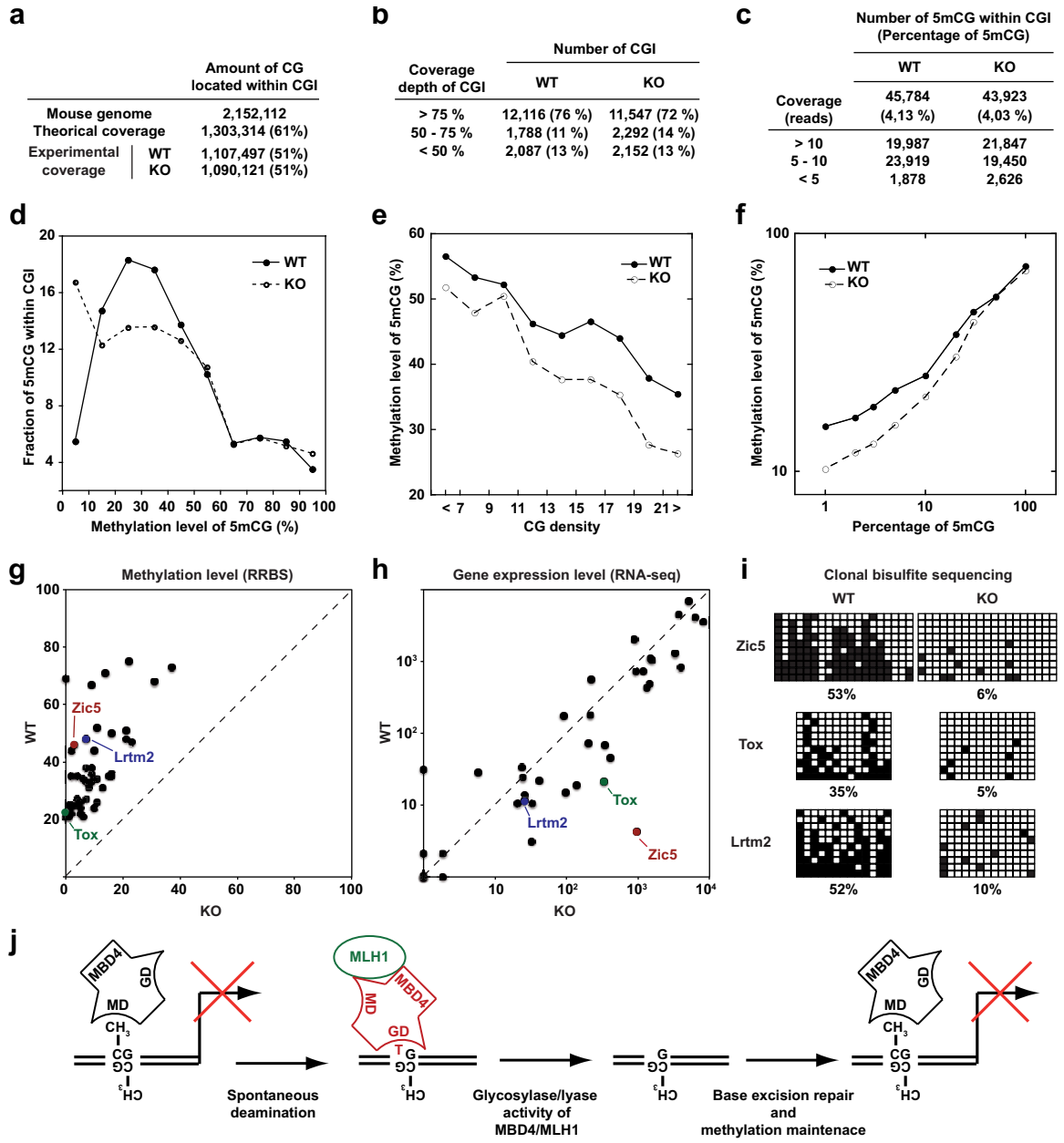
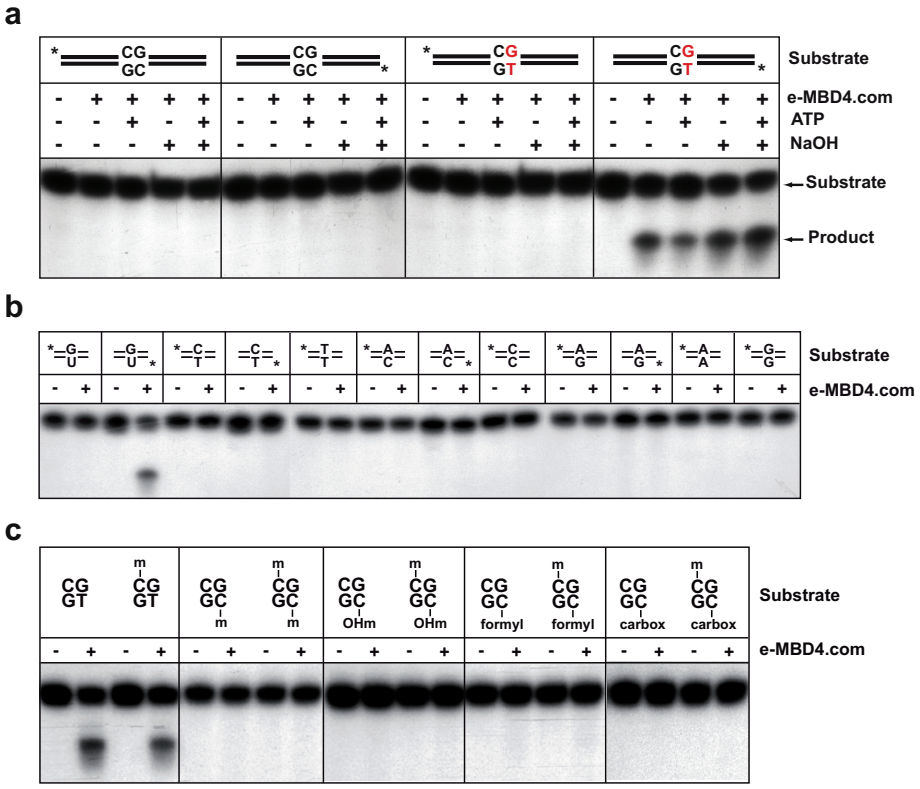


Figure 5. MBD4 preserves DNA methylation within CGI. (**a-f**) Quantification of CGI methylation level in MEFs WT or KO for *Mbd4* by RRBS. (**a-c**) Tables summarizing the number of CG (a), the coverage depth of CG (b), and the number of 5mCG (c) within CGI detected by RRBS in WT and KO cells. (**d**) Methylation level distribution of 5mCG within CGI in WT and KO cells. (**e, f**) Methylation level of 5mCG in function of the CG density (e) or in function of the percentage of 5mCG (f) of CGI. (**g-i**) Upregulation of demethylated promoters in absence of MBD4. Scatter plots comparing methylation levels of differentially methylated proximal regions (determined by RRBS, **g**) and transcription levels of

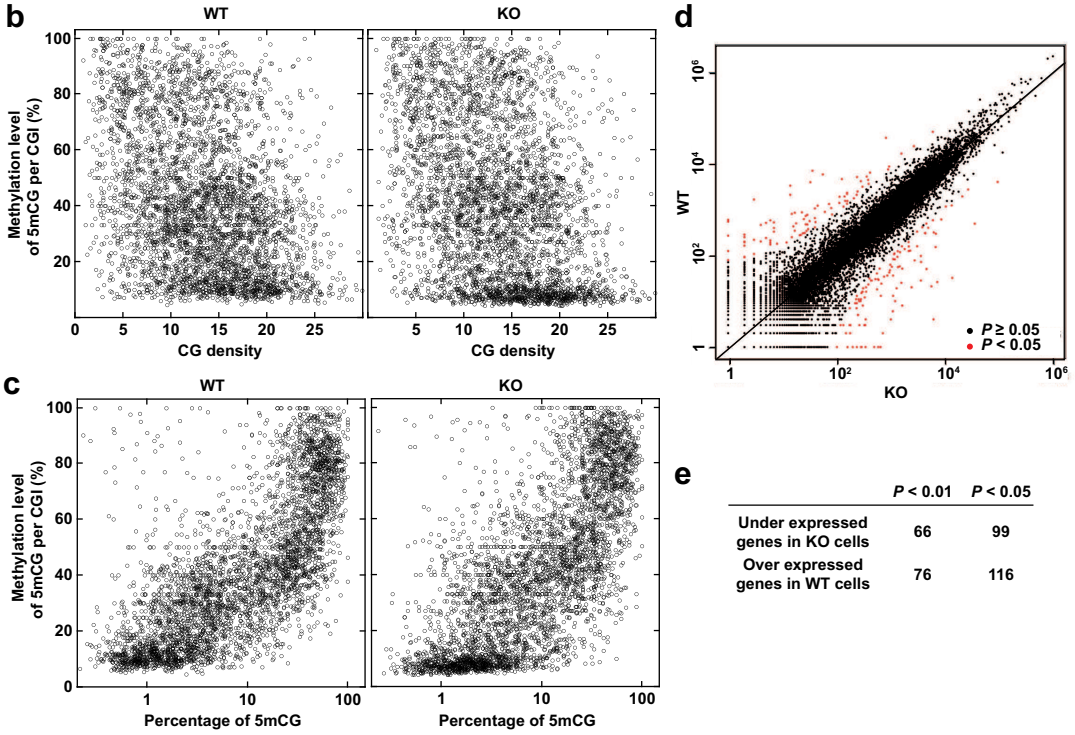
corresponding genes (quantified by RNA-sequencing, **h**) between MEFs WT or KO for *Mbd4*. Analyses were restricted to the most significantly demethylated proximal region in absence of MBD4 (methylation level KO/WT > 2, number of CG > 10, $P < 0.01$, distance to nearest TSS < 5 kb). (i) Clonal bisulfite sequencing confirmation of the hypomethylation phenotype at the proximal regions of *Zic5*, *Tox* and *Lrtm2* (i). White and black squares indicate CG and 5mCG, respectively. (**j**) Model for the function of MBD4 in vertebrates. MBD4 bind methylated CGI promoters through its methyl-binding domain (MD). If a spontaneous deamination conversion of 5-methylcytosine to thymidine occurs, MLH1 activates MBD4, which, through the glycosylase/AP lyase activity of its glycosylase domain (GD), removed the thymine base and cleaved the phosphate backbone. The generated abasic 3' cleaved site is then repaired by BER. Finally, the methylation mark is restored through the action of the DNA methylation maintenance pathway.



Supplemental Figure 1. The MBD4 complex exhibits G/T mismatch specific endonuclease activity. (a) *In vitro* glycosylase/nuclease assays. e-MBD4.com was mixed with the indicated substrates (* indicates the labelled strand), incubated for 20 minutes at 37°C and the products of the reaction were run on PAGE under denaturing conditions. Note that the generation of cut products does not require NaOH treatment. (**b, c**) e-MBD4.com were incubated with indicated substrates (* indicates the labelled strand) as described in (a). Reaction products were not treated with NaOH.

a

Sample	Insert size	Read length (bp)	Clean reads	Mapped reads	Mapped rate (%)	Uniquely mapped reads	Uniquely mapped rate (%)	Enzymes bases	Enzymes rate (%)	Bisulfite conversion rate (%)
WT	40-220	49	112,549,638	107,054,213	95.12	84,397,946	74.99	393,650,847	52.86	99.32
KO	40-220	49	118,338,540	112,910,146	95.41	88,271,057	74.59	358,717,541	62.95	99.35



e

	P < 0.01	*P* < 0.05
Under expressed genes in KO cells	66	99
Over expressed genes in WT cells	76	116

Supplemental Figure 2. MBD4 protects 5-methylcytosines *in vivo*. (**a**) Table summarizing the data obtained by RRBS after filtering and alignment of the raw reads. (**b, c**) Dot blots representing the methylation level of 5mCG by CGI in function of their CG density (number of CpG dinucleotides per 100 bp, **b**) or in function of their percentage of mCG (**c**) in MEFs WT (left pannels) and KO (right pannels) for *Mbd4*. (**d**) Scatter plot comparing global gene expression levels between WT and KO cells. (**e**) Number of significantly differentially expressed genes in absence of MBD4.

Methods

cDNA clones and construction of mutants

Full-length human cDNA clones of MBD4 (IMAGE 3534047), MLH1 (IMAGE 3451538) and PMS2 (IMAGE 7939766) were purchased from Source BioScience. Coding sequence of MBD4 was mutated using megaprimer PCR procedure to produce MBD4 R97G and MBD4 D554A mutant proteins.

Cell lines and complexes purification

MBD4 full-length cDNA was subcloned into the XhoI-NotI sites of the pOZ-N retroviral vector to produce MBD4 protein fused with N-terminal Flag- and HA-epitope tags (e-MBD4). e-MBD4 was stably expressed in HeLa cells or in MEFs by retroviral transduction (1). e-MBD4 nuclear complex (e-MBD4.com) was purified from these cells by double immunoaffinity as previously described (2). MBD4 concentration in e-MBD4.com was estimated by polyacrylamide gels silver-staining using His-tagged MBD4 protein as a standard. Identification of proteins was carried out by Taplin Biological Mass Spectrometry Facility (Harvard Medical School, Boston, MA).

For glycerol density gradient, samples were loaded onto a 4.5 mL glycerol gradient (15%-35%) and spun at 45,000 rpm in a Beckman SW60 rotor for 16 h. Fractions were collected from the bottom of the tube.

Antibodies

Antibodies employed were as follows: monoclonal antibodies anti-Flag M2-Peroxidase (Sigma), anti-TIP49B (612482, BD Transduction), anti-MLH1 (NA28,

Calbiochem), anti-PMS2 (556415, BD PharMingen), anti-TIP49A (ab51500, Abcam) ; polyclonal antibody anti-MBD4 (A-1009, Epigentek).

Purification of MBD4 recombinant protein

The His-tagged protein was cloned in pET28b vector and expressed in the BL21-CodonPlus-RIL-pLysS (Stratagene) strain. A 800 mL culture was grown in LB medium at 37°C until D_{600} of 0.5 was reached before induction with 100 µM IPTG for 2 h at 25°C. Cells were lysed in 20 mL of a buffer containing 10 mM Tri-HCl pH 7.65, 500 mM NaCl, 10 % glycerol, 0.01 % NP40, 10 mM Imidazole, 0.2 mM PMSF and protease inhibitor cocktail tablets (Roche) on ice in the presence of lysozyme at 1 mg/mL and sonicated on ice for 3 × 1 min. His-tagged proteins were immunoprecipitated from clarified supernatant with Ni-NTA-agarose (Qiagen), washed with 50 mM Imidazole and eluted with 300 mM Imidazole using a buffer containing 10 mM Tri-HCl pH 7.65, 150 mM NaCl, 10 % glycerol, 0.01 % NP40. The eluate fraction was diluted two times with sodium phosphate buffer (50 mM sodium phosphate pH 7.0, 1 mM DTT, 1 mM EDTA), incubated with SP sepharose fast flow bead (GE Healthcare), extensively washed with sodium phosphate buffer containing 300 mM NaCl and elutated with sodium phosphate buffer containing 500 mM NaCl. The eluate fraction was desalted with PD-10 Sephadex G-25 columns (GE Healthcare) equilibrated with TGEN buffer containing 150 mM NaCl. His-tagged MBD4 purified proteins were quantified using the Bradford assay with BSA as a standard.

Purification of MBD4/MLH1 and MBD4/MLH1/PMS2 complexes

Flag-tagged MBD4, His-tagged MLH1 and HA-tagged PMS2 proteins were cloned in

pFastBac vector (Invitrogen). The cloned vectors were transformed into bacterial DH10Bac competent cells for making recombinant bacmid. The recombinant bacmid was then extracted and transfected into Sf9 cells by Cellfectin II Reagent (Invitrogen). After viral amplification, SF9 cells were infected (10^6 cells per mL) with baculoviruses expressing either Flag-MBD4 alone or in combination with His-MLH1 and/or HA-PMS2 for 2 days at 27°C. Cells were harvested and resuspended in 25 mL lysis buffer containing 10 mM Tri-HCl pH 7.65, 500 mM NaCl, 10 % glycerol, 0.01 % NP40, 20 mM Imidazole, 0.2 mM PMSF and protease inhibitor cocktail tablets (Roche). The lysate was dounced 30 times, sonicated, and centrifuged for 10 min at 12,000 rpm. The clarified supernatant was incubated at 4°C with anti-FLAG M2 affinity agarose resine (SIGMA), washed 3 times with lysis buffer and 3 times with wash buffer containing 10 mM Tri-HCl pH 7.65, 500 mM NaCl, 10 % glycerol, 0.01 % NP40 and 40 mM Imidazole. The immunoprecipated proteins were eluted with Flag peptide (0.5 mg/mL). The eluted fraction was diluted 3 times with 100 mM sodium phosphate pH 7.0 and incubated with SP sepharose fast flow beads (GE Healthcare). Beads were washed 3 times with wash buffer containing 50 mM sodium phosphate pH 7.0, 100 mM NaCl, 10 % glycerol, 0.01 % NP40, 3 times with wash buffer 2 containing 50 mM sodium phosphate pH 7.0, 100 mM NaCl and eluted with 500 mM NaCl. The eluate fraction was concentrated and loaded onto a 4.5 mL glycerol gradient (10%-30%) and spun at 34,000 rpm in a Beckman SW60 rotor for 18 h. Fractions were collected from the bottom of the tube and protein containing fractions were pooled and dialyzed in buffer containing 50 mM Tri-HCl pH 7.5, 500 mM NaCl and 10 % glycerol.

DNA substrates preparation

The DNA substrates for enzymatic activity assays were prepared by annealing equimolar amounts of the corresponding synthetic oligonucleotides in a buffer containing 10 mM Tris-HCl (pH 7.5), 1 mM EDTA and 100 mM NaCl. DNA substrates were 5'-end labelled on the top or the bottom strand as indicated with [γ-^{32}P]ATP and T4 polynucleotide kinase.

Oligonucleotides used in this study.

Underlined characters indicate the position of the mismatch. 5mC, 5hmC, 5fC, 5caC and Ø indicate 5-methylcytosine, 5-hydroxymethylcytosine, 5-formylcytosine, 5-carboxylcytosine and abasic site respectively.

Name	**Sequence**
TopG	5' CTA ACG ATT GCC GTC GAG TAC CTA CGA GCC TGA TCG ATC GAT CGC TAA TGT CCG GCT AGA AGC GAT TCC GTA CGA TGC 3'
BotC	5' GCA TCG TAC GGA ATC GCT TCT AGC CGG ACA TTA GCG ATC GAT CGA TCA GGC TCG TAG GTA CTC GAC GGC AAT CGT TAG 3'
BotT	5' GCA TCG TAC GGA ATC GCT TCT AGC CGG ACA TTA GCG ATT GAT CGA TCA GGC TCG TAG GTA CTC GAC GGC AAT CGT TAG 3'
TopGmet	5' CTA ACG ATT GCC GTC GAG TAC CTA CGA GCC TGA TCG AT5mC GAT CGC TAA TGT CCG GCT AGA AGC GAT TCC GTA CGA TGC 3'
BotCmet	5' GCA TCG TAC GGA ATC GCT TCT AGC CGG ACA TTA GCG AT5mC GAT CGA TCA GGC TCG TAG GTA CTC GAC GGC AAT CGT TAG 3'
BotChmet	5' GCA TCG TAC GGA ATC GCT TCT AGC CGG ACA TTA GCG AT5hmC GAT CGA TCA GGC TCG TAG GTA CTC GAC GGC AAT CGT TAG 3'
BotCfor	5' GCA TCG TAC GGA ATC GCT TCT AGC CGG ACA TTA GCG AT5fC GAT CGA TCA GGC TCG TAG GTA CTC GAC GGC AAT CGT TAG 3'
BotCcar	5' GCA TCG TAC GGA ATC GCT TCT AGC CGG ACA TTA GCG

	AT5caC GAT CGA TCA GGC TCG TAG GTA CTC GAC GGC AAT CGT TAG 3’
TopG4met	5’ CTA ACG ATT GCC GT5mC GAG TAC CTA CGA GCC TGA T5mCG ATC GAT 5mCGC TAA TGT CCG GCT AGA AG5mC GAT TCC GTA CGA TGC 3’
BotT4met	5’ GCA TCG TAC GGA AT5mC GCT TCT AGC CGG ACA TTA G5mCG ATT GAT 5mCGA TCA GGC TCG TAG GTA CT5mC GAC GGC AAT CGT TAG 3’
TopC	5’ CTA ACG ATT GCC GTC GAG TAC CTA CGA GCC TGA TCG ATC CAT CGC TAA TGT CCG GCT AGA AGC GAT TCC GTA CGA TGC 3’
TopT	5’ CTA ACG ATT GCC GTC GAG TAC CTA CGA GCC TGA TCG ATC TAT CGC TAA TGT CCG GCT AGA AGC GAT TCC GTA CGA TGC 3’
TopA	5’ CTA ACG ATT GCC GTC GAG TAC CTA CGA GCC TGA TCG ATC AAT CGC TAA TGT CCG GCT AGA AGC GAT TCC GTA CGA TGC 3’
BotU	5’ GCA TCG TAC GGA ATC GCT TCT AGC CGG ACA TTA GCG ATU GAT CGA TCA GGC TCG TAG GTA CTC GAC GGC AAT CGT TAG 3’
BotG	5’ GCA TCG TAC GGA ATC GCT TCT AGC CGG ACA TTA GCG ATG GAT CGA TCA GGC TCG TAG GTA CTC GAC GGC AAT CGT TAG 3’
BotA	5’ GCA TCG TAC GGA ATC GCT TCT AGC CGG ACA TTA GCG ATA GAT CGA TCA GGC TCG TAG GTA CTC GAC GGC AAT CGT TAG 3’
Botabasic	5’ GCA TCG TAC GGA ATC GCT TCT AGC CGG ACA TTA GCG ATØ GAT CGA TCA GGC TCG TAG GTA CTC GAC GGC AAT CGT TAG 3’

CG/GC homoduplex substrate : TopG + BotC ; CG/GT mismatch substrate : TopG + BotT ; CG/GT mismatch hemi-methylated substrate : TopG + BotT4met or TopG4met + BotT ; CG/GT mismatch full-methylated substrate : TopG4met + BotT4met ; CG/GU mismatch substrate : TopG + BotU ; CC/GT mismatch substrate : TopC + BotT ; CT/GT mismatch substrate : TopT + BotT ; CA/GC mismatch substrate : TopA + BotC ; CC/GC mismatch substrate : TopC + BotC ; CA/GG mismatch substrate : TopA + BotG ; CA/GA mismatch substrate : TopA + BotA ; CG/GG mismatch substrate : TopG + BotG ; CG/GmetC substrate : TopG + BotCmet ; metCG/GmetC

homoduplex substrate : TopGmet + BotCmet ; CG/GhmetC homoduplex substrate : TopG + BotChmet ; metCG/GhmetC homoduplex substrate : TopGmet + BotChmet ; CG/GforC substrate : TopG + BotCfor ; metCG/GforC substrate : TopGmet + BotCfor ; CG/GcarC substrate : TopG + BotCcar ; metCG/GcarC substrate : TopGmet + BotCcar ; CG/Gabasic substrate : TopC + Botabasic.

Quantitative glycosylase/lyase assays

Reaction mixtures (10 µL) containing 20 mM Tris-HCl pH 7.65, 50 mM NaCl, 3 mM $MgCl_2$, 5 % glycerol, 1 mM DTT, 0.1 µg/µl BSA and 2 nM of end-labeled substrates was incubated 20 minutes (excepted for kinetic experiment) at 37°C with 60 nM of MBD4 (excepted as indicated). When indicated, PMS2 and MLH1 proteins were added to reaction mixture to a final concentration of 100 nM. The reaction was stopped by adding 10 µL formamide buffer (90 % formamide, 10 mM EDTA, 0.1 % blue bromophenol) and heating 5 min at 95°C before loading on a 12% denaturing polyacrylamide gel. When indicated, the reaction was pre-treated with 1 µL of 1 M NaOH 10 min at 95°C before the addition of formamide buffer. Gels were dried and quantified on a Typhoon 8600 Variable Mode Imager.

Mapping of the nicking reaction

Enzymatic activities assays were done as described above. Products of reactions together with the products of the G+A and the C+T Maxam-Gilbert cleavage reactions performed on the same substrates were loaded on a 8% denaturing polyacrylamide gel.

Isolation of MEFs

Primary MEFs WT or KO for *Mbd4* were isolated from *Mbd4*$^{+/+}$ and *Mbd4*$^{-/-}$ mice respectively, as previously described (3).

RNA-sequencing

RNA samples were purified using standard methods from subconfluent MEFs cultures. The 3′-end RNA sequencing was performed on the Illumina Hiseq 2500 as single-end 50 base reads following Illumina's instructions. Reads were mapped onto the mm9 assembly of the mouse genome by using Tophat (4) and the bowtie aligner (5). Quantification of gene expression was performed using HTSeq (http://www-huber.embl.de/users/anders/HTSeq) and gene annotations from Ensembl release 67. Read counts have been normalized across WT and KO libraries with the statistical method proposed by Anders and Huber (6) and implemented in the DESeq Bioconductor library. Resulting p-values were adjusted for multiple testing by using the Benjamini and Hochberg method (7).

Clonal bisulfite sequencing and RRBS

Genomic DNA was isolated from subconfluent primary MEFs as previsously described (8). Digested DNA (500 ng) was converted with EZ DNA Methylation-Gold Kit (Zymo Research Corporation). Primer design was accomplished using Methprimer. Bisulfite sequencing primers (5′-TTTTTTTATGAATAAGTAATTTAATAATAT-3′ and 5′-AATTTCCTAAAATCCCAAATCTCTC-3′ for *Zic5*, 5’-TTGTAGTATTTGTAGTTTGGGGTAG-3’ and 5’-AACAAATAATCCCTAATTCCCATAC for *Lrtm2*) were used to amplify the corresponding promoters. PCR included an initial incubation at 95°C for 10 minutes,

followed by 40 cycles of 95°C for 30 seconds, 52°C for 30 seconds, and 72°C for 60 seconds, followed by one cycle of 72°C for 10 minutes. The PCR products were cloned into the pCR2.1-TOPO vector using the TOPO TA cloning kit (Invitrogen) for sequencing. A total of 10 clones from each sample were sequenced at the GATC Biotech company, and the methylation status for each CpG site was determined by assessing the presence of T (unmethylated) versus C (methylated) at each CpG site.

For RRBS, bisulfite-converted genomic DNA libraries were prepared according to the previously described methods (9). Briefly, genomic DNA was digested with MspI (New England Biolabs), followed by end-repair and addition of 3′ A overhangs. Methylated adaptors (Illumina) with a 3′ T overhang were ligated to the A tailed DNA fragments. For reduced representation, 40 to 220 bp (pre-adaptor-ligation size) fragments were excised from 2% TAE agarose gels and bisulfite-converted with EZ DNA methylation Gold kit (Zymo Research). Bisulfite converted libraries were amplified by PCR and sequenced on an Illumina HiSeq 2000 sequencer with a single-ended, 49 bp run (Beijing Genomics Institute). FASTQ sequence files containing sequenced reads were obtained for both samples (MEFs WT or KO for *Mbd4*).

Data processing

After removal of the adaptor sequences, the 49 bp reads from each sample were aligned to genome reference (mm9) as well as the size-selected MspI fragments generated by our *in silico* simulation. Because of the strand specificity of DNA methylation, two rounds of alignments were carried out, i.e. the bisulfite converted reads were aligned to the genome sequences termed the “T genome” with each cytosine converted to thymine and in the meanwhile the reads were also aligned to

the genome sequences termed the “A genome” with each guanine converted to adenosine. The alignments were carried out with BAMAP aligner allowing up to two mismatches for successful mapping. Summary of the data quantity after each step of filtration is shown in Supplemental Figure 2a and both samples showed near complete bisulphite conversion of non-CpG cytosines (> 99%).

Bioinformatic analyses

The mouse CGI database was retrieved from the UCSC Genome Bioinformatics site using Table Browser program (genome : mouse ; assembly : NCBI37/mm9 ; group : Expression and Regulation ; track CpG islands). Coverage depth of CGI (Figure 5b) refers to the number of CG by CGI that were detected by RRBS divided by the total number of CG within that CGI. Coverage of 5mCG (Figure 5c) correspond to the number of reads that covered the concerned CG. The methylation level of 5mCG was determined by dividing the number of reads covering each 5mCG by the total reads covering that CG. Calculation of the methylation level of 5mCG described Figure 5 and Supplemental Figure 2 was restricted to CG within CGI covered by more than 10 reads.

Accession numbers

RNA-seq and RRBS datasets obtained with MEFs WT or KO for *Mbd4* have been deposited in Gene Expression Omnibus (GEO) under the accession number XXX.

References

1. Ouararhni K, *et al.* (2006) The histone variant mH2A1.1 interferes with transcription by down-regulating PARP-1 enzymatic activity. *Genes & development* 20(23):3324-3336.

2. Drane P, Ouararhni K, Depaux A, Shuaib M, & Hamiche A (2010) The death-associated protein DAXX is a novel histone chaperone involved in the replication-independent deposition of H3.3. *Genes & development* 24(12):1253-1265.
3. Cortellino S*, et al.* (2003) The base excision repair enzyme MED1 mediates DNA damage response to antitumor drugs and is associated with mismatch repair system integrity. *Proc Natl Acad Sci U S A* 100(25):15071-15076.
4. Trapnell C, Pachter L, & Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105-1111.
5. Langmead B, Trapnell C, Pop M, & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 10(3):R25.
6. Anders S & Huber W (2010) Differential expression analysis for sequence count data. *Genome biology* 11(10):R106.
7. Hochberg Y & Benjamini Y (1990) More powerful procedures for multiple significance testing. *Statistics in medicine* 9(7):811-818.
8. Thomassin H, Oakeley EJ, & Grange T (1999) Identification of 5-methylcytosine in complex genomes. *Methods* 19(3):465-475.
9. Wang L*, et al.* (2012) Systematic assessment of reduced representation bisulfite sequencing to human blood samples: A promising method for large-sample-scale epigenomic studies. *Journal of biotechnology* 157(1):1-6.

Combinatorial DNA methylation codes at repetitive elements

Christophe Papin[1*], Abdulkhaleg Ibrahim[1*], Stephanie Le Gras[1], Isabelle Stoll[1], Bernard Jost[1], Christian Bronner[1], Stefan Dimitrov[2] and Ali Hamiche[1,§].

[1]Département de Génomique Fonctionnelle et Cancer, Institut de Génétique et Biologie Moléculaire et Cellulaire (IGBMC), UdS, CNRS, INSERM, Equipe labélisée Ligue contre le Cancer, 1 rue Laurent Fries, B.P. 10142, 67404 Illkirch Cedex, France.

[2]INSERM/UJF, Institut Albert Bonniot, U823, Equipe labélisée Ligue contre le Cancer, Site Santé-BP 170, 38042 Grenoble Cedex 9, France.

[§]To whom correspondence should be addressed. Ali Hamiche: E-mail: hamiche@igbmc.fr.

* These authors contributed equally to this work

Running title: DNA methylation of repetitive elements.

Abstract

DNA methylation is an essential epigenetic modification, present in both unique DNA sequences and repetitive DNA elements, but its role in both normal and pathological situations remains obscure. Here, we describe a genome-wide comparative analysis of the 5mC, 5hmC, 5fC and 5caC profiles of repetitive elements in mouse embryonic fibroblasts and mouse embryonic stem cells. We provide evidence for distinct and highly specific DNA methylation/oxidation patterns of the repetitive elements in both cell types, which mainly affect CA repeats and evolutionary conserved mouse-specific transposable elements including IAP-LTRs, SINEs B1m/B2m and L1Md-LINEs. These repeated elements are clustered at specific locations in the mouse genome and we show that TDG is implicated in the regulation of their unique DNA methylation/oxidation signatures and their dynamics. Our data suggest the existence of novel epigenetic code for the most recently acquired evolutionary conserved repeats that could play a major role in cell differentiation.

Introduction

DNA methylation is an epigenetic modification essential for mammalian development (Okano et al. 1999). In mammals, the cytosine bases at position 5 in CpG dinucleotides are modified genome wide by dedicated methyl-transferases to produce 5-methylcytosine (5mC), which can be further successively oxidized by the Ten eleven translocation (TET1, 2 and 3) enzymes (Tahiliani et al. 2009; Ito et al. 2010; He et al. 2011; Ito et al. 2011) to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). Recent data showed that both 5fC and 5caC can be excised and repaired to regenerate unmodified cytosines by the concerted action of thymine-DNA glycosylase (TDG) and the base excision repair (BER) enzymes (Cortellino et al. 2011; He et al. 2011; Maiti and Drohat 2011). In mammalian genomes, the CpG dinucleotides are under-represented. Unusually dense clusters of CpG dinucleotides, called "CpG islands (CGIs), are present and overlap with the promoter regions of more than 70% of the genes (Illingworth et al. 2010; Deaton and Bird 2011). Despite their high GC content, CGIs are mainly unmethylated. The current view holds that the CpG dinucleotide remains the primary site for DNA methylation, but there is emerging evidence for non-CpG methylation in several mammalian cells and tissues, including embryonic stem cells (ESCs), induced pluripotent stem cells (iPSC), oocyte and brain (Lister et al. 2009; Laurent et al. 2010; Tomizawa et al. 2011; Xie et al. 2012; Lister et al. 2013; Ziller et al. 2013).

Over two third of the mammalian genome consists of repeated sequences (de Koning et al. 2011), including long terminal repeats (LTR), long (LINE) and short (SINE) interspersed nuclear elements, major satellites and simple repeats (Mouse Genome Sequencing et al. 2002). Previous data suggest that at least some of these repeated elements are methylated in ESCs (Shen et al. 2013) and undergo an extensive demethylation during plant development (Gehring et al. 2009). No comprehensive data are, however, available for the DNA genome-wide methylation pattern of repeated elements in differentiated mammalian cells. The biological significance of repetitive element methylation/oxidation remains elusive. Whereas the hypomethylation of repetitive elements is a recognized hallmark of cancer cells (Howard et al. 2008; Ehrlich 2009; Baba et al. 2010), how this is related to cancer development is poorly understood.

In this study, by using DNA-immunoprecipitation-sequencing (DIP-seq), we have carried out a genome-wide comparative analysis of the DNA

methylation/oxidation patterns of repetitive elements in both differentiated mouse embryonic fibroblasts (MEFs) and pluripotent ESCs. This approach allows a high coverage of repetitive elements and profiling of all CpNs within the genome (Down et al. 2008). Our data revealed distinct patterns of DNA methylation/oxidation for these two types of cells, suggesting that cell differentiation is concomitant with profound global genome-wide changes in DNA methylation of repetitive elements. The methylation profiles are dynamically regulated by TDG. The most recently acquired and most conserved lineage-specific repetitive elements showed the most striking methylation patterns and are not distributed randomly throughout the mouse genome, but instead cluster at specific loci. These data highlight a dynamic combinatorial DNA methylation code at repetitive elements and define novel DNA regulatory regions within the mouse genome.

Results

Accumulation of cytosine modifications at repeats in MEFs.

As we considered that TDG might be implicated in the dynamics of genome-wide distribution of 5mC and its oxidized forms in differentiated cells, we performed 5mC, 5hmC, 5fC and 5caC DIP-seq experiments in wild-type and *Tdg*-deficient MEFs (Supplemental Fig. 1a). Analysis of genome-wide sequencing data identified 64% to 76% of multihit reads, mapping to multiple genomic regions (Fig. 1a). The majority of these reads overlapped with the UCSC Repeat-Masker (RMSK) (Dreszer et al. 2012) sequences mainly representing repetitive elements in the murine genome. On the other hand, genome browser visualization of wig track files obtained with uniquely mapped reads showed an enrichment of each cytosine modifications at specific genomic regions (Supplemental Fig. 1b). From 77% to 90% of the called peaks overlapped with RMSK data further supporting our finding that the repetitive elements in MEFs represent the vast majority of cytosine modifications enriched region (Figure 1b). Accordingly, modified cytosines were strongly under-represented at CGIs (Supplemental Figure 1c). 5mC enrichment was found to occur at all repeats, but with a marked preference for SINEs, while 5hmC, 5fC and 5caC enrichment were found predominantly at simple repeats (Fig. 1b). In agreement, a strong overlap between 5hmC, 5fC and 5caC peaks was observed (Fig. 1c), a result that was further confirmed by the correlation matrix coefficient showing a close clustering between these three oxidized forms (Supplemental Fig. 1d-g). 5fC and 5caC peaks at simple

repeats were increased by approximately 2-fold in response to *Tdg* knockdown suggesting an involvement of TDG in the regulation of their DNA methylation/oxidation patterns. In contrast, neither 5mC nor 5hmC showed significant changes at these repeats upon *Tdg* knockdown (Fig. 1b,c).

The unexpected enrichment of cytosine modifications at repetitive elements prompted us to perform a global and systematic computational analysis of *Tdg*-deficient-dependent changes in DNA methylation patterns at each family of repeats in MEFs. We first characterized these changes at individual repeats by analyzing uniquely mapped reads (Supplemental Fig. 1d,e) and then performed a second independent analysis to include multihit reads mapping to unique repeat families (see Methods and Supplemental Fig. 1f,g). These analyses allowed extension of the results of individual repeat elements to their corresponding family.

Methylation patterns at LTRs in differentiated MEFs and pluripotent ESC cells.

In order to perform a comparative analysis of cytosine modification patterns between the repeated elements in differentiated MEFs and pluripotent ESCs, we compared our data on MEFs with reads overlapping with RMSK database from the previously published data sets for ESCs (Shen et al. 2013) that we reanalyzed independently. We first characterized the *Tdg*-deficient-dependent changes in DNA methylation at LTR retrotransposons (also known as endogenous retroviruses ERVs) in MEFs. For simplicity we will further refer to LTR retrotransposons as LTRs. RMSK database distinguishes between elements corresponding to external domains (LTR_{ext}, containing the regulatory regions of the LTR) from those corresponding to internal domains (LTR_{int}, containing the coding sequences of the proteins, necessary for the life cycle of the integrated viruses) (Fig. 2a) within the different LTR families. Bearing this in mind, we carried out independent analyses for these two regions. Heatmaps were generated using uniquely mapped reads on internal domains (LTR_{int}) with a 2 kb cutoff to eliminate truncated and degenerate LTRs (Figure 2b). The corresponding normalized density of 5mC, 5hmC, 5fC and 5caC signals is presented in Supplemental Figure 2a. Our data revealed a striking enrichment of 5mC exclusively at the external repeats of the IAP (Intracisternal A Particle) LTR subfamily (ERVK class). Of note, IAP repeats are the evolutionarily least truncated LTR subfamily (Fig. 2c) and with the highest CG content (Fig. 2d and Supplemental Fig. 2b).

Relative enrichment calculation including multihit reads (Fig. 2e left panel and Supplemental Fig. 2c left panel) further confirmed the 5mC enrichment of the IAP-LTR external regions (≈10-fold), but also revealed a moderate 5mC enrichment of ERV1 and ERVK families at their internal domains (≈2-fold). This moderate methylation was barely observed in the heatmap generated with uniquely mapped reads (Fig. 2b). We hypothesized that the high conservation score (3,500, 5,000 and 10,000 for ERV1, ERVK and IAP, respectively) of these internal domains complicates their mapping at unique loci. Our data reveal a direct correlation between the observed methylation level and the CG density of the different LTR retro-element regions (Fig. 2d, 2e left panel and Supplemental Fig. 2c left panel). Overall, we observe 5mC enrichment at every LTR region whenever the CG density exceeds the average mouse genome density (0.83). The depletion of TDG does not, however, affect this 5mC enrichment of the LTRs.

The methylation patterns of the LTR_{ext} of all LTR retrotransposons in ESCs showed, however, striking differences when compared to these in MEFs. For example in ESCs, IAPs exhibited a four fold lower enrichment for 5mC than MEFs and an enrichment for all three 5mC oxidized forms, accumulating 5fC and 5caC in response to *Tdg* knockdown (Fig.2e, upper right panel and Supplemental Fig. 2c right panel). Non-IAP ERVKs and ERV1s showed an enrichment of the LTR_{ext} for the oxidized methylated cytosines with a specific accumulation of 5caC in absence of TDG (Fig. 2e right panel and Supplemental Fig. 2c right panel). Importantly, ERVK and ERV1 internal domains (LTR_{int}) tend to be depleted in 5mC (Fig. 2e, the lower two panels). Taken as a whole, these data suggest that the evolutionary conserved LTRs harboring a high CG content are dynamically regulated by TDG in ESCs, but stably methylated during cell differentiation.

Methylation profiles of mouse-specific SINEs

SINEs are interspersed repeats that make up to 7.5% of the mouse genome (Supplemental Fig.6c) and comprise two mouse-specific families, B1m and B2m (Mouse Genome Sequencing et al. 2002). Heatmaps of 5mC/5hmC/5fC/5caC at SINEs ranked by families revealed TDG-dependent specific cytosine methylation patterns for the mouse-specific families B1m and B2m in MEFs (Supplemental Fig. 3a). Heatmaps ranked by conservation scores for the B1m and B2m families illustrated a strong correlation between the density of cytosine modification (Fig. 3a),

conservation and CG density (Fig. 3b and Supplemental Fig. 3b). Normalized density curves clearly showed that the mouse-specific SINEs are highly methylated, weakly hydroxymethylated/carboxylated and dynamically regulated by TDG (Fig. 3c). Calculation of relative enrichment using total reads further confirmed the global hypermethylation of mouse-specific SINEs over ancestral SINEs and their TDG-dependent regulation in MEFs (Fig. 3d upper panel and Supplemental Fig. 3c left panel). Comparative analysis revealed a specific hydroxymethylation of these mouse-specific SINEs in ESCs (Fig. 3d lower panel and Supplemental Fig. 3c right panel).

Since SINEs have been implicated in gene regulation and chromatin domain anchoring (Lunyak et al. 2007; Ichiyanagi 2013), we sought to determine whether these species-specific SINEs are located at specific functional loci within the mouse genome. Genome browser visualization revealed that methylated mouse-specific SINEs are clustered around CGIs (Fig. 3e upper panel and Supplemental Fig. 3d), but close to the TSS (Fig. 3e lower panel). Further calculation showed a perfect correlation between SINE conservation and proximity to TSS (Fig. 3f,g). All together, our data revealed that species-specific SINEs are preferentially integrated around TSS, hydroxymethylated in ESCs, but highly methylated and relatively weakly hydroxymethylated/carboxylated in MEFs where their methylation state is regulated in a TDG-dependent manner (Figure 3h).

DNA methylation patterns of LINEs.

LINEs are autonomous retrotransposons making up about 20% of the mouse genome (Mouse Genome Sequencing et al. 2002) (Supplemental Fig. 6c). Most of LINEs are defective due to truncation or accumulation of mutations over time (for review see refs(Edgell et al. 1987; Sookdeo et al. 2013). Indeed, only 1% of annotated LINEs are intact (length > 5 kb) and potentially active (Castro-Diaz et al. 2014). The majority of full-length LINEs is represented by the strongly conserved mouse-specific L1Md family, which also exhibits the highest CG density (Fig. 4a,b). This family contains a 5´UTR functioning as a promoter, two open reading frames, ORF1 and ORF2, and a 3´UTR containing a polyA signal. To analyze the distribution of 5mC/5hmC/5fC/5caC at these LINEs in MEFs, we divided the L1Md database into two groups according to their genomic orientation (sense or antisense strand) and normalized their lengths (see Methods for details). Tag count clustering revealed two

distinct clusters for each group (clusters 1 and 2 for positive-sense L1Md, and clusters 3 and 4 for negative-sense L1Md) (Fig. 4c). Normalized density of 5mC, 5hmC, 5fC and 5caC signals at L1Md clusters in control and *Tdg*-deficient MEFs reveal two distinct profiles (Fig. 4c,d). The first profile corresponds to L1Md containing a hypermethylated 5' UTR region (cluster 1, on positive strand, and cluster 3 on negative strand). The second profile corresponds to L1Md showing an un-methylated 5' UTR and a methylcytosine oxidation pattern along their ORF dynamically regulated by TDG (cluster 2 for positive-sense L1Md, and cluster 4 for negative-sense L1Md) (Fig. 4d). Three individual L1Md elements representative of the four described elements are visualized in (Supplemental Fig. 4a). This L1Md family-specific pattern was further validated by analyses of total mapped reads (Fig. 4e left panel and Supplemental Fig. 4b left panel).

We hypothesized that the 5'UTR hypermethylated LINEs (clusters 1 and 3) are transcriptionally inactive whereas LINEs characterized by a methylcytosine oxidation patterns along their ORFs (clusters 2 and 4) are active. Accordingly, analysis of H3K4me1, H2A.Z and Pol II distribution by ChIP-seq experiments at clusters 2 and 4 revealed a clear enrichment of these transcriptionally active marks along their ORFs strongly suggesting that transcription of LINEs is regulated by DNA methylation, i.e. LINEs are silenced by cytosine methylation of their 5'UTR whereas active LINEs are regulated by methylcytosine oxidation in a TDG-dependent manner.

We next analyzed whether the L1Md LINEs are randomly distributed throughout the genome or integrated at precise locations. We observed a high concentration of intact L1Md LINEs around gene clusters (1 L1Md/every 60 kb vs 1/180 kb for the rest of the genome) such as the Skint and Vmn2r clusters (see Supplemental Fig. 4c). Surprisingly, in ESCs L1Md were depleted in cytosine modifications (Fig. 4e right panel and Supplemental Fig. 4b right panel), suggesting that the TDG-dependent regulation of LINE transcription by methylation takes place during differentiation.

TDG-dependent methylation patterns of simple repeats.

We next investigated the DNA methylation patterns of simple repeats. Simple repeats are made up by variable numbers of successive repeating units with various lengths (for review see ref(Ellegren 2004). Heatmaps of 5mC/5hmC/5fC/5caC levels for simple repeats revealed a clear enrichment of all 5mC oxidized forms specifically in

CA repeats in MEFs (Fig. 5a). Note that the antibodies we have used are highly specific for the individual 5mC oxidized forms of the CA repeats, thus ruling out the possibility of non-specific association of the antibodies with the repeats (Supplemental Fig. 5a). Since the CA repeats exhibit a 50 fold higher CA than CG density (Supplemental Fig. 5b), the above results suggest that the cytosine within the CpA dinucleotides could also be methylated. To further validate this claim, CA repeats were ranked by conservation score and heatmaps were generated. Interestingly, a strong correlation could be seen between CA density and 5hmC>5caC>5fC densities (Fig. 5b). Normalized density curves (Fig. 5c) and genome browser views (Supplemental Fig. 5c,d) confirmed that CA dinucleotides are mainly hydroxymethylated and dynamically formyl/carboxyl-ated in a *Tdg*-dependent-deficient manner. Indeed, *Tdg* knockdown leads to a strong density accumulation of both 5fC (2 folds) and 5caC (1.6 folds) (Figure 5c). Of note, the depletion of TDG affects weakly the density of both 5mC and 5hmC (Fig. 5c, the two upper panels). Relative enrichment calculation including multihit reads for each cytosine modification at different simple repeats families confirmed that CA repeats are specifically enriched in 5mC oxidized form in both MEFs and ESCs (Fig. 5d and Supplemental Fig. 5e upper panel). Strikingly, CA repeats showed a strong 5mC enrichment in ESCs but not in MEFs (Fig. 5d). Clustered heatmap density of 5mC/5hmC/5fC levels at CA repeats in ESCs and MEFs revealed that the highly methylated CA repeats in ESCs correspond to those harboring the strongest 5mC oxidized forms in MEFs (Supplemental Fig. 5f). Together our data show that the densest CA repeats are preferentially methylated in ESCs, oxidized and dynamically regulated by TDG during differentiation.

Since CA methylation has been shown by bisulfite sequencing to occur in Drosophila and mammals mainly in the CAC trinucleotide context (Laurent et al. 2010; Lister et al. 2013; Guo et al. 2014; Takayama et al. 2014), we investigated whether this motif is also preferentially modified at simple repeats. Our analysis identified the CAC trinucleotide as the main motif targeted by 5mC oxidation at simple repeats in MEFs (Supplemental Fig. 5e lower panel).

The occurrence of 5mC oxidation at CA repeats, prompted us to analyze whether TDG could excise an oxidized cytosine introduced in a non-CpG context. In vitro assays showed that recombinant TDG efficiently excised a formylcytosine

introduced in a CpA context, but not in CpC or CpT context (Fig. 5e). Likewise, the dioxygenase activity of the Naegleria Tet-like protein has a strong preference for 5mCpG and 5mCpA (Hashimoto et al. 2014). These results further validated the occurrence of DNA methylation/oxidation at CA repeats in MEFs and ESCs.

Since CA repeats have been implicated in transcription and splicing regulation (Naylor and Clark 1990; Gebhardt et al. 1999; Pravica et al. 1999; Shimajiri et al. 1999; Gabellini 2001; Hui et al. 2003b), we sought to determine whether CA repeat modifications occurred randomly or at specific loci within the mouse genome. Our data revealed a high correlation between CA repeat density, proximity to TSS (Fig. 5f) and cytosine hydroxymethylation (Fig. 5g). Collectively, these results highlight TDG-dependent DNA methylation dynamics at conserved CA repeats that are located closer to TSS compared to degenerate CA repeats. The distinct methylation/oxidation patterns found in MEFs and ESC may reflect an active role of these modifications in shaping the transcriptional re-programming taking place during differentiation.

We finally analyzed the occurrence of cytosine modifications at major satellites and DNA transposons. Major satellites showed a unique cytosine modification pattern conserved between MEFs and ESCs, characterized by a specific 5mC, 5fC and 5caC enrichment (Supplemental Fig. 6a). DNA transposons did not show any enrichment in cytosine modifications (Supplemental Fig. 6b).

Discussion

Here, we present a genome-wide comparative analysis of DNA methylation/oxidation profiles of repetitive elements in both MEFs and ESCs. We found major differences in the DNA methylation/oxidation patterns of repetitive elements in these cells. A majority of DNA methylation/oxidation patterns is dynamically regulated by TDG and occur mainly at CA repeats and at the most recently acquired transposable elements corresponding to mouse-specific repeats with high CG content. We show that these elements are not distributed randomly throughout the mouse genome, but are clustered with respect to the TSS and hence may act as novel cis-acting regulatory elements (Supplemental Fig. 6d).

We observe enrichment methylation at every conserved repeat whenever the CG density exceeded 0.83, the average mouse genome density. For example, the IAP retroviruses, that have the highest CG density, showed the highest methylation enrichment. In ESCs, this subfamily was partially methylated and enriched in 5hmC, but was fully methylated in MEFs, which suggest their permanent inactivation during differentiation to prevent insertion mutagenesis. Accordingly, IAP transcription is constrained by methylation (Walsh et al. 1998) and LTR elements were found excluded from gene-rich regions (Medstrand et al. 2002), likely because of their potential to alter transcription. LTR families harboring an intermediate CG density such as ERV1 and non-IAP-ERVK, showed TDG-dependent oxidation dynamics specific to ESCs, while the evolutionary oldest CG-poor ERVL family escaped methylation. Collectively, our data suggest that methylation level of CG rich LTRs is highly dynamic during differentiation.

The mouse specific SINEs, B1m and B2m, are concentrated around CGIs. This peculiar localization could have profound consequences on neighboring gene expression. Accordingly, human B1 SINEs have been shown to influence the activity of downstream gene promoters, with acquisition of DNA methylation and loss of active histone marks (Estecio et al. 2012). Mouse B1m and B2m SINEs might act as boundary elements that protect CGIs against pervasive methylation and hence they could be used by ESCs (where the SINEs are not methylated) to maintain the undifferentiated state. This claim is supported by the observation of SINE hydroxymethylation in ESCs and their hypermethylation in MEFs. The hydroxymethylation could regulate the transcriptional circuit that sustains the pluripotent state before subsequent methylation silencing during differentiation.

The lineage-specific LINE L1Md showed TDG-dependent cytosine modification dynamics in MEFs. This family clusters in two classes: the first class presenting a mainly hypermethylated 5'UTR and a second-class showing TDG-dependent methylation dynamics throughout their ORFs. We hypothesize that LINEs with hypermethylated 5'UTRs are transcriptionally inactive whereas LINEs characterized by a highly dynamic 5hmC/5fC/5caC profile following transcriptional directionality, peaking at the beginning of the first ORF and diminishing toward the 3'UTR, are active. Accordingly, these L1Mds are characterized by high Pol II, H2A.Z and H3K4me1 levels. These intact LINEs also exhibited a non-random genomic distribution being concentrated around gene clusters. The biological significance of

this genomic distribution is not understood, but could be implicated in gene regulation and genome organization given that LINEs have been implicated in several fundamental processes such as differentiation and development (Speek 2001; Nigumann et al. 2002; Matlik et al. 2006; Slotkin and Martienssen 2007; Faulkner et al. 2009). Accordingly, DNA methylation dynamics at L1Md was not observed in ESCs suggesting that it takes place only during differentiation.

Another important aspect of this study is the identification of CA methylation enrichment at simple repeats. Our data show that the densest CA repeats are preferentially methylated in ESCs, but hydroxymethylated and dynamically formyl/carboxyl-ated in a TDG-dependent manner in MEFs. The biological significance of the switch from methylation to oxidation during differentiation remains unclear for the moment, but the occurrence of 5mC/5hmC on CA repeats at close distances to the TSS suggests an important role of these elements in the regulation of genome activities such as splicing regulation (Hui et al. 2003a; Hui et al. 2003b). The ability of recombinant TDG protein to excise formylcytosine exclusively in CpG and CpA contexts further validates the implication of CA repeats in genome regulation/organization. Our data support previous observations, obtained by bisulfite sequencing approaches, describing non-CG methylation in brain, oocytes, ESCs, iPSC and flies (Laurent et al. 2010; Tomizawa et al. 2011; Xie et al. 2012; Lister et al. 2013; Ziller et al. 2013) and provide clear evidence for its occurrence in the CAC motif at simple repeats.

We hypothesize that the TDG-dependent dynamic cytosine DNA methylation/oxidation process, specific to both CA repeats and the youngest lineage-specific transposable elements, may constitute a novel epigenetic code with an as yet unknown role in genome organization and function (Fig. 6). Alterations of this code could be associated with disease development. This may be particularly true for tumorigenesis, since strong hypomethylation of the repeats is observed in cancer cells (Howard et al. 2008; Ehrlich 2009; Baba et al. 2010). Since retroelement insertions are major drivers of evolutionary changes within species (Cordaux and Batzer 2009; Burns and Boeke 2012), the observed retroelement methylation dynamics could be strongly implicated in evolution.

Methods

Isolation of primary MEFs

Embryonic fibroblasts were isolated from mouse embryos at embryonic day 10.5 (genetic background C57BL/6) as previously described (Obri et al. 2014). MEFs were kept in culture for no more than 1 month.

Lentiviral knockdown of *Tdg*

shRNA targeting *Tdg* (shTDG-1, 5'-GAACGAAATATGGACGTTCAA-3' and shTDG-2, 5'-CAGGGTTCCCTGAGCTATATG-3') or the control shRNA (5'-CCTAAGGTTAAGGTTAAGTCG-3') were cloned into pLKO.1-blast vector (Addgene). To generate lentiviruses, the transducing vectors were cotransfected into 293T cells using Effectene® Transfection Reagent (Qiagen). The supernatant was harvested at 48 hr after transfection. To generate control and *Tdg*-knockdown cells, MEFs were infected with lentivirus in a 6-well plate. 24 hr after infection, blasticidine (10 μg/ml) was added to the medium (DMEM containing 10% FBS) for selecting infected cells. Cells were selected by blasticidine for 10 days and were split when necessary until being harvested.

RT-qPCR Analysis

Total RNAs were purified from MEFs using standard methods and subjected to reverse transcription using random primers (Promega) and the Superscript II reverse transcriptase (Invitrogen). Real-time quantitative PCR was done with the LightCycler 480 SYBR Green I Master kit (Roche) and the Mastercycler Realplex apparatus (Eppendorf). PCR were performed with the oligonucleotide pairs 5'-GCCAGATGTGCTCAGTTTCC-3' and 5'-CTGCCTCATAGCCTGGATCA-3' for *Tdg* and 5'-GGCTGTATTCCCCTCCATCG-3' and 5'-CCAGTTGGTAACAATGCCATGT-3' for *Actin*. Results were normalized to *Actin*.

5mC/5hmC/5fC/5caC DNA immunoprecipitation assays

DNA immunoprecipitation assays were done as previsously described(Shen et al. 2013). Briefly, 10 μg of DNA was used as input, and 2 μl of 5mC antibody (Active Motif, 39649), 4 μl of 5hmC antibody (Active Motif, 39791), 1 μl of 5fC anti-serum (Yi Zhang) and 0.5 μl of 5caC anti-serum (Yi Zhang) was used to immunoprecipitate modified DNA. DNA and antibodies were incubated at 4°C overnight in a final volume

of 500 µl of DIP buffer (10 mM sodium phosphate pH 7.0, 140 mM NaCl, 0.05% Triton X-100). The bound material was recovered after incubation with 30 µl of blocked protein G Dynabeads (beads washed three times with 1 ml of DIP buffer and incubated for 4 hr minimum with BSA 1 mg ml^{-1} and yeast tRNA 0.5 mg ml^{-1}). The beads were washed three times with 1 ml of DIP buffer, then treated overnight with RNase at 65°C in presence of 300 mM NaCl and then treated 4 hr with proteinase K at 55°C. Immunoprecipitated DNA was purified by phenol-chloroform extraction followed by ethanol precipitation. Four independent DNA immunoprecipitations were pooled for each condition before sequencing analysis.

ChIP assay

H3K4me1 and PolII ChIP experiments were performed as previsously described (Obri et al. 2014). Briefly, 50 µg of sonicated chromatin isolated from sub-confluent MEFs was immunoprecipitated using 1 µl of antibody anti-H3K4me1 (Abcam, ab8898) or 5 µg of antibody anti-Pol2 (Santa Cruz, sc-9001 X). Five independent chromatin immunoprecipitations were pooled for each antibody before sequencing analysis.

ChIP-seq, DIP-seq and computational analyses

ChIP-seq and DIP-seq were performed on an Illumina Hiseq 2500 as single-end 50 base reads following Illumina's instructions. Image analysis and base calling were performed using RTA 1.17.20 and CASAVA 1.8.2. Reads were mapped to the mouse genome (mm9) using bowtie(Langmead et al. 2009) to determine the total numbers of unmappable, multihit, and uniquely mapped reads.

Peak detection was performed using MACS (Zhang et al. 2008) (http://liulab.dfci.harvard.edu/MACS/) under settings where the input fraction was used as negative control. Peak summits detected were annotated using HOMER (http://biowhat.ucsd.edu/homer/ngs/annotation.html). Heatmaps, global clustering and quantitative comparisons of the ChIP-seq/DIP-seq data were performed using seqMINER (Ye et al. 2011) (http://bips.u-strasbg.fr/seqminer/). As reference coordinates, we used the annotated RepeatMasker (RMSK) database of mouse genome (mm9). Tag densities were collected in 50 bp sliding windows spanning 0.5 kb (for SINEs) or 1 kb (for CA repeats) of the peak/repeat element center. To normalize the LTR_{int} and the L1Md lengths, the sequences were divided in 40 bins

and the adjacent 0.5 kb sequences in 10 bins. When clustered, the collected values were subjected to k-means clustering coupled to linear-based normalization.

Generation of Wig tracks for peak visualization

To visualize peaks in the genome browser, we generated wig track files for each data set using WigMaker. We normalized tag counts in each bin to the total number of uniquely mapped reads (reads per million reads, rpm or reads per ten millions reads, rp10m, as indicated).

Repeat Analysis

Repeat analysis was performed as follow. Reads were aligned to repetitive elements in two passes. In the first pass, reads were aligned to the non-masked mouse reference genome (NCBI37/mm9) using BWA (Li and Durbin 2009) v0.6.2. Positions of the reads uniquely mapped to the mouse genome were cross-compared with the positions of the repeats extracted from UCSC (RMSK table in UCSC database for mouse genome mm9) and reads overlapping a repeat sequence were annotated with the repeat family. In the second pass, reads not mapped or multi-mapped to the mouse genome in the previous pass were aligned to RepBase(Jurka et al. 2005) v18.07 repeat sequences for rodent. Reads mapped to a unique repeat family were annotated with the family name. Finally, we summed up the read counts per repeat family of the two annotation steps.

Dot blot assay

10 ng of denatured oligos (CG-containing oligos : 5'-GCA TCG TAC GGA ATC GCT TCT AGC CGG ACA TTA GCG ATX GAT CGA TC AGG CTC GTA GGT ACT CGA CGG CAA TCG TTA G-3' or $(CA)_9$-containing oligos : 5'-CTA ACG ATT GCC GTC GCA CAC ACA XAC ACA CAC AGA TCG CTA ATG TCC GC-3'; X = C, 5mC, 5hmC, 5fC or 5caC) were spotted onto a positive charged membrane (Amersham Hybond^{TM}-XL). Membrane was then baked at 80° and blocked for 1 hour with 5% non-fat milk in TBS containing 0.1% Tween-20 (TBST). Membranes were then incubated overnight with 1:500 dilution of 5mC, 5hmC, 5fC or 5caC antibodies. After

three rounds of washing with the blocking solution, membranes were incubated with 1:20,000 dilution of HRP-conjugated anti-mouse (for 5mC) or anti-rabbit (for 5hmC, 5fC and 5caC) IgG secondary antibody. The membranes were then washed with TBST and treated with ECL.

TDG purification

The mouse His-tagged TDG protein was cloned in a pET28b vector and expressed in the BL21-CodonPlus-RIL-pLysS (Stratagene) strain. A 1 L culture was grown in LB medium at 37°C until D_{600} of 0.5 was reached before induction with 1 mM IPTG for 3 hr at 25°C. Cells were lysed in 15 mL of a buffer containing 20 mM Tri-HCl pH 7.65, 500 mM NaCl, 10% glycerol, 0.01% NP40, 20 mM Imidazole, 0.2 mM PMSF and protease inhibitor cocktail tablets (Roche) in the presence of lysozyme at 1 mg/mL and sonicated on ice. The clarified supernatant was applied to His-Tag Purification Resin (Roche), washed with 50 mM Imidazole and eluted with 300 mM Imidazole using a buffer containing 10 mM Tri-HCl pH 7.65, 150 mM NaCl, 10% glycerol, 0.01% NP40. The eluate fraction was diluted two times with sodium phosphate buffer (50 mM sodium phosphate pH 7, 1 mM DTT, 1 mM EDTA), incubated with SP sepharose fast flow bead (GE Healthcare), extensively washed with sodium phosphate buffer containing 100 mM NaCl and elutated with sodium phosphate buffer containing 500 mM NaCl. The eluate fraction was desalted with PD-10 Sephadex G-25 columns (GE Healthcare) equilibrated with TGEN buffer (20 mM Tris-HCl pH 7.65, 10% glycerol, 3 mM $MgCl_2$, 0.1 mM EDTA, 0.01% NP40).

Glycosylase assay

The DNA substrates for enzymatic activity assays were prepared by annealing equimolar amounts of the following oligonucleotides (top_78mer, 5' CTA ACG ATT GCC GTC GAG TAC CTA CGA GCC TGA TCG ATC XAT CGC TAA TGT CCG GCT AGA AGC GAT TCC GTA CGA TGC 3' ; X= G, A, T or C, and bottom_78mer, 5' GCA TCG TAC GGA ATC GCT TCT AGC CGG ACA TTA GCG AT5fC YAT CGA TCA GGC TCG TAG GTA CTC GAC GGC AAT CGT TAG 3'; Y=G, A, T or C;) in a buffer containing 10 mM Tris-HCl (pH 7.5), 1 mM EDTA and 100 mM NaCl. DNA substrates were 5'-end labelled on the bottom strand (bottom_78mer) with [γ-^{32}P]ATP and T4 polynucleotide kinase. Reaction mixtures (10 µL) containing 20 mM Tris-HCl pH 7.65, 50 mM NaCl, 5% glycerol, 1 mM DTT, 0.1 µg/µl BSA and 2 nM of

end-labeled substrates was incubated 15 minutes at 37°C with the indicated concentration of TDG (from 5 to 100 nM). The reaction was stopped by adding 1 µL of 1 M NaOH and 10 µL formamide buffer (90% formamide, 10 mM EDTA, 0.1% blue bromophenol). The mixture was heated 5 min at 95°C before loading on a 12% denaturing polyacrylamide gel.

Published datasets

To calculate normalized density of H2A.Z at L1Md in Figure 4f, we used our previously published data sets obtained in MEFs and deposited in GEO under accession number GSE51579 (Obri et al. 2014). To determine the relative enrichment for each cytosine modification at repeat families in control and *Tdg*-deficient ESCs. We downloaded data sets deposited in GEO under accession number GSE42250 (Shen et al. 2013) and performed an independent analysis of reads as described in the "repeat analysis" section.

Accession Numbers

The ChIP-seq and DIP-seq datasets obtained in MEFs have been deposited in Gene Expression Omnibus (GEO) under the accession number XXX.

Acknowledgments

We thank Irwin Davidson for critical reading of the manuscript and Yi Zhang and Shen Li for kindly providing us with anti-5fC and anti-5caC antibodies. This work was supported by institutional funds from CNRS, INSERM, Université de Strasbourg (UDS), Université de Grenoble Alpes and by grants from, INCA (INCa_4496), INCA (INCa_4454), ANR (VariZome, contract n° ANR-12-BSV8-0018-01; Nucleoplat, contract n° NT09_476241), the Association pour la Recherche sur le Cancer, La Fondation pour la Recherche Medicale, La Ligue Nationale contre le Cancer Equipe labellisée (A.H. and S.D.), the European Community's Grant agreement number 289611 (''HEM_ID'') to S.D. C.P. acknowledges the Fondation pour la Recherche

Médicale for financial support. A.I. acknowledges the Libyan Ministry of Higher Education and Scientific Research for financial support.

References

Baba Y, Huttenhower C, Nosho K, Tanaka N, Shima K, Hazra A, Schernhammer ES, Hunter DJ, Giovannucci EL, Fuchs CS et al. 2010. Epigenomic diversity of colorectal cancer indicated by LINE-1 methylation in a database of 869 tumors. *Molecular cancer* **9**: 125.

Burns KH, Boeke JD. 2012. Human transposon tectonics. *Cell* **149**(4): 740-752.

Castro-Diaz N, Ecco G, Coluccio A, Kapopoulou A, Yazdanpanah B, Friedli M, Duc J, Jang SM, Turelli P, Trono D. 2014. Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes & development* **28**(13): 1397-1409.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nature reviews Genetics* **10**(10): 691-703.

Cortellino S, Xu J, Sannai M, Moore R, Caretti E, Cigliano A, Le Coz M, Devarajan K, Wessels A, Soprano D et al. 2011. Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell* **146**(1): 67-79.

de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS genetics* **7**(12): e1002384.

Deaton AM, Bird A. 2011. CpG islands and the regulation of transcription. *Genes & development* **25**(10): 1010-1022.

Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Graf S, Johnson N, Herrero J, Tomazou EM et al. 2008. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature biotechnology* **26**(7): 779-785.

Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR et al. 2012. The UCSC Genome Browser database: extensions and updates 2011. *Nucleic acids research* **40**(Database issue): D918-923.

Edgell MH, Hardies SC, Loeb DD, Shehee WR, Padgett RW, Burton FH, Comer MB, Casavant NC, Funk FD, Hutchison CA, 3rd. 1987. The L1 family in mice. *Progress in clinical and biological research* **251**: 107-129.

Ehrlich M. 2009. DNA hypomethylation in cancer cells. *Epigenomics* **1**(2): 239-259.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5**(6): 435-445.

Estecio MR, Gallegos J, Dekmezian M, Lu Y, Liang S, Issa JP. 2012. SINE retrotransposons cause epigenetic reprogramming of adjacent gene promoters. *Mol Cancer Res* **10**(10): 1332-1342.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41**(5): 563-571.

Gabellini N. 2001. A polymorphic GT repeat from the human cardiac Na+Ca2+ exchanger intron 2 activates splicing. *Eur J Biochem* **268**(4): 1076-1083.

Gebhardt F, Zanker KS, Brandt B. 1999. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *J Biol Chem* **274**(19): 13176-13180.

Gehring M, Bubb KL, Henikoff S. 2009. Extensive demethylation of repetitive elements during seed development underlies gene imprinting. *Science* **324**(5933): 1447-1451.

Guo JU, Su Y, Shin JH, Shin J, Li H, Xie B, Zhong C, Hu S, Le T, Fan G et al. 2014. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat Neurosci* **17**(2): 215-222.

Hashimoto H, Pais JE, Zhang X, Saleh L, Fu ZQ, Dai N, Correa IR, Jr., Zheng Y, Cheng X. 2014. Structure of a Naegleria Tet-like dioxygenase in complex with 5-methylcytosine DNA. *Nature* **506**(7488): 391-395.

He YF, Li BZ, Li Z, Liu P, Wang Y, Tang Q, Ding J, Jia Y, Chen Z, Li L et al. 2011. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**(6047): 1303-1307.

Howard G, Eiges R, Gaudet F, Jaenisch R, Eden A. 2008. Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. *Oncogene* **27**(3): 404-408.

Hui J, Reither G, Bindereif A. 2003a. Novel functional role of CA repeats and hnRNP L in RNA stability. *RNA* **9**(8): 931-936.

Hui J, Stangl K, Lane WS, Bindereif A. 2003b. HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. *Nat Struct Biol* **10**(1): 33-37.

Ichiyanagi K. 2013. Epigenetic regulation of transcription and possible functions of mammalian short interspersed elements, SINEs. *Genes Genet Syst* **88**(1): 19-29.

Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr AR, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R, Bird AP. 2010. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet* **6**(9): e1001134.

Ito S, D'Alessio AC, Taranova OV, Hong K, Sowers LC, Zhang Y. 2010. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature* **466**(7310): 1129-1133.

Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, He C, Zhang Y. 2011. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**(6047): 1300-1303.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**(1-4): 462-467.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3): R25.

Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J et al. 2010. Dynamic changes in the human methylome during differentiation. *Genome Res* **20**(3): 320-331.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD et al. 2013. Global epigenomic reconfiguration during mammalian brain development. *Science* **341**(6146): 1237905.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**(7271): 315-322.

Lunyak VV, Prefontaine GG, Nunez E, Cramer T, Ju BG, Ohgi KA, Hutt K, Roy R, Garcia-Diaz A, Zhu X et al. 2007. Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* **317**(5835): 248-251.

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

Matlik K, Redik K, Speek M. 2006. L1 antisense promoter drives tissue-specific transcription of human genes. *J Biomed Biotechnol* **2006**(1): 71753.

Medstrand P, van de Lagemaat LN, Mager DL. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res* **12**(10): 1483-1495.

Mouse Genome Sequencing C Waterston RH Lindblad-Toh K Birney E Rogers J Abril JF Agarwal P Agarwala R Ainscough R Alexandersson M et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915): 520-562.

Naylor LH, Clark EM. 1990. d(TG)n.d(CA)n sequences upstream of the rat prolactin gene form Z-DNA and inhibit gene transcription. *Nucleic Acids Res* **18**(6): 1595-1601.

Nigumann P, Redik K, Matlik K, Speek M. 2002. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics* **79**(5): 628-634.

Obri A, Ouararhni K, Papin C, Diebold ML, Padmanabhan K, Marek M, Stoll I, Roy L, Reilly PT, Mak TW et al. 2014. ANP32E is a histone chaperone that removes H2A.Z from chromatin. *Nature* **505**(7485): 648-653.

Okano M, Bell DW, Haber DA, Li E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**(3): 247-257.

Pravica V, Asderakis A, Perrey C, Hajeer A, Sinnott PJ, Hutchinson IV. 1999. In vitro production of IFN-gamma correlates with CA repeat polymorphism in the human IFN-gamma gene. *Eur J Immunogenet* **26**(1): 1-3.

Shen L, Wu H, Diep D, Yamaguchi S, D'Alessio AC, Fung HL, Zhang K, Zhang Y. 2013. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. *Cell* **153**(3): 692-706.

Shimajiri S, Arima N, Tanimoto A, Murata Y, Hamada T, Wang KY, Sasaguri Y. 1999. Shortened microsatellite d(CA)21 sequence down-regulates promoter activity of matrix metalloproteinase 9 gene. *FEBS Lett* **455**(1-2): 70-74.

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* **8**(4): 272-285.

Sookdeo A, Hepp CM, McClure MA, Boissinot S. 2013. Revisiting the evolution of mouse LINE-1 in the genomic era. *Mob DNA* **4**(1): 3.

Speek M. 2001. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol Cell Biol* **21**(6): 1973-1985.

Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L et al. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**(5929): 930-935.

Takayama S, Dhahbi J, Roberts A, Mao G, Heo SJ, Pachter L, Martin DI, Boffelli D. 2014. Genome methylation in D. melanogaster is found at specific short motifs and is independent of DNMT2 activity. *Genome Res* **24**(5): 821-830.

Tomizawa S, Kobayashi H, Watanabe T, Andrews S, Hata K, Kelsey G, Sasaki H. 2011. Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-CpG methylation in oocytes. *Development* **138**(5): 811-820.

Walsh CP, Chaillet JR, Bestor TH. 1998. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature genetics* **20**(2): 116-117.

Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL, Ren B. 2012. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. *Cell* **148**(4): 816-831.

Ye T, Krebs AR, Choukrallah MA, Keime C, Plewniak F, Davidson I, Tora L. 2011. seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic acids research* **39**(6): e35.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**(9): R137.

Ziller MJ, Gu H, Muller F, Donaghey J, Tsai LT, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE et al. 2013. Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**(7463): 477-481.

Figure legends

Figure 1. Preferential accumulation of 5mC, 5hmC, 5fC and 5caC at repetitive sequences in MEFs. (a) Percentages of uniquely mapped and multihit reads. **(b)** Percentages of peaks overlapping with repetitive elements using the UCSC RepeatMasker database. **(c)** Venn diagrams showing the overlap between 5mC, 5hmC, 5fC and 5caC peaks in control (shSCR) and *Tdg*-deficient MEFs (shTDG).

Figure 2. DNA methylation profiles at the evolutionarily youngest IAP-LTR family in MEFs and ESCs. (a) RepeatMasker database distinguishes within the LTR families, elements corresponding to the external terminal repeats (LTR_{ext}) from those corresponding to the internal coding region (LTR_{int}). **(b)** Heatmaps of 5mC/5hmC/5fC/5caC levels at full-length LTRs (LTR_{int} > 2 kb) in control and *Tdg*-deficient MEFs. LTR retro-elements were ranked by families. **(c)** Distribution of LTR classes in mouse genome (total or full-length retro-elements). **(d)** Average conservation scores (black columns) and CG density (number of CG dinucleotides per 100 bp, red columns) of LTR families within both LTR_{ext} (upper panel) and LTR_{int} (lower panel) regions. **(e)** Relative enrichment for each cytosine modification in MEFs (left panel) and ESCs (right panel) for the indicated LTR families (LTR_{ext} regions, upper panel and LTR_{int} regions, lower panel).

Figure 3. DNA methylation profiles of mouse-specific SINEs (a) Heatmaps of 5mC/5hmC/5fC/5caC levels at mouse-specific SINE families B1m and B2m in control and *Tdg*-deficient MEFs. Within each family, elements were ranked by conservation scores. **(b)** The methylation level of SINE families correlates with their CG density and their evolutionarily conservation. Average conservation scores (black columns) and CG density (number of CG dinucleotides per 100 bp, red columns) of different SINE families. **(c)** Normalized density of 5mC, 5hmC, 5fC and 5caC signals at mouse-specific SINEs in control and *Tdg*-deficient MEFs. **(d)** Relative enrichment for each cytosine modification at SINEs in control and *Tdg*-deficient MEFs (upper panel) and ESCs (lower panel). Note that mouse-specific SINEs are specifically hydroxymethylated in ESCs. **(e)** Genome browser views indicating that 5mC peaks overlap with mouse-specific SINEs in control and *Tdg*-deficient MEFs (upper panel).

Hypermethylated SINEs are concentrated at close distances around CGIs (zooms 1, 2 and 3, lower panel). **(f)** Dot plot of conservation scores to TSS distances for each individual mouse-specific SINE. **(g)** Average distances to TSS of mouse-specific SINEs in function of their conservation scores. **(h)** Diagram illustrating the relationship between DNA methylation, CG density and distance to TSS for mouse-specific SINEs.

Figure 4. Distinct TDG-dependent DNA methylation patterns of full-length LINEs-L1Md in MEFs and ESCs. (a) Percentage of L1Md LINEs in mouse genome (total elements or full-length LINEs, length > 5 kb). **(b)** Average conservation scores (black columns) and CG density (number of CG dinucleotides per 100bp, red columns) of LINEs. **(c)** Heatmaps of 5mC/5hmC/5fC/5caC levels at full-length L1Md in control and *Tdg*-deficient MEFs. The L1Md database was divided in two groups in function of their orientation in mouse genome (positive or negative-sense). Tags counts clustering reveals two distinct clusters in each group (clusters 1 and 2 for positive-sense L1Md, and clusters 3 and 4 for negative-sense L1Md). **(d)** Normalized density of 5mC, 5hmC, 5fC and 5caC signals at L1Md clusters in control and *Tdg*-deficient MEFs reveal two distinct profiles. The first profile corresponds to L1Md containing a hypermethylated 5' UTR region (cluster 1, for positive-sense L1Md, and cluster 3 for negative-sense L1Md). The second profile corresponds to L1Md showing a TDG-dependent accumulation of cytosine modification along their coding sequence (cluster 2 for positive-sense L1Md, and cluster 4 for negative-sense L1Md). **(e)** Relative enrichment for each cytosine modification at different LINE families in control and *Tdg*-deficient MEFs (left panel) and ESCs (right panel). **(f)** Normalized density of H3K4me1, H2A.Z and Pol2 in MEFs at positively (upper panel) and negatively (lower panel) orientated L1Md in function of their cytosine modification pattern (hypermethylated 5' UTR, clusters 1 and 3, and hydroxymethylated coding sequence regulated by TDG, clusters 2 and 4).

Figure 5. DNA methylation patterns of simple CA repeats. (a) Heatmaps of 5mC/5hmC/5fC/5caC levels of simple repeats in control and *Tdg*-deficient MEFs showing a specific enrichment of oxidized forms of 5mC at CA repeats. Tags were counted within 1 kb around the simple repeats center. **(b)** Heatmaps of 5mC/5hmC/5fC/5caC levels at CA repeat in control and *Tdg*-deficient MEFs,

elements were ranked by conservation scores. **(c)** Normalized density of 5mC, 5hmC, 5fC and 5caC signals showing a *Tdg*-deficient-dependent accumulation of 5fC and 5caC at CA repeats in MEFs. **(d)** Relative enrichment of each cytosine modification at different simple repeat families in control and *Tdg*-deficient MEFs (upper panel) and ESCs (lower panel). The A/T family encompasses all simple repeats containing exclusively adenine and/or thymine in their repeat motif sequence. **(e)** *In vitro* glycosylase assays revealing that the recombinant TDG protein can excise formylcytosine exclusively in CpG and CpA context. **(f)** Average distances to TSS of CA repeats in function of their conservation scores. **Inset:** Dot plot of conservation scores to TSS distances for each CA repeat element. **(g)** Diagram illustrating the relationship between DNA hydroxymethylation, CA density and distance to TSS for CA repeats.

Figure 6. Combinatorial DNA methylation code at repetitive elements in both ESCs and MEFs. Schematic diagram describing DNA methylation/oxidation patterns of repetitive elements in both ESCs and MEFs, which mainly affect evolutionary conserved mouse-specific transposable elements (IAP-LTRs, SINEs B1m/B2m and L1Md-LINEs) and CA repeats. These repeated elements are clustered at specific locations in the mouse genome and TET/TDG are implicated in the regulation of their unique DNA methylation/oxidation signatures and dynamics.

Supplemental Figures

Supplemental Figure 1. Genome-wide distribution of 5mC/5hmC/5fC/5caC in control and *Tdg*-deficient MEFs. (a) RT-qPCR (upper panel) and immunoblotting (lower panel) analyses of TDG expression levels in control (shSCR) and *Tdg*-deficient (shTDG) MEFs. **(b)** Genome browser views indicating 5mC, 5hmC, 5fC and 5caC peaks in control and *Tdg*-deficient MEFs. **(c)** Normalized density of 5mC, 5hmC, 5fC and 5caC signals showing a specific depletion of cytosine modification at CGIs. Tags were counted within 5 kb around CGIs. **(d-g)** Flowchart of computational analyses used in this study (d and f) and correlation matrix for the repetitive elements (e and g), using uniquely mapped reads (d and e) and including multihits mapped reads (f and g). Heatmaps with hierarchical clustering showing Spearman's rank

correlations between all pair-wise comparisons. Spearman correlations were calculated using the raw read count across all types of repeats analyzed. Blue: shSCR. Grey: shTDG. Note that the 5hmC, 5fC and 5caC profiles were closely clustered.

Supplemental Figure 2. DNA methylation dynamics at CG-rich LTR in MEFs and ESCs. (a) Normalized density of 5mC, 5hmC, 5fC and 5caC signals at IAP LTR retro-elements (LTR_{int} > 2 kb) in control and *Tdg*-deficient MEFs. **(b)** Alignment of ten representative LTRs within ERV1 family (upper panel), IAP family (middle panel) and ERVL family (lower panel). CpG dinucleotides are highlighted in red. The average CG density of selected LTR retro-elements is indicated on the left. **(c)** Dot blot of enrichment in 5mC, 5hmC, 5fC and 5caC at each LTR family in control (blue circles) and *Tdg*-deficient (red circles) in MEFs (left panel) and ESCs (right panel). LTR_{ext} and LTR_{int} regions were analyzed separately. LTR families were sorted by classes. Note that several LTR_{ext} families of the ERV1 class are specifically hypermethylated in MEFs (orange ovals). In contrast, ESCs show a specific enrichment in 5hmC at the majority of LTR_{ext} families of the ERVK class (green ovals). Of note, we observed a strong accumulation of 5caC at these regions in absence of TDG (purple ovals).

Supplemental Figure 3. Correlation between SINE DNA methylation dynamics and evolutionarily age. (a) Heatmaps of 5mC/5hmC/5fC/5caC levels at SINEs. SINE elements were ranked by family. Tags were counted within 500 bp around the SINE center. **(b)** Alignment of ten representative SINEs B1 from the ancestral group (upper panel) and the mouse-specific group (lower panel) showing that recent SINEs B1 (mouse-specific) show a high CG density comparing to ancestral SINEs B1 (common in rodents). **(c)** Dot blot shows enrichment in 5mC, 5hmC, 5fC and 5caC of each SINE family in control (blue circles) and *Tdg*-deficient (red circles) MEFs (left panels) and ESCs (right panels). SINE families were sorted in two groups relatively to their appearance in the rodent lineage, the mouse-specific group and the ancestral group (common in rodents). Note that mouse-specific SINEs are specifically hypermethylated in MEFs but hydroxymethylated in ESCs (green ovals). **(d)** Genome browser views indicating 5mC, 5hmC, 5fC and 5caC distribution in MEFs at genomic region showing a high concentration of mouse-specific SINEs.

Supplemental Figure 4. The TDG-dependent dynamics of cytosine modifications at LINEs correlates with their evolutionarily age. (a) Genome browser views indicating 5mC, 5hmC, 5fC and 5caC peaks in control and *Tdg*-deficient MEFs at three selected L1Md elements representative of each clusters described in Figure 4. **(b)** Dot blot shows enrichment in 5mC, 5hmC, 5fC and 5caC at each LINE family in control and *Tdg*-deficient MEFs (left panels) and ESCs (right panels). LINE families were sorted in two groups relatively to their appearance in the rodent lineage (mouse-specific and ancestral). Elements from the mouse-specific LINE group were sorted according to their evolutionarily age in the mouse lineage. **(c)** L1Md elements are concentrated around gene clusters in mouse genome. Genome browser views indicating the position of full-length L1Md at representative gene clusters *Skint* (upper panel) and *Vmn2R-Olfr* (lower panel).

Supplemental Figure 5. Specific accumulation of 5hmC, 5fC and 5caC at simple repeats containing CAC motif. (a) Dot blot assays showing that 5mC, 5hmC, 5fC and 5caC antibodies specifically recognize 5mC, 5hmC, 5fC and 5caC-containing oligos in both CG and $(CA)_9$ repeat contexts. **(b)** Average CA density (blue) and CG density (red) (number of dinucleotides per 100 bp) of CA repeats, CGIs and the mouse genome. **(c)** Genome browser views showing that 5hmC peaks overlap with CA repeats. The zoom view show the sequence of a strongly hydroxymethylated CA repeat element. CpG, CpA, CpC and CpT dinucleotides are highlighted in green, red, yellow and blue, respectively. **(d)** Genome browser views showing that 5hmC, 5fC and 5caC peaks overlap with CA repeats in control and *Tdg*-deficient MEFs. **(e)** Relative enrichment for each cytosine modification at different simple repeat families in control and *Tdg*-deficient MEFs. The CA, CT, CC and CG families regroup simple repeats containing a cytosine in their repeat motif followed by a A, T, C or G, respectively (upper panel). The CAC, CAG, CAT and CAA families regroup simple repeats containing the CpA dinucleotides in their repeat motif followed by a C, G, T or A respectively (lower panel). **(f)** Heatmaps of 5mC/5hmC/5fC levels at CA repeats in control and *Tdg*-deficient ESCc and MEFs. Tag counts clustering revealed three distinct clusters according to cytosine modification density (high, middle and low enrichment). Tags were counted within 1 kb around the simple repeats center. Note that the strongly methylated CA repeats in ESCs are specifically highly enriched in 5mC oxidized form in MEFs.

Supplemental Figure 6. Similar DNA methylation patterns at major satellites and DNA transposon in MEFs and ESCs.

(a-b) Relative enrichment for each cytosine modification at major satellites (a) and DNA transposons (b) in control and *Tdg*-deficient MEFs (left panels) and ESCs (right panels). Note that major satellites show a unique cytosine modification pattern characterized by a specific 5mC, 5fC and 5caC enrichment. **(c)** Composition of interspersed repeats in the mouse genome (total elements, left panel, and families showing a specific TDG-dependent DNA methylation dynamic during differentiation, right panel).

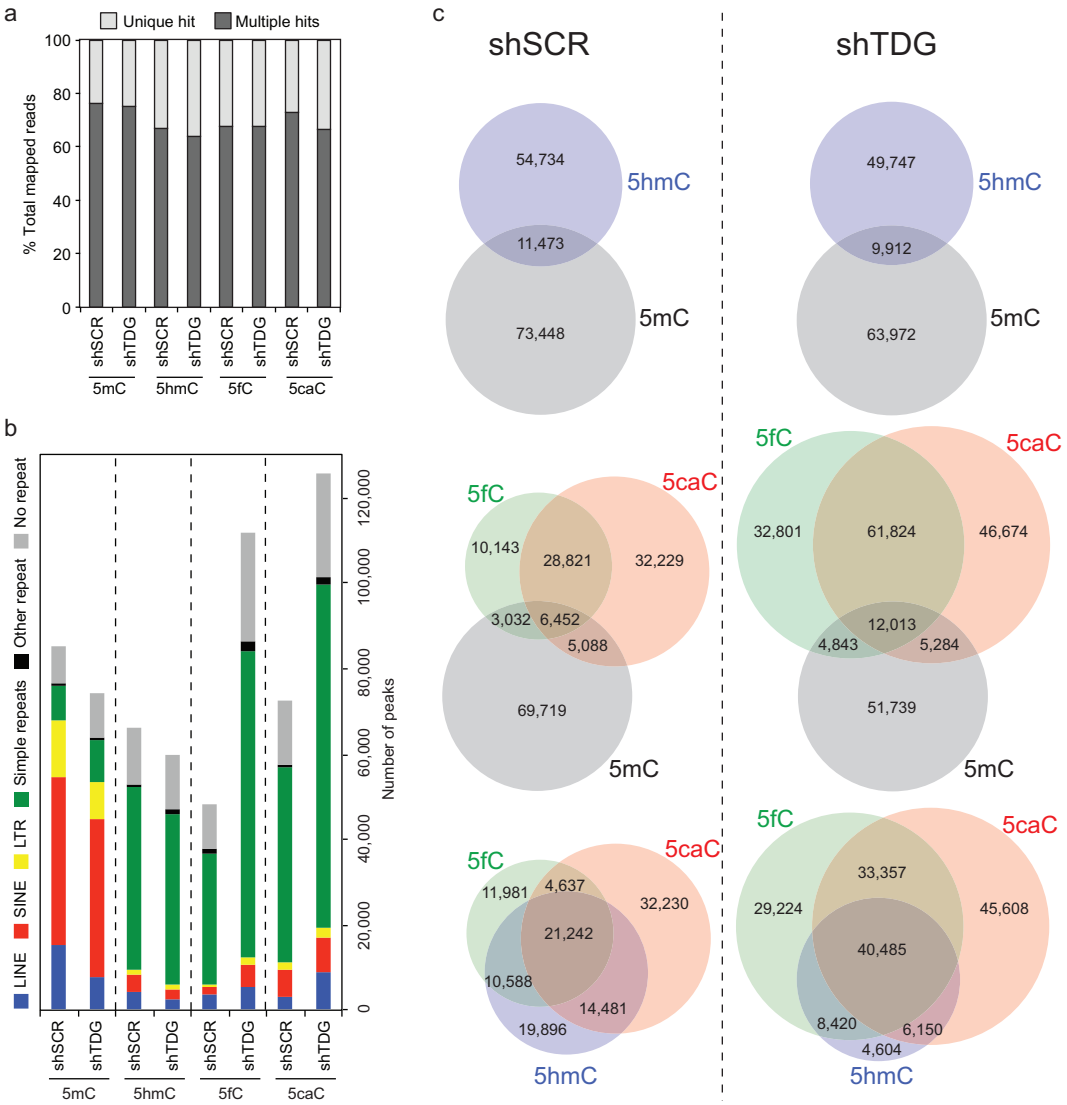

a
Unique hit
Multiple hits
% Total mapped reads
100
80
60
40
20
0
shSCR
shTDG
5mC
5hmC
5fC
5caC
b
No repeat
Other repeat
Simple repeats
LTR
SINE
LINE
Number of peaks
120,000
100,000
80,000
60,000
40,000
20,000
0
c
shSCR
shTDG
54,734
5hmC
11,473
73,448
5mC
49,747
9,912
63,972
5fC
5caC
10,143
28,821
32,229
3,032
6,452
5,088
69,719
32,801
61,824
46,674
12,013
4,843
5,284
51,739
11,981
4,637
32,230
21,242
10,588
14,481
19,896
29,224
33,357
45,608
40,485
8,420
6,150
4,604

Figure 1

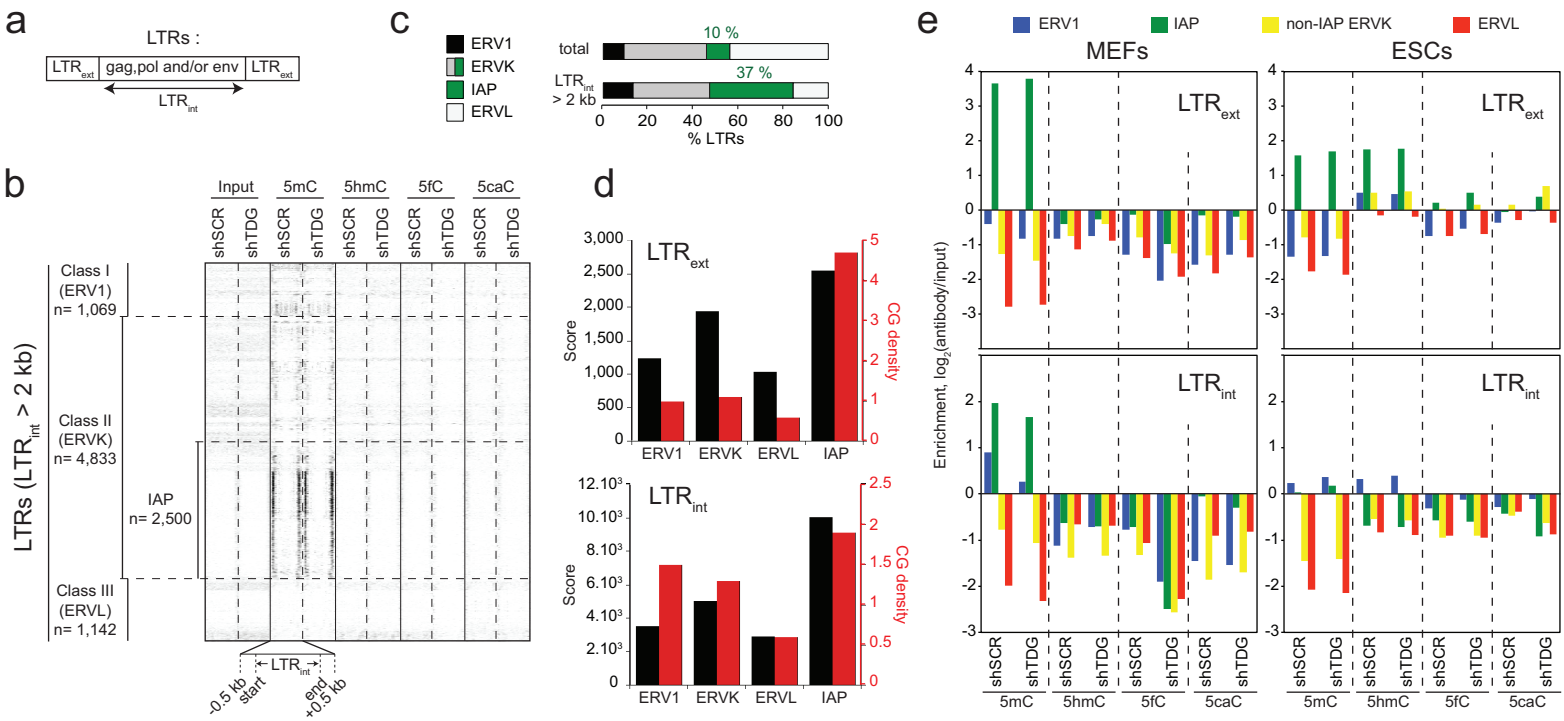

a
LTRs :
LTRext
gag,pol and/or env
LTRext
LTRint
c
ERV1
ERVK
IAP
ERVL
total
10 %
LTRint
> 2 kb
37 %
0
20
40
60
80
100
% LTRs
b
Input
5mC
5hmC
5fC
5caC
shSCR
shTDG
Class I
(ERV1)
n= 1,069
Class II
(ERVK)
n= 4,833
IAP
n= 2,500
Class III
(ERVL)
n= 1,142
LTRs (LTRint > 2 kb)
LTRint
-0.5 kb
start
end
+0.5 kb
d
LTRext
Score
CG density
ERV1
ERVK
ERVL
IAP
LTRint
e
ERV1
IAP
non-IAP ERVK
ERVL
MEFs
ESCs
LTRext
LTRint
Enrichment, log2(antibody/input)
shSCR
shTDG
5mC
5hmC
5fC
5caC

Figure 2

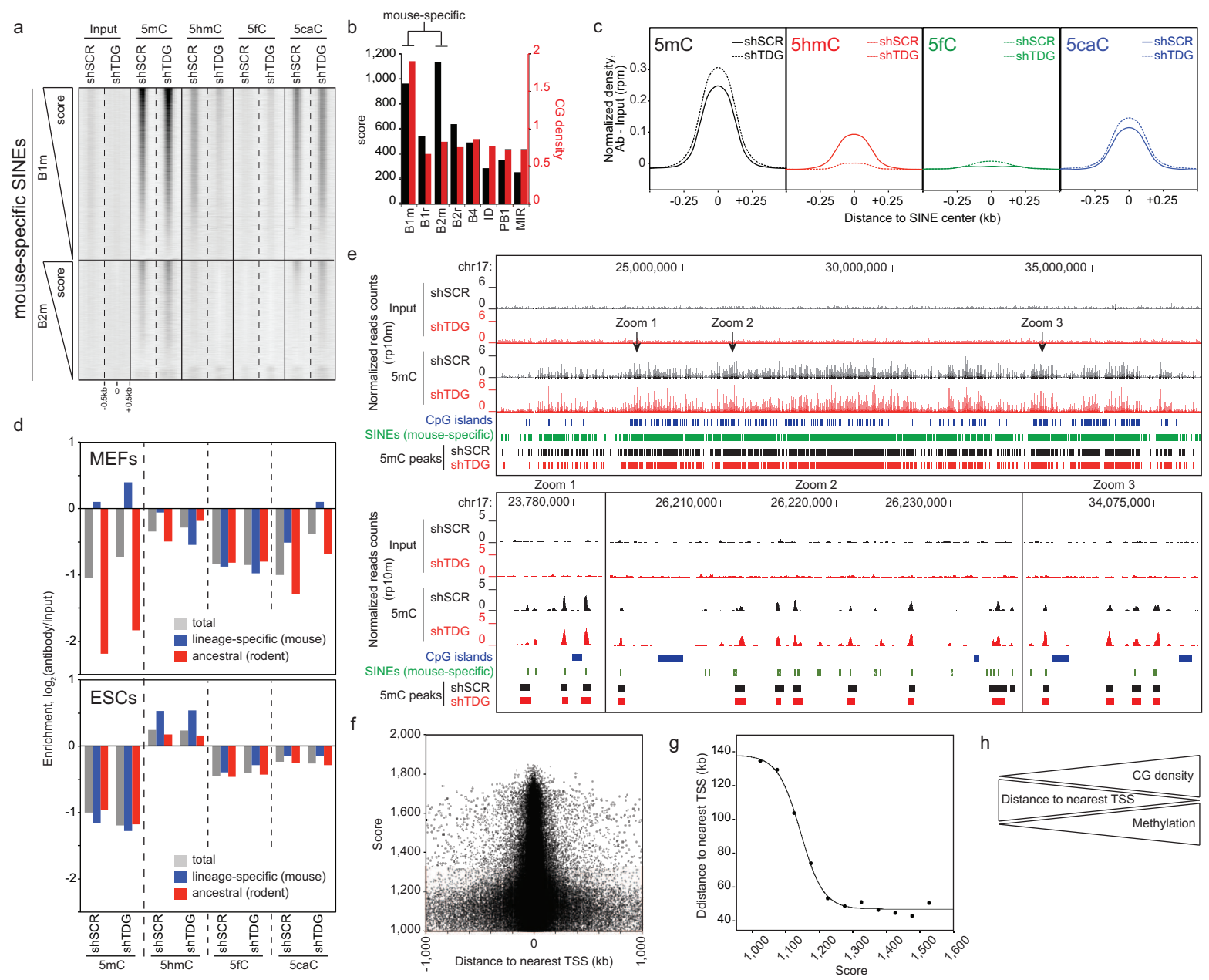

a
Input
5mC
5hmC
5fC
5caC
shSCR
shTDG
mouse-specific SINEs
B1m
B2m
score
-0.5kb
0
+0.5kb
b
mouse-specific
score
CG density
B1m
B1r
B2m
B2r
B4
ID
PB1
MIR
c
5mC
5hmC
5fC
5caC
shSCR
shTDG
Normalized density, Ab - Input (rpm)
Distance to SINE center (kb)
d
MEFs
ESCs
Enrichment, log2(antibody/input)
total
lineage-specific (mouse)
ancestral (rodent)
5mC
5hmC
5fC
5caC
e
chr17:
25,000,000
30,000,000
35,000,000
Zoom 1
Zoom 2
Zoom 3
Input
5mC
Normalized reads counts (rp10m)
CpG islands
SINEs (mouse-specific)
5mC peaks
23,780,000
26,210,000
26,220,000
26,230,000
34,075,000
f
Score
Distance to nearest TSS (kb)
g
Ddistance to nearest TSS (kb)
Score
h
CG density
Distance to nearest TSS
Methylation

Figure 3

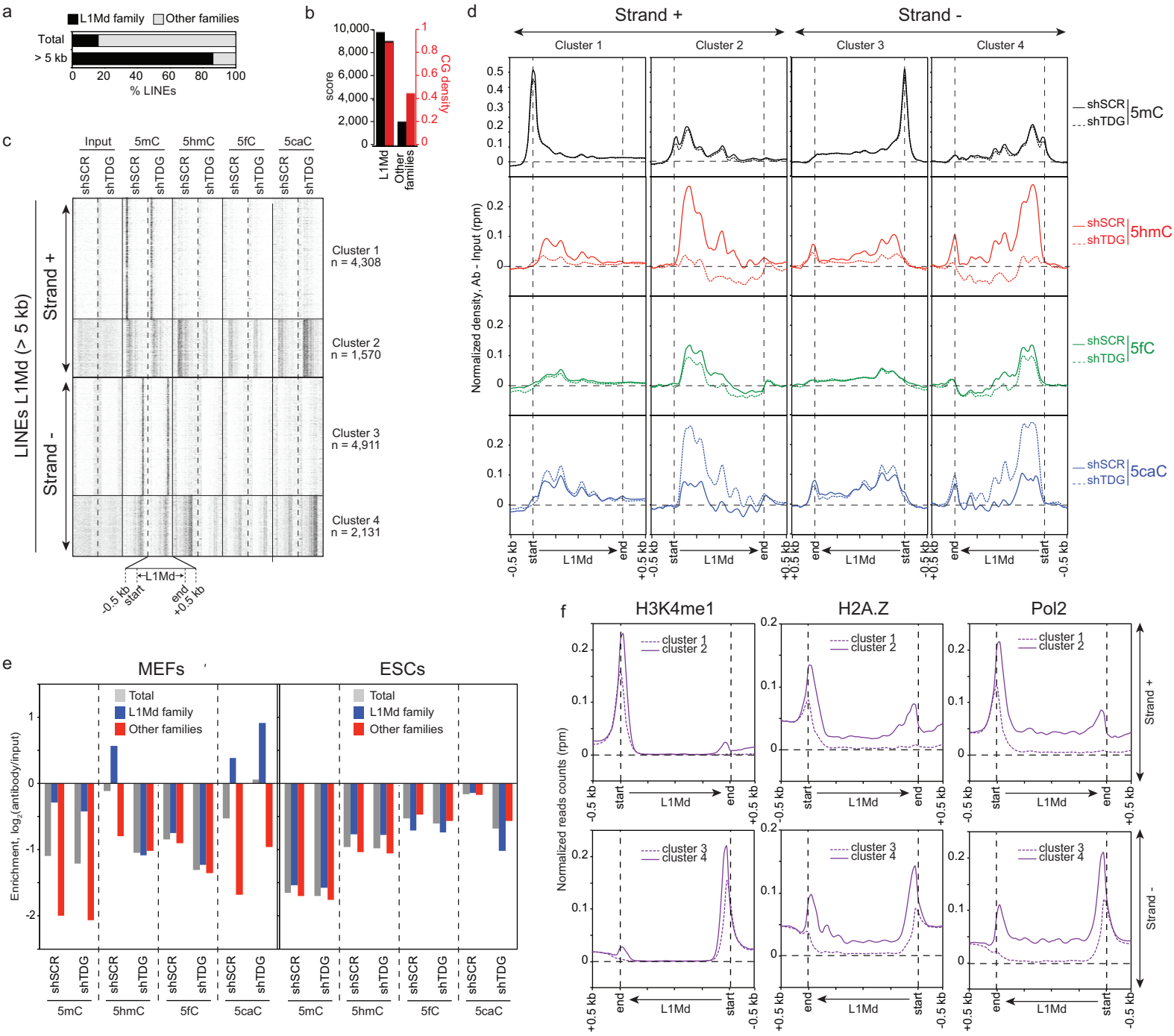

a
L1Md family
Other families
Total
> 5 kb
0 20 40 60 80 100
% LINEs
b
score
CG density
L1Md
Other families
c
Input
5mC
5hmC
5fC
5caC
shSCR
shTDG
LINEs L1Md (> 5 kb)
Strand +
Strand -
Cluster 1
n = 4,308
Cluster 2
n = 1,570
Cluster 3
n = 4,911
Cluster 4
n = 2,131
-0.5 kb
start
L1Md
end
+0.5 kb
d
Strand +
Strand -
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Normalized density, Ab - Input (rpm)
shSCR
shTDG
5mC
5hmC
5fC
5caC
e
MEFs
ESCs
Total
L1Md family
Other families
Enrichment, log2(antibody/input)
f
H3K4me1
H2A.Z
Pol2
cluster 1
cluster 2
cluster 3
cluster 4
Normalized reads counts (rpm)
Strand +
Strand -

Figure 4

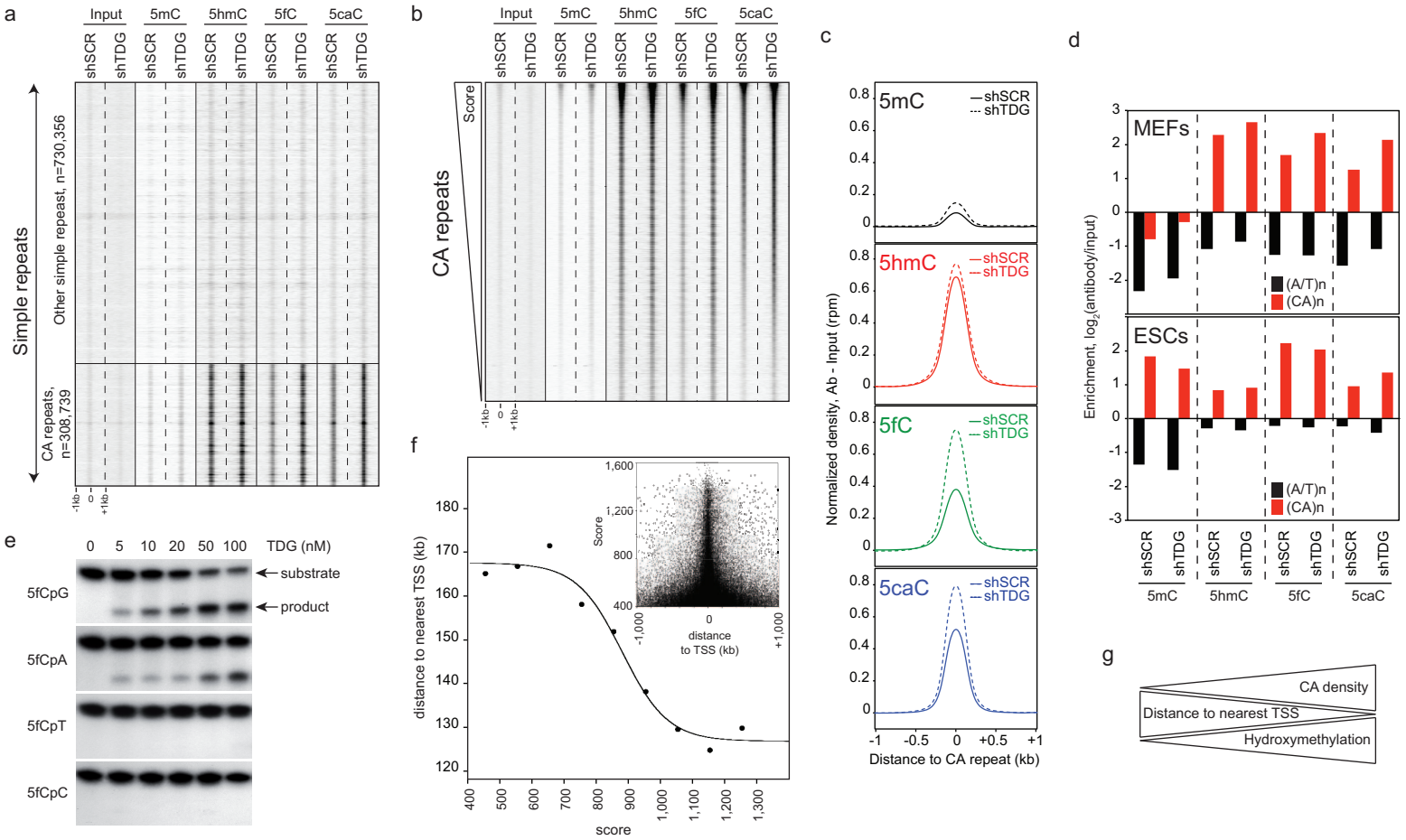

a
Input
5mC
5hmC
5fC
5caC
shSCR
shTDG
Simple repeats
Other simple repeast, n=730,356
CA repeats, n=308,739
-1kb
0
+1kb
b
Score
CA repeats
c
5mC
5hmC
5fC
5caC
shSCR
shTDG
Normalized density, Ab - Input (rpm)
Distance to CA repeat (kb)
d
MEFs
ESCs
(A/T)n
(CA)n
Enrichment, log2(antibody/Input)
e
0
5
10
20
50
100
TDG (nM)
substrate
product
5fCpG
5fCpA
5fCpT
5fCpC
f
distance to nearest TSS (kb)
score
Score
distance to TSS (kb)
g
CA density
Distance to nearest TSS
Hydroxymethylation

Figure 5

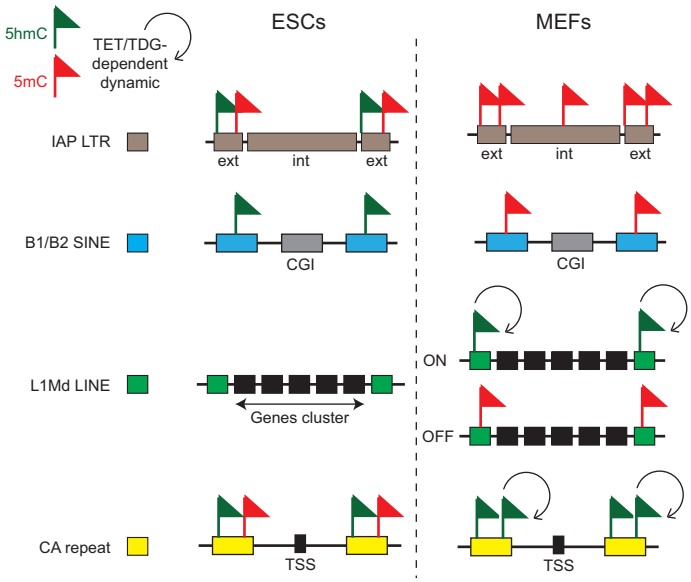
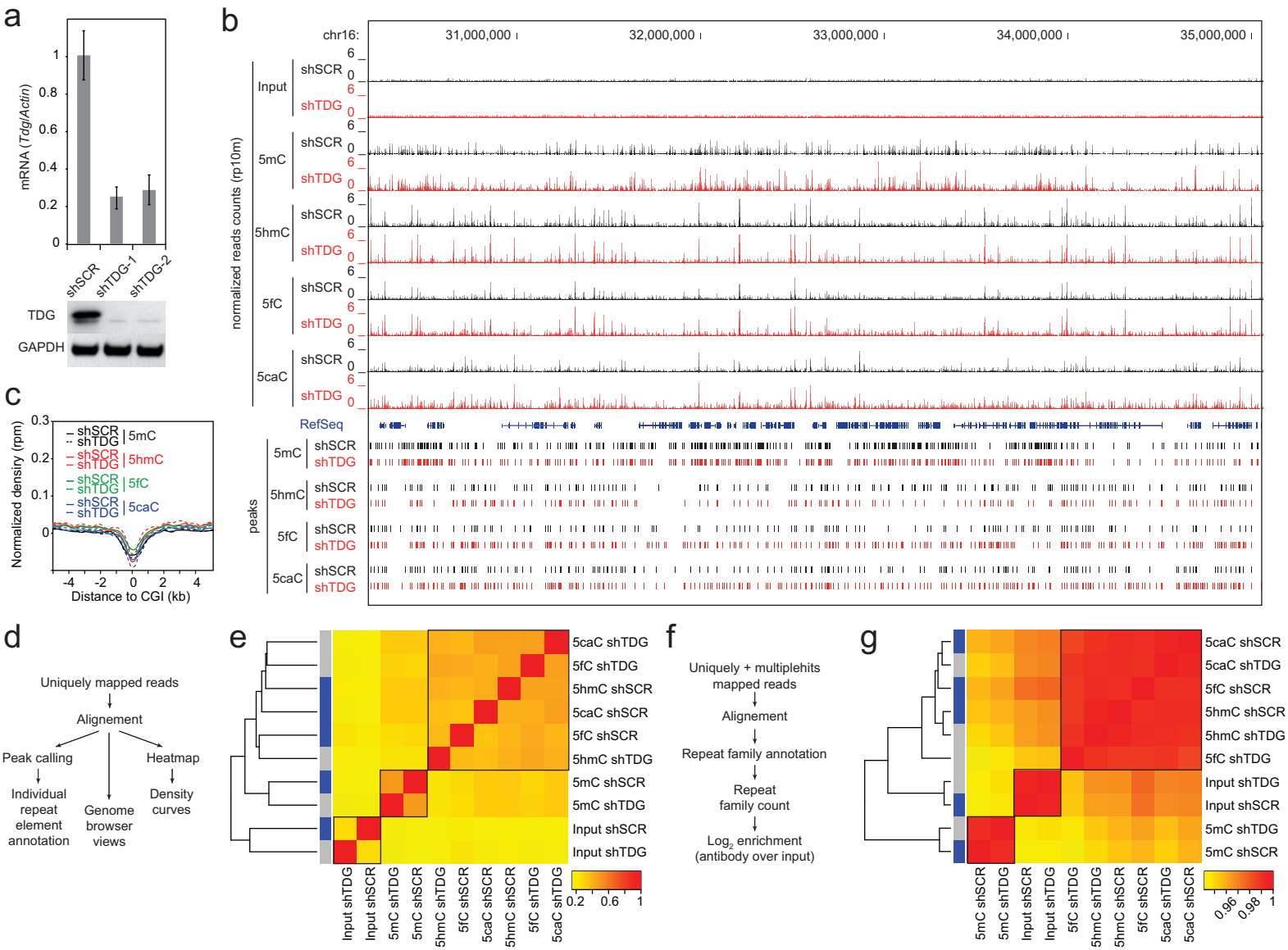
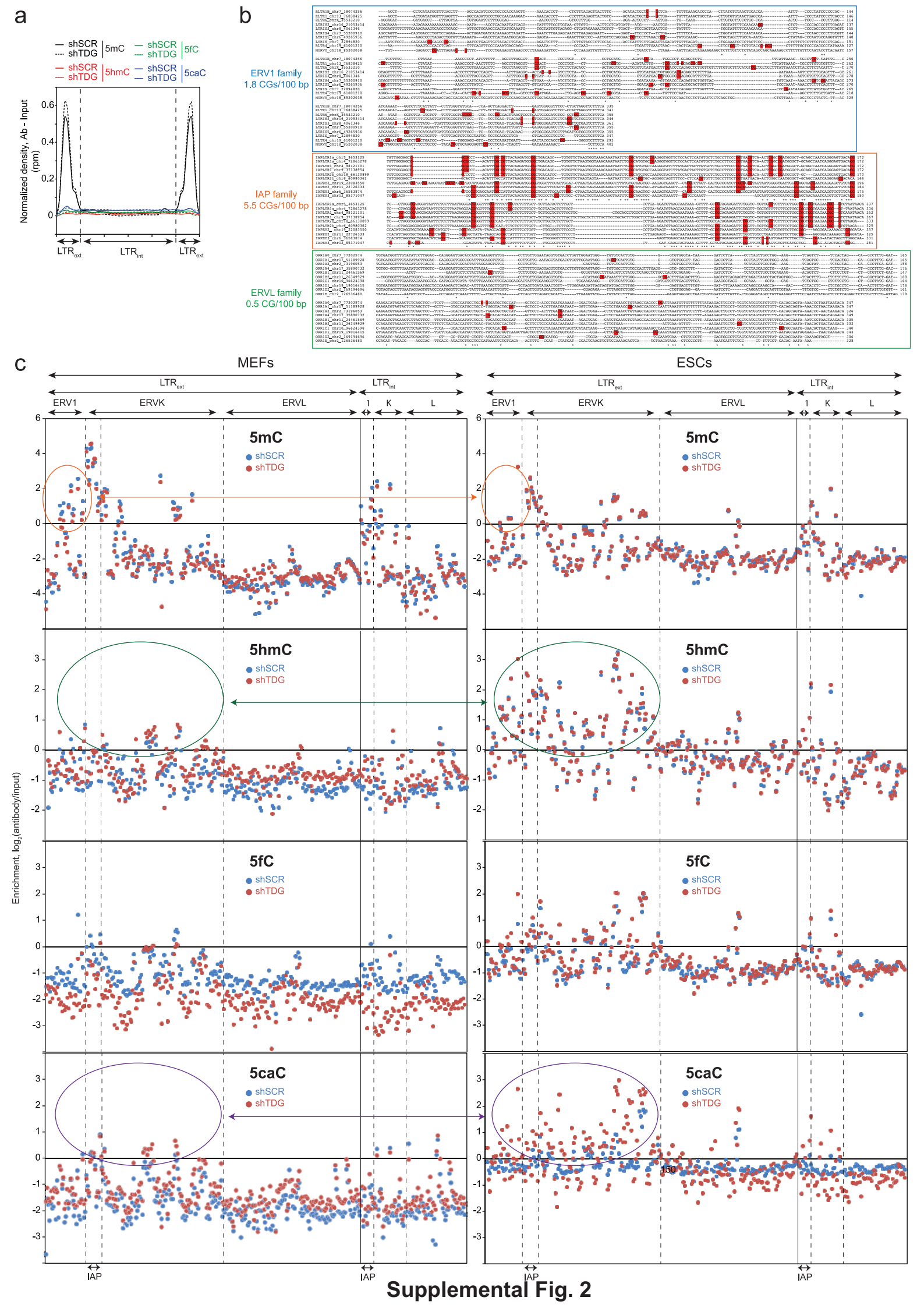

5hmC
5mC
TET/TDG-
dependent
dynamic
ESCs
MEFs
IAP LTR
ext
int
ext
ext
int
ext
B1/B2 SINE
CGI
CGI
ON
L1Md LINE
Genes cluster
OFF
CA repeat
TSS
TSS

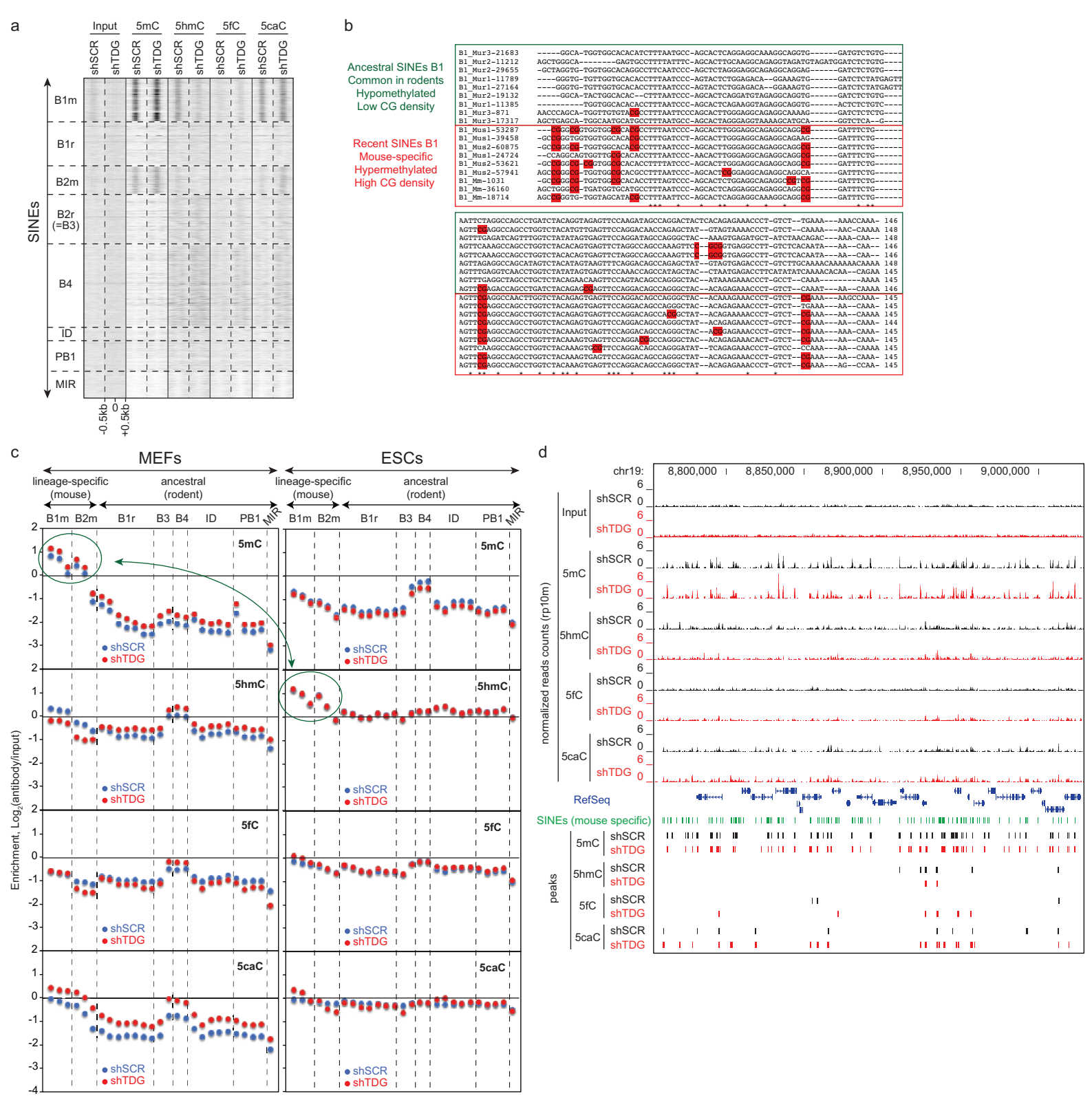
Figure 6



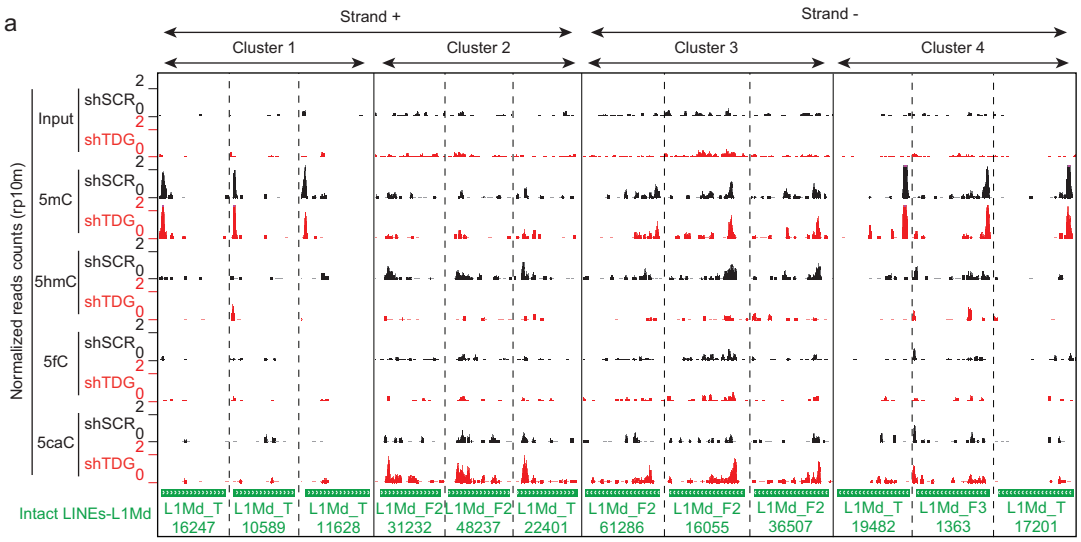
Supplemental Fig. 1

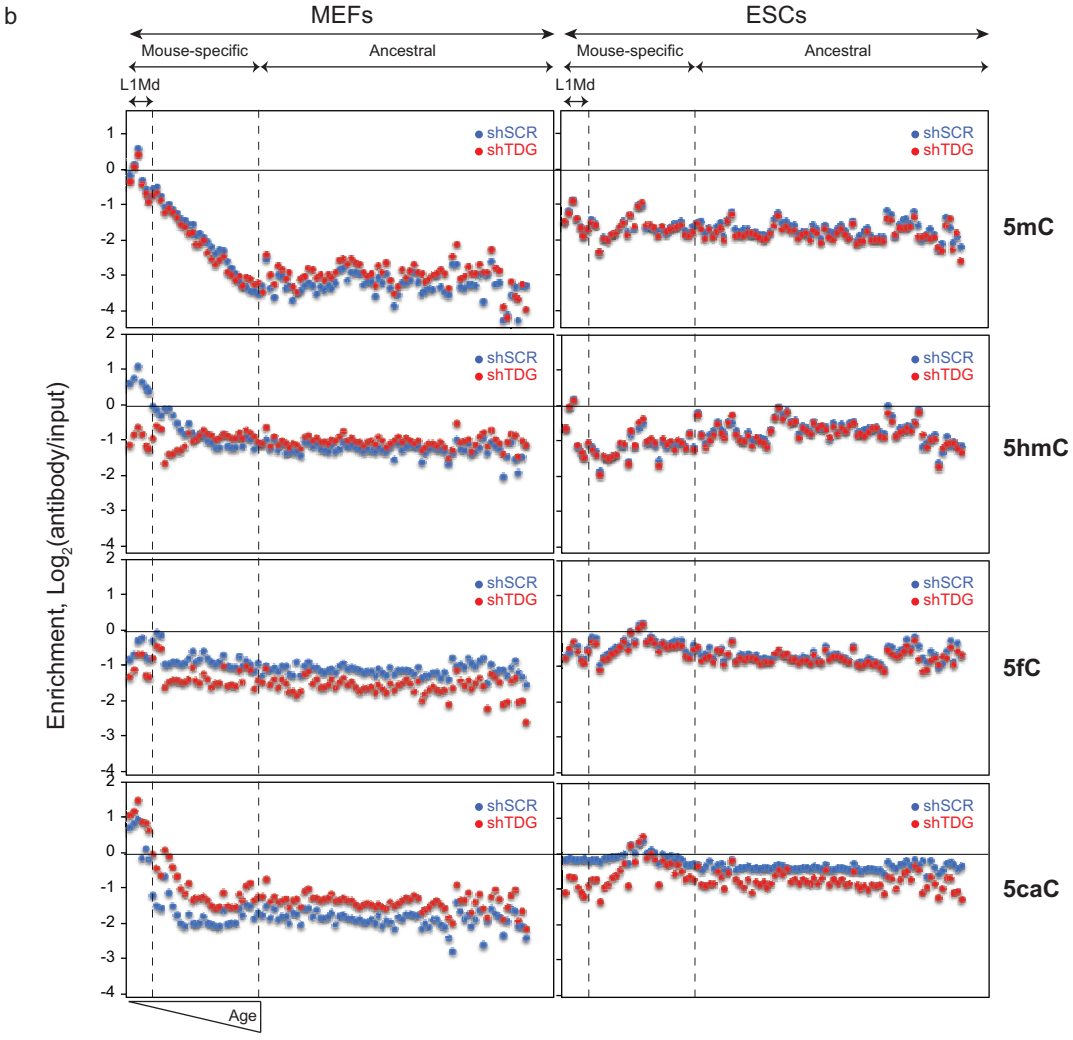


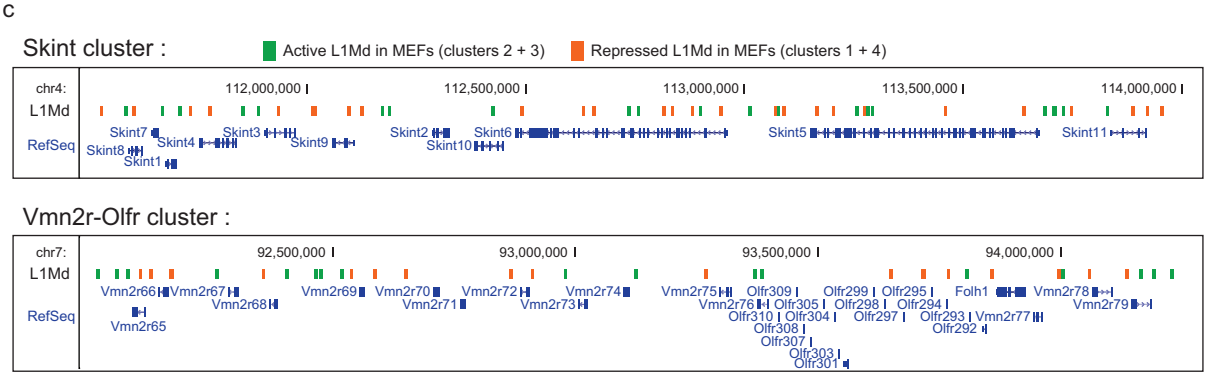
Supplemental Fig. 2



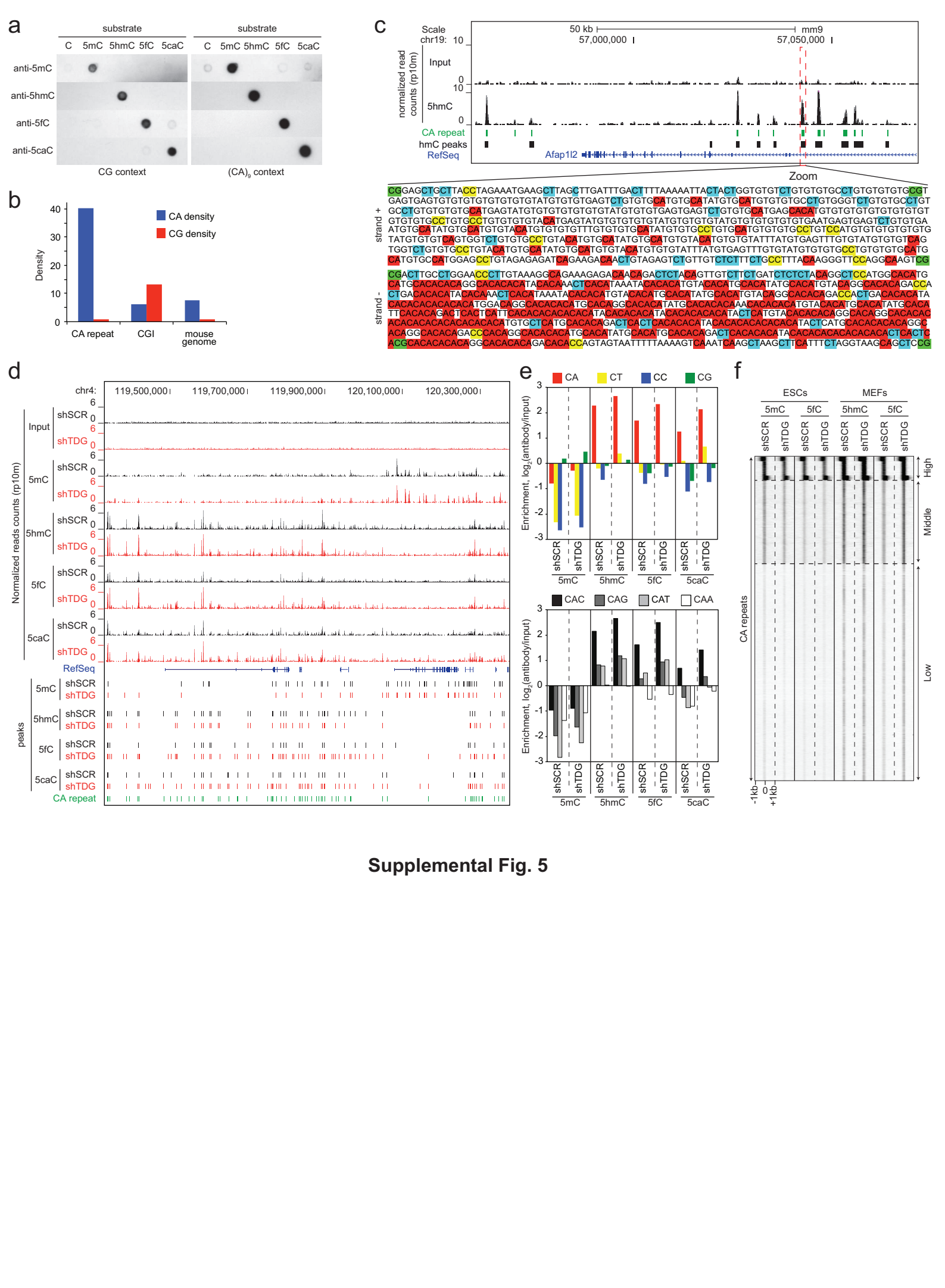
Supplemental Fig. 3





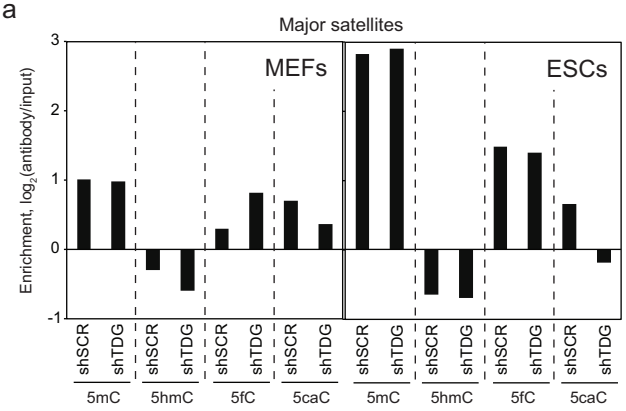


Supplemental Fig. 4

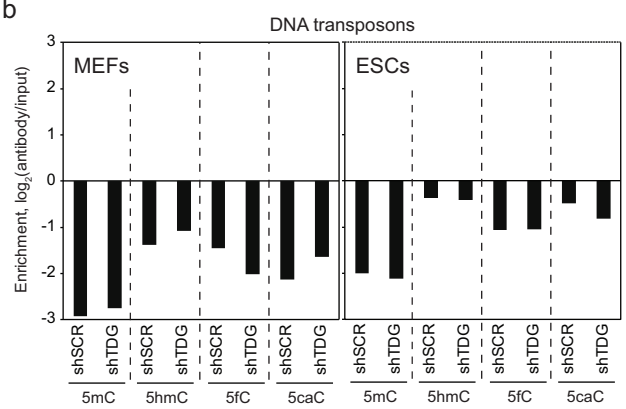


Supplemental Fig. 5

a



b



c

Family	Size (Mb)	% genome	Sub-family	Size (Mb)	% genome
LTR	273.6	10.1	IAP	19.6	0.7
LINEs	522.3	19.3	L1Md	172.4	6.4
SINEs	202.9	7.5	B1m+B2m	45.8	1.7
Simple repeats	63.4	2.3	CA repeats	19.5	0.7
Major Satellites	7	0.3			
DNA transposons	21.8	0.93			

Supplemental Fig. 6

CHAPTER 3
DISCUSSION

3. Discussion

3.1. MBD4 preserve silenced CGI-containing promoters from demethylation

In vivo, MBD4 has a protective function against mutagenesis and tumorigenesis. Mice deficient for the MBD4 gene are viable and fertile but have a predisposition to the development of intestinal tumors due to an increase in C to T transition at the methylated CpG (Millar et al. 2002, Wong et al. 2002). Consistent with these observations, between 26 and 43% of cancers, characterized by microsatellite instability (endometrial cancer, gastric cancer, colorectal cancer or pancreatic cancer) have a mutation in the Mbd4 gene (Riccio et al. 1999, Yamada et al. 2002). The anti-cancer role of MBD4 comes from its putative ability to function as a DNA methylation repair enzyme.

MBD4 is the sole mammalian protein having a methylated DNA binding domain (MBD) associated with a glycosylase domain. In vitro, MBD4 can bind a 5mC in a CpG context via its MBD domain. The methylated cytosine is sensitive to spontaneous deamination that results in producing T and thus to G/T mismatch. The unique architecture of MBD4 allows it to bind both to the substrate and the product of 5mC deamination, and it has been proposed, that the resulted T may be cleaved by the glycosylase domain of MBD4 (Hendrich et al. 1999, Petronzelli et al. 2000a). The significance of this unique domain organization remains elusive. To elucidate the importance of possessing these two domains in MBD4, and thus to decipher in more depth its role in DNA methylation dynamic, we purified the MBD4 complex to define interesting partner(s) of MBD4 with putative regulatory activities or conferring unreported activities to MBD4.

3.1.1 MBD4 interacts with mismatch repair proteins

We have shown that, in cells MBD4 binds DNA mismatch repair proteins (MMR = Mismatch Repair), the MLH1/PMS2 complex and the MSH2/MSH6 complex, and that MLH1 is the most abundant subunit in MBD4 complex. Consistently with our results, it has been earlier shown that MBD4 interacts with MLH1 and DNMT1 and accumulates at DNA damage sites (Ruzov et al. 2009). Interestingly, despite MBD4 has been reported to have a monofunctional glycosylase activity (Hendrich et al. 1999, Petronzelli et al. 2000b), our results show that the MBD4 catalyzes a bifunctional glycosylase/lyase

activity. Interestingly, while the nuclease activity of MBD4 is low and is not sensitive to methylation, the complex MBD4/MLH1 shows intense nuclease activity strongly induced by the methylation, and the biochemical analysis of MBD4 point mutants reveals that the integrity of MBD and glycosylase domains is required for this function. These activities ensure both the removal of the thymine base and cleavage of the DNA phosphate backbone at 3' end of the abasic site. Importantly, in experiments with highly purified recombinant, we could show that this activity is specific for a G/T, induced by DNA methylation, and requires the binding of MBD4 to MLH1 as enhancer of MBD4-AP lyase activity. Such an enhancer function for MLH1 has already been observed for EXO1 and PMS1 proteins, two nucleases implicated in MMR pathway in eukaryotes. Indeed, the physical interaction between MLH1 and EXO1 is required for the endonuclease function of EXO1 in MMR pathway (Dherin et al. 2009). Moreover, a recent structural study has revealed that the strictly conserved C terminus of MLH1 forms part of the PMS1 endonuclease catalytic site (Gueneau et al. 2013). All together, these data define MLH1 as a nuclease effector protein. We, therefore, do not exclude a similar pooling between some amino acids of MBD4 and some of MLH1 to form a complete part of an endonuclease catalytic site.

3.1.2 MBD4 is a bifunctional glycosylase/lyase enzyme when associated to MLH1.

DNA glycosylases, as members of the base excision repair machinery, protect DNA in cells from the damaging effects of oxidation, alkylation and deamination. Whereas the alkylation-induced damages are repaired by mono-functional glycosylases (MPG, Methylpurine glycosylase which remove 3-meA, 3-meG,7-meG and hypoxanthine), the vast majority of oxidation-induced damages is repaired by bifunctional glycosylases [OGG1 (8-OxoG DNA glycosylase 1, that remove 8oxo-G and FaPy opposite C), NTHL1 (Endonuclease III-like1) acts on Tg, FaPyG, 5-hC and 5-hU and NEHL (Endonuclease VIII-like glycosylase1) family members NEHL1, 2 and 3 , act on Tg, FaPyG, FaPyA, 8-oxoG, 5–hU, and 5–hC)] (Jacobs and Schar 2012). Note that glycosylase enzymes implicated in DNA demethylation pathway in plants (DME/ROS family, that acts on mC) show similar bifunctional activity. Monofunctional glycosylase catalyse the cleavage of the glycosidic bond between the sugar and the damaged-base, leading to an AP site. For bifunctionnal enzymes, this activity is concomitant with the cleavage of the phosphodiester bond at the 3' side of the AP site (β elimination), leading to a DNA

single-strand break. The 3'phosphor unsaturated aldehyde (3'-PUA) can be further processed by ARP (Apurinic endonuclease-redox protein) to remove the abasic sugar (δ elimination) which can be filled with a nucleotide via the action of DNA polymerase/ligase.

All enzymes known to repair the third type of deamination-induced damage such as UNG (Uracil-N glycosylase, that act on U and 5-FU), SMUG1 (Single-strand-specific monofunctional uracil DNA glycosylase 1, that act on U, 5–hmU and 5-FU), MBD4 and TDG have been described as mono-functional.

Concerning MBD4, we have observed that enzyme has a bifunctional activity conducting to an abasic site with the concomitant cleavage of the ribose backbone, thanks to the cooperation with MLH1. However, an important question still remains to be answered. How is the sugar eliminated? The presence of Pol l in the MBD4 complex would support the idea that it is the candidate enzyme that could catalyses this reaction and further generates hemi-methylated DNA, the substrate of the UHRF1/DNMT1 tandem. The response to our answer will provide evolutionary insights into how animals and plants have evolved distinct or similar DNA demethylation mechanisms for epigenetic gene regulation.

3.1.3 Function of MBD4/MMR complex in the protection of methylcytosine

During evolution, the appearance of MBD4 protein seems to have coincided with the vertebrate lineage establishment (Hendrich and Tweedie 2003). This event parallels the transition from mosaic to global DNA methylation of the genomes, and consequently would reflect the onset of a CpG-poor genomic landscape due to spontaneous deamination of 5mC and its transition into T. In this context, MBD4 appears to have an essential role during evolution in protecting a subset of promoters that necessary for cell life. But beside its role in protecting the methylated CpGs, it can participate in a DNA active demethylating mechanism together with a deaminase (via its glycosylase domain). This ambiguity is found in the literature since MBD4 has been implicated in transcriptional repression but also in the DNA demethylation (Kondo et al. 2005, Rai et al. 2008). Interestingly, like MBD4, UHRF1 also appeared with the vertebrate lineage (Bronner et al. 2007) suggesting a mutual need to ensure each respective role. Finally, we might suggest that the MBD4/MLH1/UHRF1/DNMT1 quatuor ensures faithful DNA methylation inheritance.

In this study, we showed by genome-wide methylome and transcriptome analyses that the absence of MBD4 results in alterations in both promoter methylation and transcriptional activity of a large number of genes in MEFs. Interestingly, methylation loss was preferentially observed in CGIs exhibiting both higher CG density and low methylation level. 76 genes exhibited a marked increase in transcription level in absence of MBD4. The vast majority of the hypomethylated CGI promoters were transcriptionnnally up-regulated in absence of MBD4. These observations allow us to conclude that MBD4 is designed to preserve the methylation state of CGIs at the promoters of methylation-dependent repressed genes. With this function, MBD4 is unique within the MBD class of proteins, as it has been shown that the siRNA depletion of other MBD proteins (MBD1, MBD2 or MeCP2), resulted only in derepression of the respective genes and not in demethylation of their promoters (Boeke et al. 2000, Fujita et al. 2000, Jones P. L. et al. 1998, Nan et al. 1998, Sarraf and Stancheva 2004). This functional difference with the other MBDs, stresses the role of MBD4 glycosylase domain in preserving the methylation level of the CpGs at its target gene promoters.

All these findings allow us to propose that in vivo, MBD4 is bound through its MBD to the methylated CGI-containing promoters of its target genes. In this way, MBD4, either by steric hindrance or/and by recruiting repressive complexes, keeps the promoter in a methylated state and silenced. Since MBD4 is present at high concentration at the methylated promoter, it easily excises T, via its glycosylase/AP lyase activity, in G/T mismatch that was formed as a result of the spontaneous deamination of 5mC. Then, the BER machinery further repairs the resulted “gap” and fill it with unmethylated C, resulting in a hemi-methylated DNA raising the need to methylate the replaced cytosine to render the full methylated state of these regions. Interestingly, recent study showed that DNMT1 and UHRF1 are partners of MBD4 (Laget et al. 2014, Meng et al. 2015) that indicate a very likely role of UHRF1/DNMT1 tandem in methylating the hemi-methylated DNA lesion resulted by MBD4/BER. However, we do not yet know which are the candidates that might jump in at the right moment between MBD4 and the UHRF1/DNMT1 tandem.

3.2. TDG-dependent methylation/oxidation dynamic at repetitive elements

In this work, we present a genome-wide comparative analysis of DNA methylation/oxidation profiles of repetitive elements in both MEFs and ESCs. We found

major differences in 5mC/5hmC/5fC and 5caC distributions in these cells, showing that these modifications are dynamic during differentiation. The majority of the DNA methylation/oxidation patterns are dynamically regulated by TDG and occur mainly at CA repeats and at the mouse-lineage specific retro-elements including IAP-LTRs, SINEs B1m/B2m and L1Md-LINEs, which correspond to the most conserved, CpG-rich, and recently integrated young elements. Our analysis show that these conserved lineage specific retro-elements are not distributed randomly throughout the mouse genome but are instead clustered at specific loci that could define novel DNA regulatory regions

3.2.1 Methylation dynamics at IAP LTRs

We observed methylation enrichment at every conserved repeat whenever the CG density exceeded 0.83, which is the average mouse genome density. For example, the IAP and LTRs, which have the highest CG density, showed the highest methylation enrichment. This subfamily was found partially oxidized (5mC/5hmC enriched) in ESCs and fully methylated in MEFs, which suggest its permanent inactivation during differentiation to prevent insertional mutations. Accordingly, transcription of IAP is constrained by methylation (Walsh et al. 1998), and LTR elements were found to be excluded from gene regions (Medstrand et al. 2002), likely because of their potential to alter gene transcription. LTR families harboring an intermediate CG density such as ERV1 and non-IAP-ERVK, are dynamically regulated by TDG, while the evolutionary oldest CG-poor ERVL family escaped methylation. Collectively, our data suggest that methylation level of CG rich LTRs is highly dynamic during differentiation.

3.2.2 Methylation dynamics at mouse-specific SINEs

Our analysis revealed that the lineage-specific SINEs, B1m and B2m are hydroxymethylated in ESCs and fully methylated in MEFs. Furthers calculation showed a perfect correlation between SINE conservation, modified cytosine enrichment and proximity to TSS. This suggests that lineage-specific SINEs could have profound consequences on neighboring gene expression. Accordingly, human B1 SINEs have been shown to influence the activity of downstream gene promoters, with acquisition of DNA methylation and loss of activating histone marks (Estecio et al. 2012). We proposed that B1m and B2m SINEs might act as boundary elements that protect CGIs against pervasive

methylation. SINEs hydroxymethylation in ESCs could regulate transcriptional circuit that sustains the pluripotent state before subsequent methylation silencing during differentiation.

3.2.3 Methylation dynamics at mouse-specific intact L1Md LINEs

The intact lineage-specific LINEs L1Md showed a TDG-dependent cytosine modification dynamics in MEFs. Distribution of 5mC/5hmC/5fC and 5caC at these LINEs revealed two distinct profiles in MEFs. The first profil corresponds to L1Md containing a hypermethylated 5'UTR region. The second profil corresponds to L1Md showing a highly dynamic 5hmC/5fC/5caC patterns throughout their ORFs and regulated by TDG. We hypothesise that LINEs, with hypermethylated 5'UTR, are transcriptionally inactive whereas LINEs characterized by a methylcytosine oxidation patterns along their ORFs are active. Accordingly, these L1Mds are characterized by high PolII, H2A.Z and H3K4me1 levels. These intact LINEs also exhibited a non-random genomic distribution and are concentrated around gene clusters (1 L1Md/every 60 kb vs 1/180 kb for the rest of the genome). The biological significance of this genomic distribution is not understood but this specific LINEs distribution could be implicated in gene regulation and genome organization given that LINEs have been involved in several fundamental processes such as differentiation and development (Faulkner et al. 2009, Speek 2001). Accordingly, L1Md were depleted in cytosine modifications in ESCs suggesting that the TDG-dependent regulation of LINE transcription by methylation takes place during differentiation.

3.2.4 Methylation dynamics at CA repeats

Another important aspect of this study is the identification of modified cytosine enrichment at microsatellite CA repeats. The modified cytosine enrichment level correlates with the CA density within the CA repeats, suggesting that the modifications target cytosine in the CpA repeat motif.

Non-CpG methylation has long been controversial and lacks the certitude, which may be attributed to the highly dynamic of modified bases at these regions, and because these modified bases are in asymmetric form that can not be faithfully maintained during replication. For this reason, genome-wide sequencing combined to bisulfite treatment,

make it difficult to get full information about CA methylation regarding the expected low methylation level in a CA context. Accordingly, a recent study presented a new sensitive approach to map genome-wide 5hmC and 5fC at single base resolution, by combining 5hmC-specific restriction enzyme with a 5hmC chemical labeling enrichment method (Sun et al. 2015). This approach enables detection of low-abundance modified cytosine sites. The authors detected for the first time several millions of 5hmC and 5fC sites in CpA context in ESCs, and showed that 5hmC and 5fC in non-CpG context exhibit lower abundance, more dynamically, than those in CpG context. These data were supported by a recent paper from the Greenberg lab (Gabel et al. 2015), showing that 5mCA and 5hmCA are specifically recognized and bound by MeCP2. Authors showed also that conditionally depletion of Dnmt3a in mouse brain eliminates methylation in CpA context but not in CpG context. These recently published data support previous observations, obtained by bisulfite sequencing approaches, describing non-CG methylation in brain, oocytes, ESCs, iPSC and flies (Laurent et al. 2010, Lister et al. 2013, Xie et al. 2012).

Our data show that the densest CA repeats are preferentially methylated in ESCs but hydroxymethylated and dynamically formyl/carboxyl-ated in a TDG-dependent manner in MEFs. The biological significance of the switch from methylation to oxidation during differentiation remains unclear for the moment, but the occurrence of 5mC/5hmC on CA repeats located at close distance to TSS suggests an important role of these elements in the regulation of genome activity.

Altogether our data show a specific methylation/oxidation distribution profile specific to both CA repeats and the youngest lineage-specific transposable elements. We hypothesized that the TDG-dependent methylation dynamic observed at these repetitive elements, may constitute a novel epigenetic code with as yet unknown role in genome organization and functioning. Alterations of his code could be associated with disease development. This might be particularly true for tumorogenesis, since strong hypomethylation of the repeats is observed in cancer cells (Baba et al. 2010, Ehrlich 2009, Howard et al. 2008). Since retroelements are the major drivers of evolutionary changes within species (Burns and Boeke 2012, Cordaux and Batzer 2009), the observed retroelement methylation dynamics could be strongly implicated in evolution.

CHAPTER 4
CONCLUSION and PERSPECTIVES

4. Conclusion and perspectives

Deciphering the molecular mechanisms that control the dynamics of 5mC has become a major challenge in understanding gene expression regulation that governs many biological processes such as development, genome stability and phenotype inheritance. The recent discovery of 3 new bases of DNA corresponding to the oxidized forms of 5mC profoundly changed the field and the next years will be crucial in understanding the role of 5hmC, 5fC 5caC in the biological processes mentioned above but also in diseases such as tumorigenesis and neuro-degenerative diseases. For instance, the analysis of cancer cell methylome was the subject of a considerable number of publications in recent years, which showed that aberration of the methylation profile is a main feature of a cancer cells (Ehrlich 2009). Indeed, phenotype associated with cellular transformation is characterized by a localized hypermethylation of CGI and a more genome-wide hypomethylation mainly at repeated elements (Baba et al. 2010, Howard et al. 2008). These data strongly suggest that a disturbance of the dynamics of 5mC plays an essential role in the development of cancer. Consistently, mutations of TDG and MBD4 have been shown to be involved in many cancers (Dalton and Bellacosa 2012, Riccio et al. 1999, Sjolund et al. 2013, Yamada et al. 2002).

During my thesis, we have shown in mice that TDG regulates the dynamics of methylation/demethylation at repeated sequences during differentiation. Conversely, MBD4 targets promoters that are repressed by methylation, by protecting them from deamination and might protect them also from oxidation to maintain them in a methylated state. These data define TDG and MBD4 as essential enzymes to control the dynamics of methylomes and suggest that their enzymatic activities are finely regulated. The genome wide analysis shows that TDG regulates DNA methylation at repeated sequences. However, it is not yet known how and with whom TDG can exert such regulatory activity. Towards this goal, we have already started to identify the partner protein(s) that could be regulators of its activity. We purified from HeLa cells the protein complex associated with TDG *in vivo* and we got interesting preliminary results. The analysis by mass spectrometry identified BIN1 (Box-dependent MYC-Interacting protein 1) and P15 protein as two major partners of the TDG complex. BIN1 is an essential protein in mammals, initially identified as a tumor suppressor gene through its interaction with c-Myc. BIN1 was studied primarily for its function in the cytoplasmic actin dynamics, organization of the plasma membrane and cellular polarity (Prendergast et al.

2009). Interestingly, an analysis of the protein sequence of P15 shows strong sequence homology between P15 and the catalytic domain of the protein family of cytidine deaminases (AID, APOBEC family) suggesting a potential deaminase activity of P15.

The *in vitro* analysis shows that TDG complex is able to exert endonuclease activity after having created an abasic site in a G:T, 5fC or 5caC-containig DNA substrates. Our results reveal that both TDG and MBD4 possess glycosylase and unexpected lyase activities, providing that they were associated with their co-factor, *i.e.,* MLH1 and p15, respectively. This property has never been reported neither for MBD4 nor for TDG, since the literature has described MBD4 and TDG as a monofunctional glycosylase (Hendrich et al. 1999, Wiebauer and Jiricny 1989). This was completely surprising for us, considering that the structural domains involved in the catalytic enzymatic domain for each glycosylase are completely different. The differences are more responsible for the specificity towards the substrate rather than for the catalytic reaction. We might suggest that the co-activator is participating in the recognition of the substrate and little in the reaction itself. However, so far we cannot exclude that these co-activator might participate in the formation of the catalytic site of the glycosylases. The crystallographic structures of each glycosylase bound to its partner will help us to understand the contribution of MLH1 and P15 to the functions of MBD4 and TDG.

Our data clearly show that MBD4 is involved in the repair of deaminated 5meC in methylated CGIs but it is possible that this enzyme play also a role at other genome regions. For that, it is important to extend our analysis to the whole genome of MEF MBD4$^{+/+}$ and MBD4$^{-/-}$ by WGBS (Whole Genome Bisulfite Sequencing). This data will help us to identify the methylation status of each CpG throughout the genome, particularly in repeated sequences in the presence and absence of MBD4. We want to corroborate the results obtained by DIP-seq experiments to determine the genomic distribution of 5mC, 5hmC, 5fC and 5caC. Indeed, we cannot rule out an ectopic 5mC oxidation in the absence of MBD4, which can also lead to demethylating events. These will help to identify the genome wide areas controlled by MBD4 complex, and define the molecular mechanisms by which MMR proteins regulate MBD4 activity in controlling methylation DNA.

Altogether, our results highlight new functions for two glycosylases, among which the endonuclease activity may represent an important breakthrough in the understanding of DNA methylation maintenance, protection and dynamic.

CHAPTER 5
REFERENCES

References

Alhosin M, Sharif T, Mousli M, Etienne-Selloum N, Fuhrmann G, Schini-Kerth VB, Bronner C. 2011. Down-regulation of UHRF1, associated with re-expression of tumor suppressor genes, is a common feature of natural compounds exhibiting anti-cancer properties. J Exp Clin Cancer Res 30: 41.

Arand J, et al. 2012. In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. PLoS Genet 8: e1002750.

Arita K, Ariyoshi M, Tochio H, Nakamura Y, Shirakawa M. 2008. Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. Nature 455: 818-821.

Avvakumov GV, Walker JR, Xue S, Li Y, Duan S, Bronner C, Arrowsmith CH, Dhe-Paganon S. 2008. Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. Nature 455: 822-825.

Baba Y, et al. 2010. Epigenomic diversity of colorectal cancer indicated by LINE-1 methylation in a database of 869 tumors. Mol Cancer 9: 125.

Barreto G, et al. 2007. Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation. Nature 445: 671-675.

Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK. 2011. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol 12: R10.

Bellacosa A, Cicchillitti L, Schepis F, Riccio A, Yeung AT, Matsumoto Y, Golemis EA, Genuardi M, Neri G. 1999. MED1, a novel human methyl-CpG-binding endonuclease, interacts with DNA mismatch repair protein MLH1. Proc Natl Acad Sci U S A 96: 3969-3974.

Bennett MT, Rodgers MT, Hebert AS, Ruslander LE, Eisele L, Drohat AC. 2006. Specificity of human thymine DNA glycosylase depends on N-glycosidic bond stability. J Am Chem Soc 128: 12510-12519.

Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. 2009. An operational definition of epigenetics. Genes Dev 23: 781-783.

Besser D, Gotz F, Schulze-Forster K, Wagner H, Kroger H, Simon D. 1990. DNA methylation inhibits transcription by RNA polymerase III of a tRNA gene, but not of a 5S rRNA gene. FEBS Lett 269: 358-362.

Bhattacharya SK, Ramchandani S, Cervoni N, Szyf M. 1999. A mammalian protein with specific demethylase activity for mCpG DNA. Nature 397: 579-583.

Bird A. 2002. DNA methylation patterns and epigenetic memory. Genes Dev 16: 6-21.

Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8: 1499-1504.

Bird AP, Wolffe AP. 1999. Methylation-induced repression--belts, braces, and chromatin. Cell 99: 451-454.

Blackledge NP, Klose R. 2011. CpG island chromatin: a platform for gene regulation. Epigenetics 6: 147-152.

Blackledge NP, Zhou JC, Tolstorukov MY, Farcas AM, Park PJ, Klose RJ. 2010. CpG islands recruit a histone H3 lysine 36 demethylase. Mol Cell 38: 179-190.

Boeke J, Ammerpohl O, Kegel S, Moehren U, Renkawitz R. 2000. The minimal repression domain of MBD2b overlaps with the methyl-CpG-binding domain and binds directly to Sin3A. J Biol Chem 275: 34963-34967.

Borst P, Sabatini R. 2008. Base J: discovery, biosynthesis, and possible functions. Annu Rev Microbiol 62: 235-251.

Bostick M, Kim JK, Esteve PO, Clark A, Pradhan S, Jacobsen SE. 2007. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science 317: 1760-1764.

Bronner C, Fuhrmann G, Chedin FL, Macaluso M, Dhe-Paganon S. 2010. UHRF1 Links the Histone code and DNA Methylation to ensure Faithful Epigenetic Memory Inheritance. Genet Epigenet 2009: 29-36.

Bronner C, Achour M, Arima Y, Chataigneau T, Saya H, Schini-Kerth VB. 2007. The UHRF family: oncogenes that are drugable targets for cancer therapy in the near future? Pharmacol Ther 115: 419-434.

Brown TC, Jiricny J. 1988. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. Cell 54: 705-711.

Bruniquel D, Schwartz RH. 2003. Selective, stable demethylation of the interleukin-2 gene enhances transcription by an active process. Nat Immunol 4: 235-240.

Buck-Koehntop BA, Defossez PA. 2013. On how mammalian transcription factors recognize methylated DNA. Epigenetics 8: 131-137.

Burns KH, Boeke JD. 2012. Human transposon tectonics. Cell 149: 740-752.

Campanero MR, Armstrong MI, Flemington EK. 2000. CpG methylation as a mechanism for the regulation of E2F activity. Proc Natl Acad Sci U S A 97: 6481-6486.

Carlson LL, Page AW, Bestor TH. 1992. Properties and localization of DNA methyltransferase in preimplantation mouse embryos: implications for genomic imprinting. Genes Dev 6: 2536-2541.

Cedar H, Bergman Y. 2009. Linking DNA methylation and histone modification: patterns and paradigms. Nat Rev Genet 10: 295-304.

Chahrour M, Jung SY, Shaw C, Zhou X, Wong ST, Qin J, Zoghbi HY. 2008. MeCP2, a key contributor to neurological disease, activates and represses transcription. Science 320: 1224-1229.

Chelico L, Pham P, Calabrese P, Goodman MF. 2006. APOBEC3G DNA deaminase acts processively 3' --> 5' on single-stranded DNA. Nat Struct Mol Biol 13: 392-399.

Chen CC, Wang KY, Shen CK. 2012. The mammalian de novo DNA methyltransferases DNMT3A and DNMT3B are also DNA 5-hydroxymethylcytosine dehydroxymethylases. J Biol Chem 287: 33116-33121.

Chen PY, Feng S, Joo JW, Jacobsen SE, Pellegrini M. 2011. A comparative analysis of DNA methylation across human embryonic stem cell lines. Genome Biol 12: R62.

Chen, Tsujimoto N, Li E. 2004. The PWWP domain of Dnmt3a and Dnmt3b is required for directing DNA methylation to the major satellite repeats at pericentric heterochromatin. Mol Cell Biol 24: 9048-9058.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. Nat Rev Genet 10: 691-703.

Cortazar D, Kunz C, Saito Y, Steinacher R, Schar P. 2007. The enigmatic thymine DNA glycosylase. DNA Repair (Amst) 6: 489-504.

Cortazar D, et al. 2011. Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. Nature 470: 419-423.

Cortellino S, et al. 2011. Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. Cell 146: 67-79.

Costello JF, et al. 2000. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. Nat Genet 24: 132-138.

Dalton SR, Bellacosa A. 2012. DNA demethylation by TDG. Epigenomics 4: 459-467.

Daniel JM, Spring CM, Crawford HC, Reynolds AB, Baig A. 2002. The p120(ctn)-binding partner Kaiso is a bi-modal DNA-binding protein that recognizes both a sequence-specific consensus and methylated CpG dinucleotides. Nucleic Acids Res 30: 2911-2919.

Daskalos A, Nikolaidis G, Xinarianos G, Savvari P, Cassidy A, Zakopoulou R, Kotsinas A, Gorgoulis V, Field JK, Liloglou T. 2009. Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. Int J Cancer 124: 81-87.

Defossez PA, Stancheva I. 2011. Biological functions of methyl-CpG-binding proteins. Prog Mol Biol Transl Sci 101: 377-398.

Deininger PL, Batzer MA. 2002. Mammalian retroelements. Genome Res 12: 1455-1465.

Dherin C, et al. 2009. Characterization of a highly conserved binding site of Mlh1 required for exonuclease I-dependent mismatch repair. Mol Cell Biol 29: 907-918.

Drummond JT, Bellacosa A. 2001. Human DNA mismatch repair in vitro operates independently of methylation status at CpG sites. Nucleic Acids Res 29: 2234-2243.

Eckhardt F, et al. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet 38: 1378-1385.

Ehrlich M. 2009. DNA hypomethylation in cancer cells. Epigenomics 1: 239-259.

Ehrlich M, Gama-Sosa MA, Huang LH, Midgett RM, Kuo KC, McCune RA, Gehrke C. 1982. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucleic Acids Res 10: 2709-2721.

Engel N, Tront JS, Erinle T, Nguyen N, Latham KE, Sapienza C, Hoffman B, Liebermann DA. 2009. Conserved DNA methylation in Gadd45a(-/-) mice. Epigenetics 4: 98-99.

Englander EW, Wolffe AP, Howard BH. 1993. Nucleosome interactions with a human Alu element. Transcriptional repression and effects of template methylation. J Biol Chem 268: 19565-19573.

Eskeland R, et al. 2010. Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. Mol Cell 38: 452-464.

Estecio MR, Gallegos J, Dekmezian M, Lu Y, Liang S, Issa JP. 2012. SINE retrotransposons cause epigenetic reprogramming of adjacent gene promoters. Mol Cancer Res 10: 1332-1342.

Faulkner GJ, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet 41: 563-571.

Fraga MF, Ballestar E, Montoya G, Taysavang P, Wade PA, Esteller M. 2003. The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties. Nucleic Acids Res 31: 1765-1774.

Franchini DM, Schmitz KM, Petersen-Mahrt SK. 2012. 5-Methylcytosine DNA demethylation: more than losing a methyl group. Annu Rev Genet 46: 419-441.

Fujita N, Shimotake N, Ohki I, Chiba T, Saya H, Shirakawa M, Nakao M. 2000. Mechanism of transcriptional regulation by methyl-CpG binding protein MBD1. Mol Cell Biol 20: 5107-5118.

Fujita N, Watanabe S, Ichimura T, Ohkuma Y, Chiba T, Saya H, Nakao M. 2003. MCAF mediates MBD1-dependent transcriptional repression. Mol Cell Biol 23: 2834-2843.

Gabel HW, Kinde B, Stroud H, Gilbert CS, Harmin DA, Kastan NR, Hemberg M, Ebert DH, Greenberg ME. 2015. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. Nature.

Gallinari P, Jiricny J. 1996. A new class of uracil-DNA glycosylases related to human thymine-DNA glycosylase. Nature 383: 735-738.

Gebhard C, et al. 2010. General transcription factor binding at CpG islands in normal cells correlates with resistance to de novo DNA methylation in cancer cells. Cancer Res 70: 1398-1407.

Gehring M, Huh JH, Hsieh TF, Penterman J, Choi Y, Harada JJ, Goldberg RB, Fischer RL. 2006. DEMETER DNA glycosylase establishes MEDEA polycomb gene self-imprinting by allele-specific demethylation. Cell 124: 495-506.

Gogvadze E, Buzdin A. 2009. Retroelements and their impact on genome evolution and functioning. Cell Mol Life Sci 66: 3727-3742.

Gong Z, Morales-Ruiz T, Ariza RR, Roldan-Arjona T, David L, Zhu JK. 2002. ROS1, a repressor of transcriptional gene silencing in Arabidopsis, encodes a DNA glycosylase/lyase. Cell 111: 803-814.

Greenwood C, Selth LA, Dirac-Svejstrup AB, Svejstrup JQ. 2009. An iron-sulfur cluster domain in Elp3 important for the structural integrity of elongator. J Biol Chem 284: 141-149.

Gueneau E, et al. 2013. Structure of the MutLalpha C-terminal domain reveals how Mlh1 contributes to Pms1 endonuclease site. Nat Struct Mol Biol 20: 461-468.

Guo JU, Su Y, Zhong C, Ming GL, Song H. 2011. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. Cell 145: 423-434.

Guo JU, et al. 2014. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. Nat Neurosci 17: 215-222.

Hajkova P, Jeffries SJ, Lee C, Miller N, Jackson SP, Surani MA. 2010. Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. Science 329: 78-82.

Hardeland U, Steinacher R, Jiricny J, Schar P. 2002. Modification of the human thymine-DNA glycosylase by ubiquitin-like proteins facilitates enzymatic turnover. EMBO J 21: 1456-1464.

Hardeland U, Bentele M, Jiricny J, Schar P. 2003. The versatile thymine DNA-glycosylase: a comparative characterization of the human, Drosophila and fission yeast orthologs. Nucleic Acids Res 31: 2261-2271.

Hashimoto H, Horton JR, Zhang X, Bostick M, Jacobsen SE, Cheng X. 2008. The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. Nature 455: 826-829.

Hashimoto H, Liu Y, Upadhyay AK, Chang Y, Howerton SB, Vertino PM, Zhang X, Cheng X. 2012. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. Nucleic Acids Res 40: 4841-4849.

Hawkes NA, et al. 2002. Purification and characterization of the human elongator complex. J Biol Chem 277: 3047-3052.

He YF, et al. 2011. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science 333: 1303-1307.

Heintzman ND, et al. 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459: 108-112.

Hendrich B, Bird A. 1998. Identification and characterization of a family of mammalian methyl-CpG binding proteins. Mol Cell Biol 18: 6538-6547.

Hendrich B, Tweedie S. 2003. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. Trends Genet 19: 269-277.

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

Hendrich B, Guy J, Ramsahoye B, Wilson VA, Bird A. 2001. Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development. Genes Dev 15: 710-723.

Hermann A, Schmitt S, Jeltsch A. 2003. The human Dnmt2 has residual DNA-(cytosine-C5) methyltransferase activity. J Biol Chem 278: 31717-31721.

Holliday R, Pugh JE. 1975. DNA modification mechanisms and gene activity during development. Science 187: 226-232.

Hotchkiss RD. 1948. The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. J Biol Chem 175: 315-332.

Howard G, Eiges R, Gaudet F, Jaenisch R, Eden A. 2008. Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. Oncogene 27: 404-408.

Howell CY, Bestor TH, Ding F, Latham KE, Mertineit C, Trasler JM, Chaillet JR. 2001. Genomic imprinting disrupted by a maternal effect mutation in the Dnmt1 gene. Cell 104: 829-838.

Ichiyanagi K, et al. 2011. Locus- and domain-dependent control of DNA methylation at mouse B1 retrotransposons during male germ cell development. Genome Res 21: 2058-2066.

Ichiyanagi T, Ichiyanagi K, Miyake M, Sasaki H. 2013. Accumulation and loss of asymmetric non-CpG methylation during male germ-cell development. Nucleic Acids Res 41: 738-745.

Iguchi-Ariga SM, Schaffner W. 1989. CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. Genes Dev 3: 612-619.

Illingworth RS, Bird AP. 2009. CpG islands--'a rough guide'. FEBS Lett 583: 1713-1720.

Inoue A, Zhang Y. 2011. Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. Science 334: 194.

Iqbal K, Jin SG, Pfeifer GP, Szabo PE. 2011. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. Proc Natl Acad Sci U S A 108: 3642-3647.

Ito S, D'Alessio AC, Taranova OV, Hong K, Sowers LC, Zhang Y. 2010. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature 466: 1129-1133.

Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, He C, Zhang Y. 2011. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science 333: 1300-1303.

Jacobs AL, Schar P. 2012. DNA glycosylases: in DNA repair and beyond. Chromosoma 121: 1-20.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13: 484-492.

Jones PL, Veenstra GJ, Wade PA, Vermaak D, Kass SU, Landsberger N, Strouboulis J, Wolffe AP. 1998. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat Genet 19: 187-191.

Kim JK, Esteve PO, Jacobsen SE, Pradhan S. 2009a. UHRF1 binds G9a and participates in p21 transcriptional regulation in mammalian cells. Nucleic Acids Res 37: 493-505.

Kishigami S, Van Thuan N, Hikichi T, Ohta H, Wakayama S, Mizutani E, Wakayama T. 2006. Epigenetic abnormalities of the mouse paternal zygotic genome associated with microinsemination of round spermatids. Dev Biol 289: 195-205.

Kohli RM, Zhang Y. 2013. TET enzymes, TDG and the dynamics of DNA demethylation. Nature 502: 472-479.

Kokura K, Kaul SC, Wadhwa R, Nomura T, Khan MM, Shinagawa T, Yasukawa T, Colmenares C, Ishii S. 2001. The Ski protein family is required for MeCP2-mediated transcriptional repression. J Biol Chem 276: 34115-34121.

Kondo E, Gu Z, Horii A, Fukushige S. 2005. The thymine DNA glycosylase MBD4 represses transcription and is associated with methylated p16(INK4a) and hMLH1 genes. Mol Cell Biol 25: 4388-4396.

Kriaucionis S, Heintz N. 2009. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science 324: 929-930.

Lachner M, O'Carroll D, Rea S, Mechtler K, Jenuwein T. 2001. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature 410: 116-120.

Laget S, Miotto B, Chin HG, Esteve PO, Roberts RJ, Pradhan S, Defossez PA. 2014. MBD4 cooperates with DNMT1 to mediate methyl-DNA repression and protects mammalian cells from oxidative stress. Epigenetics 9: 546-556.

Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409: 860-921.

Laurent L, et al. 2010. Dynamic changes in the human methylome during differentiation. Genome Res 20: 320-331.

Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet 11: 204-220.

Lee JH, Skalnik DG. 2005. CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. J Biol Chem 280: 41725-41731.

Lewis JD, Meehan RR, Henzel WJ, Maurer-Fogy I, Jeppesen P, Klein F, Bird A. 1992. Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. Cell 69: 905-914.

Li CJ. 2013. DNA demethylation pathways: recent insights. Genet Epigenet 5: 43-49.

Li E. 2002. Chromatin modification and epigenetic reprogramming in mammalian development. Nat Rev Genet 3: 662-673.

Li E, Bestor TH, Jaenisch R. 1992. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell 69: 915-926.

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

Lister R, et al. 2013. Global epigenomic reconfiguration during mammalian brain development. Science 341: 1237905.

Liu WM, Schmid CW. 1993. Proposed roles for DNA methylation in Alu transcriptional repression and mutational inactivation. Nucleic Acids Res 21: 1351-1359.

Liu X, Gao Q, Li P, Zhao Q, Zhang J, Li J, Koseki H, Wong J. 2013. UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9. Nat Commun 4: 1563.

Loenarz C, Schofield CJ. 2011. Physiological and biochemical aspects of hydroxylations and demethylations catalyzed by human 2-oxoglutarate oxygenases. Trends Biochem Sci 36: 7-18.

Ma DK, Jang MH, Guo JU, Kitabatake Y, Chang ML, Pow-Anpongkul N, Flavell RA, Lu B, Ming GL, Song H. 2009. Neuronal activity-induced Gadd45b promotes epigenetic DNA demethylation and adult neurogenesis. Science 323: 1074-1077.

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

Maiti A, Morgan MT, Pozharski E, Drohat AC. 2008. Crystal structure of human thymine DNA glycosylase bound to DNA elucidates sequence-specific mismatch recognition. Proc Natl Acad Sci U S A 105: 8890-8895.

Manvilla BA, Maiti A, Begley MC, Toth EA, Drohat AC. 2012. Crystal structure of human methyl-binding domain IV glycosylase bound to abasic DNA. J Mol Biol 420: 164-175.

Martinowich K, Hattori D, Wu H, Fouse S, He F, Hu Y, Fan G, Sun YE. 2003. DNA methylation-related chromatin remodeling in activity-dependent BDNF gene regulation. Science 302: 890-893.

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

Medstrand P, van de Lagemaat LN, Mager DL. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res 12: 1483-1495.

Meng H, Harrison DJ, Meehan RR. 2015. MBD4 interacts with and recruits USP7 to heterochromatic foci. J Cell Biochem 116: 476-485.

Metivier R, et al. 2008. Cyclical DNA methylation of a transcriptionally active promoter. Nature 452: 45-50.

Millar CB, Guy J, Sansom OJ, Selfridge J, MacDougall E, Hendrich B, Keightley PD, Bishop SM, Clarke AR, Bird A. 2002. Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. Science 297: 403-405.

Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, Bibel M, Schubeler D. 2008. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol Cell 30: 755-766.

Morgan HD, Dean W, Coker HA, Reik W, Petersen-Mahrt SK. 2004. Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming. J Biol Chem 279: 52353-52360.

Morgan MT, Bennett MT, Drohat AC. 2007. Excision of 5-halogenated uracils by human thymine DNA glycosylase. Robust activity for DNA contexts other than CpG. J Biol Chem 282: 27578-27586.

Mudbhary R, et al. 2014. UHRF1 overexpression drives DNA hypomethylation and hepatocellular carcinoma. Cancer Cell 25: 196-209.

Nan X, Meehan RR, Bird A. 1993. Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. Nucleic Acids Res 21: 4886-4892.

Nan X, Ng HH, Johnson CA, Laherty CD, Turner BM, Eisenman RN, Bird A. 1998. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature 393: 386-389.

Neddermann P, Jiricny J. 1993. The purification of a mismatch-specific thymine-DNA glycosylase from HeLa cells. J Biol Chem 268: 21218-21224.

Neddermann P, Gallinari P, Lettieri T, Schmid D, Truong O, Hsuan JJ, Wiebauer K, Jiricny J. 1996. Cloning and expression of human G/T mismatch-specific thymine-DNA glycosylase. J Biol Chem 271: 12767-12774.

Ng HH, Jeppesen P, Bird A. 2000. Active repression of methylated genes by the chromosomal protein MBD1. Mol Cell Biol 20: 1394-1406.

Ng HH, Zhang Y, Hendrich B, Johnson CA, Turner BM, Erdjument-Bromage H, Tempst P, Reinberg D, Bird A. 1999. MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. Nat Genet 23: 58-61.

Okada Y, Yamagata K, Hong K, Wakayama T, Zhang Y. 2010. A role for the elongator complex in zygotic paternal genome demethylation. Nature 463: 554-558.

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

Okano M, Bell DW, Haber DA, Li E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99: 247-257.

Olsen LC, Aasland R, Wittwer CU, Krokan HE, Helland DE. 1989. Molecular cloning of human uracil-DNA glycosylase, a highly conserved DNA repair enzyme. EMBO J 8: 3121-3125.

Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. Nature 284: 604-607.
Oswald J, Engemann S, Lane N, Mayer W, Olek A, Fundele R, Dean W, Reik W, Walter J. 2000. Active demethylation of the paternal genome in the mouse zygote. Curr Biol 10: 475-478.

Otani J, Arita K, Kato T, Kinoshita M, Kimura H, Suetake I, Tajima S, Ariyoshi M, Shirakawa M. 2013. Structural basis of the versatile DNA recognition ability of the methyl-CpG binding domain of methyl-CpG binding domain protein 4. J Biol Chem 288: 6351-6362.

Pacaud R, Brocard E, Lalier L, Hervouet E, Vallette FM, Cartron PF. 2014. The DNMT1/PCNA/UHRF1 disruption induces tumorigenesis characterized by similar genetic and epigenetic signatures. Sci Rep 4: 4230.

Petronzelli F, Riccio A, Markham GD, Seeholzer SH, Genuardi M, Karbowski M, Yeung AT, Matsumoto Y, Bellacosa A. 2000a. Investigation of the substrate spectrum of the human mismatch-specific DNA N-glycosylase MED1 (MBD4): fundamental role of the catalytic domain. J Cell Physiol 185: 473-480.

Petronzelli F, Riccio A, Markham GD, Seeholzer SH, Stoerker J, Genuardi M, Yeung AT, Matsumoto Y, Bellacosa A. 2000b. Biphasic kinetics of the human DNA repair protein MED1 (MBD4), a mismatch-specific DNA N-glycosylase. J Biol Chem 275: 32422-32429.

Popp C, Dean W, Feng S, Cokus SJ, Andrews S, Pellegrini M, Jacobsen SE, Reik W. 2010. Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature 463: 1101-1105.

Prendergast GC, Muller AJ, Ramalingam A, Chang MY. 2009. BAR the door: cancer suppression by amphiphysin-like genes. Biochim Biophys Acta 1795: 25-36.

Rai K, Huggins IJ, James SR, Karpf AR, Jones DA, Cairns BR. 2008. DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45. Cell 135: 1201-1212.

Rakyan VK, Preis J, Morgan HD, Whitelaw E. 2001. The marks, mechanisms and memory of epigenetic states in mammals. Biochem J 356: 1-10.

Ramsahoye BH, Biniszkiewicz D, Lyko F, Clark V, Bird AP, Jaenisch R. 2000. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. Proc Natl Acad Sci U S A 97: 5237-5242.

Ratnam S, Mertineit C, Ding F, Howell CY, Clarke HJ, Bestor TH, Chaillet JR, Trasler JM. 2002. Dynamics of Dnmt1 methyltransferase expression and intracellular localization during oogenesis and preimplantation development. Dev Biol 245: 304-314.

Razin A, Riggs AD. 1980. DNA methylation and gene function. Science 210: 604-610.
Riccio A, et al. 1999. The DNA repair gene MBD4 (MED1) is mutated in human carcinomas with microsatellite instability. Nat Genet 23: 266-268.

Riggs AD. 1975. X inactivation, differentiation, and DNA methylation. Cytogenet Cell Genet 14: 9-25.

Robertson KD. 2005. DNA methylation and human disease. Nat Rev Genet 6: 597-610.

Rougier N, Bourc'his D, Gomes DM, Niveleau A, Plachot M, Paldi A, Viegas-Pequignot E. 1998. Chromosome methylation patterns during mammalian preimplantation development. Genes Dev 12: 2108-2113.

Ruzov A, Shorning B, Mortusewicz O, Dunican DS, Leonhardt H, Meehan RR. 2009. MBD4 and MLH1 are required for apoptotic induction in xDNMT1-depleted embryos. Development 136: 2277-2286.

Ruzov A, Dunican DS, Prokhortchouk A, Pennings S, Stancheva I, Prokhortchouk E, Meehan RR. 2004. Kaiso is a genome-wide repressor of transcription that is essential for amphibian development. Development 131: 6185-6194.

Sancar A, Lindsey-Boltz LA, Unsal-Kacmaz K, Linn S. 2004. Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. Annu Rev Biochem 73: 39-85.

Santos F, Hendrich B, Reik W, Dean W. 2002. Dynamic reprogramming of DNA methylation in the early mouse embryo. Dev Biol 241: 172-182.

Sarraf SA, Stancheva I. 2004. Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly. Mol Cell 15: 595-605.

Sasai N, Matsuda E, Sarashina E, Ishida Y, Kawaichi M. 2005. Identification of a novel BTB-zinc finger transcriptional repressor, CIBZ, that interacts with CtBP corepressor. Genes Cells 10: 871-885.

Scharer OD, Jiricny J. 2001. Recent progress in the biology, chemistry and structural biology of DNA glycosylases. Bioessays 23: 270-281.

Schiesser S, Hackner B, Pfaffeneder T, Muller M, Hagemeier C, Truss M, Carell T. 2012. Mechanism and stem-cell activity of 5-carboxycytosine decarboxylation determined by isotope tracing. Angew Chem Int Ed Engl 51: 6516-6520.

Serandour AA, et al. 2012. Dynamic hydroxymethylation of deoxyribonucleic acid marks differentiation-associated enhancers. Nucleic Acids Res 40: 8255-8265.

Sharif J, et al. 2007. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. Nature 450: 908-912.

Shen L, Wu H, Diep D, Yamaguchi S, D'Alessio AC, Fung HL, Zhang K, Zhang Y. 2013. Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. Cell 153: 692-706.

Singer MF. 1982. SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. Cell 28: 433-434.

Sjolund AB, Senejani AG, Sweasy JB. 2013. MBD4 and TDG: multifaceted DNA glycosylases with ever expanding biological roles. Mutat Res 743-744: 12-25.

Speek M. 2001. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol 21: 1973-1985.

Steinacher R, Schar P. 2005. Functionality of human thymine DNA glycosylase requires SUMO-regulated changes in protein conformation. Curr Biol 15: 616-623.

Stock JK, Giadrossi S, Casanova M, Brookes E, Vidal M, Koseki H, Brockdorff N, Fisher AG, Pombo A. 2007. Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. Nat Cell Biol 9: 1428-1435.

Sun Z, Dai N, Borgaro JG, Quimby A, Sun D, Correa IR, Jr., Zheng Y, Zhu Z, Guan S. 2015. A Sensitive Approach to Map Genome-wide 5-Hydroxymethylcytosine and 5-Formylcytosine at Single-Base Resolution. Mol Cell 57: 750-761.

Tahiliani M, et al. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science 324: 930-935.

Takai D, Jones PA. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A 99: 3740-3745.

Thomson JP, et al. 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature 464: 1082-1086.

Tomizawa S, Kobayashi H, Watanabe T, Andrews S, Hata K, Kelsey G, Sasaki H. 2011. Dynamic stage-specific changes in imprinted differentially methylated regions during early mammalian development and prevalence of non-CpG methylation in oocytes. Development 138: 811-820.

Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13: 36-46.

Vassetzky NS, Kramerov DA. 2013. SINEBase: a database and tool for SINE analysis. Nucleic Acids Res 41: D83-89.

Walsh CP, Chaillet JR, Bestor TH. 1998. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat Genet 20: 116-117.

Wang H, An W, Cao R, Xia L, Erdjument-Bromage H, Chatton B, Tempst P, Roeder RG, Zhang Y. 2003. mAM facilitates conversion by ESET of dimethyl to trimethyl lysine 9 of histone H3 to cause transcriptional repression. Mol Cell 12: 475-487.

Waters TR, Swann PF. 1998. Kinetics of the action of thymine DNA glycosylase. J Biol Chem 273: 20007-20014.

Watt F, Molloy PL. 1988. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. Genes Dev 2: 1136-1143.

Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schubeler D. 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet 37: 853-862.

Wiebauer K, Jiricny J. 1989. In vitro correction of G.T mispairs to G.C pairs in nuclear extracts from human cells. Nature 339: 234-236.

Wong E, et al. 2002. Mbd4 inactivation increases Cright-arrowT transition mutations and promotes gastrointestinal tumor formation. Proc Natl Acad Sci U S A 99: 14937-14942.

Wossidlo M, Nakamura T, Lepikhov K, Marques CJ, Zakhartchenko V, Boiani M, Arand J, Nakano T, Reik W, Walter J. 2011. 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. Nat Commun 2: 241.

Wu H, Zhang Y. 2014. Reversing DNA methylation: mechanisms, genomics, and biological functions. Cell 156: 45-68.

Wu H, D'Alessio AC, Ito S, Xia K, Wang Z, Cui K, Zhao K, Sun YE, Zhang Y. 2011. Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature 473: 389-393.

Wu SC, Zhang Y. 2010. Active DNA demethylation: many roads lead to Rome. Nat Rev Mol Cell Biol 11: 607-620.

Wyszynski M, Gabbara S, Bhagwat AS. 1994. Cytosine deaminations catalyzed by DNA cytosine methyltransferases are unlikely to be the major cause of mutational hot spots at sites of cytosine methylation in Escherichia coli. Proc Natl Acad Sci U S A 91: 1574-1578.

Xie W, Barr CL, Kim A, Yue F, Lee AY, Eubanks J, Dempster EL, Ren B. 2012. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. Cell 148: 816-831.

Yamada T, Koyama T, Ohwada S, Tago K, Sakamoto I, Yoshimura S, Hamada K, Takeyoshi I, Morishita Y. 2002. Frameshift mutations in the MBD4/MED1 gene in primary gastric cancer with high-frequency microsatellite instability. Cancer Lett 181: 115-120.

Yoder JA, Walsh CP, Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13: 335-340.

Yoon JH, Iwai S, O'Connor TR, Pfeifer GP. 2003. Human thymine DNA glycosylase (TDG) and methyl-CpG-binding protein 4 (MBD4) excise thymine glycol (Tg) from a Tg:G mispair. Nucleic Acids Res 31: 5399-5404.

Zhang L, Lu X, Lu J, Liang H, Dai Q, Xu GL, Luo C, Jiang H, He C. 2012. Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA. Nat Chem Biol 8: 328-330.

Zhang Y, Ng HH, 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 Dev 13: 1924-1935.

Ziller MJ, et al. 2011. Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. PLoS Genet 7: e1002389.

Zingg JM, Shen JC, Yang AS, Rapoport H, Jones PA. 1996. Methylation inhibitors can increase the rate of cytosine deamination by (cytosine-5)-DNA methyltransferase. Nucleic Acids Res 24: 3267-3275.

I. INTRODUCTION

Chez les mammifères, la méthylation est une marque épigénétique ciblant le carbone 5 des cytosines principalement dans un contexte CpG, produisant une methylcytosine (5mC). La majorité des CpGs (70 à 80%) chez les mammifères sont méthylés. Bien que les 5mC soient retrouvées principalement au niveau des séquences répétées, l'étude du méthylome de ces régions a été longtemps ignorée au profit de régions supposées plus importantes fonctionnellement comme les îlots CpGs (CGI). Les CGI sont des régions très denses en CpG, localisées au niveau des promoteurs de 70% des gènes.

I. 1 Les îlots CpG, plateforme pour la régulation de l'expression génique

Les méthylcytosines (5mC) sont fortement sensibles à une déamination spontanée conduisant à la formation d'une thymine (1,2) (T). Par conséquent, le génome des vertébrés est caractérisé par une forte diminution en dinucléotides CpG. Ce paysage génomique fortement méthylé et pauvre en CpG est toutefois ponctué par des régions denses en CpG et globalement non méthylées, les îlots CpG (CGI=CpG Island). Les CGI sont situés sur des loci associés à une activité transcriptionnelle (3,4). Ils semblent donc avoir été maintenus au cours de l'évolution par une pression de sélection sous-jacente à un rôle majeur dans la régulation de l'expression génique. Chez les mammifères, 70% des promoteurs de gènes codant ont un CGI. Ces CGI sont reconnus par des protéines à doigt de zinc de type ZF-CxxC, comme les protéines CFP1, KDM2A et TET1.

La protéine CFP1 lie les CGI et y recrute le complexe à activité methyltransférase SET1, qui catalyse la tri-méthylation de la lysine 4 de l'histone H3 (5-8). La modification H3K4me3 sert en suite de plateforme au recrutement des protéines à domaine PHD (Plant Homeo Domain) impliquées dans l'initiation de la transcription tel que TFIID, ING4 ou encore NURF (9). De la même façon, KDM2A lie spécifiquement les CGI et catalyse la déméthylation de H3K36me2 (10), modification reconnue par les complexes à activité histone déacétylase HDAC qui inhibent la transcription. Ainsi, les protéines à domaine ZF-CxxC coopèrent pour former au niveau des CGIs une architecture chromatienne unique ,enrichie en H3K4me3 et dépourvue en H3K36me2, qui favorise l'initiation de la transcription.

Les CGIs sont également reconnus par la protéine TET1 qui, à l'inverse des protéines CFP1 et KDM2A, est impliquée dans le recrutement des complexes répresseurs de la famille polycomb PRC1 et PRC2 (11). TET1 recrute PRC2 qui catalyse la tri-méthylation de la lysine 27

de l'histone H3. La modification H3K27me3 est ensuite reconnue par PRC1, qui inhibe l'élongation de la transcription en favorisant l'ubiquitinylation de l'histone H2A et la compaction de la chromatine (12-14). Notez que ce mécanisme de répression est indépendant de la méthylation des CGI.

L'analyse de la distribution de la méthylation à l'échelle du génome entier (ou méthylome) a montré qu'au sein des cellules somatiques, une fraction des CGI se méthyle de façon tissu-spécifique au cours du développement (15-18). La majorité des promoteurs de type CGI qui acquièrent la méthylation au cours de la différenciation cellulaire, est déjà réprimée par les protéines polycomb, au sein des ESCs (19). La méthylation de l'ADN n'est donc pas un événement initiateur de la répression transcriptionnelle, mais agit plutôt comme un marqueur épigénétique héritable qui maintient une répression dans le temps. Cette mise sous silence se fait principalement par l'intermédiaire de deux mécanismes. D'une part, la méthylation des cytosines prévient la liaison des protéines CFP1 et KDM2A et ainsi, inhibe la mise en place d'une chromatine favorable à l'initiation de la transcription. D'autre part, les 5mC sont reconnues spécifiquement par les protéines de la famille MBD (Methyl-CpG Binding Domain) qui recrutent des enzymes de modification des histones impliquées dans la répression transcriptionnelle (19)

Malgré le role important dans la regulation de la transcription les CGIs ne contiennent que 7% des CpG et sont globalement dépourvus de méthylation. A l'inverse, les séquences répétées, fortement méthylées dans des cellules saines, sont spécifiquement ciblées par un processus de déméthylation dans des cellules cancéreuses. Les séquences répétées couvrent près de la moitié du génome des mammifères. Parmi elles, on distingue les éléments transposables (SINEs, LINEs, LTRs et transposons) des répétitions en tandem (appelées également satellites) faites de séquences successivement répétées. Selon la taille de l'unité de répétition et le nombre de répétitions, on distingue les satellites mineurs, les satellites majeurs et les microsatellites.

I. 2 Déméthylation de l'ADN par excision de base

Plusieurs études ont mis en évidence dans des cellules somatiques une déméthylation rapide d'une fraction des CGIs en réponse à différents stimuli environnementaux, conduisant à une réactivation des gènes affectés (20-24). Ces observations sont à l'origine des nombreux efforts de la communauté scientifique pour identifier les enzymes impliquées dans ce processus. La déméthylation active de l'ADN peut être accomplie par une glycosylase qui excise directement une 5mC. Bien que de nombreuses données biochimiques et génétiques supportent l'utilisation de ce mécanisme chez la plante (avec la famille ROS1) (25), aucune glycosylase

spécifique des 5mC n'a été identifiée chez les mammifères. Toutefois sur la base d'études biochimiques, il a été proposé que les glycosylases MBD4 et TDG, capables de cliver les produits de la déamination et de l'oxydation des 5mC, seraient impliquées dans ce processus. Deux voies mécanistiques ont alors été proposées pour déméthyler l'ADN (Fig. I). Dans une première hypothèse, le processus s'initie par la déamination d'une 5mC en T conduisant à la formation d'un mésappariement G/T dans un contexte CpG. Le mésappariement G/T est ensuite reconnu par l'activité glycosylase de MBD4 ou TDG qui clive spécifiquement la thymidine mésappariée (26-29). Dans un second mécanisme, les 5mC subissent plusieurs étapes d'oxydation par les protéines TET conduisant à la formation de 5fC et de 5caC, qui sont reconnus et clivés par TDG. Cette voie a été proposée suite aux résultats de deux études récentes montrant que TDG peut cliver avec la même affinité un mésappariement G/T, une 5fC ou une 5caC dans un contexte CpG (30,31). Dans les deux hypothèses, l'activité de MBD4 ou de TDG se traduit par la formation d'un site abasique nécessitant l'intervention d'une enzyme à activité AP-lyase pour finaliser la réaction de clivage. En l'absence de cette AP-lyase putative, les deux modèles restent incomplets et controversés.

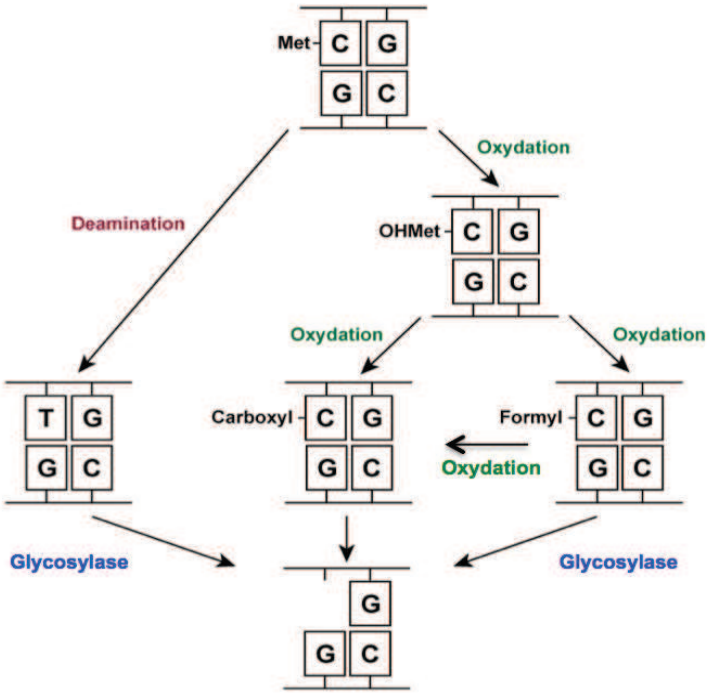


Figure I : Schématisation des deux voies hypothétiques de déméthylation de l'ADN chez les mammifères.

II. RESULTATS

II. 1 Déchiffrage d'un code de méthylation au niveau des séquences répétées.

Les modifications 5fC et 5caC sont spécifiquement reconnues et clivées par la protéine TDG et sont donc potentiellement des marques dénotant un processus actif de déméthylation. Nous avons caractérisé la distribution des cytosines modifiées (5mC, 5hmC, 5fC et 5caC) par des expériences d'immunoprécipitation d'ADN suivit de séquençage haut-débit (DIP-seq), dans des MEFs. De manière frappante, nos résultats révèlent un enrichissement en cytosines modifiées spécifiquement au niveau des séquences répétées (**Fig. 1**). Alors que les 5mC ciblent l'ensemble des séquences répétées avec une préférence pour les SINEs, les formes oxydées (5hmC, 5fC et 5caC) sont retrouvées préférentiellement au niveau des microsatellites (Simple repeats). Nous avons ainsi décidé d'effectuer une étude comparative systématique et globale de la distribution des cytosines modifiées au niveau des séquences répétées dans des cellules totipotentes (ESCs=embryonic stem cells) et différenciées (MEFs).

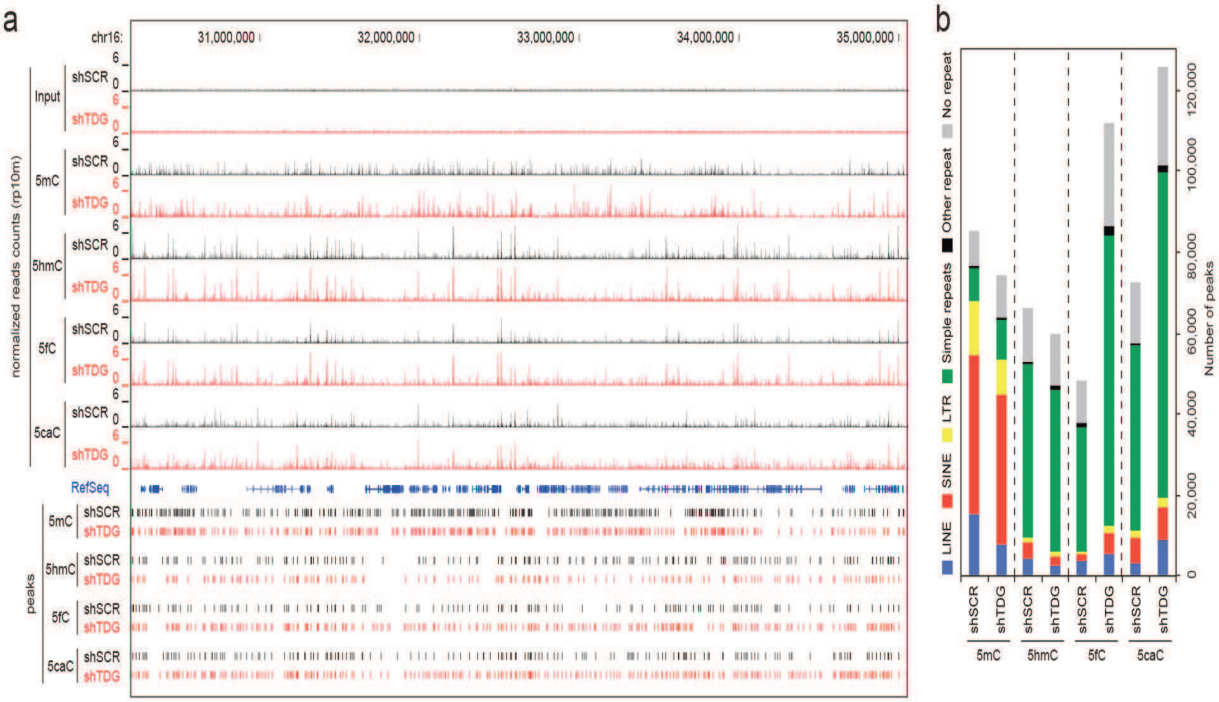


Figure 1 : Accumulation des cytosines modifiées au niveau des séquences répétées dans les MEFs. a. Visualisation par le navigateur UCSC de la distribution des 5mC, 5hmC, 5fC et 5caC ainsi que les pics d'enrichissement correspondant identifiés par le logiciel MACS. **b.** Annotation des pics d'enrichissement par le logiciel HOMER. Notez qu'entre 75 et 90% des pics correspondent à des régions répétées.

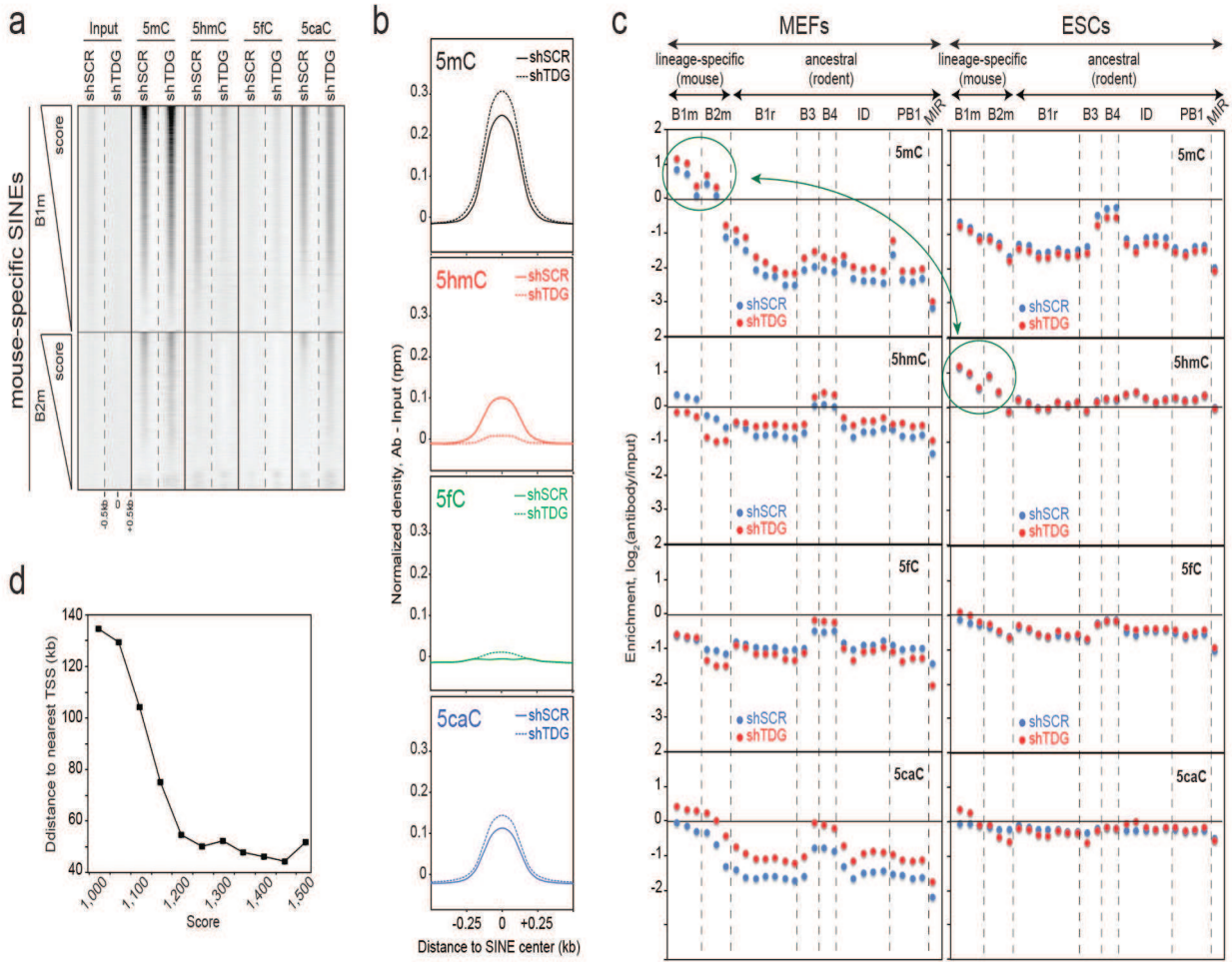


Figure 2 : TDG régule la méthylation des SINEs évolutivement les plus récents lors de la différentiation. a-b. Heatmap (a) et densité moyenne (b) des 5mC, 5hmC, 5fC et 5caC au niveau des SINEs B1m et B2m dans les MEFs. Les éléments ont été classés par conservation (score) au sein de chaque famille (a). **c.** Analyse comparative de l'enrichissement en cytosines modifiées pour chaque famille de SINEs dans les MEFs et les ESCs. Notez que les SINEs spécifiques du lignage souris sont exclusivement enrichis en 5hmC dans les ESCs (ronds verts). **d.** Courbe représentant la distance moyenne au TSS le plus proche des SINEs B1m et B2m en fonction de leur conservation (score).

La distribution des 5mC, 5hmC, 5fC et 5caC au niveau des SINEs dans les MEFs (**Fig. 2**) révèle une dynamique de ces modifications, régulée par TDG, uniquement au niveau des SINEs de la famille B1m et B2m (**Fig. 2a**). Les calculs de densité révèlent un fort enrichissement en 5mC, et un léger enrichissement en 5hmC et 5caC régulé par TDG, indiquant un processus actif de méthylation/déméthylation au niveau des SINEs B1m et B2m. De manière intéressante, ces familles sont conservées, riches en CG et sont évolutivement les plus récentes (spécifiques du lignage souris). Dans les ESCs, les SINEs B1m et B2m sont également les seuls à être modifié mais sont spécifiquement enrichis en 5hmC (**Fig. 2c**). L'analyse de la distribution génomique des SINEs B1m et B2m montrent que les SINEs spécifiques de la lignée souris sont d'autant plus conservés qu'ils sont proches d'un TSS (Transcription Start Site), suggérant une fonction

importante dans la régulation de l'expression génique (**Fig. 2d**). En conclusion, ces résultats mettent en évidence une vague de méthylation/déméthylation au niveau des SINEs récemment intégrés, induite par la différentiation et régulée par TDG.

Cette dynamique a été également observée pour les LTRs et les LINEs. De manière similaire aux résultats obtenus pour les SINEs, la régulation TDG-dépendante par la méthylation est observée uniquement au niveau des rétro-éléments évolutivement les plus récents (spécifiques de la lignée souris), la famille L1Md pour les LINEs et la famille IAP pour les LTRs. Ces familles regroupent la majorité des rétro-transposons pleine-taille et donc potentiellement mobiles. Nous avons montré que les IAPs sont enrichis en 5mC et 5hmC dans les ESCs mais uniquement au niveau des extrémités répétées des LTRs. Dans les MEFs, les IAP accumulent la méthylation tout le long du rétro-élément, suggérant une inactivation transcriptionnelle par la méthylation pendant la différentiation.

Par ailleurs, on observe une dynamique des cytosines modifiées dépendante de TDG au niveau des LINEs de la famille L1Md spécifiquement dans les MEFs (**Fig. 3**). Dans ces cellules, 70% des L1Md ont un promoteur hyperméthylé (**Fig. 3a**, cluster 1) et 30% sont hydroxyméthylés le long de leur séquence codante (**Fig. 3a**, cluster 2). En absence de TDG, l'enrichissement en 5hmC disparaît au profit de 5caC, mettant en évidence une dynamique de méthylation/déméthylation. La présence de H2A.Z et de la PolII le long de la séquence codante des L1Md appartenant au cluster 2, suggère qu'ils sont actifs en transcription, alors que ceux appartenant au cluster 1 sont réprimés par la méthylation (**Fig. 3b**). En conclusion, ces résultats montrent que l'activité transcriptionnelle des LINEs au sein des cellules différenciées, est régulée par des cycles de méthylation/déméthylation dépendants de TDG.

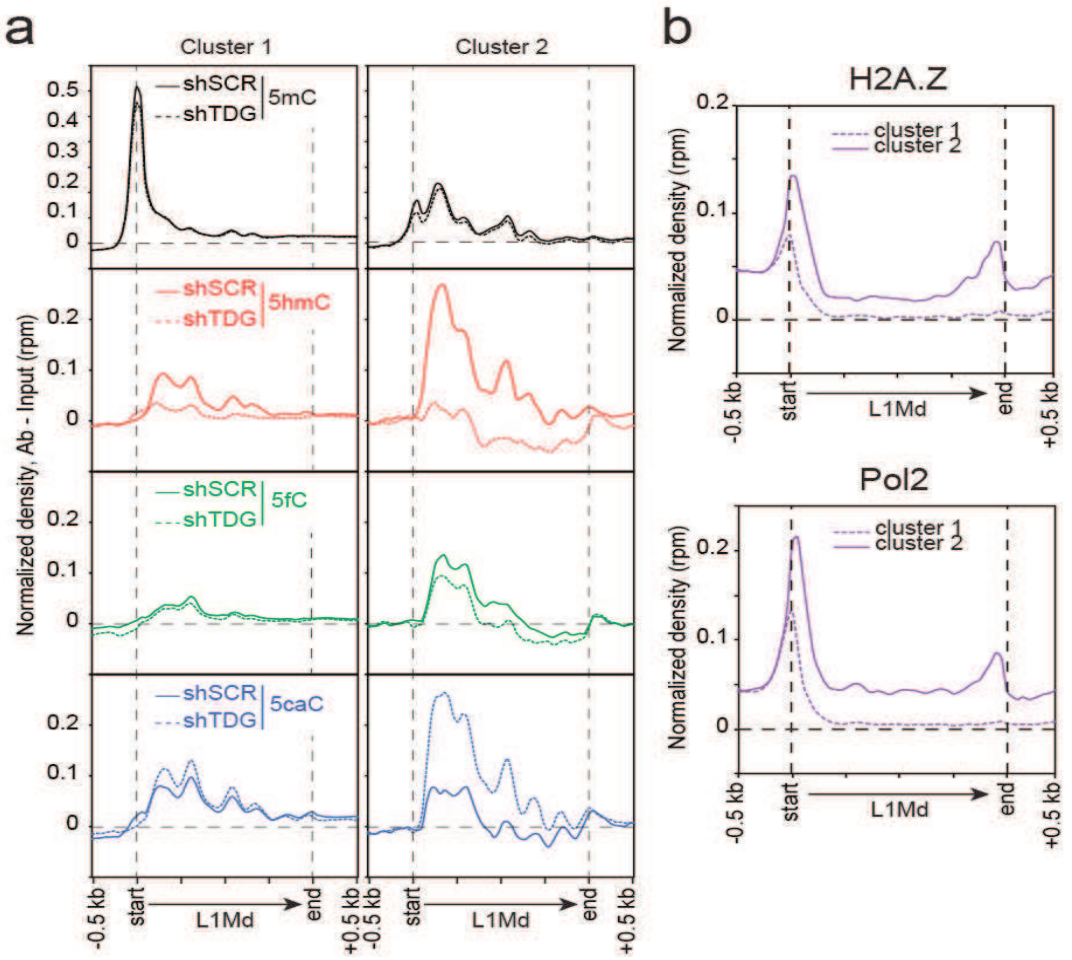


Figure 3 : TDG régule la méthylation des LINEs actifs en transcription. a. Distribution des 5mC, 5hmC, 5fC et 5caC au niveau des LINEs L1Md dans les MEFs. Le cluster 1 regroupe les LINEs dont le promoteur est hyperméthylé. Le cluster 2 contient les LINEs enrichis en 5hmC le long de leur séquence codante, et régulés par TDG. **b.** Distribution de H2A.Z et Pol2 déterminée par ChIPseq le long des LINEs L1Md.

Une autre donnée importante de cette étude est l'enrichissement en 5hmC, 5fC et 5caC observé au niveau des microsatellites dans les MEFs (**Fig. 4**). Nous avons montré que ces modifications ciblent spécifiquement les répétitions de type CA (CA repeat) de manière proportionnelle à leur densité en CpA (**Fig. 4a**), suggérant que l'oxydation des méthylcytosines cible les dinucléotides CpA. De manière importante, en absence de TDG on observe une accumulation spécifique des modifications 5fC et 5caC (**Fig. 4b**). De plus, TDG clive *in vitro* une formylcytosine uniquement dans un contexte CpG ou CpA (**Fig. 4c**). Dans les ESCs, les répétitions CA sont principalement enrichies en 5mC et 5fC et ne sont pas régulées par TDG (**Fig. 4d**). L'analyse de la distribution génomique des répétitions CA montre qu'une répétition CA est d'autant plus proche d'un TSS qu'elle est dense en CpA. Ensemble, ces résultats

suggèrent que les microsatellites de type CA subissent des cycles de méthylation/déméthylation régulés par TDG pendant la différentiation cellulaire.

Dans ce travail nous avons mis en évidence un code de méthylation au niveau des séquences répétées, dynamique au cours de la différentiation et régulé par TDG. Ces modifications ciblent spécifiquement les répétitions CA ainsi que les rétro-éléments spécifiques du lignage souris, qui correspondent aux éléments les plus conservés, denses en CpG, et récemment intégrés au sein de l'espèce. Nous proposons que la dynamique des rétro-éléments contrôlée par la méthylation, qui exerce une force évolutive majeure au sein des espèces, soit aussi un mécanisme fondamental de régulation de la transition d'une cellule d'un état totipotent vers un état différencié. Cette étude fait l'objet d'un article que je signe comme premier auteur et actuellement soumis pour publication (Papin et al, Submitted_1)[1].

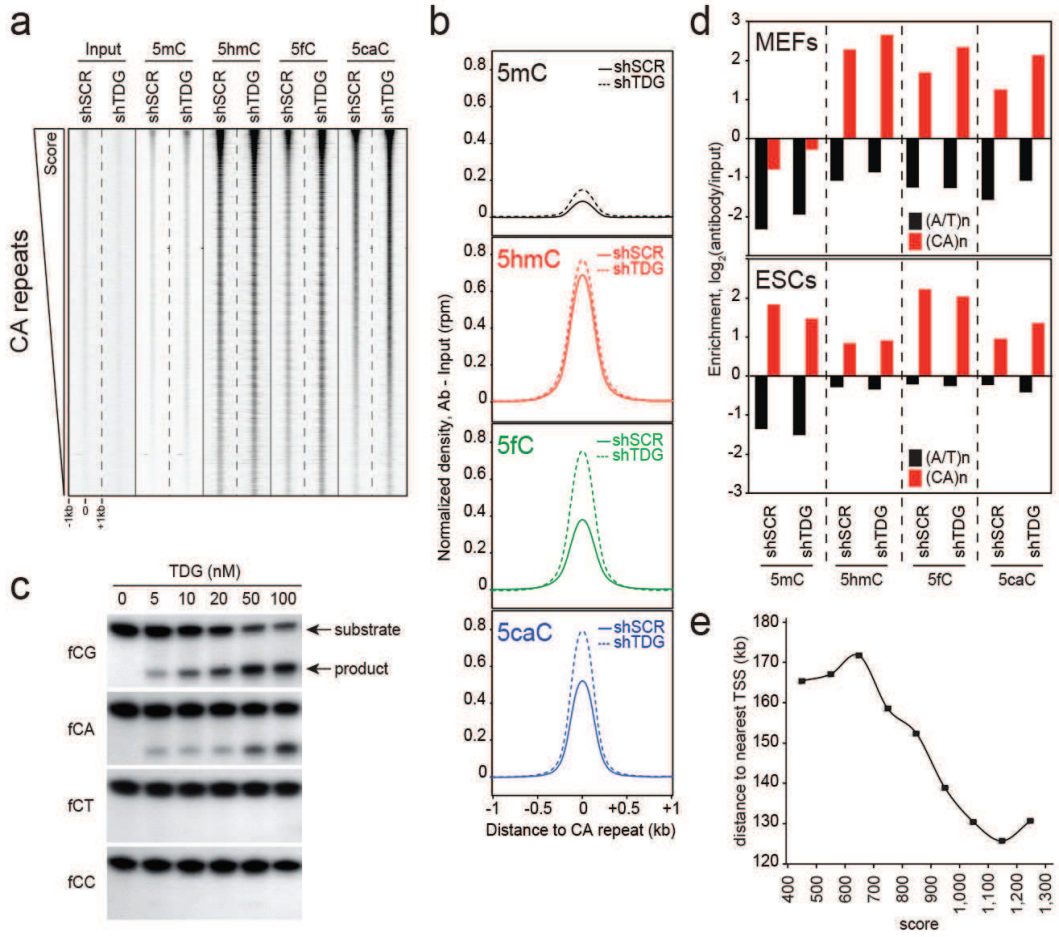


Figure 4 : TDG régule la méthylation des répétitions CA au sein des cellules différentiées. **a.** Heatmap (a) et densité moyenne (b) représentant le niveau de 5mC, 5hmC, 5fC et 5caC au niveau des répétitions CA dans des MEFs. Les éléments ont été classés par densité en CpA (score). **c.** Test glycosylase *in vitro* montrant que TDG clive une formylcytosine uniquement dans un contexte CpG ou CpA. **d.** Analyse comparative (MEFs *vs* ESCs) de l'enrichissement en cytosines modifiées au niveau des répétitions de type A/T ou CA. Notez l'enrichissement en 5mC exclusivement dans les MEFs.**e.** Courbe représentant la distance moyenne au TSS le plus proche des répétitions CA en fonction de leur densité en CpA (score).

II. 2 Régulation épigénétique de l'activité transcriptionnelle par MBD4

MBD4 est particulièrement intrigante puisqu'elle est la seule protéine des mammifères possédant un domaine de liaison à l'ADN méthylé (MBD, Methyl Binding Domain) associé à un domaine glycosylase (**Fig. 5**). *In vitro*, la protéine MBD4 a une activité glycosylase monofonctionnelle qui clive spécifiquement la liaison N-glycosidique d'une thymidine mésappariée à une guanine (G/T), conduisant à la formation d'un site abasique (**Fig. 6**).

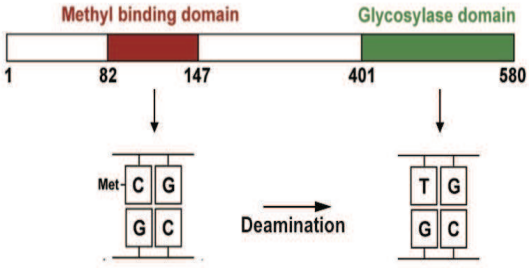


Figure 5 : Représentation schématique de la protéine MBD4. MBD4 peut lier une méthylcytosine par son domaine MBD (en rouge) mais également un mésappariement G/T par son domaine glycosylase (en vert).

La déamination d'une 5mC en T produisant un mésappariement G/T, l'architecture unique de MBD4 lui permet de lier à la fois le substrat et le produit de la déamination d'une 5mC (**Fig. 5**). Par conséquent MBD4, en liant les 5mC peut les protéger de la déamination, mais en clivant un G/T peut aussi participer à un mécanisme actif de déméthylation de l'ADN en association avec une déaminase.

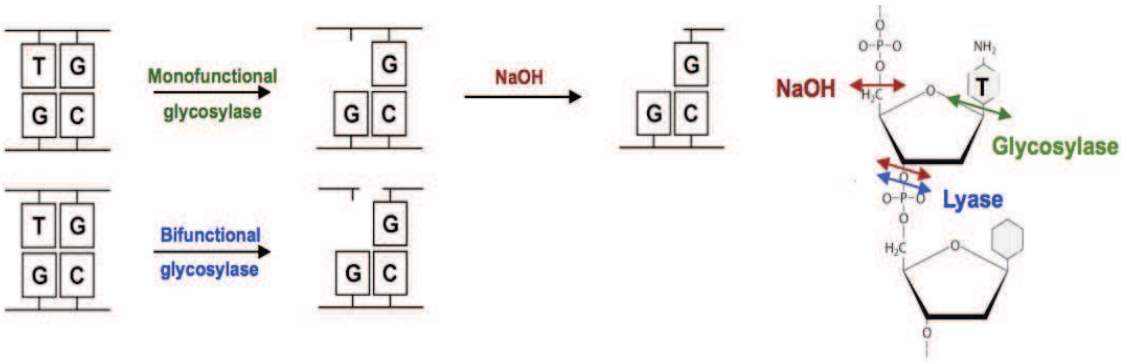


Figure 6 : Biochimie des glycosylases. Les glycosylases monofonctionnelles rompent la liaison N-glycosidique (en vert) entre la base et le sucre formant un site abasique. Pour visualiser une activité monofonctionnelle sur gel dénaturant, le squelette d'ADN doit être clivé par un traitement NaOH qui clive les liaisons phosphodiester en 5' et 3' du site abasique (en rouge). Pour les enzymes bifonctionnelles, l'activité glycosylase est couplée à une activité lyase (en bleu) qui coupe directement la liaison phosphodiester en 3' du site abasique.

Afin de clarifier la fonction de cette enzyme dans la dynamique des 5mC, il est essentiel d'identifier les protéines associées à MBD4 *in vivo*. Dans ce but, j'ai purifié le complexe MBD4 à partir de cellules HeLa. L'analyse par spectrométrie de masse de ce complexe montre que MBD4 est associé aux protéines PMS2, MLH1, MSH2 et MSH6, quatre protéines impliquées dans la réparation des mésappariements d'ADN (MMR, MisMatch Repair). Les tests enzymatiques *in vitro* montrent que le complexe MBD4/MMR possède une activité bifonctionnelle glycosylase/lyase spécifique d'un G/T et dirigée par la méthylation (**Fig. 7**). L'analyse biochimique de mutants ponctuels de MBD4 révèle que l'intégrité des domaines MBD et glycosylase est requise pour cette fonction.

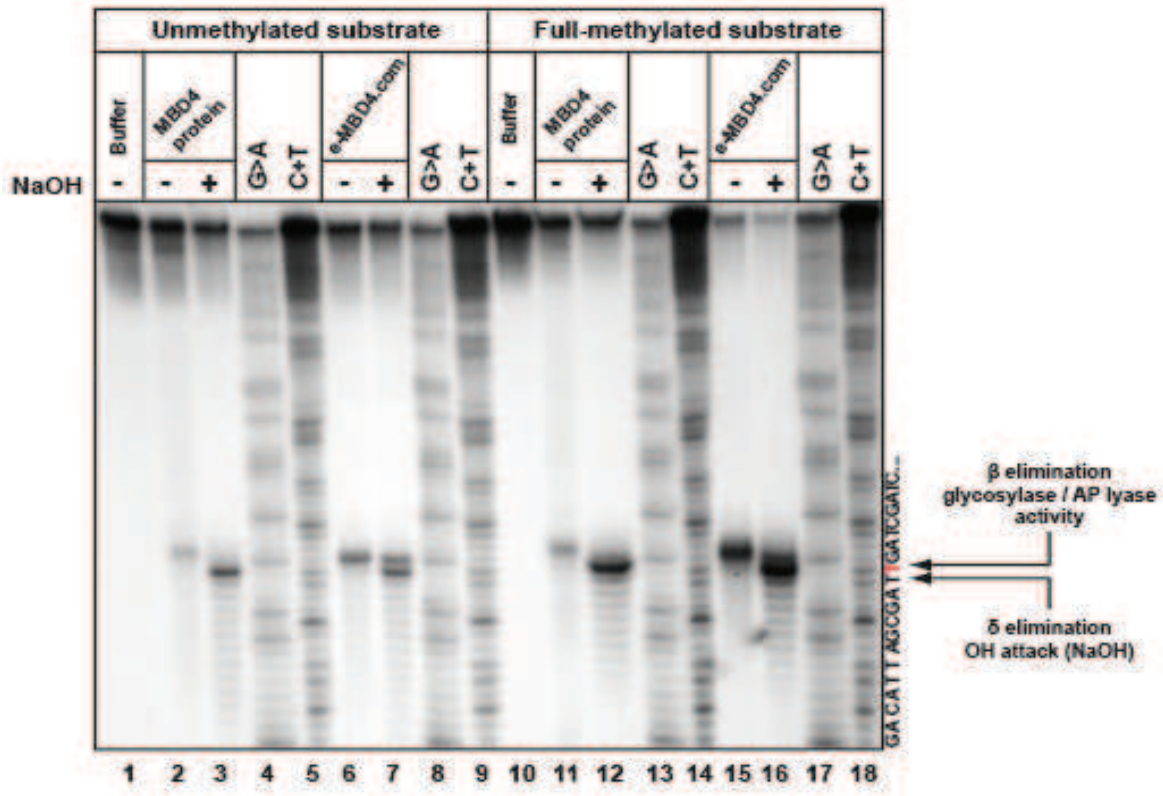


Figure 7 : Le complexe MBD4/MMR possède une activité bifonctionnelle glycosylase/lyase induite par la méthylation. La protéine MBD4 ou le complexe MBD4/MMR ont été incubés avec un substrat contenant un mésappariement G/T, méthylé ou non-méthylé. Les produits de la réaction ont été traités ou pas au NaOH, dénaturés à 95°C et séparés par électrophorèse sur gel dénaturant. Les pistes annotées G>A et C+T correspondent aux produits d'une réaction de séquençage Maxam-Gilbert. Notez qu'en absence de NaOH, la réaction de clivage se fait en 3' du T mésapparié (en rouge), ce qui est caractéristique d'une activité lyase (voir Fig. 6).

Il est important de rappeler que d'après la littérature, la protéine MBD4 recombinante possède une activité glycosylase monofonctionnelle non induite par la méthylation. Nos données suggèrent donc un rôle activateur et/ou enzymatique des protéines du MMR dans l'activité bifonctionnelle du complexe MBD4. Pour vérifier cette hypothèse, nous avons purifié la protéine recombinante MBD4, le dimère PMS2/MLH1 et le dimère MSH2/MSH6, et vérifié que la protéine MBD4 interagit physiquement avec les protéines du MMR. *In vitro*, une faible activité nucléase n'est détectée qu'avec la protéine MBD4 suggérant que l'activité enzymatique du complexe natif MBD4/MMR est catalyzée par MBD4 et régulée par les protéines du MMR.

La protéine MLH1 étant la sous-unité du complexe MBD4 la plus abondante dans nos analyses par spectrométrie de masse, nous nous sommes demandés si MLH1 pouvait agir comme activateur de MBD4 au sein du complexe MBD4/MMR. En accord avec cette hypothèse, j'ai pu reconstituer un complexe MBD4/MLH1 (**Fig. 8A**). De manière intéressante, alors que l'activité nucléase de MBD4 est faible et n'est pas sensible à la méthylation, le complexe MBD4/MLH1 montre une activité nucléase intense fortement induite par la méthylation (**Fig. 8B-C**). L'absence d'activité enzymatique détectée avec le complexe purifié à partir d'un mutant catalytique de MBD4 (MBD4 D554A/MLH1), montre que la réaction de coupure est catalysée par MBD4 mais dépendante de sa liaison à MLH1.

Par ailleurs, aucune activité nucléolytique n'est observée avec les différentes protéines purifiées sur un substrat contenant un site abasique (**Fig. 8B**), excluant la présence d'une endonucléase contaminante. En conclusion, nous avons reconstitué à partir de protéines recombinantes le complexe minimal ayant les mêmes propriétés enzymatiques que le complexe MBD4 natif.

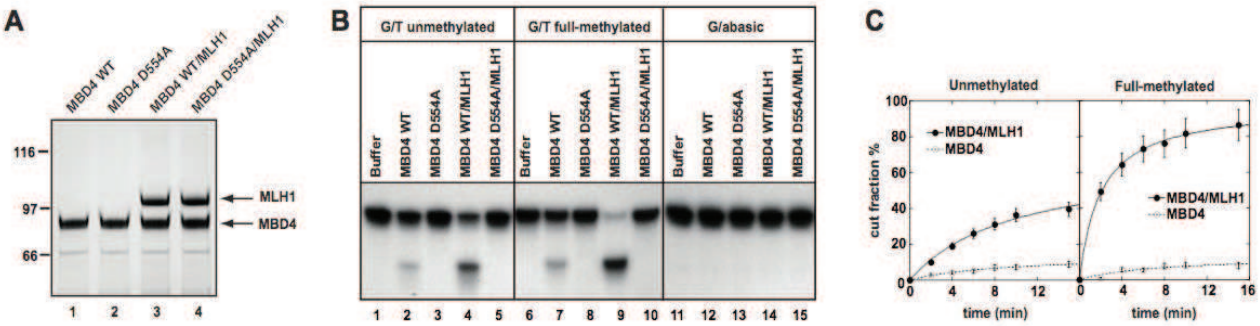


Figure 8 : MLH1 active MBD4 *in vitro*. A. Purification des protéines MBD4 et des complexes MBD4/MLH1 séparés sur gel SDS-PAGE. **B.** L'activité enzymatique des protéines indiquées a été analysée comme décrit Fig. 3. Les produits de la réaction n'ont pas été traités au NaOH. **C.** Cinétique de coupure de la protéine MBD4 seule ou associée à MLH1 sur un substrat non-méthylé (à gauche) ou méthylé (à droite). La quantification montre clairement que l'activité nucléase de MBD4 induite par la méthylation, est dépendante de sa liaison à MLH1.

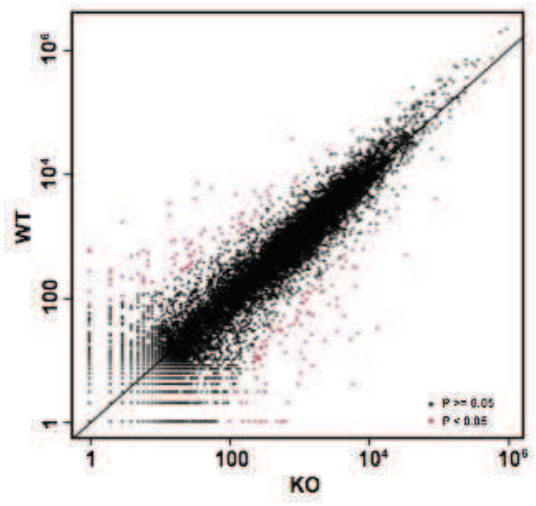


Figure 9 : MBD4 est un régulateur de l'expression génique. Analyse comparative du transcriptome des MEF Mbd4$^{+/+}$ (WT) et des MEF Mbd4$^{-/-}$ (KO). Les gènes significativement dérégulés sont indiqués en rouge (P < 0,05).

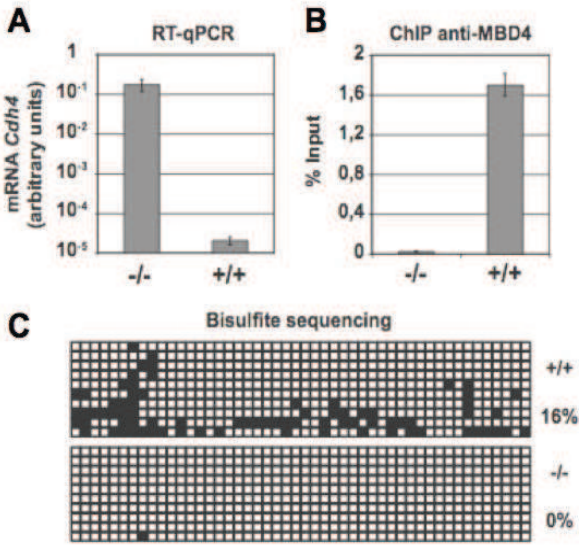


Figure 10 : MBD4 maintient les promoteurs de type CGI dans un état réprimé en les protégeant de la déméthylation. Exemple du gène *Cdh4*. -/- indique les MEF KO pour MBD4, +/+ indique les MEF WT pour MBD4. **A. RT-qPCR.** *Cdh4* est surexprimé en absence de MBD4. **B. Immunoprécipitation de chromatine.** MBD4 lie le promoteur de *Cdh4*. **C. Séquençage après traitement au bisulfite.** Chaque carré blanc indique un CpG non-méthylé, chaque carré noir un CpG méthylé. Chaque ligne représente un clone. Le CGI de *Cdh4* est déméthylé en absence de MBD4.

L'ensemble de ces résultats montre que la protéine MBD4 est spécialement conçue pour réparer des mésappariements G/T dans un contexte riche en 5mC. Dans le génome des vertébrés, des régions denses en 5mC sont retrouvées au niveau des promoteurs contenant un CGI, réprimés par la méthylation. Une des fonctions de MBD4 *in vivo* pourrait donc être de protéger les CGI méthylés de la déamination. Pour vérifier cette hypothèse, j'ai d'abord mis en évidence une fonction régulatrice de MBD4 dans l'expression génique. En effet, l'analyse par RNA-seq du transcriptome de MEFs isolées à partir d'embryons de souris sauvages (WT) ou dépourvues du gène *Mbd4* (KO), montre que 215 gènes sont significativement dérégulés ($P < 0,05$) en absence de MBD4 (**Fig. 9**). J'ai montré que les gènes réprimés ne sont pas directement régulés par MBD4. En effet les promoteurs réprimés ne sont pas liés par MBD4 dans la situation sauvage, et ne montrent pas de variation significative de leur taux de méthylation en absence de MBD4. A l'inverse, les promoteurs des gènes surexprimés montrent une perte de la méthylation au niveau de leur CGI en absence de MBD4 (le cas du gène *Cdh4* est donné à titre d'exemple **Fig. 10**).

Nous proposons que MBD4, en liant les 5mC par son domaine MBD, protège les promoteurs de type CGI réprimés par la méthylation. Si une déamination a lieu, MBD4 clive par son domaine glycosylase le mésappariement G/T, conduisant à la formation d'un site abasique clivé à son extrémité 3'. Cet intermédiaire est un signal d'activation de la voie de réparation par excision de base (BER=Base Excision Repair), qui en association avec la voie de maintenance de la méthylation, réincorpore une méthylcytosine.

En conclusion, ce travail a mis en évidence une interaction entre la protéine MBD4 et les protéines du MMR. De manière importante, la protéine MBD4 montre une activité bifonctionnelle glycosylase/lyase dépendante de sa liaison à MLH1, et induite par la méthylation de l'ADN. *In vivo*, MBD4 maintient les promoteurs de type CGI dans un état réprimé en les protégeant de la déméthylation. Cette étude fait l'objet d'un article que je signe comme premier auteur et actuellement soumis pour publication (Papin et al, Submitted_2)[2].

BIBLIOGRAPHIE

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

2 Coulondre, C., Miller, J. H., Farabaugh, P. J. & Gilbert, W. Molecular basis of base substitution hotspots in Escherichia coli. *Nature* **274**, 775-780 (1978).

3 Illingworth, R. S. *et al*. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS genetics* **6**, e1001134, doi:10.1371/journal.pgen.1001134 (2010).

4 Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes & development* **25**, 1010-1022, doi:10.1101/gad.2037511 (2011).

5 Lee, J. H. & Skalnik, D. G. CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *The Journal of biological chemistry* **280**, 41725-41731, doi:10.1074/jbc.M508312200 (2005).

6 Voo, K. S., Carlone, D. L., Jacobsen, B. M., Flodin, A. & Skalnik, D. G. 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. *Molecular and cellular biology* **20**, 2108-2121 (2000).

7 Thomson, J. P. *et al*. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464**, 1082-1086, doi:10.1038/nature08924 (2010).

8 Blackledge, N. P. & Klose, R. CpG island chromatin: a platform for gene regulation. *Epigenetics : official journal of the DNA Methylation Society* **6**, 147-152 (2011).

9 Blackledge, N. P. *et al*. CpG islands recruit a histone H3 lysine 36 demethylase. *Molecular cell* **38**, 179-190, doi:10.1016/j.molcel.2010.04.009 (2010).

10 Wu, H. *et al*. Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* **473**, 389-393, doi:10.1038/nature09934 (2011).

11 Stock, J. K. *et al*. Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. *Nature cell biology* **9**, 1428-1435, doi:10.1038/ncb1663 (2007).

12 Zhou, W. *et al*. Histone H2A monoubiquitination represses transcription by inhibiting RNA polymerase II transcriptional elongation. *Molecular cell* **29**, 69-80, doi:10.1016/j.molcel.2007.11.002 (2008).

13 Eskeland, R. *et al*. Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Molecular cell* **38**, 452-464, doi:10.1016/j.molcel.2010.02.032 (2010).

14 Mohn, F. *et al*. Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Molecular cell* **30**, 755-766, doi:10.1016/j.molcel.2008.05.007 (2008).

15 Schilling, E. & Rehli, M. Global, comparative analysis of tissue-specific promoter CpG methylation. *Genomics* **90**, 314-323, doi:10.1016/j.ygeno.2007.04.011 (2007).

16 Shen, L. *et al*. Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS genetics* **3**, 2023-2036, doi:10.1371/journal.pgen.0030181 (2007).

17 Weber, M. *et al*. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature genetics* **37**, 853-862 (2005).

18 Buck-Koehntop, B. A. & Defossez, P. A. On how mammalian transcription factors recognize methylated DNA. *Epigenetics : official journal of the DNA Methylation Society* **8**, 131-137, doi:10.4161/epi.23632 (2013).

19 Clouaire, T. & Stancheva, I. Methyl-CpG binding proteins: specialized transcriptional repressors or structural components of chromatin? *Cellular and molecular life sciences : CMLS* **65**, 1509-1522, doi:10.1007/s00018-008-7324-y (2008).

20 Bruniquel, D. & Schwartz, R. H. Selective, stable demethylation of the interleukin-2 gene enhances transcription by an active process. *Nature immunology* **4**, 235-240, doi:10.1038/ni887 (2003).

21 Martinowich, K. *et al*. DNA methylation-related chromatin remodeling in activity-dependent BDNF gene regulation. *Science* **302**, 890-893, doi:10.1126/science.1090842 (2003).

22 Kangaspeska, S. *et al*. Transient cyclical methylation of promoter DNA. *Nature* **452**, 112-115, doi:10.1038/nature06640 (2008).

23 Metivier, R. *et al*. Cyclical DNA methylation of a transcriptionally active promoter. *Nature* **452**, 45-50, doi:10.1038/nature06544 (2008).

24 Kersh, E. N. *et al*. Rapid demethylation of the IFN-gamma gene occurs in memory but not naive CD8 T cells. *Journal of immunology* **176**, 4083-4093 (2006).

25 Wu, S. C. & Zhang, Y. Active DNA demethylation: many roads lead to Rome. *Nat Rev Mol Cell Biol* **11**, 607-620, doi:nrm2950 [pii] 10.1038/nrm2950 (2010).

26 Hendrich, B., Hardeland, U., Ng, H. H., Jiricny, J. & Bird, A. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature* **401**, 301-304, doi:10.1038/45843 (1999).

27 Petronzelli, F. *et al*. Biphasic kinetics of the human DNA repair protein MED1 (MBD4), a mismatch-specific DNA N-glycosylase. *The Journal of biological chemistry* **275**, 32422-32429, doi:10.1074/jbc.M004535200 (2000).

28 Neddermann, P. & Jiricny, J. The purification of a mismatch-specific thymine-DNA glycosylase from HeLa cells. *The Journal of biological chemistry* **268**, 21218-21224 (1993).

29 Wiebauer, K. & Jiricny, J. In vitro correction of G.T mispairs to G.C pairs in nuclear extracts from human cells. *Nature* **339**, 234-236, doi:10.1038/339234a0 (1989).

30 He, Y. F. *et al*. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303-1307, doi:10.1126/science.1210944 (2011).

31 Maiti, A. & Drohat, A. C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites. *The Journal of biological chemistry* **286**, 35334-35338, doi:10.1074/jbc.C111.284620 (2011).

Papin C, Ibrahim A, Legras S, Stoll I, Bronner C, Jost B, Shen L, Zhang Y, Hamiche A (Submitted_1) Combinatorial DNA methylation code at repetitive sequences. *Submitted*

Papin C, Ibrahim A, Ouararhni K, Obri A, Grigoriev M, Bellacosa A, Dimitrov S, Hamiche A (Submitted_2) The methyl-directed nuclease activity of MBD4/MLH1 protein complex is required to protect silenced promoters from demethylation. *Submitted*