PREFACE

DNA methylation is the most prominent epigenetic modification in the genome of higher eukaryotes. In mammals, DNA methylation mainly occurs at the C5 position of symmetrically arranged cytosines in CpG dinucleotides, and approximately 60% to 90% of these CpG sites are modified. Appropriate DNA methylation is a prerequisite for normal development and is involved in various processes such as gene repression, imprinting, X-chromosome inactivation, suppression of repetitive genomic elements and carcinogenesis. The importance of such processes is exemplified by the essential requirement for DNA methyltransferases (DNMT1, DNMT3A, and DNMT3B) in embryonic and early mammalian development. The biology of DNA methylation has recently attracted attention because of the recent findings that DNA methylation can be reversed actively in mammalian cells and that the 5-methylcytosine bases can be further modified by hydroxyl group to yield hydroxymethylcytosine bases.

This thesis containing collected papers and discussion of my studies at Department of Molecular Engineering, Graduate School of Engineering, Kyoto University during April 2006 – March 2012. The aim of this thesis is revealing the molecular details of DNA methylation. PART I provides introductions of above theme. PART II reveals the selective binding of the Histone N-terminal tail recognition by *de novo* DNA methyltransferase DNMT3A by the crystal structure of the ADD domain of DNMT3A, the broad binding specificity of Methyl CpG Binding domain of MBD4 by MBD_{MBD4} crystal structures in complex with a series of DNA fragments containing modified CpG sequences and the dynamic nature of the hydroxymethylcytosine bases in mouse embryonic stem cells by biochemical assays. Finally, the summaries and conclusions of this thesis are described in PART III.

ACKNOWLEDGES

These works were performed under the direction of Professor Masahiro Shirakawa. I would like to express my graduate for his guidance and discussion throughout these works. I also thank him for providing me a comfortable working environment and am proud of the challenge to these themes. I wish to express his sincere gratitude to Associate Professors Hidehito Tochio and Mariko Ariyoshi for their guidance and constructive discussions through the course of these studies. I also wish to express my sincere appreciations to Assistant Professor Kyohei Arita for his help and advice.

It should be emphasized that the studies in this thesis have required the cooperation with a number of groups of investigation. I wish to express my gratitude to Dr. Susumu Inamoto (Institute of Molecular and Cellular Biosciences, Tokyo University) and Mr. Toshiyuki Nankumo (Department of Supramolecular Biology, Yokohama City University) for structural study of DNMT3A. I also wish thank to Professor Shoji Tajima, Associate Professor Isao Suetake and Assistant Professor Hironobu Kimura (Institute for Protein Research, Osaka University) for his helps and collaborations of in biochemical assays. I gratefully thank Dr. Naotaka Sekiyama (Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School) and Dr. Shin Isogai (Department of Molecular Engineering, Kyoto University) for their helps and discussions in NMR experiments. I wish to express my gratitude to Professor I. Hamachi and Mr. Shohei Fujishima (Department of Synthetic Chemistry and Biological Chemistry, Kyoto University), Professor Akio Ohjida (Department of Medico-Pharmaceutical Sciences, Kyushu University) for their support in ITC measurements. The X-ray diffraction experiments have been performed under the approval of the Photon Factory Program Advisory Committee.

These studies would not have been possible without help of the members of Biomolecular Function Chemistry Group. I am grateful to Dr. Kohsuke Inomata, Dr. Shin Isogai, Dr. Naoko Iwaya, Messrs. Ryuji Igarashi, Daichi Morimoto, Erik Walinda, Satoshi Ito, Syuhei Murayama, Akira Morito, Naotaka Tutumi, Yuichi Ohnoki, Shingo Sotoma, Nobuyuki Ogane, Shinji Oda, Ryugo Kawase, Kotaro Shimokawa, Takahiro Shirai, Shigeyuki Mori, Tatuya Kurimoto, Gennosuke Komiya, Hiromitu Murakami, Shintaro Ryu and Mses. Yoko Imai and Ayumi Okuda.

March, 2012

Kyoto Japan

Junji Otani

LIST OF PUBLICATIONS

PART II

CHAPTER 1.

Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX–DNMT3–DNMT3L domain

Junji Otani, Toshiyuki Nankumo, Kyohei Arita, Susumu Inamoto, Mariko Ariyoshi and Masahiro Shirakawa

EMBO reports, **2009**, 10, 1235–1241

CHAPTER 2.

The structural basis of the versatile DNA recognition by the Methyl CpG Binding Domain of MBD4

Junji Otani, Kyohei Arita, Mariko Kinoshita, Hironobu Kimura, Isao Suetake, Shoji Tajima, Mariko Ariyoshi and Masahiro Shirakawa

Nature structural and molecular biology (to be submitted)

CHAPTER 3.

De novo DNA methylation is balanced by 5-hydroxymethylcytosine-mediated passive DNA demethylation in mouse embryonic stem cells

Junji Otani, Isao Suetake, Hironobu Kimura, Mariko Ariyoshi, Masahiro Shirakawa and Shoji Tajima

Manuscript in preparation

CONTENTS

PART I

GENERAL INTRODUCTION

Epigenetics

In biology, epi-genetics is the study of heritable changes in gene expression or cellular phenotype caused by mechanisms other than changes in the underlying DNA nucleotide sequence[1]. The propagation of epigenetic states during DNA replication is critical for maintaining gene expression patterns across cell generations[2]. Phenotypically diverse, but genetically identical, cells within a multi-cellular organism originate from a single cell, the zygote. During development, the cells derived from this zygote will divide and differentiate along multiple developmental pathways until reaching their final cell fates. A different subset of their common genetic information are expressed by a different type of cells.

The minimal repeating unit of chromatin in eukaryotes is the nucleosome, which is composed of approximately 147 bp of DNA wrapped 1.7 times around an octamer of histones containing two each of H2A, H2B, H3, and H4 (Figure 1)[3,4]. Differential gene expression in alternate cell-types is dependent, in part, upon chemical modifications to DNA and histones (Figure 2)[5,6]. The chemical modifications of histones or DNA may change the surface properties of nucleosome and/or the interaction with other protein factors resulting in the change in higher order structure of chromatin which is associated with the accessibility of transcription machinery and other enzymes thereby influence the state of gene expression[7-9]. When the DNA itself is replicated, these modification patterns must also be replicated in order for epigenetic states to be inherited. Dysregulation of the epigenetic cellular machinery is increasingly being recognized as a cause of human diseases such as cancers and cardiovascular diseases[10-12].

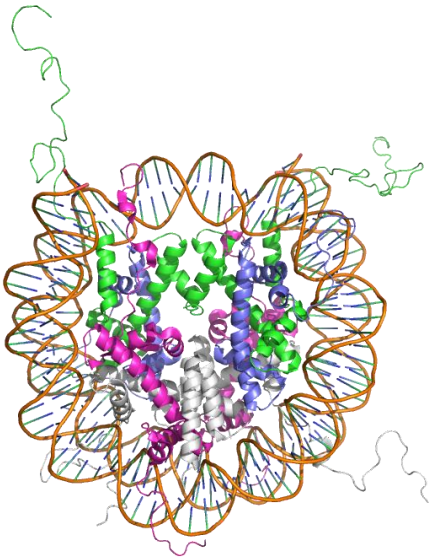


Figure 1. Crystal structure of the nucleosome core particle (PDB code: 1KX5). Two copies of histones H3, H4, H2A and H2B, colored in green, blue, red and white, form octamer wrapped by the 147bp DNA fragment.

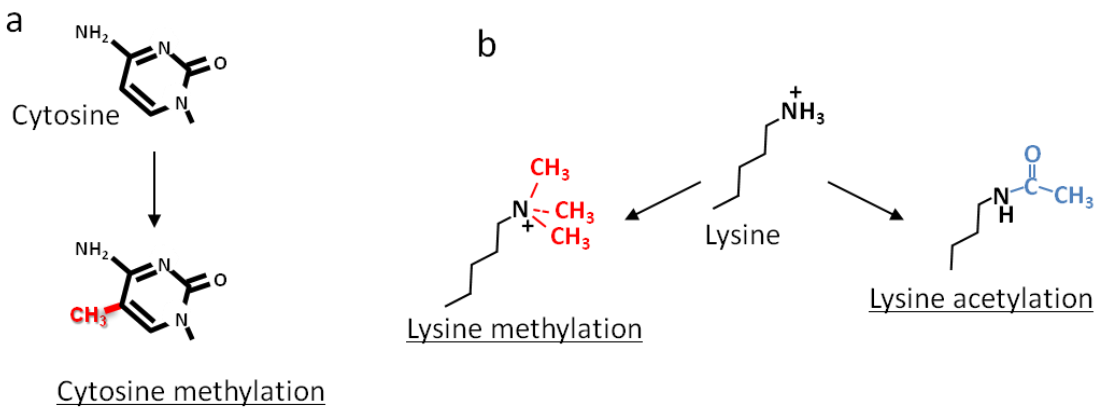


Figure 2. Examples of chemical modification of chromatin. (a) DNA methylation occurs at the C5 position of the cytosine ring. (b) The terminal amino group of the lysine side chain can be methylated and acetylated. Lysine methylation modifies protein-protein interactions and lysine acetylation alters surface property of the protein cancelling the positive charge of the lysine residue.

DNA methylation

DNA methylation is the most prominent epigenetic modification in the genome of higher eukaryotes[13,14]. In mammals, DNA methylation mainly occurs at the C5 position of symmetrically arranged cytosines in CpG dinucleotides, and approximately 60% to 90% of these CpG sites are modified[15]. Appropriate DNA methylation is a prerequisite for normal development and is involved in various processes such as gene repression, imprinting, X-chromosome inactivation, suppression of repetitive genomic elements and carcinogenesis[16]. In mammalian genomes, there are three enzymatically active mammalian DNA methyltransferases, DNMT1, DNMT3A and DNMT3B, and one related regulatory protein, DNMT3L, which lacks catalytic activity[17]. DNMT1 is primarily a maintenance methyltransferase that preserves methylation patterns during cell division. It localizes to DNA replication foci during S phase, at which it preferentially methylates hemimethylated CpG dinucleotides through its interaction with UHRF1 which can recognize the hemimethylated CpG sites[18-21]. DNMT3A and 3B are responsible for *de novo* methylation which establish the genome-wide distribution of DNA methylation during early stages of embryonic development[17,22]. DNMT3 activities are targeted by their intrinsic DNA sequence preference and through interaction with sequence-specific transcription factors or chromatin-interacting proteins[23-26]. Although DNMT3L shows no methyltransferase activity, the domain structure of the protein is similar to those of DNMT3A and DNMT3B (Figure 3) and it is indispensable for the *de novo* methylation of most imprinted loci in germ cells[27]. DNMT3L both stabilizes the conformation of the active-site loop of DNMT3A, to enhance *de novo* methylation, and increases the binding of S-adenosylmethionine which is a methyl donor in the methyltransfer reaction[27]. Histone modifications also contribute significantly to guide

DNA methyltransferases[28]. It is known that there is a correlation between DNA methylation and histone H3 lysine9 methylation which is a mark for condensed heterochromatin[29], and a inverse correlation between DNA methylation and histone H3 lysine4 methylation which is a mark for active, open-state chromatin[30,31]. The importance of DNA methylation is exemplified by the essential requirement for DNA methyltransferases (DNMT1, DNMT3A, and DNMT3B) in embryonic and early mammalian development[32,33].

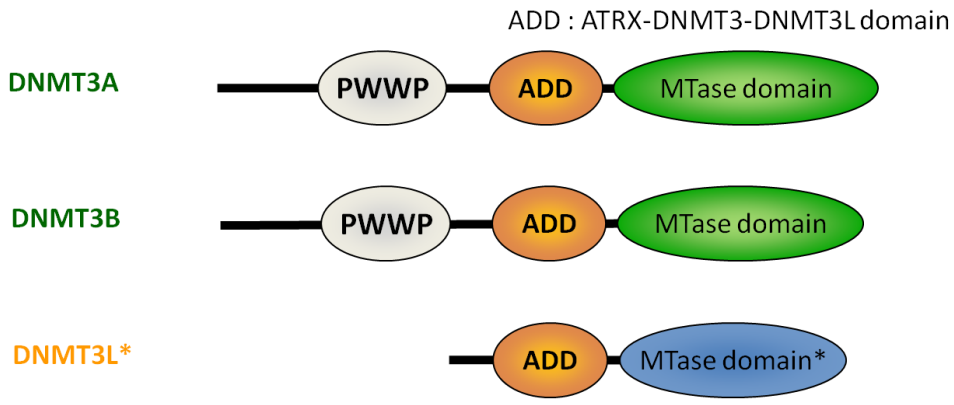


Figure 3. Domain structures of *de novo* DNA methyltransferases. DNMT3A and DNMT3B shares PWWP - proline-tryptophan-tryptophan-proline, ADD – ATRX-DNMT3-DNMT3L and MTase - methyltransferase domains, whereas DNMT3L lacks PWWP domain. The MTase domain of DNMT3L lacks critical residues for the activity thereby act as a regulatory protein.

Methylated CpG binding proteins

There are two general mechanisms by which DNA methylation inhibits gene expression: first, modification of cytosine bases can inhibit the association of some DNA-binding factors with their cognate DNA recognition sequences[34]; and second, proteins that recognize methyl-CpG can elicit the repressive potential of methylated DNA[35,36]. Currently, five Methyl CpG Binding Domain (MBD) proteins, MeCP2, MBD1, MBD2, MBD3 and MBD4 are known to share a MBD domain which consists of an α-helix and a three-stranded anti-parallel β-sheet (Figure 4)[6]. With the exception of MBD3, all MBD family members bind methylated CpG dinucleotide specifically[6]. Three structurally unrelated zinc finger proteins, ZBTB33, ZBTB4 and ZBTB38, are also known to recognize methylated CpG sequence (Figure 4)[6]. Methyl-CpG-binding proteins use transcriptional co-repressor molecules to silence transcription and to modify surrounding chromatin, providing a link between DNA methylation and chromatin remodelling and modification[37-42].

MeCP2 is the first identified MBD protein member that binds specifically to methylated DNA[43]. Mutations in the MeCP2 gene were later found to be the cause of an autism spectrum disorder, Rett syndrome which is a neurodevelopmental disorder that affects girls [44]. In addition to binding methylated DNA, MeCP2 associates with various co-repressor complexes such as Sin3a, NCoR, and c-Ski at the sites of its occupance[37,45]. When targeted to promoter DNA, MeCP2 causes strong transcriptional repression, associated with histone deacetylase activity and chromatin condensation[37,38]. MeCP2 is known to have *in vitro* activity to induce compaction of chromatin structure, in addition to its role to recruit chromatin modifier enzymes[46]. MeCP2 has multiple chromatin interacting elements including MBD domain and a methylation-independent DNA

binding site(s) and thus it can act as a bridge between DNA fragments or DNA and nucleosome to induce condensed chromatin structure[46]. MBD1 forms transient protein complex with the histone H3 lysine9 methyltransferase enzyme SETDB1 during S-phase coupling recognition of DNA methylation to modification of the surrounding chromatin by histone methylation. MBD2 and MBD3 are included in a large protein complex known as NuRD (Nucleosome Remodelling and Histone Deacetylation) which contains chromatin remodelling ATPase Mi-2, HDAC1 and HDAC2 histone deacetylases as well as other proteins[41,42]. MBD4 is unique among MBD family proteins because it has glycosylase domain at its C-terminus. MBD4 plays important roles in DNA mismatch repair to excise thymine base of T/G mismatch base pair resulting from deamination of 5-methylcytosine base (5mC) and in transcriptional repression through recruitment of Sin3A and HDAC1[47,48].

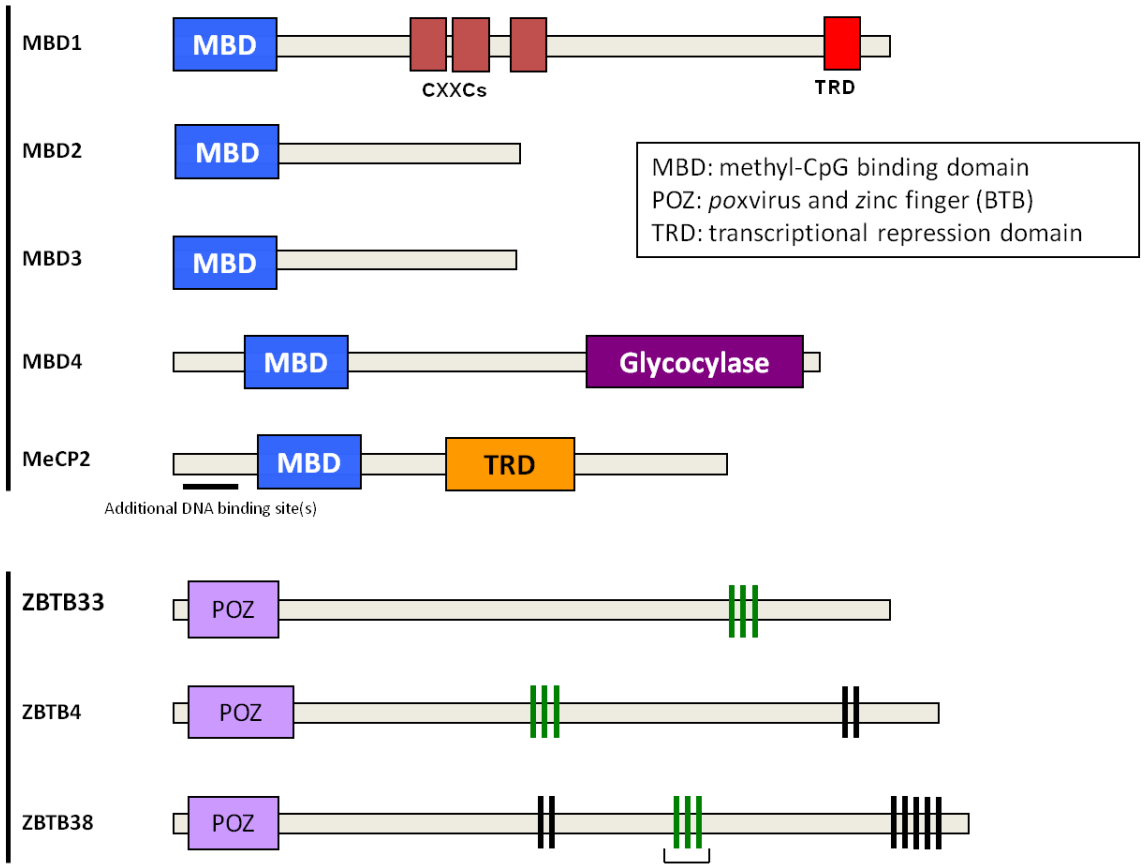


Figure 4. The domain structures of Methyl CpG binding proteins. MBD family proteins contain a conserved MBD domain. MBD1 has extra DNA binding motifs which can bind to unmodified CpG sequence. MBD1 and MeCP2 have a TRD where the amino acid sequence is not conserved. ZBTB33, ZBTB4 and ZBTB38 contain a conserved POZ domain involved in protein-protein interactions and three C2H2 zinc finger motifs, two of which are essential for binding to methylated DNA.

Discovery of 5-hydroxymethylcytosine

Recently, 5-hydroxymethylcytosine (hmC), the 6th base of DNA, was discovered as the product of the hydroxylation of 5mC by the ten-eleven translocation (TET) oncogene family members in the mouse cerebellum and embryonic stem cells[49,50]. Coincident with critical roles for DNA methyltransferases in the regulation of pluripotency, TET dependent hydroxylation also contributes to the maintenance of pluripotency in embryonic stem cells[51,52]. All three mouse TET enzymes have been shown to oxidize 5mC to hmC *in vitro* and *in vivo*[51,53], and the presence of hmC depends on pre-existing 5mC *in vivo*[54,55], suggesting that this is the only route for the synthesis of genomic hmC. Because hmC seems to be stable, it may function like other modifications by altering local chromatin structure or contributing to the recruitment or exclusion of other factors that influence transcription. For example, MBD proteins bind to methylated DNA, but do not recognize densely hydroxymethylated DNA[56,57]. It is also possible that the TET proteins may facilitate passive demethylation in dividing cells as hmC is not recognized by maintenance DNA methyltransferase, DNMT1[58]. Furthermore, TET enzymes have been shown to be able to catalyse further oxidation of hmC to yield 5-formyl- and 5-carboxylcytosine (Figure 5) in mouse ES cells and these bases can be excised by Thymine DNA Glycosylase (TDG)[59-61]. Alternatively, hmC may be deaminated by cytidine deaminases, AID/APOBECs and the resulting 5-hydroxymethyluracil base mispaired with guanine is excised by T/G mismatch glycosylases, TDG and MBD4[62-64]. Discovery of these new epigenetic modifications has led to the hypothesis that hmC may be an intermediate in the removal of 5mC, although this remains controversial.

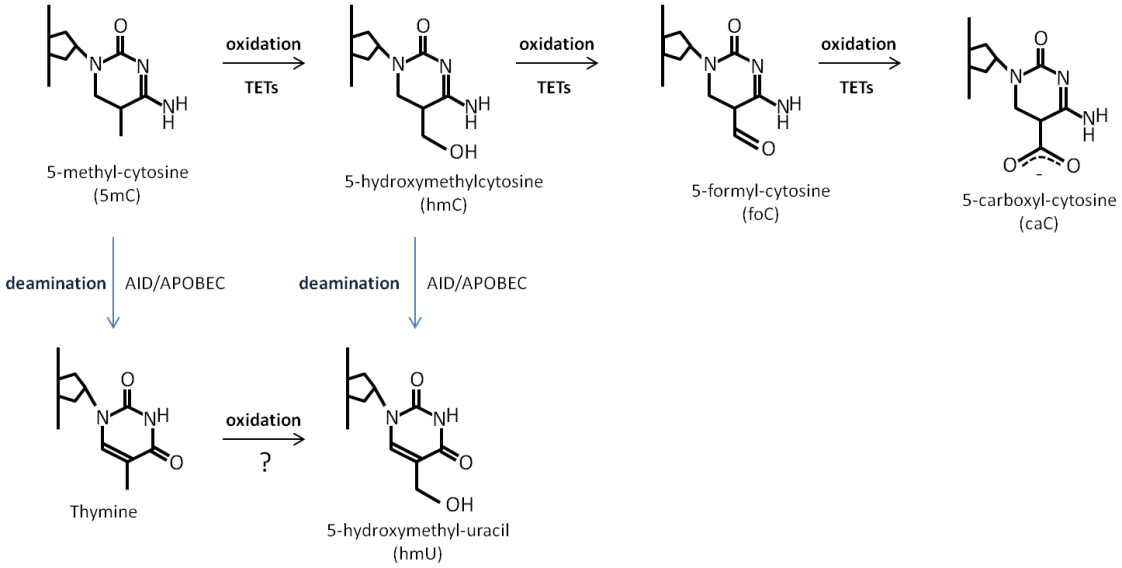


Figure 5. Oxidative modifications of cytosine and thymine.

Dynamic changes in DNA methylation during vertebrate development

Although stable and inheritable in somatic cells, global DNA methylation patterns are dynamic during the mammalian life cycle. Global remodeling of DNA methylation occurs twice in mammals, during germ cell development and preimplantation development[65,66]. The first erasure of DNA methylation marks takes place during germ cell development, when the imprinted marks are reset. This involves a wave of remethylation which is needed to establish the parental imprints. The second demethylation event takes place during preimplantation development and does not affect imprinted regions[67]. When compared to the oocyte genome, the sperm genome is highly methylated, which correlates well with its inactive chromatin state and compact structure[65]. Immunohistochemistry and bisulfate conversion experiments in mice showed that the male pronucleus gets rapidly demethylated shortly after fertilization[68,69], while the maternal genome displays a slow but progressive drop in DNA methylation levels consistent with passive demethylation. Recent reports showed the specific accumulation of hmC by the act of TET3 followed by replication dependent loss of hmC in the paternal pronucleus of the zygote, indicating the critical role of TET3 and hmC in the acute loss of 5mC in paternal genome during preimplantation development[70-73]. TET-mediated 5mC oxidation may be involved also in DNA demethylation of germ cells because high level expression of TET1 is detected coincidentally with DNA demethylation in primordial germ cells[74].

References

1. Wolffe, A.P. & Matzke, M.A. Epigenetics: regulation through repression. *Science* **286**, 481-6 (1999).
2. Margueron, R. & Reinberg, D. Chromatin structure and the inheritance of epigenetic information. *Nat Rev Genet* **11**, 285-96 (2010).
3. Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F. & Richmond, T.J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-60 (1997).
4. Davey, C., Sargent, D., Luger, K., Maeder, A. & Richmond, T. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J Mol Biol* **319**, 1097-113 (2002).
5. Struhl, K. Histone acetylation and transcriptional regulatory mechanisms. *Genes Dev* **12**, 599-606 (1998).
6. Bogdanović, O. & Veenstra, G.J. DNA methylation and methyl-CpG binding proteins: developmental requirements and function. *Chromosoma* **118**, 549-65 (2009).
7. Vermaak, D., Ahmad, K. & Henikoff, S. Maintenance of chromatin states: an open-and-shut case. *Curr Opin Cell Biol* **15**, 266-74 (2003).
8. Shogren-Knaak, M. et al. Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science* **311**, 844-7 (2006).
9. Robinson, P. et al. 30 nm Chromatin Fibre Decompaction Requires both H4-K16 Acetylation and Linker Histone Eviction. *J Mol Biol* (2008).
10. Baylin, S.B. & Jones, P.A. A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Cancer* **11**, 726-34 (2011).

11. Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet* **8**, 286-98 (2007).

12. Ordovás, J.M. & Smith, C.E. Epigenetics and cardiovascular disease. *Nat Rev Cardiol* **7**, 510-9 (2010).

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

14. Li, E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet* **3**, 662-73 (2002).

15. Ehrlich, M. et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res* **10**, 2709-21 (1982).

16. Robertson, K.D. & Wolffe, A.P. DNA methylation in health and disease. *Nat Rev Genet* **1**, 11-9 (2000).

17. Goll, M. & Bestor, T. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* **74**, 481-514 (2005).

18. Avvakumov, G.V. et al. Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. *Nature* **455**, 822-5 (2008).

19. Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y. & Shirakawa, M. Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature* **455**, 818-21 (2008).

20. Hashimoto, H. et al. The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. *Nature* **455**, 826-9 (2008).

21. Sharif, J. et al. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* **450**, 908-12 (2007).

22. Denis, H., Ndlovu, M.N. & Fuks, F. Regulation of mammalian DNA

methyltransferases: a route to new mechanisms. *EMBO Rep* **12**, 647-56 (2011).

23. Suzuki, M. et al. Site-specific DNA methylation by a complex of PU.1 and Dnmt3a/b. *Oncogene* **25**, 2477-88 (2006).

24. Brenner, C. et al. Myc represses transcription through recruitment of DNA methyltransferase corepressor. *EMBO J* **24**, 336-46 (2005).

25. Fuks, F., Burgers, W.A., Godin, N., Kasai, M. & Kouzarides, T. Dnmt3a binds deacetylases and is recruited by a sequence-specific repressor to silence transcription. *EMBO J* **20**, 2536-44 (2001).

26. Handa, V. & Jeltsch, A. Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. *J Mol Biol* **348**, 1103-12 (2005).

27. Jia, D., Jurkowska, R., Zhang, X., Jeltsch, A. & Cheng, X. Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature* **449**, 248-51 (2007).

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

29. Lehnertz, B. et al. Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Curr Biol* **13**, 1192-200 (2003).

30. Hodges, E. et al. High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. *Genome Res* **19**, 1593-605 (2009).

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

32. Li, E., Bestor, T.H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915-26 (1992).

33. Okano, M., Bell, D., Haber, D. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247-57 (1999).

34. Watt, F. & Molloy, P.L. Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes Dev* **2**, 1136-43 (1988).

35. Boyes, J. & Bird, A. DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. *Cell* **64**, 1123-34 (1991).

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

37. Jones, P.L. et al. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. *Nat Genet* **19**, 187-91 (1998).

38. Nan, X. et al. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386-9 (1998).

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

40. Sarraf, S.A. & Stancheva, I. 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 (2004).

41. Zhang, Y. et al. Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes Dev* **13**, 1924-35 (1999).

42. Wade, P.A. et al. Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nat Genet* **23**, 62-6 (1999).

43. Lewis, J.D. et al. Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell* **69**, 905-14 (1992).

44. Guy, J., Cheval, H., Selfridge, J. & Bird, A. The role of MeCP2 in the brain. *Annu Rev Cell Dev Biol* **27**, 631-52 (2011).

45. Kokura, K. et al. The Ski protein family is required for MeCP2-mediated transcriptional repression. *J Biol Chem* **276**, 34115-21 (2001).

46. Nikitina, T. et al. Multiple modes of interaction between the methylated DNA binding protein MeCP2 and chromatin. *Mol Cell Biol* **27**, 864-77 (2007).

47. 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**, 301-4 (1999).

48. Kondo, E., Gu, Z., Horii, A. & Fukushige, S. The thymine DNA glycosylase MBD4 represses transcription and is associated with methylated p16(INK4a) and hMLH1 genes. *Mol Cell Biol* **25**, 4388-96 (2005).

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

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

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

52. Koh, K.P. et al. Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell Stem Cell* **8**,

200-13 (2011).

53. Ko, M. et al. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **468**, 839-43 (2010).

54. Ficz, G. et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398-402 (2011).

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

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

57. Valinluck, V. et al. Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2). *Nucleic Acids Res* **32**, 4100-8 (2004).

58. Valinluck, V. & Sowers, L.C. Endogenous cytosine damage products alter the site selectivity of human DNA maintenance methyltransferase DNMT1. *Cancer Res* **67**, 946-50 (2007).

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

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

61. Maiti, A. & Drohat, A.C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: Potential implications for active demethylation of CpG sites. *J Biol Chem* (2011).

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

63. Rai, K. et al. DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45. *Cell* **135**, 1201-12 (2008).

64. Guo, J.U., Su, Y., Zhong, C., Ming, G.L. & Song, H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell* **145**, 423-34 (2011).

65. Morgan, H.D., Santos, F., Green, K., Dean, W. & Reik, W. Epigenetic reprogramming in mammals. *Hum Mol Genet* **14 Spec No 1**, R47-58 (2005).

66. Wu, H. & Zhang, Y. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. *Genes Dev* **25**, 2436-52 (2011).

67. Mann, M.R. & Bartolomei, M.S. Epigenetic reprogramming in the mammalian embryo: struggle of the clones. *Genome Biol* **3**, REVIEWS1003 (2002).

68. Mayer, W., Niveleau, A., Walter, J., Fundele, R. & Haaf, T. Demethylation of the zygotic paternal genome. *Nature* **403**, 501-2 (2000).

69. Oswald, J. et al. Active demethylation of the paternal genome in the mouse zygote. *Curr Biol* **10**, 475-8 (2000).

70. Iqbal, K., Jin, S.G., Pfeifer, G.P. & Szabó, P.E. Reprogramming of the paternal genome upon fertilization involves genome-wide oxidation of 5-methylcytosine. *Proc Natl Acad Sci U S A* **108**, 3642-7 (2011).

71. Wossidlo, M. et al. 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat Commun* **2**, 241 (2011).

72. Gu, T.P. et al. The role of Tet3 DNA dioxygenase in epigenetic reprogramming

by oocytes. *Nature* **477**, 606-10 (2011).

73. Inoue, A. & Zhang, Y. Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science* **334**, 194 (2011).

74. Hajkova, P. et al. Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. *Science* **329**, 78-82 (2010).

PART II

STRUCTURAL ANALYSIS OF EPIGENETIC MARK READERS

CHAPTER 1

Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain.

Abstract

DNMT3 proteins are *de novo* DNA methyltransferases responsible for the establishment of DNA methylation patterns in mammalian genomes. Here we have determined the crystal structures of the ADD domain of DNMT3A in an unliganded form and in a complex with the N-terminal tail of histone H3. Combined with the results of biochemical analysis, the complex structure indicates that DNMT3A recognizes the unmethylated state of Lys4 in histone H3. This finding suggests that the recruitment of DNMT3A onto chromatin, and thereby *de novo* DNA methylation, is mediated by recognition of the histone modification state by its ADD domain. Furthermore, our biochemical and NMR data demonstrate a mutually exclusive binding of the ADD domain of DNMT3A and the chromo domain of HP1α to the H3 tail. These results imply that *de novo* DNA methylation by DNMT3A requires alteration of chromatin structure.

Introduction

DNA methylation is one of the major epigenetic marks associated with a repressed chromatin state and gene silencing, and it regulates various physiological events such as X chromosome inactivation, embryogenesis and genomic imprinting (Bird, 2002). In mammals, cytosine methylation at the C5 position in CpG dinucleotides is the only covalent modification of genomic DNA under physiological condition. The *de novo* establishment of DNA methylation patterns in early mammalian development involves the DNMT3 family members DNMT3A and 3B and the DNMT3-like non-enzymatic regulatory factor, DNMT3L (Goll & Bestor, 2005). DNMT3L has been shown to recognize unmethylated Lys4 of histone H3 (H3K4me0) through its ADD (ATRX-DNMT3-DNMT3L) domain, suggesting that the DNMT3A:DNMT3L complex is targeted to chromatin containing H3K4me0 (Ooi *et al*, 2007). Genome-wide analyses have also shown that the chromatin regions tagged with methylated H3K4 are protected from CpG methylation (Okitsu & Hsieh, 2007; Weber *et al*, 2007).

In mammals, *de novo* DNA methylation in pericentric heterochromatin requires tri-methylation of Lys9 in H3 (H3K9me3) by H3K9 histone methyltransferases (HMTases), Suv39h1/2, and the subsequent binding of heterochromatin protein 1 (HP1) to H3K9me3 (Lehnertz *et al*, 2003). On the other hand, DNA methylation in euchromatic regions has been suggested to involve HMTase G9a, which catalyzes mono- and dimethylation of H3K9 (El Gazzar *et al*, 2008). Recently, G9a has been shown to interact directly with DNMT3A/B and to facilitate *de novo* DNA methylation in early embryonic gene promoters independently of its HMTase activity (Epsztejn-Litman *et al*, 2008; Tachibana *et al*, 2008). Thus, the mechanism of DNMT3 recruitment for *de novo* methylation may be different for heterochromatin and euchromatin regions, and a

complex scheme of interplay between *de novo* DNA methylation and histone modifications has recently emerged. However, little is yet known about molecular mechanisms regulating the hierarchical order and/or cooperative action of these epigenetic traits.

Here, we report the crystal structures of the ADD domain from DNMT3A (ADD_{3A}) in a ligand-free form and in complex with the peptide derived from the N-terminal tail of histone H3 (H3-tail), and thereby provide structural insight into recognition of histone H3 by the ADD domain at atomic resolution. Together with biochemical data, the crystal structures have revealed that ADD_{3A} specifically binds to H3K4me0. Our finding suggests that the ADD domains of the DNMT3 family play a decisive role in blocking DNMT activity in areas of the genome with chromatin containing methylated H3K4. In addition, our biochemical and NMR data have demonstrated that ADD_{3A} competes with the chromo domain of HP1α ($CD_{HP1\alpha}$) for binding to the H3-tail. Considering the previous observations of recruitment of DNMTs to chromatin mediated by HP1 proteins (Fuks *et al*, 2003; Jackson *et al*, 2002), our results imply that ADD_{3A} may promote the local chromatin structure conversion required for the catalytic activity of DNMT3A by temporarily displacing HP1α.

Results

Crystal structure of the DNMT3A ADD domain

The crystal structure of the ligand-free ADD_{3A} was determined using the multiple-wavelength dispersion method with intrinsic zinc atoms at a resolution of 2.3 Å (Fig. 1A; supplementary Fig. S1A). The electron densities for four residues in a loop region (residues 577-580) and the C-terminal five residues (residues 610-614) were not observed, indicating structural disorder in these regions. The overall structure of ADD_{3A} is similar to the ADD domains of DNMT3L (ADD_{3L}) and ATRX (Argentaro *et al*, 2007; Ooi *et al*, 2007), and is composed of two C_4 type zinc fingers: GATA-1 and plant homeo domains (PHDs) type fingers (supplementary Fig. S1B). The histone binding surface character of DNMT3L seemed well conserved in the structure of ADD_{3A} (supplementary Fig. S1C).

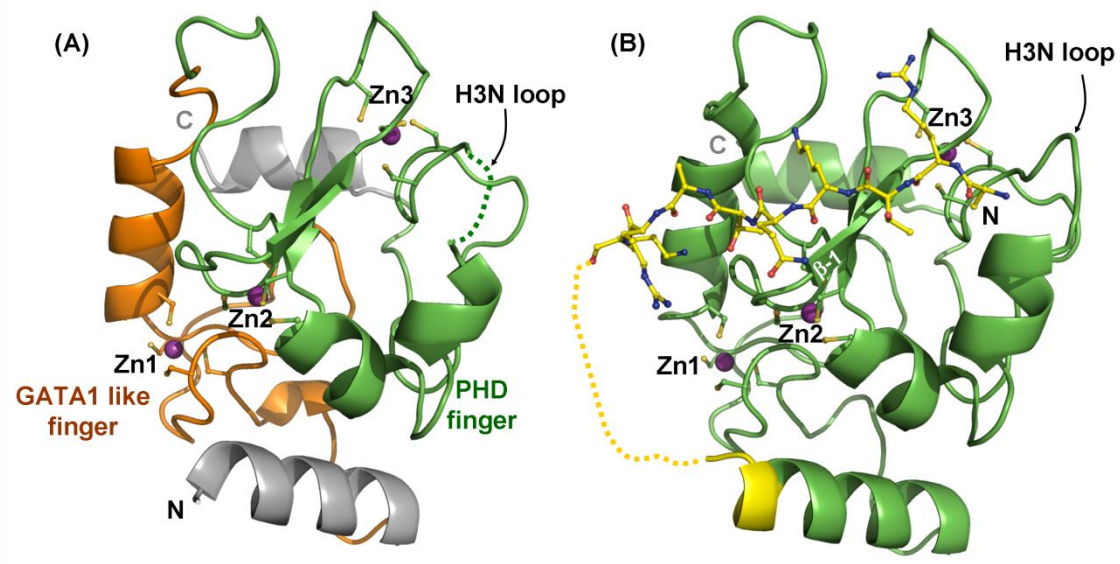


Fig. 1 | Overall structures of ADD_{3A} in the unliganded and H3-bound forms. **(A)** Ribbon representation of the unliganded ADD_{3A}. Three zinc ions and cysteine residues forming CxxC motifs are shown as magenta spheres and stick models, respectively. GATA-1-like and PHD fingers are shown in orange and green, respectively. The green dotted line represents the disordered H3N loop. **(B)** Crystal structure of ADD_{3A} (green) bound to the H3-tail (yellow). H3-peptide (residues 1-9) is shown as a ball-and-stick model. The disordered linker peptide that connects the C-terminus of the peptide and the N-terminus of ADD_{3A} is represented by a yellow dotted line. The H3N loop is shown in red.

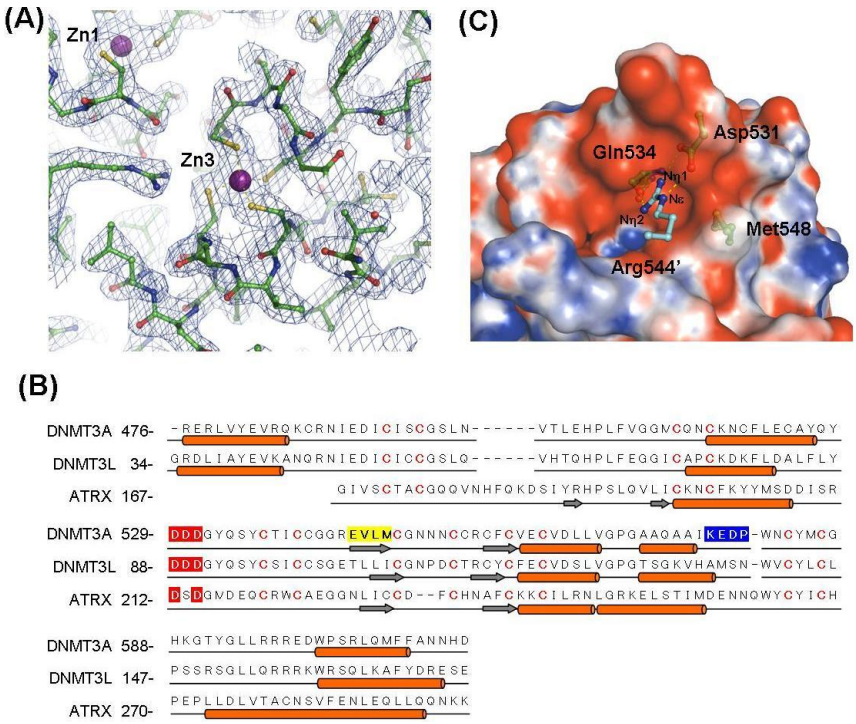


Fig. S1 | Structure of ligand-free ADD_{3A}. **(A)** Refined $2F_O$-F_C map contoured at 1.0 σ around Zn-3 coordinated by two CxxC motifs in the free-form structure of ADD_{3A}. **(B)** Multiple sequence alignment of ADD domains. Secondary structural elements of each ADD domain are indicated below the sequences. α-helices and β-strands are shown by orange cylinders and gray arrows, respectively. Acidic residues forming the H3K4 binding pockets are highlighted in red. Cysteine residues coordinating zinc ions are indicated in red. The H3N loop and β1 residues of ADD_{3A} are highlighted in blue and yellow, respectively. **(C)** Close-up view of the H3K4 binding pocket. The guanidino group of Arg544' from a symmetry-related molecule occupies the acidic pocket consisting of three aspartic acid residues, Asp529, Asp 530 and Asp 531. The side chain carbonyl group of Asp531 forms hydrogen bonds with the Nη1 and Nε atoms of Arg544'. The main chain carbonyl group of Gln534 forms hydrogen bonds with the Nη2 of Arg544'. The residues that form the H3K4 binding pocket has also been observed in ADD_{3L}.

Histone H3 binding of the DNMT3A ADD domain depends on the H3K4 methylation state

We examined binding of ADD_{3A} to the H3-tail using isothermal titration calorimetry (ITC). ADD_{3A} binds to peptides with the sequences of the N-terminal 19 and 10 amino acids of histone H3 (designated as $H3_{1-19}$ and $H3_{1-10}$, respectively) with high affinities. The dissociation constants, K_D, are 0.26 μM and 0.75 μM for $H3_{1-19}$ and $H3_{1-10}$, respectively (Table 1). To analyze the effect of the methylation of H3K4 on the interaction between ADD_{3A} and the H3-tail, we introduced analogues of non-, di- and tri-methylated lysine at the position of Lys4 in $H3_{1-19}$ according to a recently developed method (Simon *et al*, 2007), and measured their affinities for ADD_{3A} using ITC. The affinity of the $H3_{1-19}$ peptide harboring a H3K4me0 analogue for ADD_{3A} (K_D = 0.57 ± 0.05 μM) was slightly smaller than that of the native $H3_{1-19}$ (Fig. 2A; Table 1). Nevertheless, the analogue peptide seemed to mimic the native peptide well (methods). On the other hand, the H3K4me2 and H3K4me3 analogue peptides exhibited approximately ten-fold lower affinities for ADD_{3A}, with K_D values of 4.02 ± 0.3 μM and 6.76 ± 0.14 μM, respectively (Figs 2B & 2C; Table 1). Hence, ADD_{3A} specifically recognizes the non-methylated state of H3K4. Weaker binding to the H3-tail containing methylated H3K4 analogue is in agreement with previous observations showing protection of H3K4 methylated regions from DNA methylation (Okitsu & Hsieh, 2007; Weber *et al*, 2007).

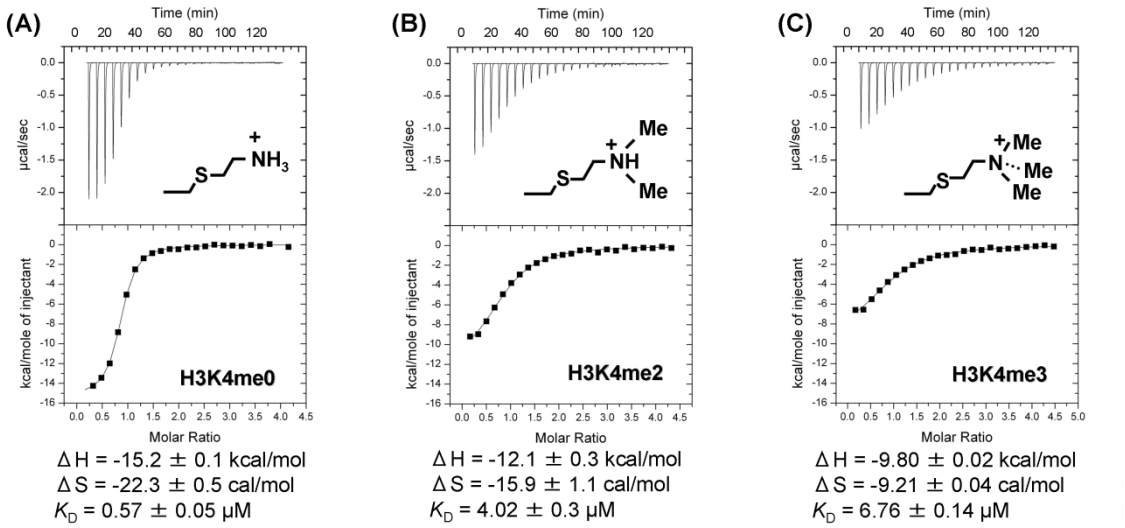
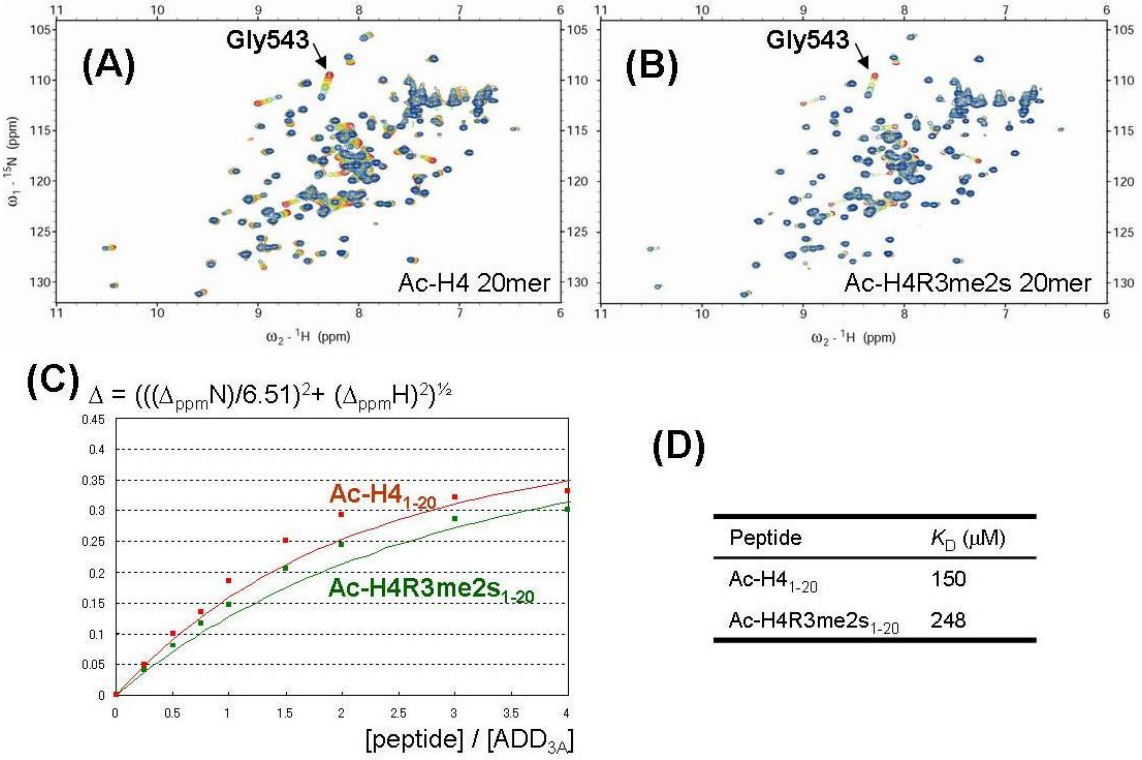


Fig. 2 | Calorimetric study of the binding of ADD_{3A} to $H3_{1\text{-}19}$ peptides containing **(A)** non-, **(B)** di- or **(C)** tri-H3K4 methylated analogues. ΔS, ΔH and K_D values are means of two experiments using different peptide and protein concentrations.

Binding experiments with H4 peptides

During preparation of this manuscript, a paper was published reporting that symmetric methylation of Arg3 in histone H4 (H4R3me2s) and its subsequent recognition by ADD_{3A} are required for DNA methylation at the human *β-globin* locus, leading to silencing of the embryo-specific genes (Zhao *et al*, 2009). Then, we have analyzed the interaction between ADD_{3A} and N-terminally acetylated histone H4 peptide (residues 1-20) with and without symmetrical dimethylation on H4R3 (H4R3me0; H4R3me2s) using ITC and NMR. Zhao and co-workers demonstrated an interaction between ADD_{3A} and N-terminally acetylated residues 1-20 of H4 using a pull-down experiment (Zhao *et al*, 2009). Although a small amount of heat emission was detected when ADD_{3A} was added to the H4R3me0 peptide solution in a buffer containing 25 mM HEPES-NaOH, 100 mM NaCl and 0.1 mM TCEP at pH 7.4, the heat emitted was so small that we could not determine the binding affinity (data not shown). For the H4R3me2s peptide, no significant heat was detected when ADD_{3A} was added (data not shown). We also analyzed the interaction of ADD_{3A} with the H4R3me0 and H4R3me2s peptides using NMR titration experiments in a buffer containing 10 mM Tris-HCl, 50 mM NaCl and 2 mM DTT at pH 7.0. In contrast to the titration of the H3 peptide, which showed slow exchange (supplementary Fig. S5A), several cross peaks exhibited chemical shift changes in the fast exchange regime on the NMR time scale, suggesting that both of the H4 peptides bind to ADD_{3A} more weakly than does the H3 peptide (supplementary Fig. S2). The K_D values of the H4R3me0 and H4R3me2s peptides were estimated to be 150 μM and 250 μM, respectively, by monitoring the chemical shifts of the main chain amide group of Gly543. In addition, we measured the NMR spectrum of ^{15}N-labeled ADD_{3A} in the presence of 4 equimolar H4R3me0 or H4R3me2s peptide. The

measurements were made under buffer conditions containing 420 mM NaCl, similar to those in the pull-down binding experiment of Zhao *et al.* The spectra were very similar to the spectrum recorded without H4 peptide in buffer containing 420 mM NaCl (data not shown). Taken together, our ITC and NMR experiments suggest that both the H4R3me0 and H4R3me2s peptides bind much more weakly to ADD_{3A} than the H3 peptide does. Zhao *et al.* used N-terminally acetylated H3 and H4 tails in their pull-down experiments. Our data suggest that N-terminal acetylation of the H3 tail largely impedes binding to ADD_{3A} (Table 1).



Peptide	K_D (μM)
Ac-H4$_{1-20}$	150
Ac-H4R3me2s$_{1-20}$	248

Fig. S2 | Binding experiment with N-terminus acetylated H4 tail without and with symmetrical dimethylation on Arg3 (Ac-H4 and Ac-H4R3me2s) using ^{1}H-^{15}N correlation spectra. ^{15}N-labeled ADD$_{3A}$ at the concentration of 100 μM was titrated by unlabeled Ac-H4 **(A)** or Ac-H4R3me2s peptide **(B)** from the peptide stock solutions at the concentration of 1.0 μM. The ^{1}H-^{15}N correlation spectra were recorded at the peptide to protein ratio of 0, 0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0 and 4.0 (red to blue). Almost identical peak shifts were observed regardless of the methylation status of H4R3. Nonlinear least-square fitting of the gradual peak shift of the backbone amide of Gly543 to the equation,

$$\Delta = \Delta_{\max} \times \left\{ \frac{\left([protein]_0 + [peptide]_0 + K_D\right) - \sqrt{\left([protein]_0 + [peptide]_0 + K_D\right)^2 - 4[protein]_0 \times [peptide]_0}}{2 \times [protein]_0} \right\}$$

where $\Delta = \sqrt{\left(\frac{\Delta_{ppm}N}{6.51}\right)^2 + \left(\Delta_{ppm}H\right)^2}$ was used to estimate K_D values of the interaction between ADD$_{3A}$ and H4 peptides (C,D).

Table 1. Dissociation constants of the interaction between ADD_{3A} and H3/H4-tail peptides determined by ITC measurements or NMR titration experiments.

Method	Protein	Peptide	K_D value (μM)
ITC	ADD_{3A}	$H3_{1-19}$	0.26
ITC	ADD_{3A}	$H3_{1-10}$	0.75
ITC	ADD_{3A}	[1]$H3K4me0_{1-19}$	[2]0.57 ± 0.05
ITC	ADD_{3A}	[1]$H3K4me2_{1-19}$	[2]4.0 ± 0.3
ITC	ADD_{3A}	[1]$H3K4me3_{1-19}$	[2]6.8 ± 0.1
ITC	ADD_{3A}	[1]$H3K9me2_{1-19}$	0.23
ITC	ADD_{3A}	[1]$H3K9me3_{1-19}$	[2]0.25 ± 0.01
ITC	ADD_{3A}	$H3_{1-19}$ K9C	0.50
ITC	ADD_{3A}	$H3R2me2a_{1-15}$	0.74
ITC	ADD_{3A}	Ac-$H3_{1-19}$	Not detected
NMR	ADD_{3A}	Ac-$H3_{1-19}$	128
ITC	ADD_{3A}	Ac-$H4_{1-20}$	Not detected
NMR	ADD_{3A}	Ac-$H4_{1-20}$	150
ITC	ADD_{3A}	Ac-$H4R3me2s_{1-20}$	Not detected
NMR	ADD_{3A}	Ac-$H4R3me2s_{1-20}$	250
ITC	ADD_{3L}	$H3_{1-19}$	[2]3.4 ± 0.87
ITC	$CD_{HP1\alpha}$	[1]$H3K9me3_{1-19}$	[2]1.3 ± 0.05
NMR	$CD_{HP1\alpha}$	[1]$H3K9me3_{1-19}$	[3]3.6 ± 1.3
NMR	$CD_{HP1\alpha}$	[1]$H3K9me2_{1-19}$	[3]7.0 ± 3.2

[1]Methylated lysine analogue peptide, [2]Means of two experiments (±mean deviations)

[3]Values obtained from competitive NMR binding experiment (see methods)

Recognition of non-methylated H3K4 by the ADD domain of DNMT3A

We also determined the crystal structure of ADD_{3A} bound to the H3-tail (Fig. 1B). In the unliganded crystal structure, the N-terminus of ADD_{3A} is located in the vicinity of a putative histone binding groove (Fig. 1A). We then linked a peptide consisting of the N-terminal 20 residues of histone H3 to the N terminus of ADD_{3A} by expressing an H3 peptide-ADD_{3A} fusion protein (methods), obtained crystals suitable for X-ray diffraction experiments, and determined the crystal structure of the linked complex at 2.3 Å resolution. The unambiguous electron density map allowed us to build a model of the H3 peptide except for the amino acid residues from 10 to 16.

As commonly observed in PHD fingers:H3-tail complexes, the extended H3 peptide fits into a shallow groove on the PHD finger motif in ADD_{3A}, resulting in 721 $Å^2$ of total buried surface area (Fig. 3A and supplementary Fig. S3). The segment from Arg2 to Thr6 forms intermolecular main chain hydrogen bonds with strand β1 of ADD_{3A} (residues 541-545), resulting in a continuous three-stranded anti-parallel β-sheet. Upon binding to the H3-tail, a conformational change is induced in the H3 N-terminus recognition loop (H3N loop; residues 577-580) of ADD_{3A}, which is disordered in the unliganded form. The H3N loop is fixed by hydrogen bonds between the terminal amino group of the H3-tail and main chain carbonyl oxygen atoms of Ala575, Ile576 and Glu578 of ADD_{3A} (Fig. 3B; supplementary Fig. S4). Over 100-fold decreased affinity of N-terminus acetylated H3 peptide to ADD_{3A} indicated the importance of the N-terminus recognition (Table 1). The overall structure of the ADD_{3A}-H3 complex is similar to that of the ADD_{3L}-H3 complex, and the recognition mode of unmodified H3K4 is well conserved (Fig. 3C; supplementary Figs. S3A, S3B and S4). However, the affinity of ADD_{3A} for the H3 peptide is about 10-fold higher than that of ADD_{3L} (Table 1). The

higher H3-binding affinity of ADD_{3A} may be attributable to the substitution of two residues, Ile576 and Met548, that are located at the histone interface of ADD_{3A}. More detailed comparison is given in supplementary discussion.

In the ADD_{3A}-H3 tail structure, the side chain of H3R2 is fully exposed to solvent, as was previously observed in the DNMT3L-H3 complex structure (Ooi *et al*, 2007), implying that the methylation of H3R2 does not affect the affinity between H3 and ADD_{3A} (supplementary Fig. S8). We examined the effect of asymmetric dimethylation of H3R2, which is depleted from active mammalian promoter and negatively correlated with H3K4me3 (Guccione *et al*, 2007), on the binding of H3 to ADD_{3A} using ITC. The $H3_{1-15}$ peptide carrying asymmetrically dimethylated H3R2 bound to ADD_{3A} with an affinity of 0.74 μM. This value is similar to the K_D value for the unmodified H3 peptide, indicating that the ADD_{3A}-H3 interaction is independent of the methylation status of H3R2.

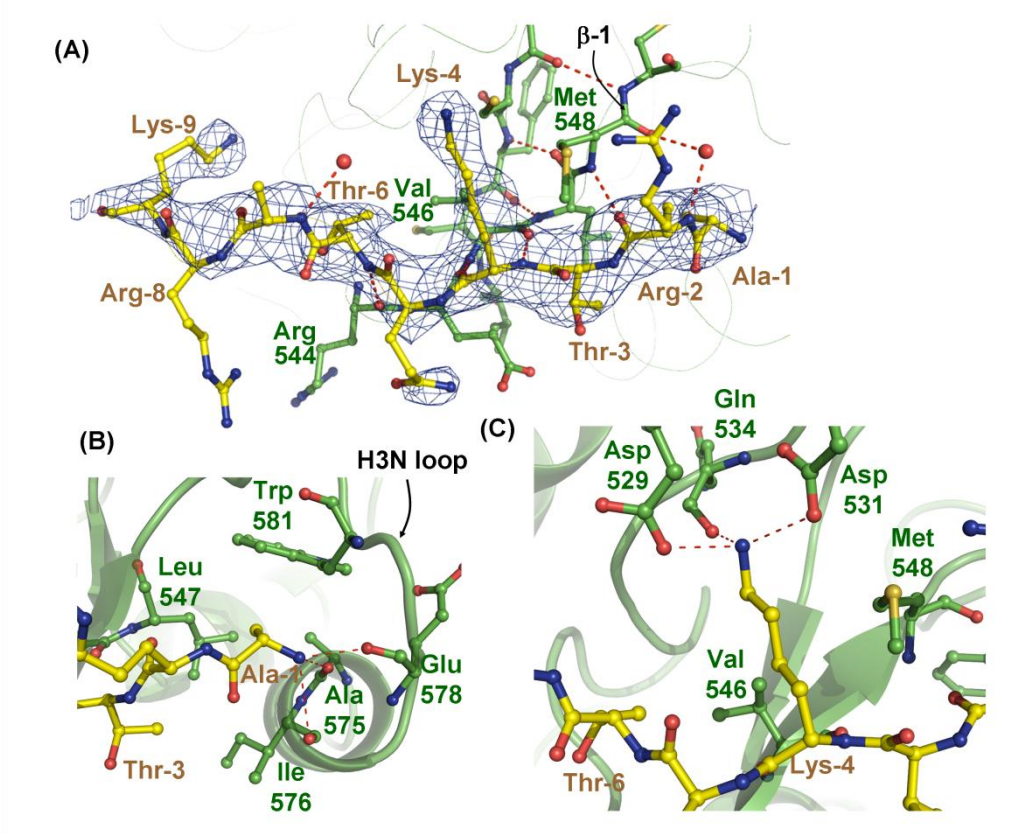


Fig. 3 | Recognition of non-methylated H3K4 by ADD_{3A}. Red dotted lines indicate hydrogen bonds (<3.3 Å) between H3 and ADD_{3A} residues. ADD_{3A} and H3 residues are represented as ball-and-stick models in green and yellow, respectively. **(A)** H3 peptide is shown with a F_o-F_c omit map contoured at 2.0 σ. **(B)** Close-up view of the recognition of the N-terminus of the H3-tail by ADD_{3A}. **(C)** Close-up view of the Lys4 binding pocket of ADD_{3A}.

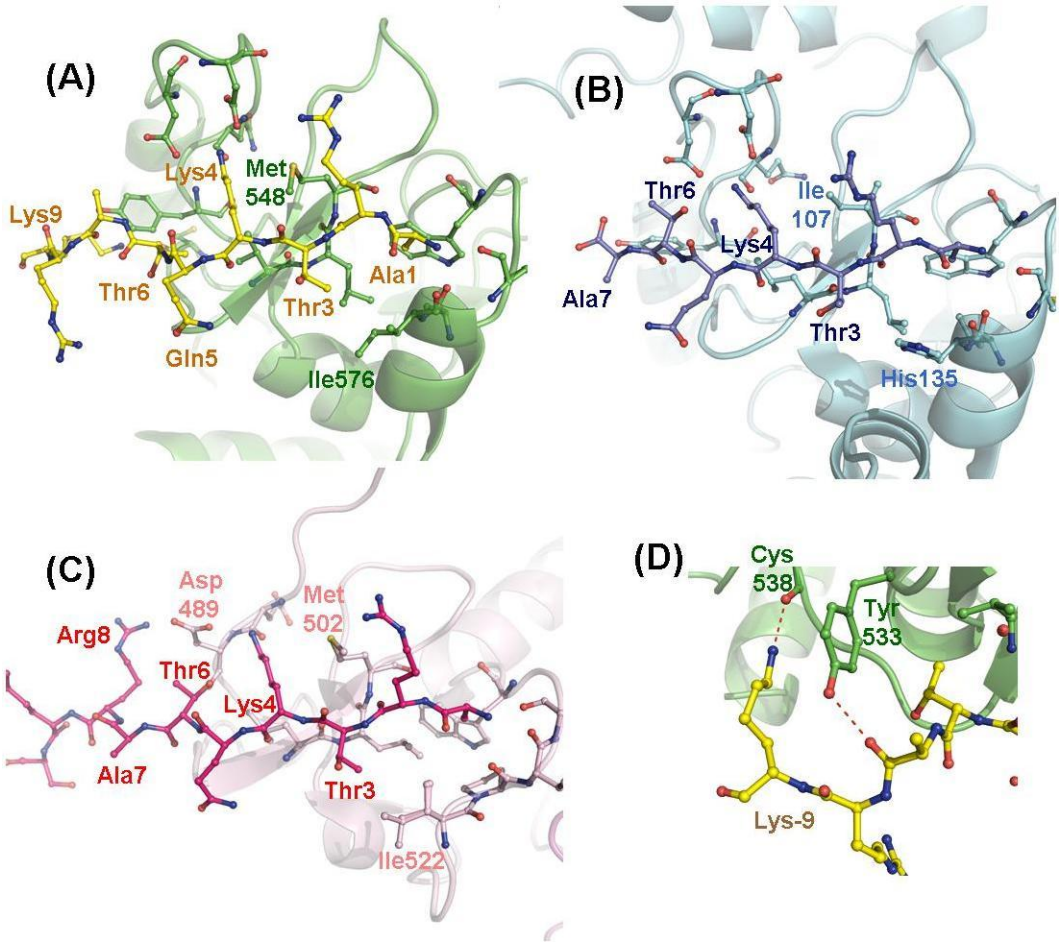


Fig. S3 | Structural comparison between the histone H3 tail bound to ADD_{3A} **(A)**, ADD_{3L} **(B)** and the PHD of BHC80 (PHD_{BHC80}) **(C)** (PDB code: 3A1B, 2PVC and 2PUY, respectively). The ribbon and stick representation of protein models of ADD_{3A}, ADD_{3L} and PHD_{BHC80} are shown in green, cyan and pink, respectively. The H3 tail peptide bound to these proteins are shown as stick models in yellow, blue and red, respectively. The regions from H3A1 to H3Q5 of the H3 peptide bound to ADD_{3A} adopts the similar conformation to those bound to ADD_{3L} and, PHD_{BHC80} which recognize non-methylated H3K4 (Lan *et al*, 2007; Ooi *et al*, 2007), while the conformation of H3 peptide from H3T6 to H3K9 observed in the ADD_{3A}-H3 complex is slightly deviated from that in the PHD_{BHC80} -H3 complex. **(D)** Close-up view of the Lys9 contact site in the ADD_{3A}-H3 complex structure.

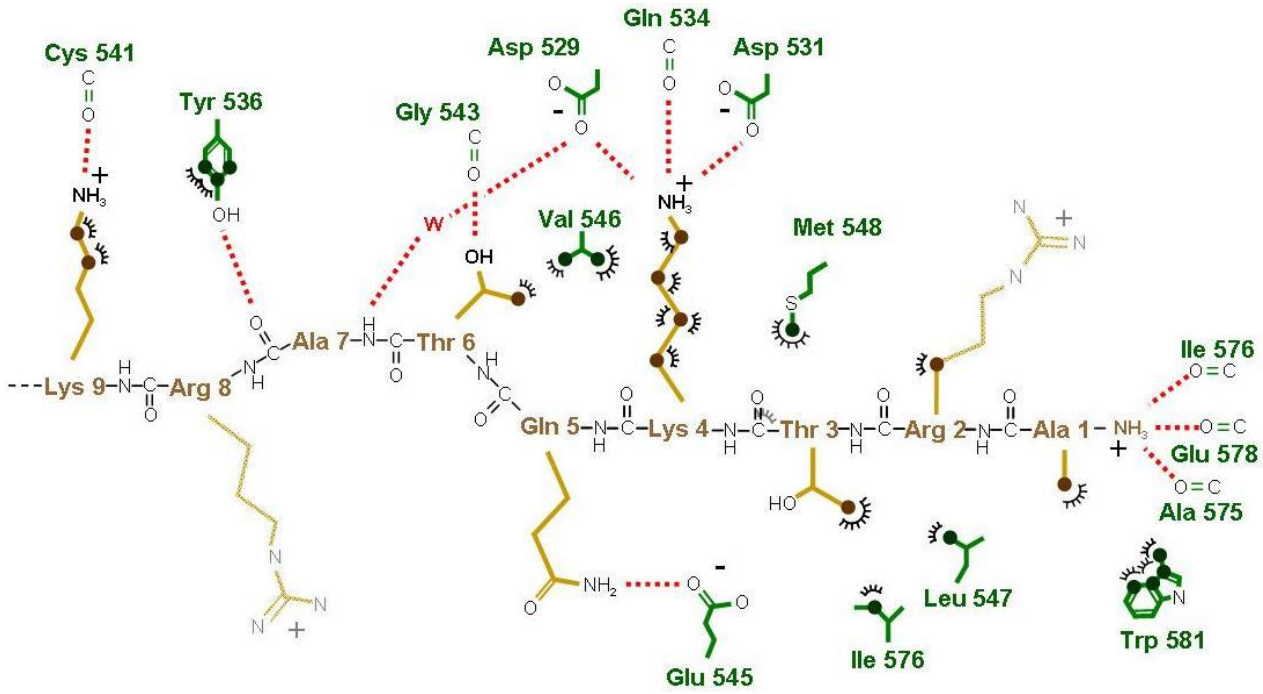


Fig. S4 | Schematic diagram of the ADD_{3A}:H3-tail interaction. Amino acid residues of ADD_{3A} involved in the H3-tail interaction are indicated. The main chain hydrogen bonds forming the inter-molecular β-sheet are not shown. Hydrogen bonds (<3.3 Å) between ADD_{3A} and the H3-tail are shown in red dotted lines, van der Waals contacts (<4.0 Å) are represented by semicircular arcs. 'W' represents a water molecule.

Interaction surface of ADD_{3A} with the H3-tail in unlinked complex

To examine whether the mode of interaction between ADD_{3A} and the H3-tail in the fusion protein is the same as that of the non-fused ADD_{3A}:H3-tail complex, we performed a chemical shift perturbation experiment. ^{15}N-labeled ADD_{3A} was titrated with unlabeled $H3_{1-19}$ peptide, and chemical shift changes of the main chain amides of ADD_{3A} were observed in ^{1}H-^{15}N correlation spectra (supplementary Fig. S5A). Of 135 cross peaks whose assignments were obtained, 37 showed chemical shift changes larger than 0.1 ppm in ^{1}H or 0.5 ppm in ^{15}N . These residues are largely confined to the groove of ADD_{3A}, in which the H3-tail fits in the crystal structure of the H3-tail: ADD_{3A} fusion protein. Thus, the resulting chemical shift perturbations are consistent with the crystal structure, suggesting that the interactions observed between ADD_{3A} and the H3-tail in the crystal are not artificially generated by the link between them (supplementary Fig. S5B). Almost identical chemical shift perturbations were observed in the NMR titration experiment with the 10mer H3 peptide, $H3_{1-10}$. During the titration of the H3-tail, all the shifted resonances of ADD_{3A} were in the slow exchange regime on the NMR time scale, which is consistent with the tight binding shown by the ITC experiments. Collectively, the structural features observed in the fusion complex are highly likely to reflect the interaction between ADD_{3A} and the H3-tail in solution.

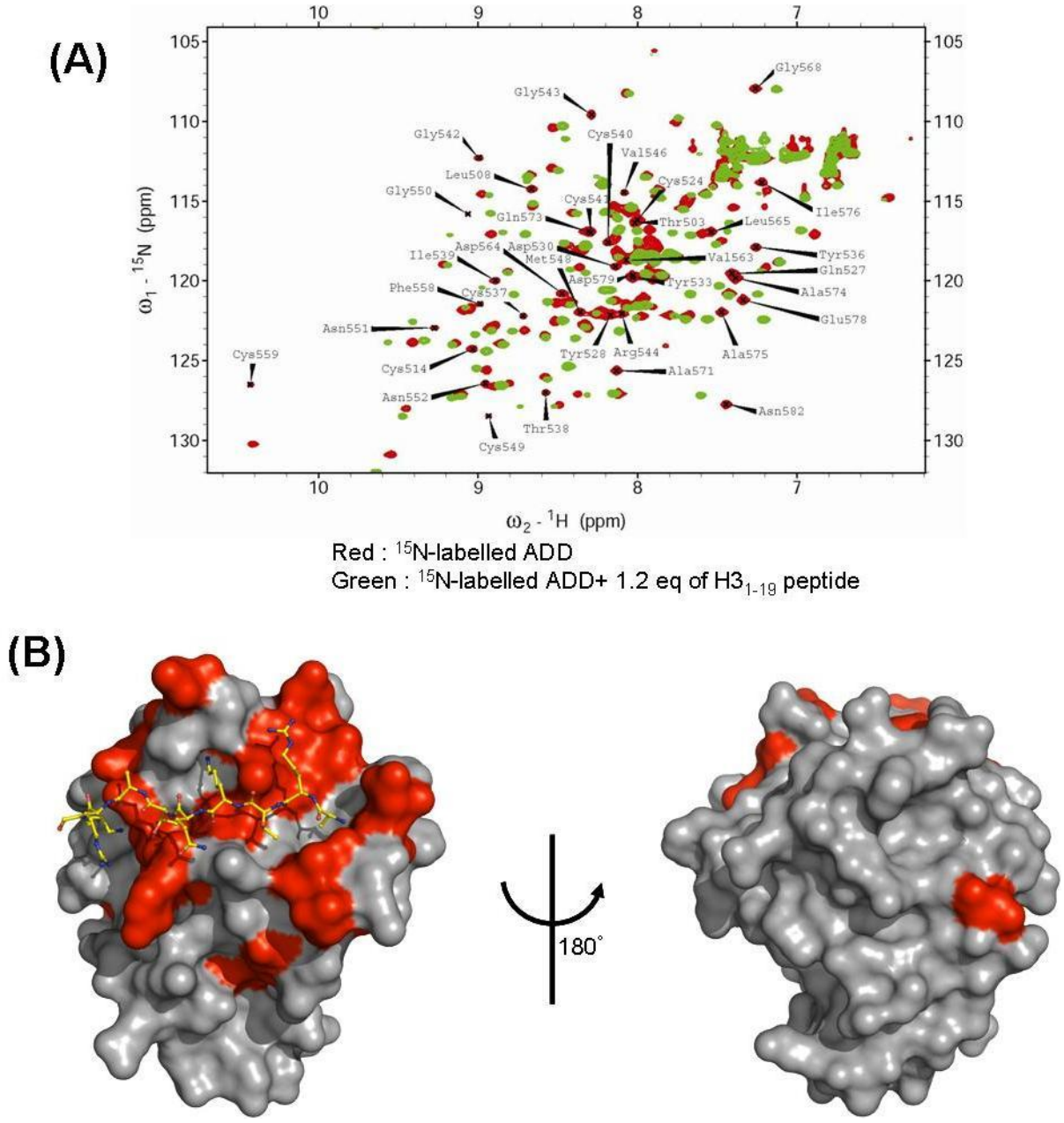


Fig. S5 | ^{1}H-^{15}N chemical shift perturbation upon H3 peptide binding. **(A)** ^{1}H-^{15}N HSQC spectra of free (red) and H3-bound ADD_{3A} (green). Residues that undergo chemical shift changes larger than 0.1 ppm in ^{1}H or 0.5 ppm in ^{15}N upon H3 binding are indicated with labels. **(B)** Surface representation of ADD_{3A}. Surface residues that exhibited large chemical shifts upon binding, labeled in A, are highlighted in red. The H3 peptide bound to ADD_{3A} in the fusion crystal is shown as stick model.

Mutually exclusive binding of DNMT3A and HP1 to the H3 tail

Several studies have demonstrated that DNMTs are recruited to chromatin through the interactions with chromo shadow domains (CSD) of HP1 proteins (Fuks *et al*, 2003; Honda & Selker, 2008). The chromo domain (CD) of HP1 proteins, which specifically binds to methylated H3K9, is indispensable for maintenance of condensed chromatin structure in heterochromatin regions. The model of the ternary complex of ADD_{3A}, $CD_{HP1\alpha}$ and H3-tail, which was built based on the crystal structures of the ADD_{3A}:H3 and $CD_{HP1\alpha}$:H3 (PDB entry; 3FDT) complexes (supplementary Fig. S6), suggested that binding of $CD_{HP1\alpha}$ to the H3-tail blocks ADD_{3A} binding, and *vice versa*, through steric occlusion of the H3 segment from Gln5 to Thr6. The H3 main chain from Ala1 to Thr6 is covered with ADD_{3A}, while $CD_{HP1\alpha}$ binds to the main chain from Gln5 to Arg8.

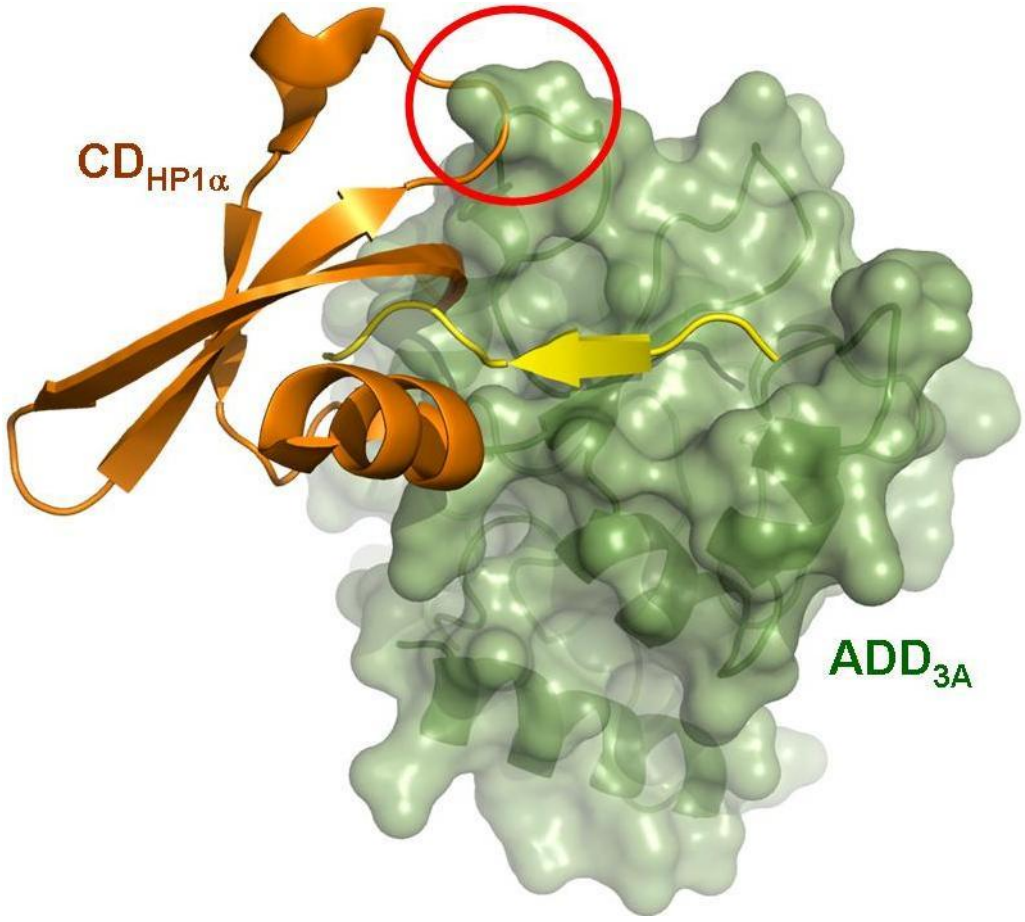


Fig. S6 | The model of the ternary complex of ADD_{3A} (green), $CD_{HP1\alpha}$ (orange) and H3-tail (yellow). The model was built by superimposing the Cα atoms of Gln5 and Thr6 of the H3 tail in the ADD_{3A}:H3-tail complex to the corresponding atoms in the $CD_{HP1\alpha}$: H3-tail complex (PDB code; 3FDT). A red circle indicates the most significant structural collision between two proteins.

We next examined whether ADD_{3A} and $CD_{HP1\alpha}$ bind simultaneously or exclusively to $H3_{1-19}$ containing H3K9me2 or H3K9me3 analogue using GST pull-down assays. Both the complexes of $CD_{HP1\alpha}$ and ADD_{3A} with K9me2/3 H3-tail were observed (Fig. 4A, lane 5; Fig. 4B, lane 5), but the ADD_{3A}: $CD_{HP1\alpha}$:H3-tail ternary complex was not (Fig. 4A, lanes 6-9; Fig. 4B, lanes 6-9). A two-fold molar excess of ADD_{3A} efficiently inhibited binding of K9me2 $H3_{1-19}$ to $CD_{HP1\alpha}$ (Fig. 4A, lanes 7). However, a ten-fold molar excess of $CD_{HP1\alpha}$ was needed to compete for approximately half of the peptide bound to ADD_{3A} (Fig. 4B, lane 9). The mutually exclusive binding of $CD_{HP1\alpha}$ and ADD_{3A} to the H3 peptide was also demonstrated by a competitive NMR titration experiment. The ^{1}H-^{15}N correlation spectrum of ^{15}N-labeled $CD_{HP1\alpha}$ in the presence of the non-labeled K9me3 H3 peptide exhibited a cross peak pattern indicating $CD_{HP1\alpha}$:H3-tail complex formation (Fig. 4C, left panel). Upon titration of non-labeled ADD_{3A}, the cross peak pattern shifted back to that of unliganded $CD_{HP1\alpha}$ in a slow exchange regime of an NMR time scale (Fig. 4C, right panel). In the presence of two-fold molar excess of ADD_{3A}, most of cross peaks of ^{15}N-$CD_{HP1\alpha}$ are superimposable onto those of the free-form protein (supplementary Fig. S7A). The absence of direct interaction between ADD_{3A} and $CD_{HP1\alpha}$ was confirmed by adding unlabeled ADD_{3A} into ^{15}N-labeled $CD_{HP1\alpha}$ (supplementary Fig. S7B). Reciprocally, ^{1}H-^{15}N correlation experiments using ^{15}N-labeled ADD_{3A} showed that H3 binding of ADD_{3A} was perturbed by $CD_{HP1\alpha}$, albeit less efficiently. Additions of 21- and 10-fold molar excesses of $CD_{HP1\alpha}$ were required for dissociation of approximately 50% of the complexes of ADD_{3A} with the H3K9me2 and H3K9me3, respectively (Fig. 4D & 4E). The K_D values of $CD_{HP1\alpha}$ for H3K9me3 and H3K9me2 were estimated to be 3.6 ± 1.3 μM and 7.0 ± 3.2 μM, respectively, from the signal intensity ratio between the H3-bound and unliganded ^{15}N-labeled ADD_{3A}

domains observed in the titration experiment. While the interaction between ADD_{3A} and CSD of HP1β has been previously observed in GST pull-down assay (Fuks *et al*, 2003), the ADD_{3A}:H3-tail complex formation was not inhibited by addition of $CSD_{HP1\alpha}$ in our NMR titration experiments and GST pull-down competition assays (data not shown).

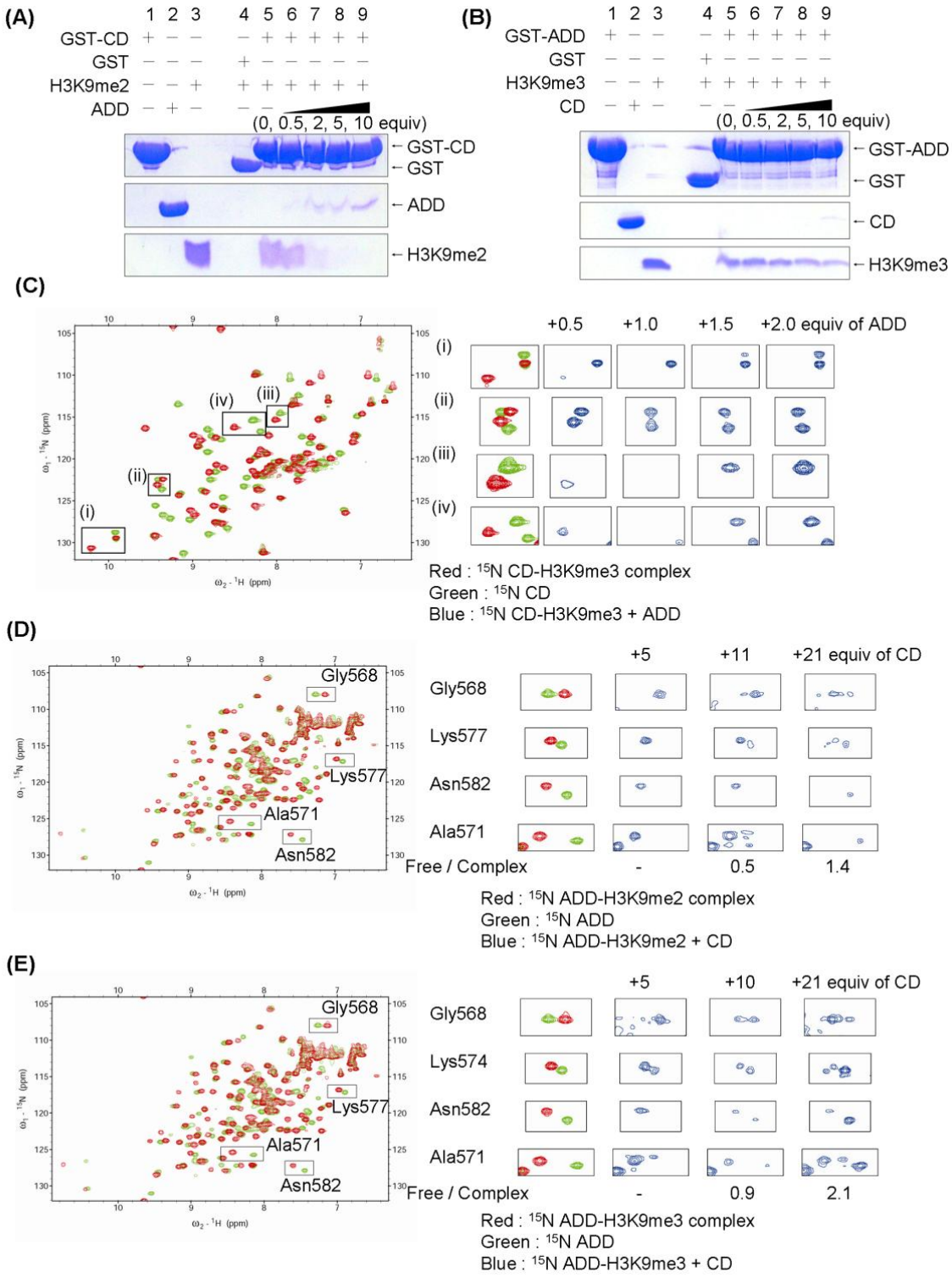

(A)
1 2 3 4 5 6 7 8 9
GST-CD
GST
H3K9me2
ADD
(0, 0.5, 2, 5, 10 equiv)
← GST-CD
← GST
← ADD
← H3K9me2
(B)
1 2 3 4 5 6 7 8 9
GST-ADD
GST
H3K9me3
CD
(0, 0.5, 2, 5, 10 equiv)
← GST-ADD
← GST
← CD
← H3K9me3
(C)
(i)
(ii)
(iii)
(iv)
+0.5 +1.0 +1.5 +2.0 equiv of ADD
Red : 15N CD-H3K9me3 complex
Green : 15N CD
Blue : 15N CD-H3K9me3 + ADD
(D)
Gly568
Lys577
Ala571
Asn582
+5 +11 +21 equiv of CD
Free / Complex - 0.5 1.4
Red : 15N ADD-H3K9me2 complex
Green : 15N ADD
Blue : 15N ADD-H3K9me2 + CD
(E)
Gly568
Lys574
Asn582
Ala571
+5 +10 +21 equiv of CD
Free / Complex - 0.9 2.1
Red : 15N ADD-H3K9me3 complex
Green : 15N ADD
Blue : 15N ADD-H3K9me3 + CD

Fig. 4 | Competitive binding of ADD_{3A} and $CD_{HP1\alpha}$ to the H3-tail. **(A)** GST pull-down assay analyzed by SDS-PAGE and CBB staining. The GST-$CD_{HP1\alpha}$ bound to K9me2 H3 peptide in the absence of ADD_{3A} (lane 5) and was competed out by adding ADD_{3A} (lanes 6-9). **(B)** Reciprocal experiment of (A) using GST-ADD_{3A}, the K9me3 H3 peptide and $CD_{HP1\alpha}$. **(C)** ^{1}H-^{15}N HSQC spectra of the ^{15}N-$CD_{HP1\alpha}$ in the unliganded (green) and K9me3 H3 bound (red) forms are superimposed (*left panel*). The $CD_{HP1\alpha}$:K9me3-H3 complex was titrated with unlabeled ADD_{3A} (*right panel*). The cross peak positions of ^{15}N-$CD_{HP1\alpha}$ shifted from those of the H3-bound form to those of the free form during the titration. Selected spectral regions (i)-(iv) are magnified. **(D), (E)** Reciprocally, ^{15}N-labeled ADD_{3A} complexed with H3K9me2 (D) or H3K9me3 peptide (E) was titrated by unlabeled $CD_{HP1\alpha}$.

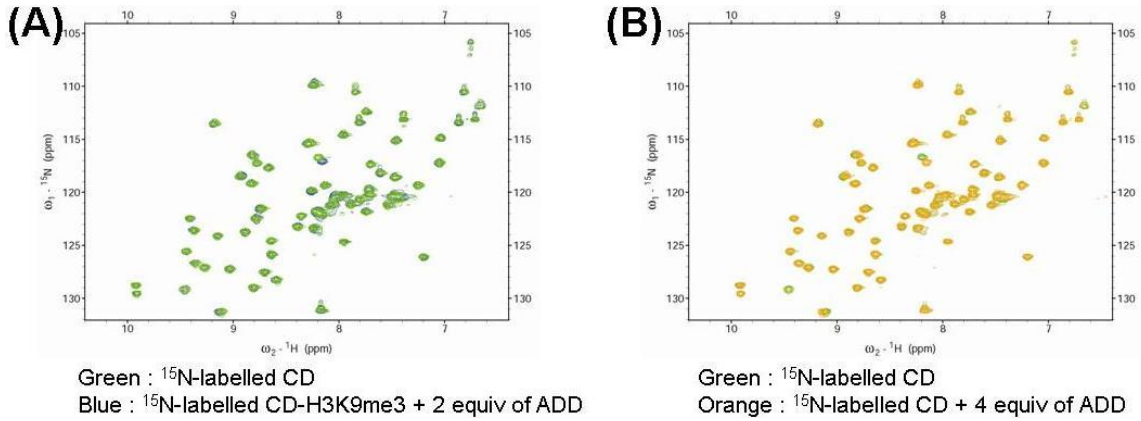


Fig. S7 | NMR binding experiments. **(A)** The spectrum of $CD_{HP1\alpha}$ (green) and $CD_{HP1\alpha}$ in the presence of 1.25-molar equiv of K9me3-H3 and 2-molar equiv of ADD_{3A} (blue). Dissociation of $CD_{HP1\alpha}$: K9me3H3-tail complex by ADD_{3A} was observed. **(B)** The spectrum of $CD_{HP1\alpha}$ (green) and $CD_{HP1\alpha}$ in the presence of 4-molar excess of ADD_{3A} (orange). No direct interaction between ADD_{3A} and $CD_{HP1\alpha}$ was observed.

Discussion

Our structural and biochemical data clearly demonstrated that the DNMT3A ADD domain specifically interacts with non-methylated H3K4, implying a role of targeting DNMT activity to regions of chromatin lacking methylated H3K4. On the other hand, HP1 proteins have previously been suggested to promote DNA methylation in heterochromatic regions by interacting with both methylated H3K9 and DNMTs. Therefore, it is intriguing that ADD_{3A} and $CD_{HP1\alpha}$ are not able to bind simultaneously to the H3-tail: Our NMR and GST pull-down analyses indicate that ADD_{3A} can effectively dissociate $CD_{HP1\alpha}$ from the H3-tail, which is probably due to higher H3-tail binding affinity of ADD_{3A} (K_D, 0.25 μM) over $CD_{HP1\alpha}$ (K_D, 1.3 μM) (Table 1). DNMT enzymatic activity has been demonstrated to be inhibited by nucleosomal structure and linker histones *in vitro* despite the association of DNA methylation with repressive chromatin *loci*, indicating that temporal local loosening of condensed chromatin might be required for DNA methylation (Takeshima *et al*, 2008). The HP1 proteins are thought to be essential for the maintenance of heterochromatin structure (Schulze & Wallrath, 2007), as suggested by the observations that a point mutation within the H3-tail binding surface of CD of HP1 causes loosening of heterochromatin structure (Cryderman *et al*, 1998) and that tethering of HP1 to euchromatic regions leads to their heterochromatinization in *Drosophila* (Li *et al*, 2003). Recently, disruption of the interaction between the HP1β CD and K9me3 H3-tail has been shown to alter chromatin structure and facilitate H2AX phosphorylation in response to DNA damage (Ayoub *et al*, 2008). Thus, it is assumed that the dissociation of HP1 from the H3-tail may cause loosening of chromatin structures. HP1 proteins have been shown to be dynamically mobile and they are continuously exchanged in chromatin in the cells

(Schmiedeberg *et al*, 2004). Therefore, once DNMT3A has been recruited to the chromatin complex, it is possible that ADD_{3A} competitively binds to H3 tails harboring K9me2 or me3, displacing the HP1 proteins. Nevertheless, HP1 proteins that are released from the H3 tail may stay in the chromatin complex through interactions between their chromoshadow domains and chromatin-binding proteins (Eskeland *et al*, 2007). Together with the abovementioned observation suggesting that relatively loose chromatin structure is required for DNMT enzymatic activity, it is tempting to speculate that H3-tail binding of ADD_{3A} may facilitate dissociation or rearrangement of HP1, which causes local loosening of heterochromatin structure at sites of *de novo* methylation.

The affinity of ADD_{3A} to the K4me0 H3-tail is markedly higher than that of the ADD from DNMT3L (K_D, 3.4 μM) (Table1), despite high sequence identity (54-55%). The higher H3-binding affinity of ADD_{3A} may be attributable to the substitution of two residues, Ile576 and Met548, that are located at the histone interface of ADD_{3A} (supplementary Figs. S3A and S3B). One of these residues, Met548 of ADD_{3A}, is positioned at the center of the H3 binding cleft. This residue has a larger contact surface (96.18 $Å^2$) with the H3 peptide than the corresponding residue in the ADD_{3L}-H3 complex, i.e., Ile107 in the ADD_{3L}-H3 structure (59.73 $Å^2$, Ooi *et al*, 2007). The other residue, Ile576 of ADD_{3A}, forms a hydrophobic pocket with Leu547, Ala575 and Trp581, which makes hydrophobic contacts with the methyl groups of Ala1 and Thr3 of histone H3. In contrast, ADD_{3L} has His135 in place of Ile576 of ADD_{3A} and seems to pack less tightly against Thr3 of histone H3 (Ooi *et al*, 2007). These contacts may result in ADD_{3A} binding more strongly to the H3 tail than does ADD_{3L}.

The affinity of the DNMT3L ADD domain for the H3-tail is comparable to those of

CDs of HP1α, HP1β and HP1γ for the K9me3 H3-tail (K_D; 13 μM, 3 μM, and 7 μM, respectively (Fischle *et al*, 2005); Table 1 for $CD_{HP1\alpha}$). Thus, although the ADD domains of both DNMT3A and DNMT3L reside in a same DNMT3A:DNMT3L complex, they may have functionally distinct roles.

In the present structure of the ADD_{3A}-H3 tail complex, Lys9 makes contacts with ADD_{3A}. However, no electron densities were observed for Arg8 and Lys9 of H3 in the ADD_{3L}-H3 tail complex (Ooi *et al*, 2007; supplementary Fig. S3D). We assume that the ADD_{3A} contacts that are formed by the side chain of Lys9 are not caused by the covalent link between the N-terminus of ADD_{3A} and the C-terminus of H3 peptide, because the ^{1}H and ^{15}N chemical shifts of the main chain amides of Cys541 and Tyr536 (with which Lys9 makes the contacts in the crystal structure) changed on binding to the H3 peptide in the NMR titration experiment (supplementary Fig. S5). However, the contribution of these contacts to the affinity between ADD_{3A} and the H3 tail seem to be relatively small because substitution of Lys9 with cysteine or methylated lysine analogues resulted in no significant change in the affinity (Table 1).

Several PHD fingers that bind to H3K4me3, such as the fingers of BPTF and ING2, have acidic residues that interact electrostatically with unmodified H3R2 on the H3 binding surface. (Li *et al*, 2006; Peña *et al*, 2006; supplementary Fig. S8D). The methylation of H3R2 has been reported to inhibit the interactions of these fingers with H3 (Iberg *et al*, 2008). However, the ADD_{3A}-H3 interaction is independent of the methylation status of H3R2. ADD_{3A} lacks acidic residues at the corresponding binding site for H3R2 and instead possesses a tyrosine residue, Tyr533 (supplementary Fig. S8A). This structural feature is conserved in the ADD domains of DNMT3A and DNMT3L (supplementary Fig. S8B), and is similar to the PHD finger of RAG2, which

recognizes methylated H3R2 through a π-π stacking interaction with its Tyr445 (Ramón-Maiques *et al*, 2007; supplementary Fig. S8C).

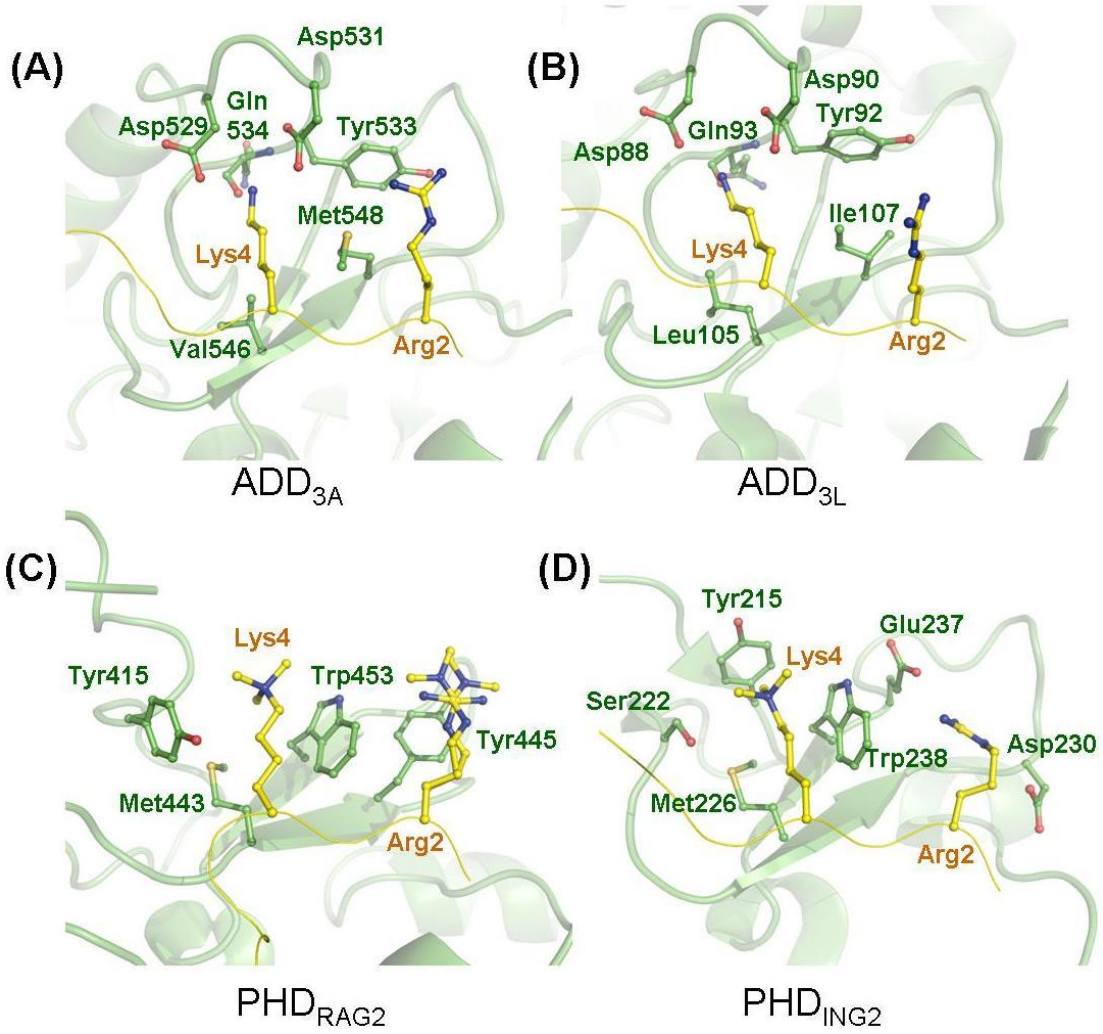


Fig. S8 | H3R2 binding pockets of PHD fingers. In the structure of ADD_{3A}-H3 complex **(A)**, the side chain of H3R2 is fully exposed to the solvent as is observed in the structure of DNMT3L-H3 complex (**B**, PDB code: 2PVC) and the methylation on this residue seems not to affect the interaction between ADD_{3A} and H3 tail. The PHD finger of RAG2 was reported to bind to the H3 tail carrying both methylated H3R2 and H3K4 with slightly higher affinity than to the tail carrying only methylated H3K4 (Ramón-Maiques *et al*, 2007). In the structure of PHD_{RAG2} complexed with methylated H3R2 (**C**, PDB code: 2V86), methylated side chain of H3R2 stacks with aromatic side chain of Tyr445. In contrast, it was proposed that the methylation of H3R2 generally

impedes binding of effector proteins that recognize H3K4me3 (Iberg *et al*, 2008). In the PHD_{ING2}-H3 structure (**D**, PDB code: 2G6Q), the side chain of H3R2 makes electrostatic contacts with acidic residues of PHD_{ING2}. This type of H3R2 recognition is widely seen in the PHD finger-H3 complex structures and seems to be disrupted by methylation on H3R2.

Methods

Protein Expression and Purification.

A DNA fragment encoding ADD_{3A} was amplified by PCR and cloned into a bacterial expression vector pGEX6P-1 (GE Healthcare Biosciences) containing a N-terminal Glutathione-S-transferase (GST) tag. ADD_{3A} was overexpressed in *E. coli* strain BL21(DE3). Cells were grown at 37 °C in Luria–Bertani medium (LB) containing 50 μg/ml ampicillin and 20 μM zinc acetate to an optical density of 0.5–0.6 at 660 nm, and then induced with 0.2 mM isopropyl β-d-thiogalactoside (IPTG) for 15 hours at 18 °C. For preparation of ^{15}N-labeled or ^{15}N,^{13}C- double labeled ADD_{3A}, M9 minimal media containing 0.5 g/L $^{15}NH_4Cl$ or 0.5 g/L $^{15}NH_4Cl$ and 1 g/L ^{13}C glucose was used instead of LB media. The following steps were carried out at 4°C. Cells were harvested by centrifugation, and lysed by sonication in 50 mM Tris-HCl pH 8.0 buffer containing 300 mM NaCl, 1 mM dithiothreitol (DTT), 5% glycerol, 0.1% Triton X-100, 1 mM phenylmethylsulfonyl fluoride (PMSF) and 20 μM zinc acetate. The clarified lysate was loaded onto Glutathione Sepharose 4 Fast Flow beads (GE Healthcare) after the debris was removed by centrifugation. ADD_{3A} was eluted from the beads by releasing from the GST tag with PreScission protease (GE Healthcare Biosciences). Subsequently, the eluted protein was concentrated and loaded onto a HiLoad 16/60 Superdex 75 (GE Healthcare Biosciences) column equilibrated with 10 mM Tris–HCl buffer pH 8.0 containing 150 mM NaCl and 2 mM DTT. Purified protein was concentrated to 10 mg ml^{-1} using an Amicon Ultra 3,000 cut-off membrane concentrator (Millipore).

The DNA fragment encoding an N-terminal fragment of histone H3 (amino-acid residues 1-20) fused with ADD_{3A} (amino-acid residues 476-614) was cloned into a pGEX4T-3 vector (GE Healthcare Bio-Sciences) engineered for protein expression with

an N-terminal GST and small ubiquitin-like modifier-1 (SUMO-1) fusion tag, generating a construct with no additional residues at the N-terminus of the histone H3-tail after removal of the tag by SUMO specific protease SENP2. The H3-tail:ADD_{3A} fusion protein was overexpressed in *E. coli* strain BL21(DE3) and purified by GST affinity column chromatography. The GST-SUMO tag was cleaved off with SENP2 on the glutathione beads, resulting in elution of the H3-tail:ADD_{3A} fusion protein. The eluted protein was further purified in the same manner as ADD_{3A}. The $CD_{HP1}\alpha$ (residues 15-78) was bacterially expressed as a GST-SUMO1-fusion protein. GST-SUMO1- $CD_{HP1\alpha}$ was digested by SENP2. GST fused $CD_{HP1\alpha}$ was prepared for GST pull-down assays. For preparation of isotope-labeled protein samples, M9 minimal media was used instead of LB media. The ADD of DNMT3L (residues 35-174) was expressed as a GST-SUMO1 fusion protein, and was purified in the same manner as ADD_{3A}.

H3 peptide preparation and in vitro binding assay.

An N-terminal peptide derived from histone H3 (residues 1-19) with an additional tryptophan residue at its C-terminus, which allowed us to determine the peptide concentration by measuring absorbance at 280 nm, was cloned into a modified pGEX4T-3 vector. The H3 peptide was expressed in BL21(DE3) for 3 hours at 37 °C induced by 0.5 mM IPTG. After cell lysis and GST affinity purification, the GST-SUMO tag was removed by digestion with SENP protease and the peptide was purified using acetone precipitation and reversed phase HPLC on a C18 column with an acetonitrile gradient in the presence of 0.05% trifluoroacetic acid.

H3 peptide analogues harboring non-, di- or trimethylated lysine were prepared by alkylation of cysteine residues as described previously (Simon *et al*, 2007).We treated

100 μM peptide carrying the K4C or K9C mutation with 10 mM DTT for 1 hour, followed by alkylation of the cysteine using distinct conditions for each methylation variant. To prepare the non-methylated lysine analogue, we treated the peptide with 200 mM (2-chloroethyl) ammonium chloride in 1M Tris-HCl (pH=8.5) at 37 °C for 4 hours. The reaction was then stopped by adding 0.7 M β-mercaptoethanol. The yield of the reaction was 80~90%. For the di- and trimethylated lysine analogues, we used the same conditions as those reported by Simon *et al.*, and the yield was almost 100%. The product peptides were then separated from the unreacted peptides using reverse-phase HPLC and used in the ITC experiments. All peptide samples were analyzed by mass spectrometry performed on an ABI Voyager Elite MALDI-TOF (Applied Biosystems).

The structure of the ADD_{3A}-H3 tail complex shows that most parts of the side chain of H3K4 make contacts with ADD_{3A}, through hydrophilic contacts at the terminal amino group and through hydrophobic contacts at the alkyl chain moiety. Therefore, it might be possible that the approximately two-fold smaller binding affinity of the H3 peptide harboring the cysteine-derived lysine analogue to ADD_{3A} (Table 1) is at least partly due to the weaker packing of the lysine analogue caused by the sulfur atom at the β position.

We purchased the following histone peptides from Toray Research Center (Tokyo, Japan) which are chemically synthesized: $H3_{1-10}$, H3R2me2a$_{1-15}$, Ac-$H3_{1-19}$, Ac-$H4_{1-20}$ and Ac-H4R3me2s$_{1-20}$.

Isothermal titration calorimetry (ITC) measurements were performed in 25 mM HEPES-NaOH buffer (pH 7.4) containing 100 mM NaCl and 0.1 mM TCEP on a MicroCal VP-ITC instrument at 25 °C. Protein solutions were exchanged into the ITC measurement buffer by gel-filtration chromatography or dialysis. Lyophilized histone H3

peptides were dissolved in the same buffer. The peptide solution at 10-20 μM in the calorimetric cell was titrated with protein solution at 200-600 μM. Binding constants were calculated by fitting the data using the ITC data analysis module of Origin 7.0 (OriginLab Corporation).

Data collection and structure determination.

Crystals of ADD_{3A} were obtained in a hanging drop consisting of 2.0 μl protein solution mixed with 1.0 μl reservoir solution containing 17% PEG 8,000, 100 mM Tris-HCl pH 8.5 and 200 mM $MgSO_4$. Crystals of H3-fused ADD_{3A} were grown under the conditions containing 10% PEG 2000 monomethyl ether and 100 mM Bis-Tris pH 5.5. X-ray diffraction data sets were collected on the beamline BL-5A at the Photon Factory, Japan, at a temperature of 100K in cryoprotectant solution containing 20% ethylene glycol. X-ray diffraction data sets were processed with the program HKL2000 (Otwinowski, 1997). The structure of unliganded ADD_{3A} was solved by the multiwavelength anomalous dispersion method using the programs SOLVE and RESOLVE (Terwilliger, 2000; Terwilliger & Berendzen, 1999). The model was built using the program COOT (Emsley & Cowtan, 2004) and refined against the data collected at a wavelength of 1.0 Å using REFMAC (Murshudov *et al*, 1997) from the CCP4 suite, yielding a crystallographic R factor of 20.9% and a free R factor of 25.0% to 2.3 Å. The structure of the H3-tail:ADD_{3A} fusion protein was solved by a molecular replacement method with the program Molrep from the CCP4 suite (Vagin & Teplyakov, 1997), using the free-form structure of ADD_{3A} as a search model and refined against data to 2.3 Å resolution with a crystallographic R factor of 19.2% and free R factor of 22.0%. The final model contains residues 1-9 and 17-20 of the H3 tail, residues 476-610 of

ADD_{3A}, 37 water molecules and seven ethyleneglycol molecules. The H3 residues 17-20 folded as an extension of the N-terminal α-helix of ADD_{3A}. Considering the crystal packing and connectivity in the poor electron density for the H3 residues 10-16, it is inferred that ADD_{3A} interacts with the H3 peptide within the same polypeptide chain in the crystal of the fusion protein. The diffraction and refinement data are summarized in supplementary Tables S1 and S2. The stereochemical quality of the final models was assessed using PROCHECK (Laskowski *et al*, 1993). All figures of protein molecules were produced using PyMOL (W. L. DeLano; http://www.pymol.org). Buried surface area analysis was carried out using the PISA server (Krissinel & Henrick, 2007).

NMR spectroscopy.

NMR spectra were recorded at 19 °C in 10 mM Tris-HCl, pH 7.0, 50 mM NaCl, 1 mM DTT, 10%(V/V) D_2O on a Bruker AVANCE 700. Spectra were processed using NMRpipe and analyzed by using Sparky (http://www.cgl.ucsf.edu/home/sparky/). The comparison of ^{1}H-^{15}N cross peaks of backbone amides of ADD_{3A} or $CD_{HP1\alpha}$ between in the free and the H3-tail bound forms was carried out based on assumption of minimum chemical shift perturbation. Backbone resonances of ADD_{3A} in the unliganded state were assigned by using 3D HNCO, HN(CA)CO, HNCA, HN(CO)CA, HNCACB, and CBCA(CO)NH. Peptide titration experiments were performed by stepwise addition of concentrated H3 peptide stock solutions into 400 μl samples of 100 μM ^{15}N-labeled ADD_{3A}. The NMR competition assay was performed as described previously (Suzuki *et al*, 2008). ^{15}N-labeled ADD_{3A} at the concentration of 90-100 μM was first titrated by unlabeled K9me2-H3 or K9me3-H3 peptide at 110-120 μM and then titrated by unlabeled 1.8-2.2 mM CD_{HP1}. Reciprocally, ^{15}N-labeled CD_{HP1} at 100 μM was first

titrated by unlabeled H3K9me3 peptide to 1.25 molar equiv and then titrated by unlabeled 510 μM ADD_{3A}. The K_D values of the interactions between $CD_{HP1\alpha}$ and K9me2-H3 or K9me3-H3 peptide were calculated from the peak height ratio of H3-bound to the free-form ^{15}N-ADD_{3A} in the spectra of the ^{15}N-labeled ADD_{3A} in the presence of H3K9me2 or H3K9me3 peptide and 10 or 21 molar equiv of $CD_{HP1\alpha}$. The assignments of ^{1}H-^{15}N cross peaks of backbone amides of ADD_{3A} in the H3-tail complex were obtained based on assumption of minimum chemical shift perturbation from those of unliganded protein.

GST pull-down assay.

GST pull-down assays were carried out with Glutathione sepharose 4FF (GE Healthcare) in the same buffer used in the ITC measurements. GST-fused $CD_{HP1\alpha}$ (residues 15-78) and GST-ADD_{3A}, each attached to the glutathione beads, were incubated with 25 μM H3K9me2 or H3K9me3 analog peptides in the presence of 0 to 250 μM ADD_{3A} and $CD_{HP1\alpha}$, respectively. The beads were extensively washed with the binding buffer containing 0.2% Triton X-100 and analyzed by 15% SDS polyacrylamide gel electrophoresis followed by CBB stain.

Coordinates.

The structures of ADD_{3A} and the H3-tail:ADD_{3A} complex have been deposited in the Protein Data Bank (accession codes 3A1A and 3A1B).

Table S1 Crystallographic data and data collection statistics

Crystallographic data and Data collection statistics						
X-ray resource	PF BL-5					
Detector	ADSC Q315					
Crystal			ADD_{3A}			$H3\text{-}ADD_{3A}$
	Peak	Edge	High remote	Low remote		
Wave-length (Å)	1.28221	1.28333	1.25730	1.29001	1.00000	1.00000
Space group	$P6_122$	←	←	←	←	$P4_12_12$
Resolution (Å)[2]	50-(2.55)2.46	←	←	←	50-(2.38)2.3	50-(2.38)2.3
Total observations	64689	67026	64522	64971	81100	293430
Unique reflections	5410	5414	5421	5415	6567	12159
Completeness (%)[2]	99.1 (92.0)	99.1 (92.0)	99.1 (91.6)	99.0 (90.6)	98.9 (99.2)	98.3 (86.9)
Rmerge (%)[1, 2]	0.085 (0.263)	0.075 (0.264)	0.087 (0.298)	0.077 (0.308)	0.063 (0.410)	0.082 (0.428)
Redundancy[2]	12.5 (9.3)	12.4 (8.6)	11.9 (7.5)	12.0 (6.7)	12.3 (12.5)	24.1 (15.5)
I / σ<I>	13.2	12.9	11.8	12.2	16.4	15

[1] $Rmerge = \Sigma h \Sigma i \mid I(h)i - \langle I(h)\rangle \mid / \Sigma h \Sigma i\, I(h)i$,where $I(h)$ is the intensity of reflection h, Σh is the sum of all measured reflections and Σi is the sum of i measurements of reflection.

[2] Numbers in parentheses are the values for the highest resolution shell of each data set.

Table S2 Refinement statistics

Refinement Statistics		
Crystal	ADD_{3A}	H3-ADD_{3A}
Resolution range (Å)	28.8- (2.36)2.30	32.7- (2.35)2.29
*R*work (%)[1, 2]	20.9 (20.5)	19.2 (23.2)
*R*free (%)[1, 2]	25 (28.5)	22 (26)
R.M.S. deviations		
Bond length (Å)	0.012	0.013
Bond angle (°)	1.316	1.501
Ramachandran plot		
most favored (%)	89.4	94.5
additional arrowed (%)	10.6	5.5

[1] *R*work and *R*free = (Σ*hkl* ||*F*o| - |*F*c||) / Σ*hkl* |*F*o|, where the free reflections (5% of the total used) were held aside for *R*free throughout refinement.

[2] Numbers in parentheses are the values for the highest resolution shell of each data set.

References

Argentaro A, Yang J, Chapman L, Kowalczyk M, Gibbons R, Higgs D, Neuhaus D, Rhodes D (2007) Structural consequences of disease-causing mutations in the ATRX-DNMT3-DNMT3L (ADD) domain of the chromatin-associated protein ATRX. *Proc Natl Acad Sci U S A* **104:** 11939-11944

Ayoub N, Jeyasekharan A, Bernal J, Venkitaraman A (2008) HP1-beta mobilization promotes chromatin changes that initiate the DNA damage response. *Nature* **453:** 682-686

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

Cryderman D, Cuaycong M, Elgin S, Wallrath L (1998) Characterization of sequences associated with position-effect variegation at pericentric sites in Drosophila heterochromatin. *Chromosoma* **107:** 277-285

El Gazzar M, Yoza B, Chen X, Hu J, Hawkins G, McCall C (2008) G9a and HP1 couple histone and DNA methylation to TNFalpha transcription silencing during endotoxin tolerance. *J Biol Chem* **283:** 32198-32208

Emsley P, Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60:** 2126-2132

Epsztejn-Litman S, Feldman N, Abu-Remaileh M, Shufaro Y, Gerson A, Ueda J, Deplus R, Fuks F, Shinkai Y, Cedar H, Bergman Y (2008) De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes. *Nat Struct Mol Biol* **15:** 1176-1183

Eskeland R, Eberharter A, Imhof A (2007) HP1 binding to chromatin methylated at H3K9 is enhanced by auxiliary factors. *Mol Cell Biol* **27:** 453-465

Fischle W, Tseng B, Dormann H, Ueberheide B, Garcia B, Shabanowitz J, Hunt D, Funabiki H, Allis C (2005) Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. *Nature* **438:** 1116-1122

Fuks F, Hurd P, Deplus R, Kouzarides T (2003) The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. *Nucleic Acids Res* **31:** 2305-2312

Goll M, Bestor T (2005) Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* **74:** 481-514

Guccione E, Bassi C, Casadio F, Martinato F, Cesaroni M, Schuchlautz H, Lüscher B, Amati B (2007) Methylation of histone H3R2 by PRMT6 and H3K4 by an MLL complex are mutually exclusive. *Nature* **449:** 933-937

Honda S, Selker E (2008) Direct interaction between DNA methyltransferase DIM-2 and HP1 is required for DNA methylation in Neurospora crassa. *Mol Cell Biol* **28:** 6044-6055

Iberg A, Espejo A, Cheng D, Kim D, Michaud-Levesque J, Richard S, Bedford M (2008) Arginine methylation of the histone H3 tail impedes effector binding. *J Biol Chem* **283:** 3006-3010

Jackson J, Lindroth A, Cao X, Jacobsen S (2002) Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature* **416:** 556-560

Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. *J Mol Biol* **372:** 774-797

Lan F, Collins R, De Cegli R, Alpatov R, Horton J, Shi X, Gozani O, Cheng X, Shi Y (2007) Recognition of unmethylated histone H3 lysine 4 links BHC80 to LSD1-mediated gene repression. *Nature* **448:** 718-722

Laskowski R, Macarthur M, Moss D, Thornton J (1993) Procheck – a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* **26:** 283-291

Lehnertz B, Ueda Y, Derijck A, Braunschweig U, Perez-Burgos L, Kubicek S, Chen T, Li E, Jenuwein T, Peters A (2003) Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Curr*

Biol **13:** 1192-1200

Li H, Ilin S, Wang W, Duncan E, Wysocka J, Allis C, Patel D (2006) Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature* **442:** 91-95

Li Y, Danzer J, Alvarez P, Belmont A, Wallrath L (2003) Effects of tethering HP1 to euchromatic regions of the Drosophila genome. *Development* **130:** 1817-1824

Murshudov G, Vagin A, Dodson E (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr* **53:** 240-255

Okitsu C, Hsieh C (2007) DNA methylation dictates histone H3K4 methylation. *Mol Cell Biol* **27:** 2746-2757

Ooi S, Qiu C, Bernstein E, Li K, Jia D, Yang Z, Erdjument-Bromage H, Tempst P, Lin S, Allis C, Cheng X, Bestor T (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448:** 714-717

Otwinowski ZaM, W. (1997) Processing of X-Ray Diffraction Data Collected in Oscillation Mode. In *Methods Enzymol.*, Sweet CWCJaRM (ed), Vol. 276, pp 307-326. New York: Academic Press

Peña P, Davrazou F, Shi X, Walter K, Verkhusha V, Gozani O, Zhao R, Kutateladze T (2006) Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2. *Nature* **442:** 100-103

Ramón-Maiques S, Kuo A, Carney D, Matthews A, Oettinger M, Gozani O, Yang W (2007) The plant homeodomain finger of RAG2 recognizes histone H3 methylated at both lysine-4 and arginine-2. *Proc Natl Acad Sci U S A* **104:** 18993-18998

Schmiedeberg L, Weisshart K, Diekmann S, Meyer Zu Hoerste G, Hemmerich P (2004) High- and low-mobility populations of HP1 in heterochromatin of mammalian cells. *Mol Biol Cell* **15:** 2819-2833

Schulze S, Wallrath L (2007) Gene regulation by chromatin structure: paradigms

established in Drosophila melanogaster. *Annu Rev Entomol* **52:** 171-192

Simon M, Chu F, Racki L, de la Cruz C, Burlingame A, Panning B, Narlikar G, Shokat K (2007) The site-specific installation of methyl-lysine analogs into recombinant histones. *Cell* **128:** 1003-1012

Suzuki C, Garces R, Edmonds K, Hiller S, Hyberts S, Marintchev A, Wagner G (2008) PDCD4 inhibits translation initiation by binding to eIF4A using both its MA3 domains. *Proc Natl Acad Sci U S A* **105:** 3274-3279

Tachibana M, Matsumura Y, Fukuda M, Kimura H, Shinkai Y (2008) G9a/GLP complexes independently mediate H3K9 and DNA methylation to silence transcription. *EMBO J* **27:** 2681-2690

Takeshima H, Suetake I, Tajima S (2008) Mouse Dnmt3a preferentially methylates linker DNA and is inhibited by histone H1. *J Mol Biol* **383:** 810-821

Terwilliger T (2000) Maximum-likelihood density modification. *Acta Crystallogr D Biol Crystallogr* **56:** 965-972

Terwilliger T, Berendzen J (1999) Automated MAD and MIR structure solution. *Acta Crystallogr D Biol Crystallogr* **55:** 849-861

Vagin A, Teplyakov A (1997) MOLREP: an Automated Program for Molecular Replacement. *Journal of Applied Crystallography* **30:** 1022-1025

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

Zhao Q, Rank G, Tan Y, Li H, Moritz R, Simpson R, Cerruti L, Curtis D, Patel D, Allis C, Cunningham J, Jane S (2009) PRMT5-mediated methylation of histone H4R3 recruits DNMT3A, coupling histone and DNA methylation in gene silencing. *Nat Struct Mol Biol* **16**, 304 - 311

CHAPTER 2

The structural basis of the versatile DNA recognition by the Methyl CpG Binding Domain of MBD4.

Abstract

MBD4 is one of the Methyl-CpG Binding Domain proteins which works in transcriptional repression and DNA repair as a glycosylase to remove thymine base within the T/G mismatch base pairs. Unlike other MBD domain proteins, MBD4 is known to be able to bind to the T/G mismatch DNA generated by spontaneous deamination of 5-methylated cytosine. We have shown that the MBD domain of MBD4 can bind to the fully methylated CpG sequence and its deamination product with similar affinity and provided the structural basis for the broad substrate specificity of the MBD domain of MBD4. The flexible DNA binding surface of MBD4 can accommodate 5-hydroxymethylated cytosine base which has recently become a focus of interest in the field of epigenetics.

Introduction

DNA methylation is the most prominent epigenetic modification in the genome of higher eukaryotes[1,2]. In mammals, DNA methylation mainly occurs at the C5 position of symmetrically arranged cytosines in CpG dinucleotides, and approximately 60% to 90% of these CpG sites are modified[3]. Appropriate DNA methylation is a prerequisite for normal development and is involved in various processes such as gene repression, imprinting, X-chromosome inactivation, suppression of repetitive genomic elements and carcinogenesis[4].

DNA methylation sites recruit mediator proteins including methyl-CpG binding domain (MBD) proteins in complex with chromatin-modifying enzymes[5]. MBD4 belongs to the MBD family and takes part in DNA repair as a thymine glycosylase to excise a thymine base of the T/G mismatch, a product of spontaneous deamination of 5-methylated cytosine (^{5m}C) of the ^{5m}C/G pair. In line with this function, mutations in MBD4 have been found in various human carcinomas associated with microsatellite instability[6] and MBD4-/- mice showed the increased frequency of C to T transitions at CpG sites[7]. Unlike other MBD family members, MBD4 binds preferentially to the T/G mismatch containing CpG sites which are generated after ^{5m}C deamination of the fully methylated CpG sequence (^{5m}CG/TG)[8]. Although the three-dimensional configuration of the MBD domain and how they recognize fully methylated CpG sites is known by the structural studies of MBD1, MBD2 and MeCP2[9-11], how MBD4 recognizes the ^{5m}CG/TG sequence remains unexplained.

Recent evidence suggests that DNA methylation can be reversed actively in mammalian cells and, although the mechanism has not been fully characterized, the pathway seems to include a base excision step of the mismatch base pair or the

oxidatively modified cytosine bases[12-17]. MBD4 and Thymine DNA Glycosylase (TDG) are the only known proteins capable of excising the thymine base from the T/G mismatch, and TDG is standing in the center of attention because TDG has been reported to act as the glycosylase for the oxidized cytosine bases in addition to the mismatch thymine bases[18,19].

Here we report the MBD domain of mouse MBD4 (MBD_{MBD4}) can bind to the DNA fragment containing symmetrically methylated CpG ($^{5m}CG/^{5m}CG$) or its deamination product ($^{5m}CG/TG$) with comparable affinity and provide the structural basis for the diverse DNA binding substrate specificity of MBD_{MBD4} with the crystal structures of MBD_{MBD4}-$^{5m}CG/^{5m}CG$ and MBD_{MBD4}-$^{5m}CG/TG$ complexes. The structures indicated the flexible nature of the DNA binding surface of MBD_{MBD4} enables the diverse substrate binding. In addition, we found that MBD_{MBD4} binding is relatively compatible with the oxidation of ^{5m}C base because of the flexible DNA binding surface.

Results

Dual binding specificity of MBD_{MBD4} for $^{5m}CG/^{5m}CG$ and $^{5m}CG/TG$ sites.

$MBD_{\mathbf{MBD4}}$ from mice consisting of residues from 69 to 136, which are conserved well among the MBD family proteins (Figure 1a), was generated for biochemical and structural analyses. The DNA binding properties of $MBD_{\mathbf{MBD4}}$ was examined quantitatively by using isothermal titration calorimetry measurements. The concentrated MBD solution is titrated against 14-bp double-stranded DNA oligomers containing a single CpG site in various modification or mismatch states. In agreement with previous reports[8], MBD_{MBD4} strongly binds $^{5m}CG/TG$ with a dissociation constant (K_D) of 98.8 nM, but neither interaction with non-metylated CG/CG nor $^{5m}CG/CG$ was observed. The affinity of MBD_{MBD4} for the $^{5m}CG/^{5m}CG$ site (K_D, 97.5 nM) was similar to the value for the $^{5m}CG/TG$ (Table 1). These results clearly demonstrate that MBD_{MBD4} is able to recognize both of the $^{5m}CG/^{5m}CG$ and the $^{5m}CG/TG$ sequences. In contrast, MBD_{MBD1} showed 5-fold stronger affinity for the $^{5m}CG/^{5m}CG$ site (K_D, 72.5 nM) than that for the $^{5m}CG/TG$ sequence (K_D, 458 nM).

The binding specificities of MBD_{MBD4} and MBD_{MBD1} were further assessed by the competitive binding assays, in which ^{32}P-labeled $^{5m}CG/^{5m}CG$ oligomer DNA was competed with the non-labeled $^{5m}CG/^{5m}CG$ or $^{5m}CG/TG$ fragment. The non-labeled $^{5m}CG/TG$ efficiently competed off the binding of MBD_{MBD4} to the $^{5m}CG/^{5m}CG$ oligomer, whereas the interaction between MBD_{MBD1} and the $^{5m}CG/^{5m}CG$ site was not abrogated under the same conditions (Figure S1). An approximately 1.5-fold higher amount of the non-labeled $^{5m}CG/^{5m}CG$ fragment was required for obtaining the competitive effect equivalent to the non-labeled $^{5m}CG/TG$. It is consistent with the similar affinities of MBD_{MBD4} for the $^{5m}CG/TG$ site and the fully methylated site, which were estimated by

our ITC experiments. Thus, $MBD_{\mathbf{MBD4}}$ is characterized as a unique MBD family protein by its dual DNA binding ability in spite of the well-conserved amino acid sequences of the MBD domain (Figure 1a).

Table 1. Thermodynamic parameters obtained by ITC experiments.

DNA		MBD4	MBD1
^{5m}CG/^{5m}CG	K_D (nM)	97.5 ± 76	72.5 ± 11
	ΔH (kJ/mol)	-7.69 ± 0.93	-54.7 ± 5.1
	$-T\Delta S$ (kJ/mol)	-32.8 ± 2.5	13.9 ± 5.3
^{5m}CG/TG	K_D (nM)	98.8 ± 42	458 ± 92
	ΔH (kJ/mol)	-18.5 ± 1.7	-46.7 ± 1.3
	$-T\Delta S$ (kJ/mol)	-21.7 ± 0.71	10.4 ±0.86
^{5m}CG/^{hm}CG	K_D (nM)	162 ± 58	1040 ± 422
	ΔH (kJ/mol)	-5.0 ± 1.3	-49.5 ± 1.2
	$-T\Delta S$ (kJ/mol)	-33.9 ± 2.3	15.1 ± 0.1
CG/TG	K_D (nM)	213 ± 58	4025 ± 45
	ΔH (kJ/mol)	-11.5 ±0.3	-43.2 ± 1.0
	$-T\Delta S$ (kJ/mol)	-26.7 ± 1.0	12.3 ± 1.1
^{hm}UG/^{5m}CG	K_D (nM)	287 ± 78	
	ΔH (kJ/mol)	-15.0 ±1.3	Not performed
	$-T\Delta S$ (kJ/mol)	-22.3 ± 1.8	
^{5m}CG/CG	K_D (nM)		3080 ± 830
	ΔH (kJ/mol)	Not detected	-36.8 ± 0.9
	$-T\Delta S$ (kJ/mol)		5.2 ± 1.6
CG/CG	K_D (nM)		33000 ± 6900
	ΔH (kJ/mol)	Not detected	-15.3 ± 1.9
	$-T\Delta S$ (kJ/mol)		-10.3 ± 2.3

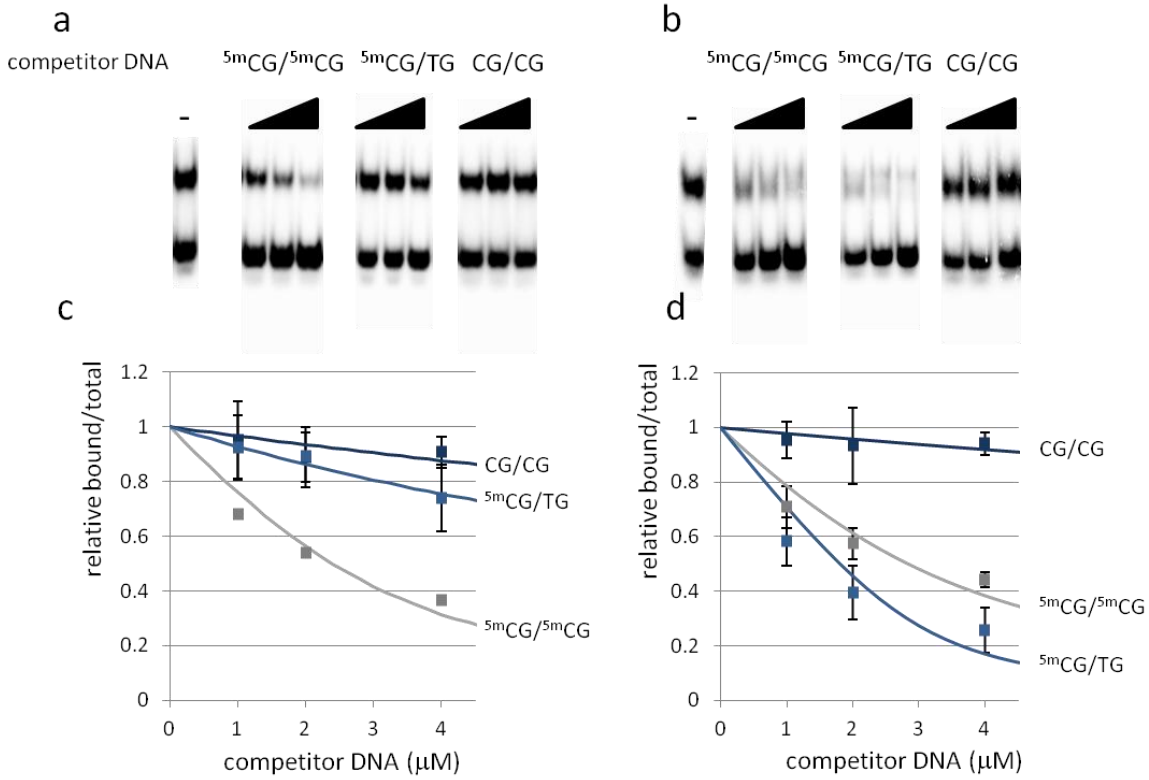


Figure S1. Competitive electrophoretic mobility shift assay. (**a**, **b**) The representative autoradiographic images of the competition assay of MBD_{MBD1} (a) and MBD_{MBD4} (b). Each of the competitor DNA sequence is indicated above. (**c**, **d**) The bound to total ratios are quantified from gel band density and the values relative to the control lanes are plotted against the amount of competitor DNA. Each dots represent the average of three independent experiments using MBD_{MBD1} (c) or MBD_{MBD4} (d). The solid lines represent the fitting curves using the Morrison's equation.

Crystal structures of MBD_{MBD4} in the complex with methylated CpG and its deamination product.

It is noteworthy that the key residues for the methylated CG recognition in MBD_{MBD1} and MBD_{MeCP2} are almost completely conserved in MBD_{MBD4} (Figure 1a). Nevertheless, MBD_{MBD4} possesses the unique dual DNA binding property. To unravel the structural elements that enable the dual binding ability of MBD4, we solved the crystal structures of MBD_{MBD4} in complex with the DNA fragments carrying the ^{5m}CG/TG or ^{5m}CG/^{5m}CG sequence.

The crystal structures of the MBD_{MBD4} - ^{5m}CG/TG complex were determined in two crystal forms at atomic resolutions: the MBD_{MBD4} bound to 14-bp and 11-bp DNA were crystallized in orthorhombic $C222_1$ and triclinic P1 forms, and were solved at 2.0 Å and 2.5 Å, respectively (Figures 1b & 1c). MBD_{MBD4} shares an overall fold, consisting of one α-helix (α1) and three β-strands (β1-3), with other MBD family proteins such as MBD_{MeCP2}, MBD_{MBD1} and MBD_{MBD2} [9-11]. The conserved hydrophobic core residues are indeed well superimposed between MBD_{MBD4} and MBD_{MeCP2} (Figures 1a & S2). The orthorhombic crystal contains one protein-DNA complex in an asymmetric unit, while two protein and one DNA molecules are contained in the triclinic crystal. In the $C222_1$ form, the C-terminal parts of MBD_{MBD4} (residues: 121~136) were swapped between a pair of symmetry-related molecules, resulting in swapped dimer formation linked through a disulfide bond (Figure 1b). However, the dimer formation is not observed either in the P1 form crystal structure (Figure 1c) or in our gel-filteration experiments (data not shown). The swapped dimer is therefore likely to be caused by a crystallographic artifact. Structural comparison of the MBD domains between two crystal forms suggests that the C-terminal swapping does not affect the DNA binding

surface. The structural perturbation is observed only on the opposite side of the DNA interface, where a helical turn of the C-terminus of the α1 helix (residues 116~118) is melted (Figure S3).

The T/G mismatch DNA fragment bound to MBD_{MBD4} adopts canonical B-form conformation in both crystal structures. In the $C222_1$ crystal structure, the ^{5m}CG/TG site is recognized by MBD_{MBD4} from a major groove side as previously observed in the methylated DNA complexes of MBD1, MBD2 and MeCP2[9-11]. In the triclinic crystal structure, one of MBD_{MBD4} molecules in an asymmetric unit binds to the ^{5m}CG/TG site in the conserved manner, whereas another protein molecule interacts with the junction of symmetry related two DNA fragments continuously linked through base stacking interactions in a head-to-tail configuration as shown in Figure 1b and supplementary figure S19. The latter protein-DNA interaction suggests the non-specific DNA binding mode of MBD4 as described below.

Furthermore, MBD_{MBD4} in complex with fully methylated CpG sequence, ^{5m}CG/^{5m}CG was crystallized in the $C222_1$ form, and its crystal structure was determined at 2.2 Å resolution. The overall structural of the ^{5m}CG/^{5m}CG complex, including the C-terminal swapping, is essentially similar to that observed in the orthorhombic crystal of the ^{5m}CG/TG complex (Figure S4).

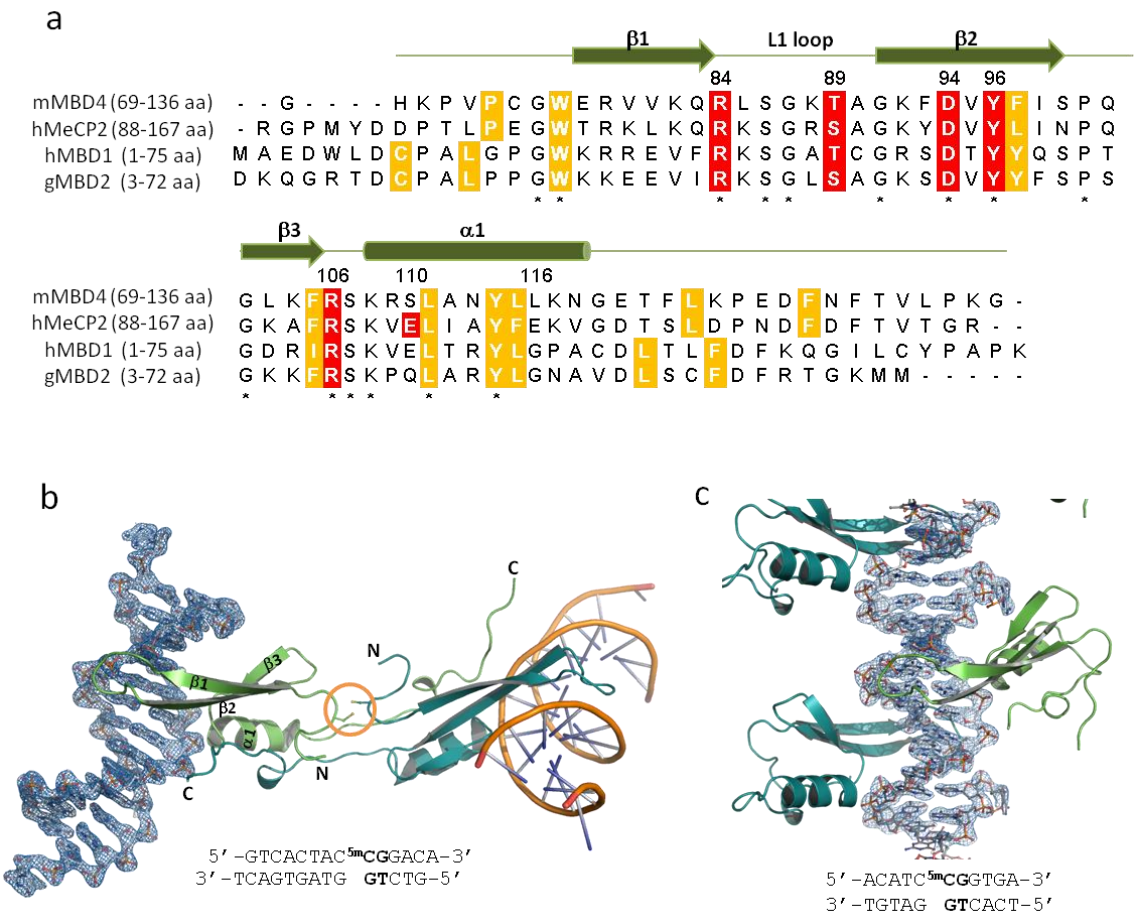


Figure 1. (**a**) The amino acid sequence alignment of structurally known MBD domains of mouse MBD4, human MeCP2, human MBD1 and chicken MBD2. The asterisks indicate that the residues are conserved among these MBD domains. The amino acid residues, highlighted in yellow, form hydrophobic core of the domain and the residues in red are involved in the recognition of the methylated CpG base pair.

(**b**, **c**) The overall structures of MBD_{MBD4} complexed with the 14- or 11-bp DNA fragment carrying ^{5m}CG/TG mismatch sequence. The blue mesh is the $2mF_O$-DF_C electron density map contoured at 1.5 σ around the DNA molecule.

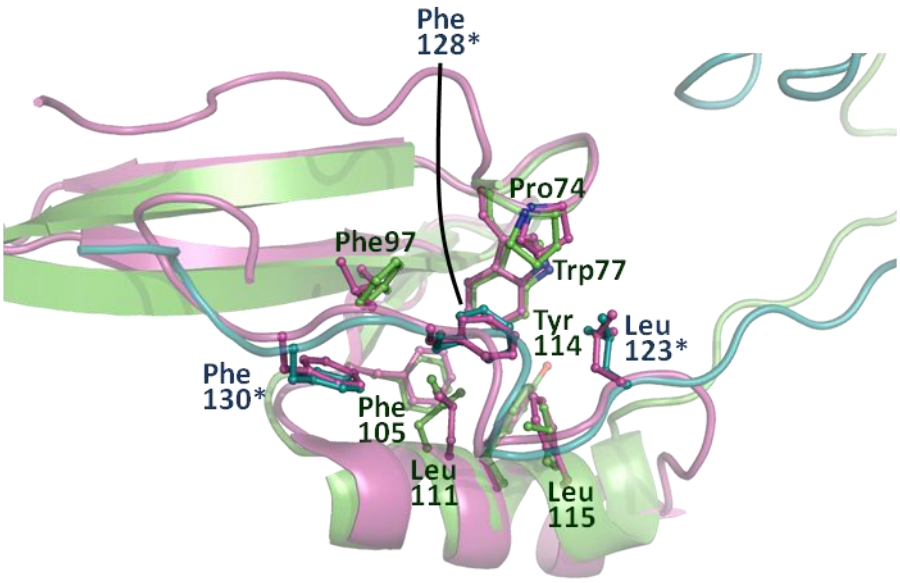


Figure S2. The structural alignment of the amino acid residues in the hydrophobic core of the MBD_{MBD4} and MBD_{MeCP2}. MBD_{MBD4} in the complex with 14-bp ^{5m}CG/TG fragment is shown in green and the C-terminal part of the domain which comes from the symmetry related molecule is shown in cyan. MBD_{MeCP2} is represented in magenta. The hydrophobic core side chains of MBD_{MBD4}, Pro74, Trp77, Phe97, Phe105, Leu111, Tyr114, Leu115, Leu123, Phe128 and Phe130 and the corresponding amino acid residues of MBD_{MeCP2} are shown as a stick model.

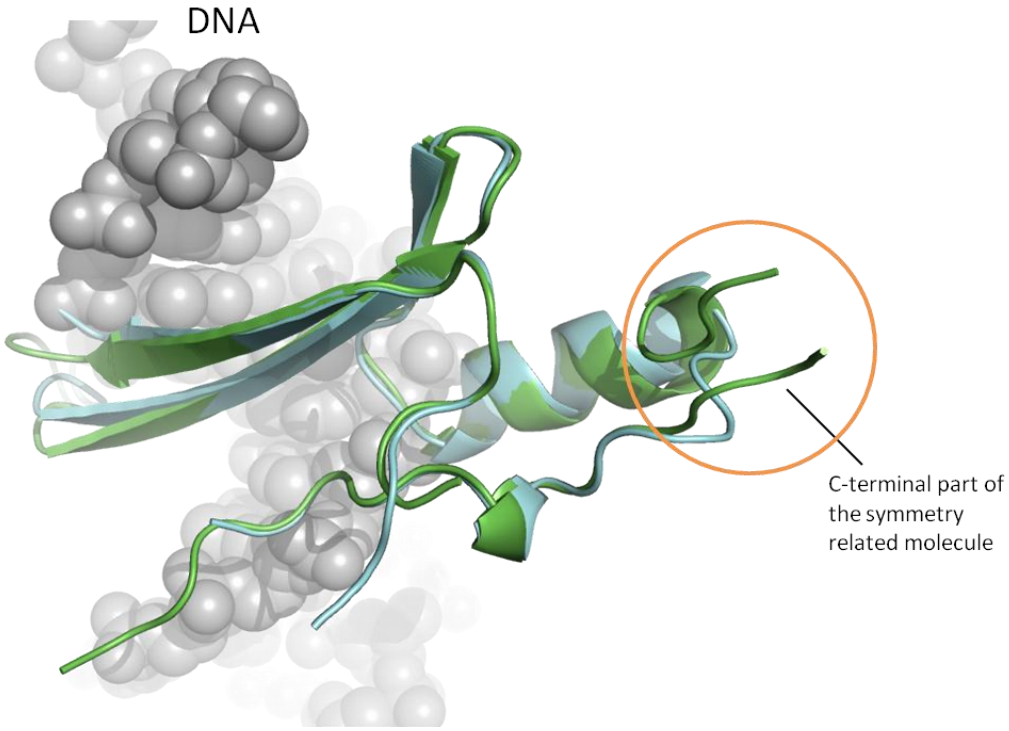


Figure S3. The structural comparison of MBD_{MBD4} structures with or without C-terminal swapping. The MBD_{MBD4} molecule in the P1 crystal (without swapping), shown in blue ribbon model, is overlaid onto the green ribbon model of the MBD_{MBD4} molecule in the $C222_1$ crystal (with swapping). The exchange of peptide chain occurs at the loop between Asn118 and Leu124. The structural perturbation because of the swapping artifact can be seen at the limited region on the other side of the DNA binding interface (orange circle).

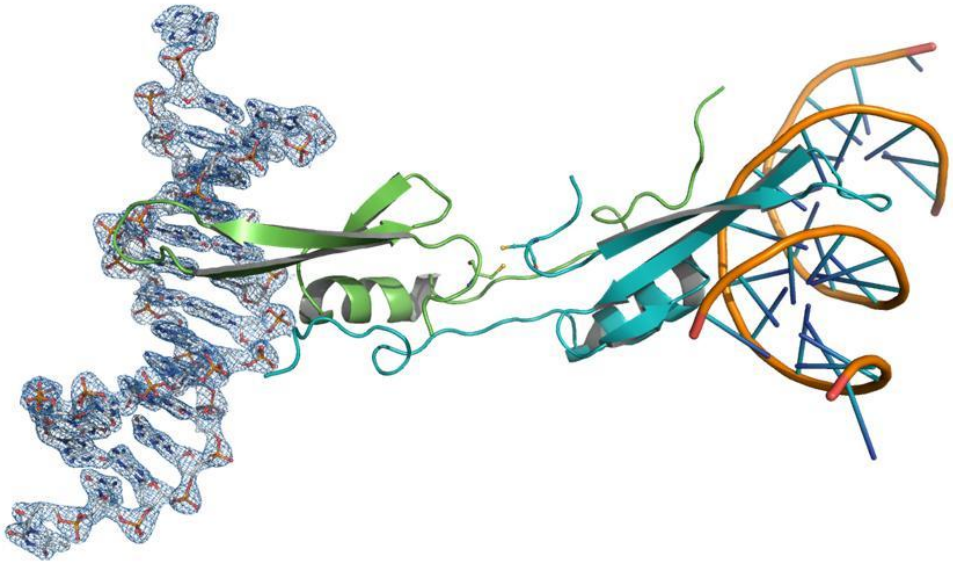


Figure S4. The overall structure of MBD_{MBD4} complexed with the 14-bp DNA fragment carrying $^{5m}CG/^{5m}CG$ sequence. The blue mesh is the $2mF_O$-DF_C electron density map contoured at 1.5 s around the DNA molecule. The C-terminal swapping and a disulphide bond are conserved with the orthorhombic crystal structure of MBD_{MBD4}-$^{5m}CG/TG$ complex.

MBD_{MBD4} recognizes the ^{5m}CG/^{5m}CG sequence with conserved arginine fingers.

The overall ^{5m}CG/^{5m}CG recognition mode of MBD_{MBD4} is essentially analogous to those of MBD_{MeCP2}, MBD_{MBD1} and MBD_{MBD2}. The arginine fingers of MBD4, Arg84 and Arg106, which are completely conserved in the MBD family (Figure 1a), recognize symmetrically arranged guanine bases in the ^{5m}CG/^{5m}CG sequence in the essentially identical manner to other MBD proteins. These Arg residues, Arg84 and Arg106 are hereafter termed as Arg fingers-1 and -2. A guanidino group of Arg finger-1 donates hydrogen bonds with the O6 and N7 atoms of the guanidine base in the lower strand, while Arg finger-2 recognizes the guanine base in the upper strand through the same hydrogen bonding pattern (Figure 2a). The aliphatic side chains of these arginine fingers each make Van der Waals contacts to the 5-methyl group of the neighboring methyl-cytosine. In addition, the main chain carbonyl group of Arg finger-2 also forms the CHO type hydrogen bond with 5-methyl group of the ^{5m}C base in the lower strand (Figure 2a).

The position of Arg finger-1 is fixed by the intra-molecular interaction with the conserved acidic residue, Asp 94. The side chain carbonyl group of Asp94 forms salt bridges with the guanidino group of the finger-1, resulting in the arginine side chain conformation suitable for recognition of the ^{5m}CG sequence. Asp94 also forms a CHO type hydrogen bond with 5-methyl group of the neighboring ^{5m}C base (Figure 2a). In contrast, Arg finger-2 lacks an acidic residue corresponding to Asp94 that locks the arginine side chain. Thus, MBD4 possesses flexibility of Arg finger-2 and asymmetric structural property in two fingers, while the Arg fingers in MBD_{MeCP2} are both locked by acidic residues (Figure 2b).

The most significant structural difference in $MBD4_{MBD}$ is observed in the orientation

of the conserved tyrosine residue, Tyr96 locating in the DNA binding surface. The corresponding tyrosine residues of the MBD_{MeCP2}, MBD_{MBD1} are MBD_{MBD2} are restrictedly oriented towards the ^{5m}C base in the lower strand through hydrophobic interactions with side chains of Lys107, Arg18, and Glu20, respectively (Figures 2c, 2d and S5). In the crystal structure of MBD_{MeCP2}-DNA complex, the side chain of Tyr123 indeed recognizes the 5-methyl group of the ^{5m}C base via a water-mediated CHO type hydrogen bond. In contrast, Tyr96 of MBD_{MBD4} is flipped almost vertical to make water-mediated interactions with the phosphate backbone of the lower DNA strand (Figure 2c). The aromatic side chain of Tyr96 is stabilized by the stacking interaction with the compact hydrophobic side chain of Val80 that is replaced by the aliphatic side chains of Lys, Arg or Glu in other MBD structures. In consequence, a more open vacant space is generated in the MBD4-DNA interface compared with other MBD proteins. Such a space generated in the vicinity of the ^{5m}C base in the lower strand is filled with ordered water molecules making a network. Of these water molecules, three are found within a distance (~4.2 Å) to make Van der Waals interactions or the CHO type hydrogen bond with the 5-methyl group of the lower strand ^{5m}C base (Figure S6). The ^{5m}C base in the upper strand is also surrounded by three water molecules (Figure S7). It is notable that positions of these hydration water molecules are well conserved between our structure and the crystal structure of the MBD_{MeCP2}-DNA complex, implying functional importance of water mediated interaction for the ^{5m}C recognition (Figure S6 & S7).

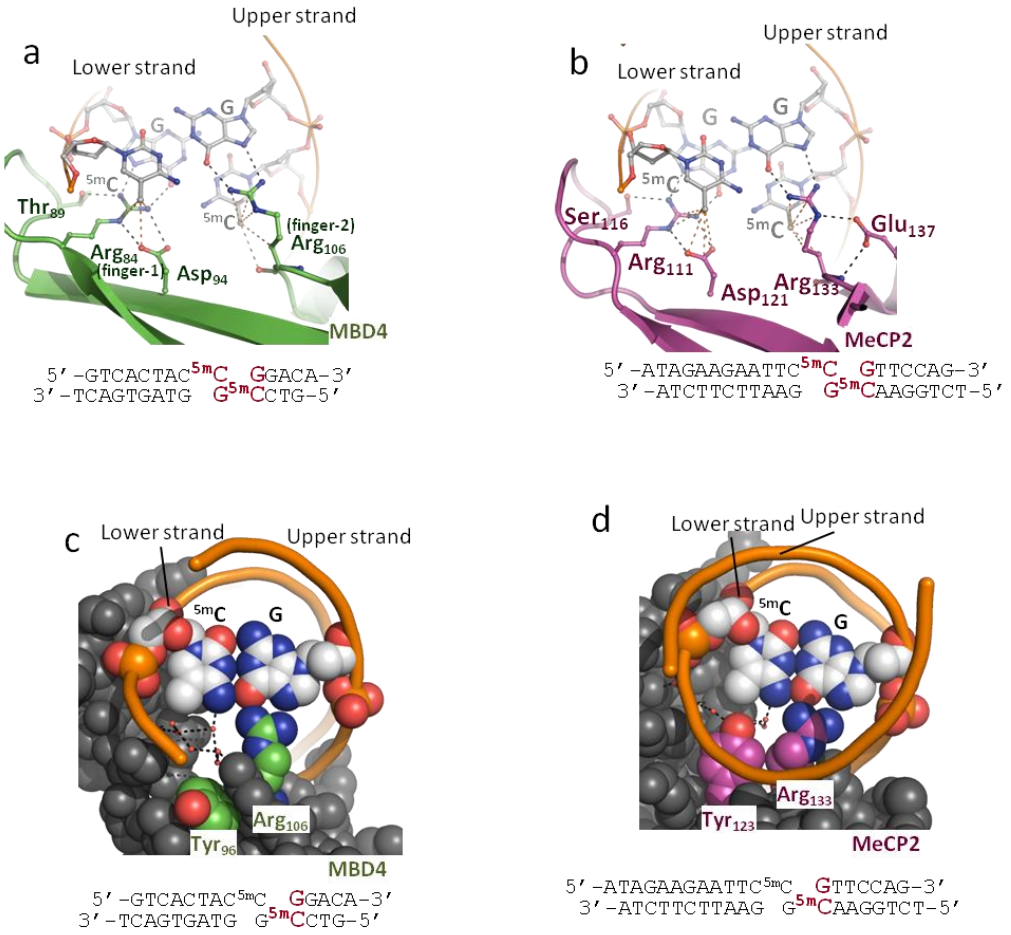


Figure 2. (**a**, **b**) The close up views of the DNA binding interface of the MBD_{MBD4}-$^{5m}CG/^{5m}CG$ complex (a) or MBD_{MeCP2}-$^{5m}CG/^{5m}CG$ complex (b). The CpG bases (shown in red in the DNA sequence, below) and amino acid residues involved in the base recognition are shown in the stick models. The dotted line in black and orange indicate the hydrogen bond (~3.2 Å) and the van der Waals interaction (~4.2 Å), respectively. The arginine finger-1 or finger-2 contact with lower or upper strand ^{5m}CpG step, respectively. (**c**, **d**) Space filling representation of the protein-DNA interface of the MBD_{MBD4}-$^{5m}CG/^{5m}CG$ complex (c) or MBD_{MeCP2}-$^{5m}CG/^{5m}CG$ complex (d). The ordered water molecules are shown as red balls and hydrogen bonds between water molecules are shown as black dotted lines. There is a cavity on the MBD_{MBD4}-DNA interface because of the difference in the side chain conformation of Tyr96.

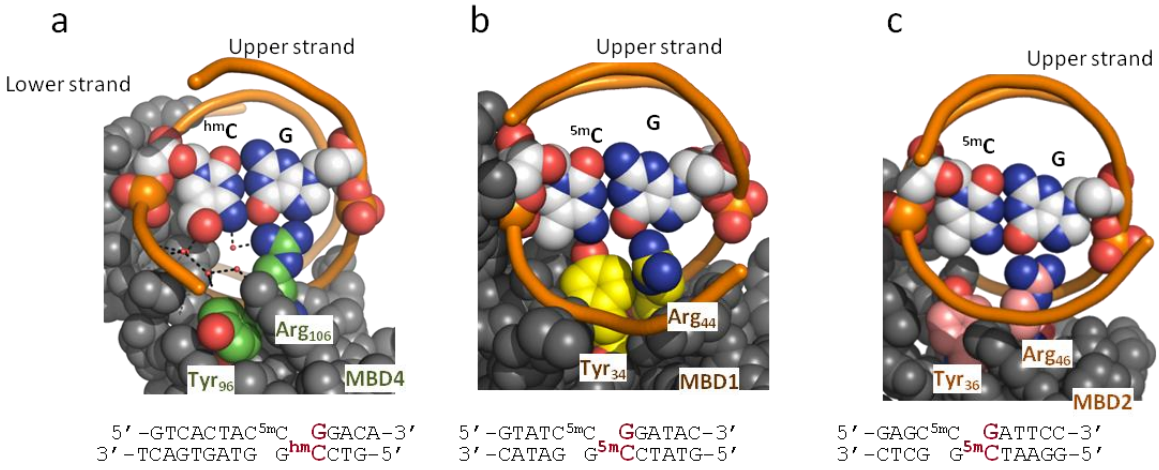


Figure S5. The space filling model representation of the MBD_{MBD4}-$^{5m}CG/^{hm}CG$ (a), MBD_{MBD1}-$^{5m}CG/^{5m}CG$ (b) or MBD_{MBD2}-$^{5m}CG/^{5m}CG$ complex structure (c) as shown in figure 3.

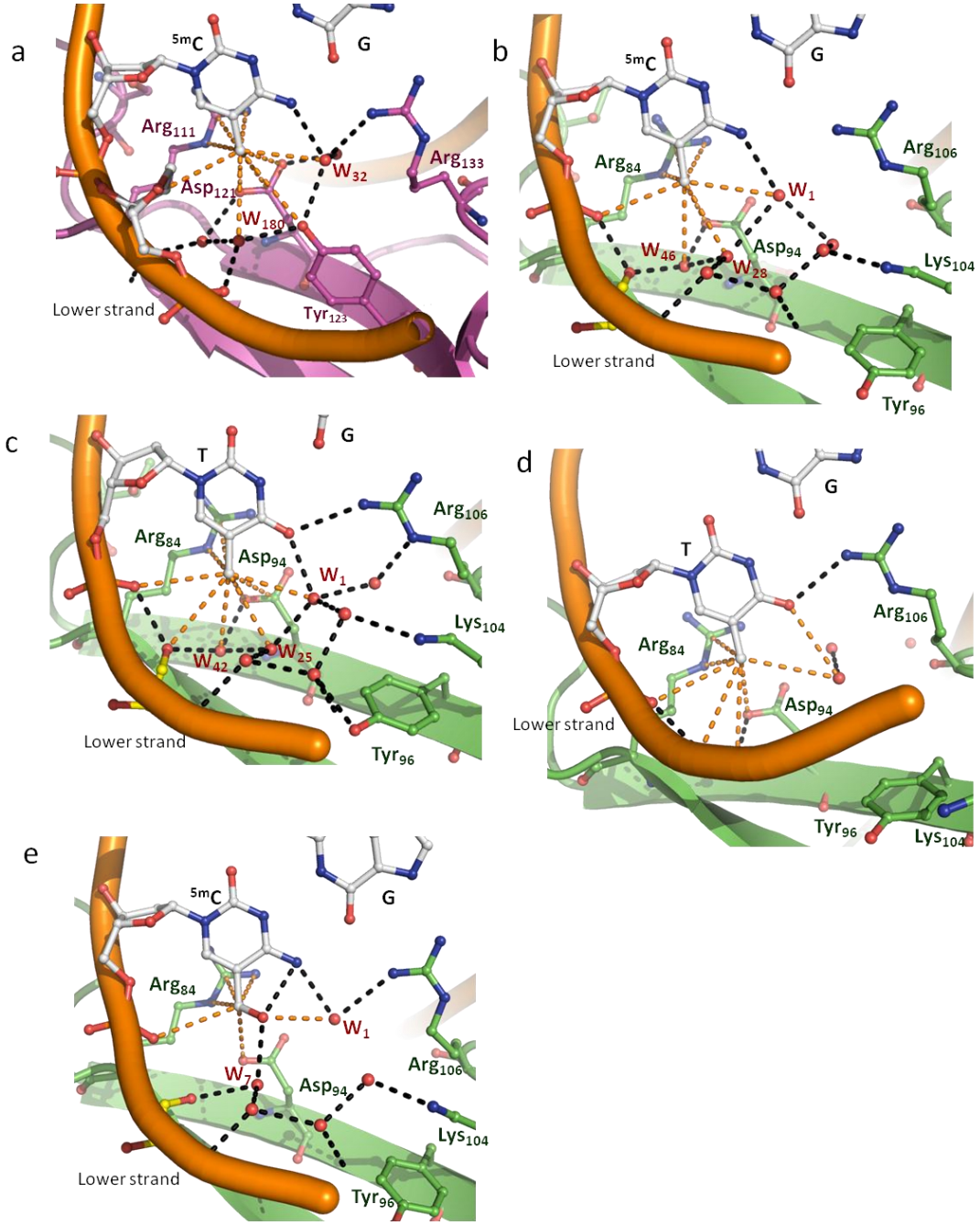


Figure S6. The close up views of the lower strand pyrimidine base recognition in the MBD_{MeCP2}-$^{5m}CG/^{5m}CG$ complex (a) or MBD_{MBD4} in complex with $^{5m}CG/^{5m}CG$ (b), 14bp-$^{5m}CG/TG$ (c), 11bp-$^{5m}CG/TG$ (d) or $^{5m}CG/^{hm}CG$ (e). Hydrogen bonds (~3.2 Å) and van der Waals contacts or CH-O type hydrogen bonds (~4.2 Å) are represented in the black and orange dotted lines, respectively.

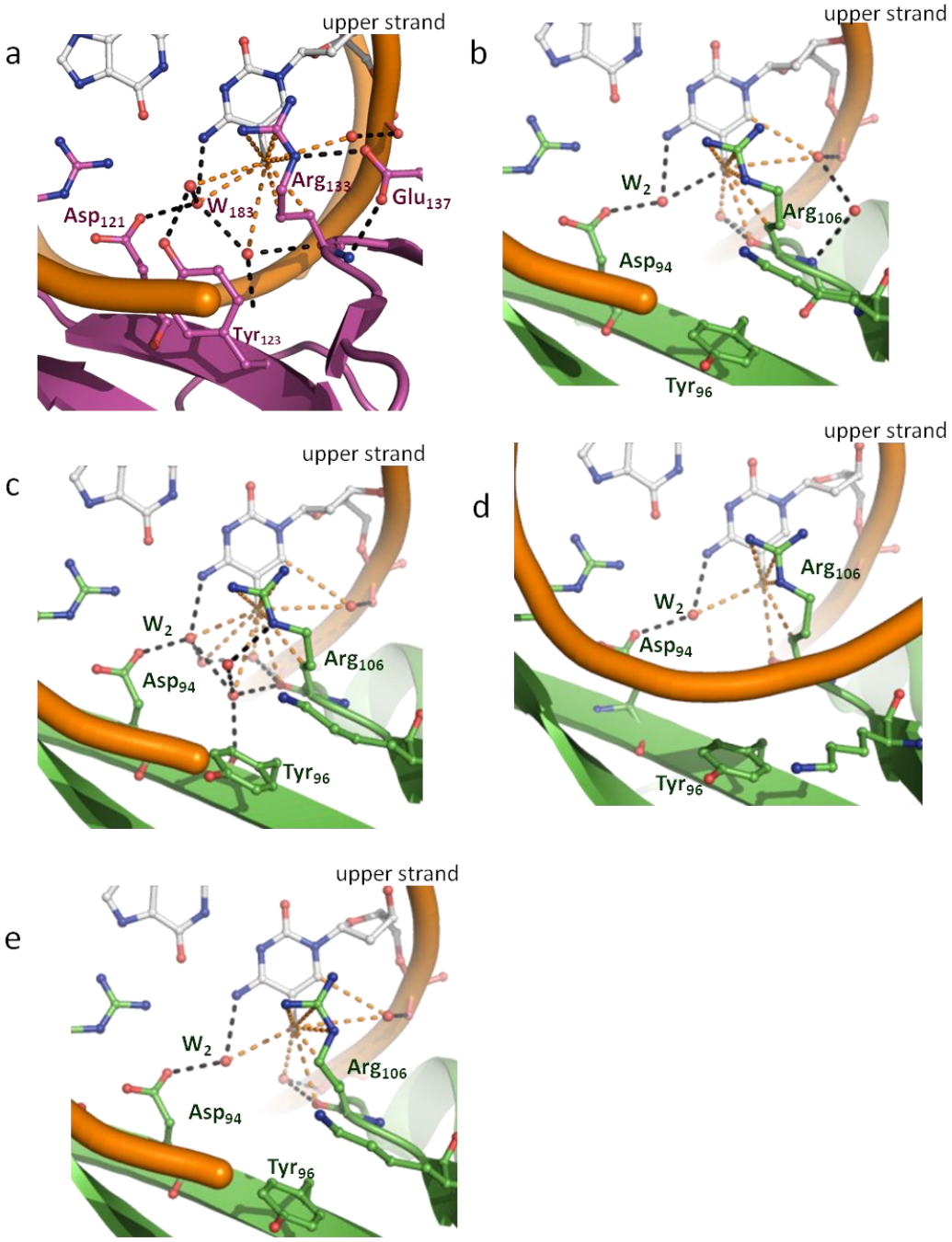


Figure S7. The close up view of the upper strand ^{5m}C base recognition in the MBD_{MeCP2}-^{5m}CG/^{5m}CG complex (a) or MBD_{MBD4} in complex with ^{5m}CG/^{5m}CG (b), 14bp-^{5m}CG/TG (c), 11bp-^{5m}CG/TG (d) or ^{5m}CG/^{hm}CG (e). Hydrogen bonds (~3.2 Å) and van der Waals contacts or CH-O type hydrogen bonds (~4.2 Å) are represented in the black and orange dotted lines, respectively.

Recognition of ^{5m}CG/TG by the flexible DNA binding surface of MBD$_{MBD4}$

The hydrogen bonding pattern of the T/G mismatch base pair in the MBD$_{MBD4}$-^{5m}CG/TG complex is identical to that observed in the crystal structure of T/G mismatch DNA itself (PDB entry: 113D)[20] (Figure 3a & 3d). The T/G mismatch still allows two hydrogen bonds between bases and an overall shape similar to that in Watson-Crick base pairings as shown in figure 3d. Because the base stacking interaction with neighboring pairs is unaffected[21], the entire DNA binding mode common to MBDs is retained in the complex of MBD4 with the mismatch ^{5m}CG/TG DNA. However, the mismatch thymine base is shifted by 1~2 Å toward the major groove side of the DNA duplex, that is, towards the vacant space in the protein-DNA interface, which is unique to MBD4 (Figure 3b).

The recognition mode of the ^{5m}C base of ^{5m}CG/TG in the upper strand is essentially same as observed in the MBD$_{MBD4}$-^{5m}CG/^{5m}CG complex (Figures S7 & S8). Although Arg finger-2 recognizes the guanine in the T/G mismatch through the same hydrogen bonding pattern as observed in the ^{5m}CG/^{5m}CG complex, its orientation is shifted by 0.8 Å to make the additional hydrogen bond with the carbonyl group at the 4th position of the thymine ring that protrudes toward the protein interface (Figures 3b & 3d), which might account for the about 1.5-fold higher selectivity of MBD$_{MBD4}$ for the ^{5m}CG/TG site over the ^{5m}CG/^{5m}CG (Figure S1). The 5-methyl group of the thymine base is recognized via contacts with Arg84, Asp94 and water molecules in the similar manner to the lower strand ^{5m}C recognition in the ^{5m}CG/^{5m}CG crystal structure (Figures 2a, 3a and S7). Except for the movement of Arg finger-2, there is no significant change in the protein structure in comparison with the full methylated CG complex. Thus, the flexibility of the Arg106 side chain attributed to the lack of the intra-molecular lock

seems to be favorable for T/G mismatch recognition. It is obvious that the Glu137 residue of MBD_{MeCP2} or the corresponding acidic residue in MBD1 inhibits the Arg finger-2 to adapt conformation for T/G mismatch recognition (Figure 2b). Consistently, MBD_{MBD1} showed the significantly weaker binding to $^{5m}CG/TG$ than $^{5m}CG/^{5m}CG$ (Table1 and Figure S1).

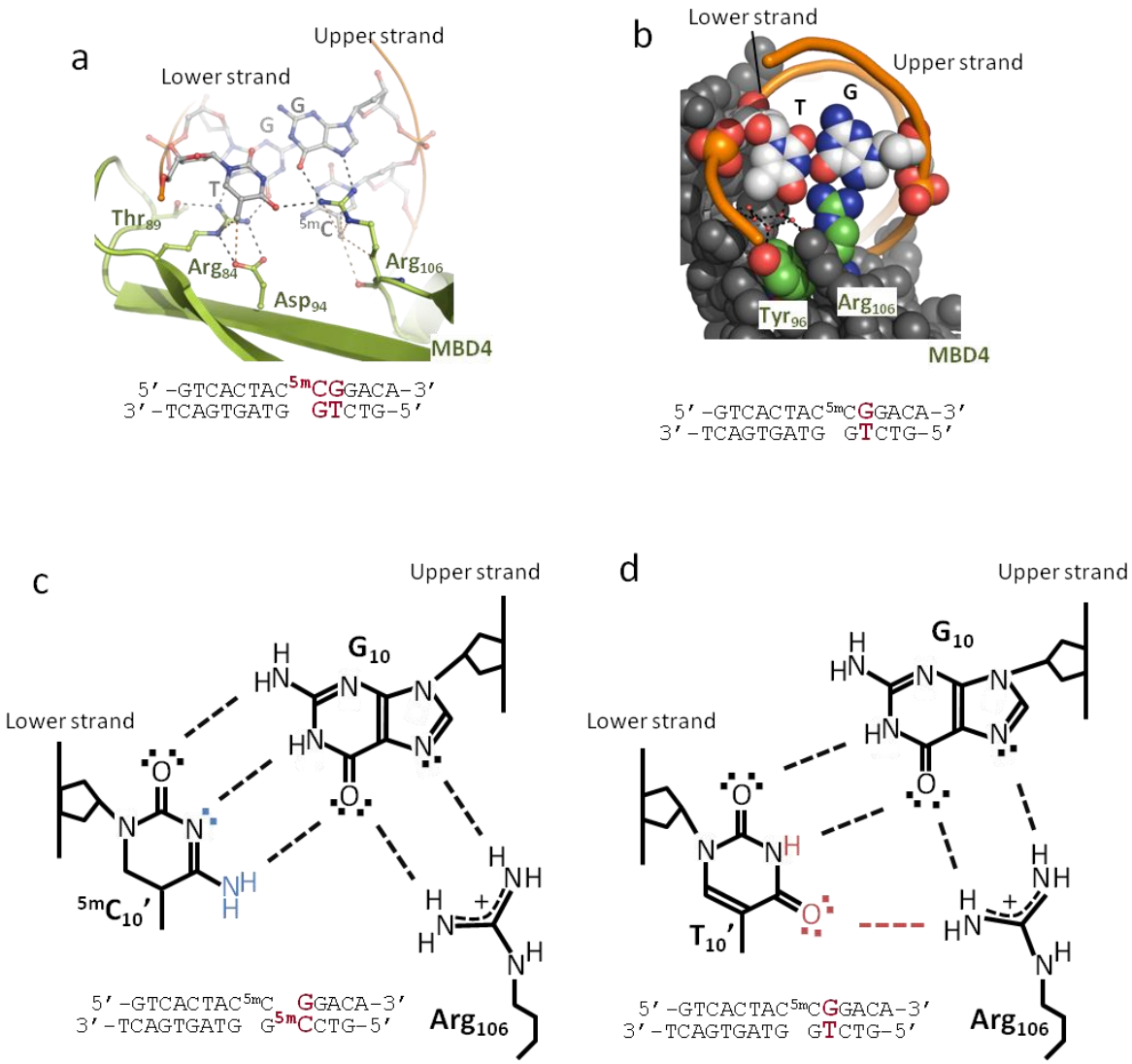


Figure 3. (**a**) The close up view of the DNA binding interface of the MBD_{MBD4}-^{5m}CG/TG complex similar to the figures 2a and 2b. The overall recognition mode of ^{5m}CG/TG sequence by MBD_{MBD4} is similar to that of ^{5m}CG/^{5m}CG sequence. (**b**) Space filling representation of the MBD_{MBD4}-^{5m}CG/TG interface similar to the figures 2c and 2d. The ordered water molecules are shown as red balls and hydrogen bonds between these water molecules are shown as black dotted lines. (**c**, **d**) Schematic diagrams of hydrogen bonding pattern of ^{5m}C/G (c) and T/G mismatch base pair (d). The structural elements specific to the ^{5m}C or thymine base is highlighted in blue or red, respectively.

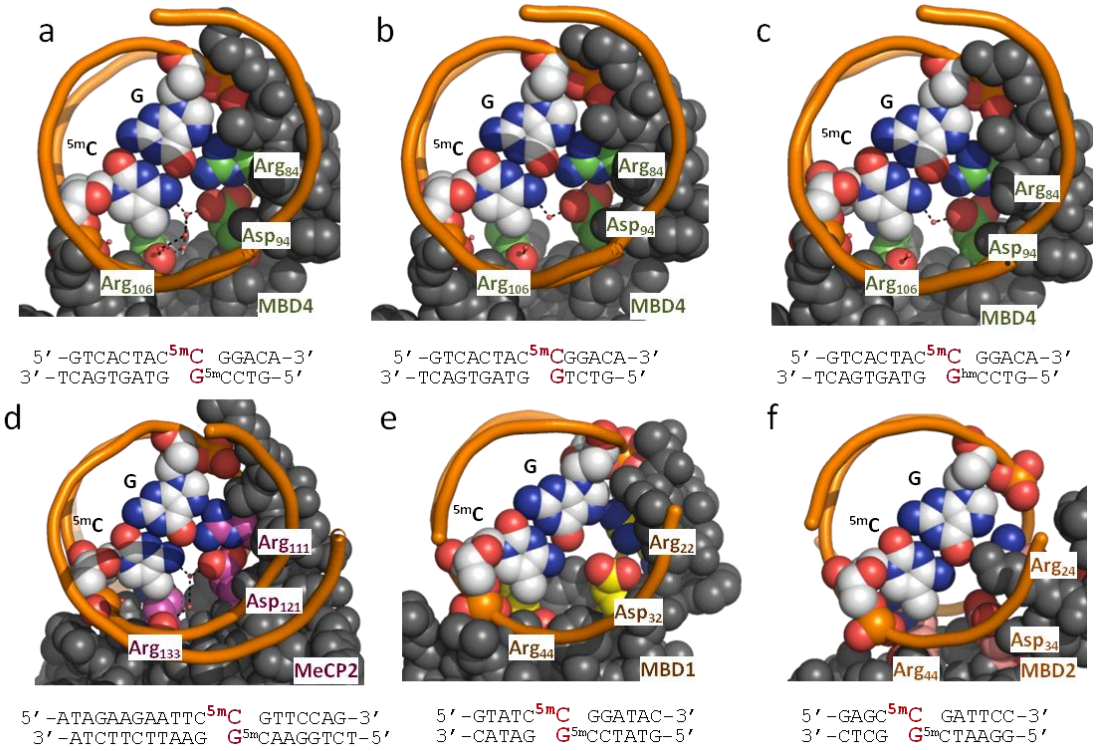


Figure S8. The space filling model representation of the MBD_{MBD4}-$^{5m}CG/^{5m}CG$ (a), MBD_{MBD4}-$^{5m}CG/TG$ (b), MBD_{MBD4}-$^{5m}CG/^{hm}CG$ (c), MBD_{MeCP2}-$^{5m}CG/^{5m}CG$ (d), MBD_{MBD1}-$^{5m}CG/^{5m}CG$ (e) or MBD_{MBD2}-$^{5m}CG/^{5m}CG$ complex structure (f) viewed from the opposed angle of that of figure 3 and supplementary figure S4. The upper strand ^{5m}C recognition mode is conserved among all structurally known MBD domain proteins.

Non specific DNA binding mode of MBD_{MBD4}.

The non-specific DNA binding mode of MBD4, which is observed in the crystal structure of the MBD4-^{5m}CG/TG complex in the triclinic form, implies a sliding mode of MBD_{MBD4} on DNA in prior to recognition of the target sequence. In the non-specific complex, MBD_{MBD4} also binds DNA from the major groove side. The positive end of helix dipole from α1 helix is placed in the major groove and capped by the phosphate group of the DNA backbone. The L1 loop, connecting β1 and β2, also contributes to hold the phosphate backbone through making extensive electrostatic contacts involving residues Leu85~Thr89. Such a phosphate backbone recognition is also observed in the specific complex (Figure S9). On the contrary, the base pair recognition by two arginine fingers is not retained in the non-specific complex, and the dynamic movement of Arg finger-2 is of the most interest. In the non-specific complex, the side chain of Arg finger-2, which is responsible for recognition of the guanine base in the target sequence, is positioned to make a hydrogen bond with a phosphate backbone atom reinforcing the binding to DNA duplex (Figure S9b). The unique flexibility of Arg finger-2 in MBD4 presumably facilitates the non-specific DNA interaction. In agreement with the structural observations, MBD_{MBD4} showed stronger binding to non-modified CpG than MBD_{MBD1} (Figure S10). Furthermore, Arg finger-1, which is locked by Asp94, makes a close contact with the 5' methyl group of a thymine base in a non-target base pair. Arg finger-1 might sense the 5-methyl group of ^{5m}C or thymine base to find out the target sequence during sliding on DNA molecule (Figure S9b).

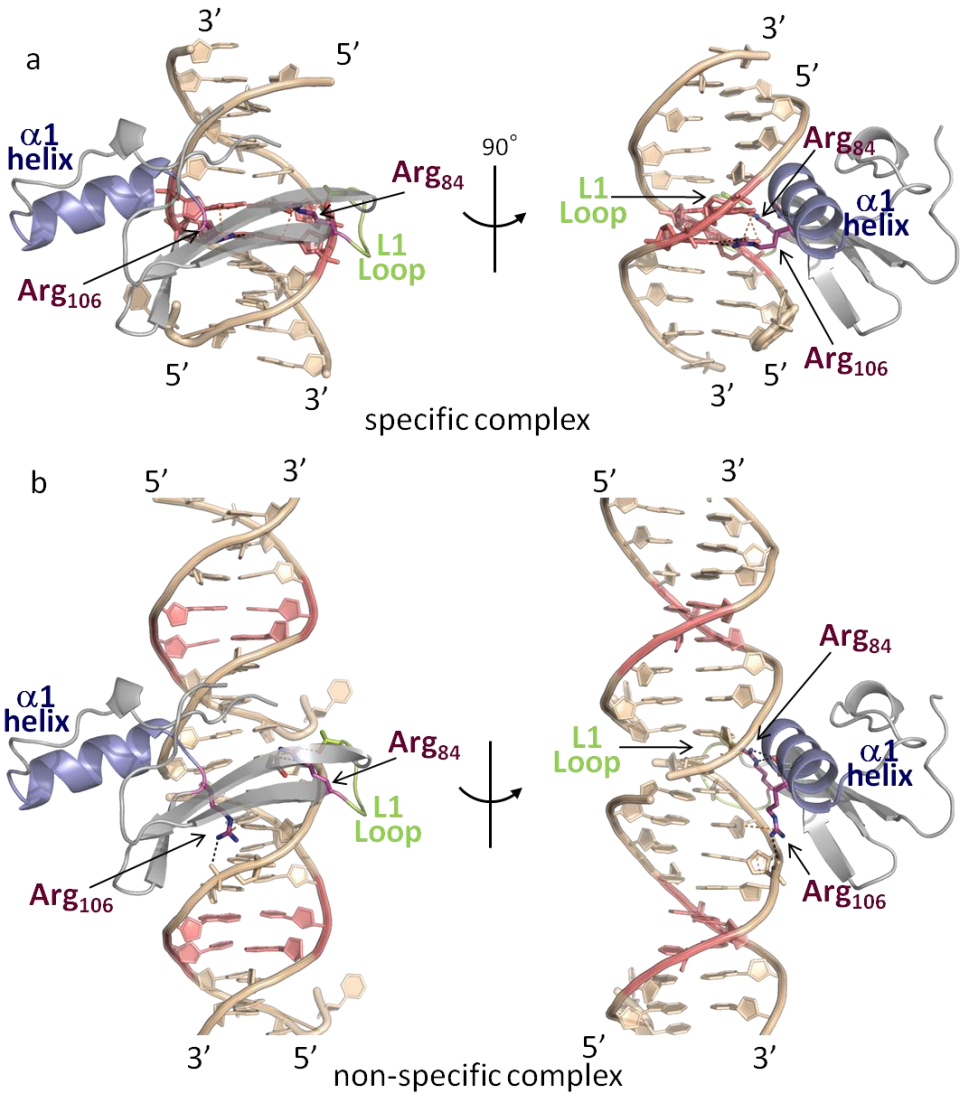


Figure S9. The triclinic crystal structure of MBD_{MBD4} in complex with 11-bp ^{5m}CG/TG in the space group of P1. (a) The specific complex binding to the ^{5m}CG/TG site (colored in red) essentially in the same manner as in the orthorombic crystal. (b) The non-specific complex at the junction of the two symmetry-related DNA fragments. The a1 helix and the L1 loop are colored in blue and green, respectively. The two arginine side chains are shown as a stick model in purple. The hydrogen bonds and van der Waals interactions which involve Arg84 or Arg106 are depicted as black and orange dotted lines, respectively.

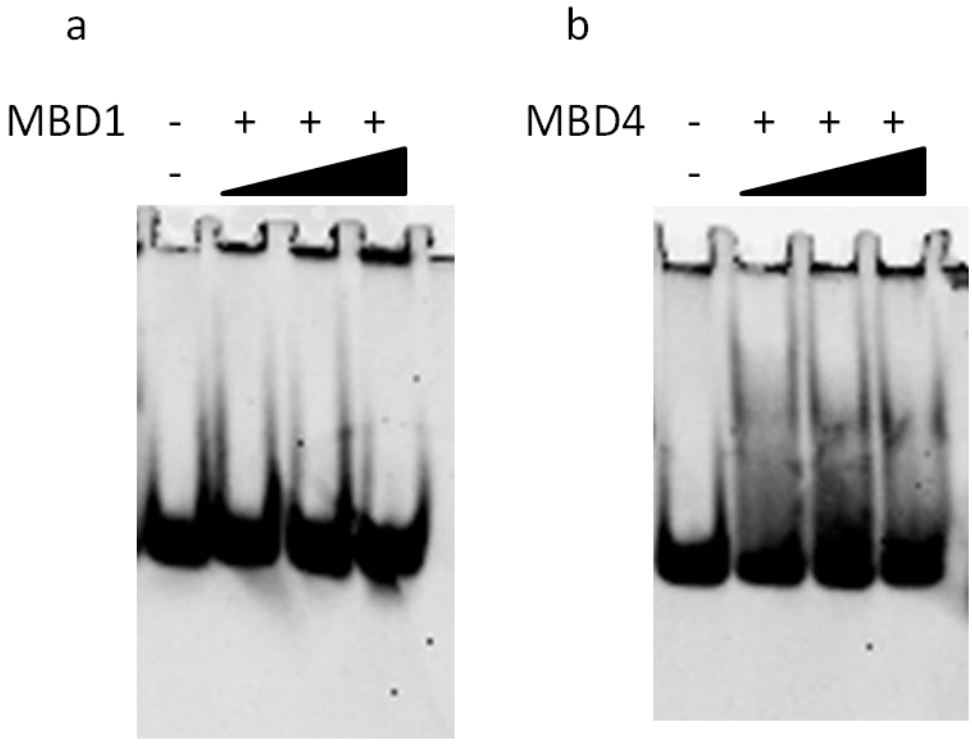


Figure S10. (**a**, **b**) Electrophoretic mobility shift assay of MBD1 (a) and MBD4 (b). 1 mM of each MBD domain protein, 3 mM of 14-bp non-modified CpG DNA and 100, 200 or 400 ng of non-specific competitor, poly dI-dC are mixed, incubated for 15 minutes at 4 ˚C and separated by the 7.5% acrylamide gel electrophoresis followed by GelGreen staining (Cosmo Bio Co. Ltd.).

The cavity of MBD_{MBD4} generated by the Tyr96 flipping can accommodate hydroxymethylcytosine.

The vacant space harboring the water molecule network observed in the protein-DNA interface implied that MBD_{MBD4} would possibly bind more bulky modifications at the 5th position of cytosine than a methyl group. Next we carried out competitive gel mobility shift assays to examine the binding of MBD_{MBD4} to methylated CpG containing a 5-hydroxymethylcytosine base (^{hm}C), 5-hydroxymethyluracil base (^{hm}U), 5-formylcytosine (^{fo}C) or 5-carboxylcytosine (^{ca}C) in the lower strand. As shown in Figure 4b, the complex between the MBD domain and ^{32}P-labeled ^{5m}CG/^{5m}CG duplex was competed off with non-labeled ^{5m}CG/^{hm}CG, ^{5m}CG/^{hm}UG and ^{5m}CG/^{fo}CG fragments. The interactions of MBD_{MBD4} with ^{5m}CG/^{hm}CG, ^{5m}CG/^{hm}UG and ^{5m}CG/^{fo}CG are estimated approximately 2- to 3-fold weaker than that with ^{5m}CG/^{5m}CG by comparison of the required amount of the competitor fragment to reach the same degree of competition (Figure 4d). The ^{5m}CG/^{ca}CG and ^{hm}CG/^{hm}CG fragments showed less tight binding to MBD_{MBD4} than the other oxidative CpG sequences used in this study (Figures 4b & S11). Much less competition of the MBD_{MBD1}-^{5m}CG/^{5m}CG complex with these fragments was observed, suggesting more strict specificity of MBD1 towards full methylated CpG (Figure 4a). The affinities of MBD_{MBD4} for the ^{5m}CG/^{hm}CG and ^{5m}CG/^{hm}UG containing 14-bp DNA duplexes were also estimated based on ITC measurements. Consistent with the results of the competitive EMSA, MBD_{MBD4} was shown to bind ^{5m}CG/^{hm}CG and ^{5m}CG/^{hm}UG with K_D value of 162 and 287 nM, respectively, which were 2- or 3-fold lower than the values for ^{5m}CG/^{5m}CG (K_D, 97.5 nM) and ^{5m}CG/TG (K_D, 98.8 nM) (Table.1). The affinity of MBD_{MBD1} for ^{5m}CG/^{hm}CG (K_D, 1.04 μM) is 10-fold weaker than that for ^{5m}CG/^{5m}CG (K_D, 72.5 nM) as indicated in

Table 1. Combined with the structural data, these data suggest the binding tolerance of MBD_{MBD4} to the further oxidative modification introduced in methylated CpG sequences as far as ^{5m}C in the upper strand is maintained.

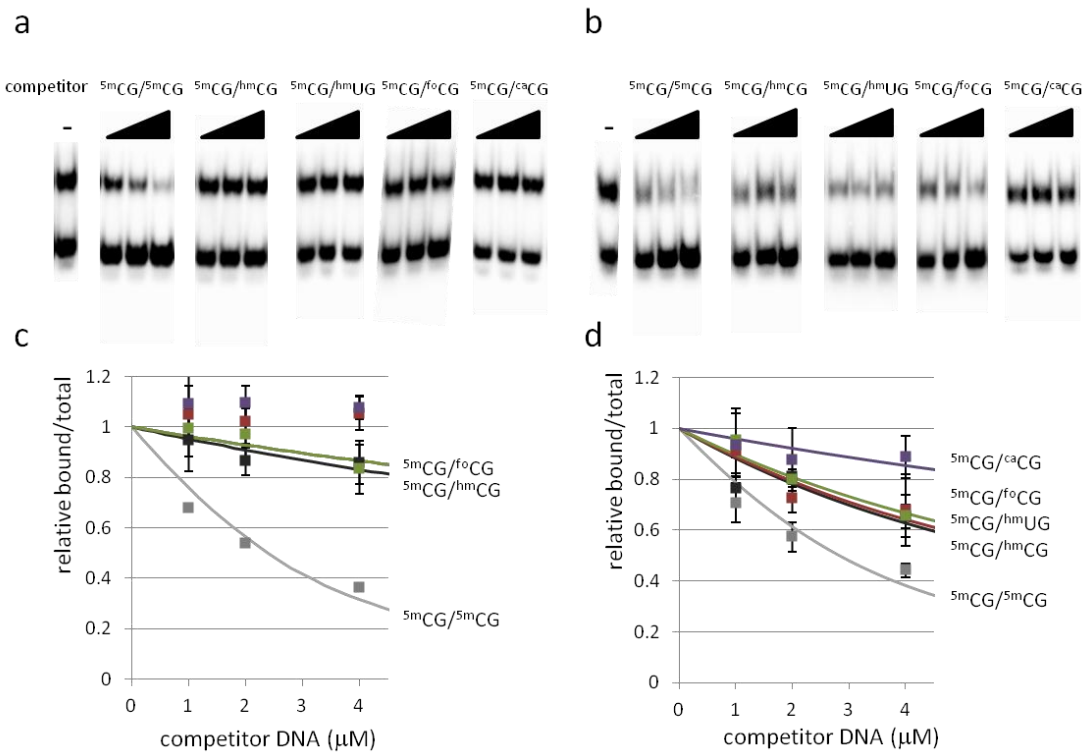


Figure 4. Competitive electrophoretic mobility shift assay. (**a**, **b**) The representative autoradiographic images of the competition assay of MBD_{MBD1} (a) and MBD_{MBD4} (b). Each of the competitor DNA sequence is indicated above. (**c, d**) The bound to total ratios are quantified from gel band density and the values relative to the control lanes are plotted against the amount of competitor DNA. Each dots represent the average of three independent experiments using MBD_{MBD1} (c) or MBD_{MBD4} (d) with standard deviation. The solid lines represent the fitting curves using the Morrison's equation. The values for the experiment using competitor fragment $^{5m}CG/^{ca}CG$ or $^{5m}CG/^{hm}UG$ for the MBD_{MBD1}-$^{5m}CG/^{5m}CG$ complex could not fit to the equation because no competitive effect was observed for these competitor sequences.

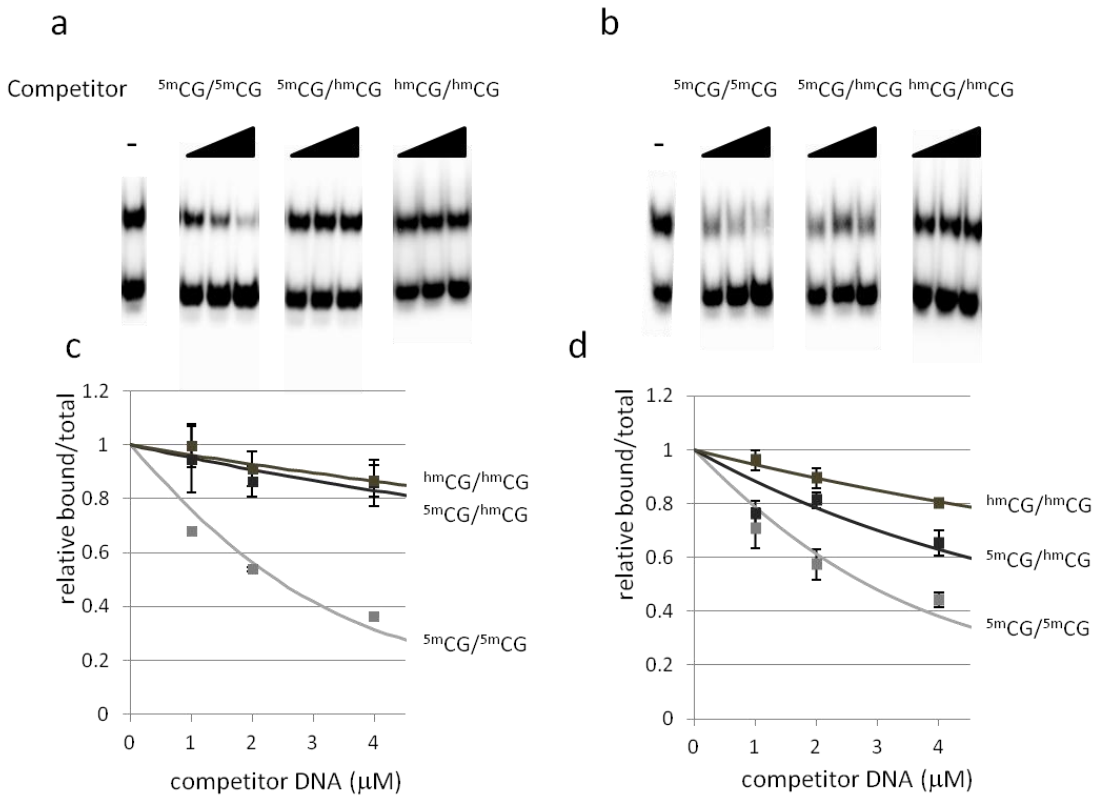


Figure S11. Competitive electrophoretic mobility shift assay using ^{5m}CG/^{5m}CG, ^{5m}CG/^{hm}CG or ^{hm}CG/^{hm}CG sequence as a competitor DNA as shown in figure 4. (**a**, **b**) The representative autoradiographic images of the competition assay of MBD_{MBD1} (a) and MBD_{MBD4} (b). Each of the competitor DNA sequence is indicated above. (**c, d**) The bound to total ratios are quantified from gel band density and the values relative to the control lanes are plotted against the amount of competitor DNA. Each dots represent the average of three independent experiments using MBD_{MBD1} (c) or MBD_{MBD4} (d) with standard deviation. The solid lines represent the fitting curves using the Morrison's equation.

To further unravel molecular basis for the versatile DNA binding ability of MBD_{MBD4}, we determined the crystal structure of MBD_{MBD4} bound to the 14-bp DNA oligomer containing $^{5m}CG/^{hm}CG$ at 2.4 Å resolution. The hydroxylation of 5-methyl group does not perturb either normal hydrogen bonding pattern in the C/G base pair or the interaction with MBD_{MBD4}. Unambiguous electron density for the hydroxyl group of ^{hm}C indicates confined rotational movement of the hydroxylmethyl moiety against the pyrimidine ring (Figure 5a). Interestingly, the hydroxyl group makes an intra-base hydrogen bond with the amino group at the 4^{th} position in addition to a hydrogen bond with a water molecule in the hydrogen bonding network of water molecules (Figure 5a & S6). As indicated in Figure 5b, the 5-hydroxymethyl moiety also donates CHO type hydrogen bonds to the carbonyl of Asp94 and the phosphate group of the DNA backbone. These CHO-type hydrogen bonds and the covalent bond with the hydroxyl group are in a tetrahedral coordination around the carbon atom in the 5-hydroxymetyl moiety. Thus, the positional preference of the hydroxyl group is ensured by the intra-base hydrogen bond and the tetrahedral configuration of the 5-hydroxylmethyl moiety despite of the close contacts with the neighboring cytosine base at the 5' side (Figure 5b).

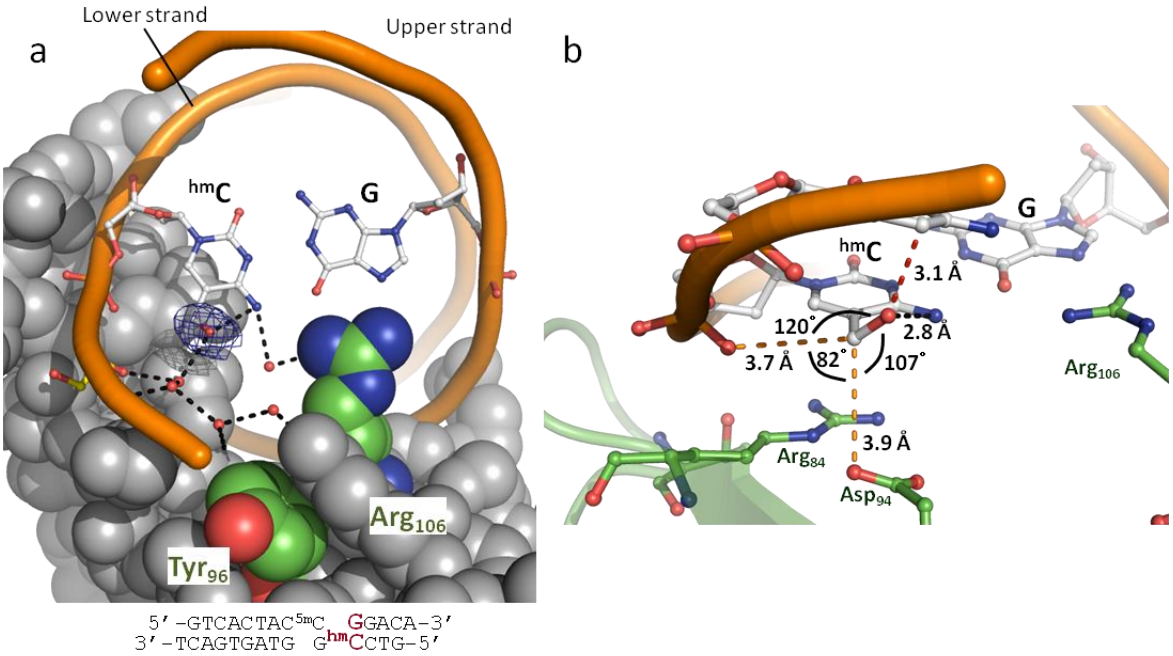


Figure 5. (**a**) The structure of the MBD_{MBD4}-$^{5m}CG/^{hm}CG$ complex. MBD_{MBD4} is represented as space filling model. The DNA molecule is shown as ribbon model and the $^{hm}C/G$ base pair is represented as stick model. The blue mesh is the mF_O-DF_C simulated annealed omit map of the hydroxyl group of the ^{hm}C base contoured at 3.0 s. The ethylene glycol molecule is shown in yellow stick model. Water molecules are represented in red small balls. Black dotted lines indicate the hydrogen bonds (~3.2 Å) (**b**) The tetrahedral configuration around the carbon atom at 5th position of the cytosine ring. The black, orange or red dotted line represents a hydrogen bond, a CHO type hydrogen bond or an unfavorable close contact.

We were not able to obtain crystals of either MBD_{MBD4}-$^{5m}CG/^{hm}UG$ or MBD_{MBD4}-$^{5m}CG/^{fo}CG$ complex, and therefore investigated their structural aspect by a model building approach. The models of the MBD_{MBD4}-$^{5m}CG/^{hm}UG$ and MBD_{MBD4}-$^{5m}CG/^{fo}CG$ complexes were built based on the structures of the MBD_{MBD4}-^{5m}CG/TG and MBD_{MBD4}-$^{5m}CG/^{5m}CG$ complexes as described in Methods (Figure S12). Our models suggest that the oxygen atom of the ^{hm}U base or the ^{fo}C base is accommodated into the DNA binding surface of MBD_{MBD4} without any steric hindrance although the close contact with 5' neighboring base is expected for the MBD_{MBD4}-$^{5m}CG/^{hm}UG$ complex (Figure S12).

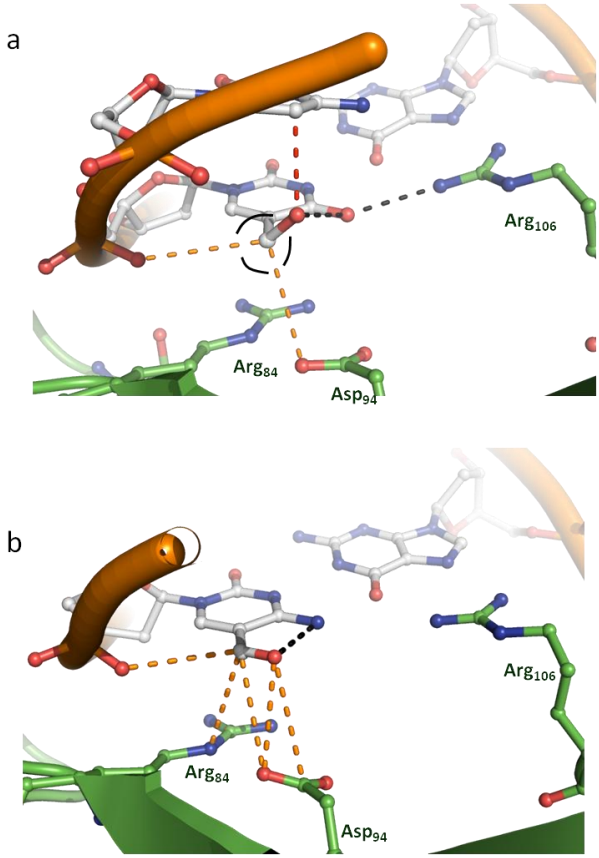


Figure S12. The structural models of the MBD_{MBD4}-$^{5m}CG/^{hm}UG$ and MBD_{MBD4}-$^{5m}CG/^{fo}CG$ complexes. (a) The ^{hm}C base from the MBD_{MBD4}-$^{5m}CG/^{hm}CG$ complex is overlaid onto the thymine base of the MBD_{MBD4}-$^{5m}CG/TG$ complex structure. The tetrahedral coordination of the 5-methyl moiety is expected to be conserved. The CH-O type hydrogen bonds donated by the methyl moiety (orange dotted lines), hydrogen bonds (black dotted lines) and a close contact with the 5' neighboring base (red dotted line) are expected. (b) The ^{fo}C base from the small molecular crystal is overlaid onto the lower strand ^{5m}C base of the MBD_{MBD4}-$^{5m}CG/^{5m}CG$ complex structure. The intra-base hydrogen bond (black dotted line) of the ^{fo}C base is reported in reference 41. The CH-O type hydrogen bonds or van der Waals contacts (orange dotted lines) between the ^{fo}C base and Arg84 or Asp94 are expected.

Glycosylase activity of MBD4 towards ^{hm}U in addition to the mismatch thymine.

We examined the glycosylase activity of the full-length MBD4 protein for mismatch, deamination and oxidation products in the context of the ^{5m}CG/^{5m}CG sequence. The glycosylase activity was assessed by NaOH cleavage of the resulting apyrimidinic (AP) site. Consistent with previous reports[8,22], we observed the significant digestion band for the strand containing either T or ^{hm}U, which is in a mismatch wobble base pair (Figure 6a). The bases of ^{5m}C, ^{hm}C, ^{fo}C and ^{ca}C, each of which is in the canonical Watson-Crick base pairs were not removed by MBD4, whereas the human TDG exhibited the activity toward ^{fo}C and ^{ca}C in addition to T and ^{hm}U bases (Figure 6b).

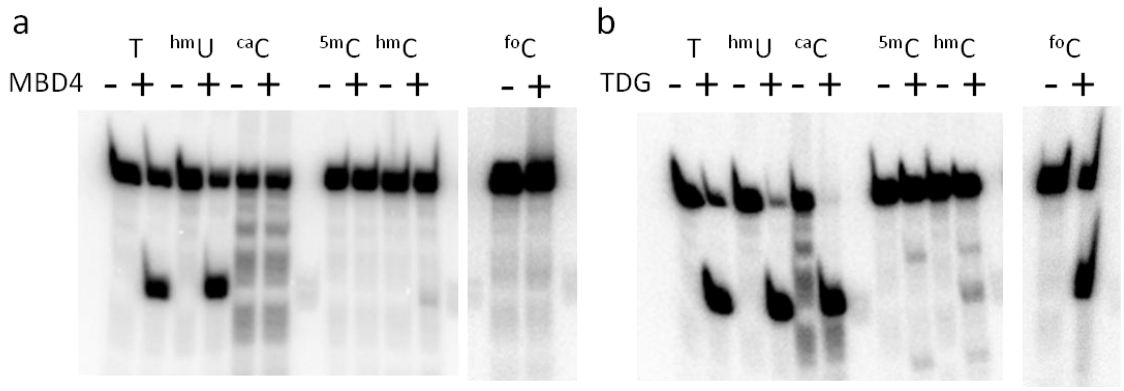


Figure 6. (**a**, **b**) The glycosylase assays toward a series of pyrimidine bases, T, ^{hm}U, ^{ca}C, ^{5m}C, ^{hm}C and ^{fo}C, in the context of ^{5m}CG/XpG sequence DNA using full-length MBD4 (a) or TDG (b) as an enzyme. The nucleotide sequences of the substrates for the assay are listed in Table S1.

Discussion

MBD4 is a multi-functional protein that plays important roles in DNA mismatch repair as a G/T mismatch glycosylase and in transcriptional repression through recruitment of Sin3A and HDAC1[8,23]. Our quantitative binding experiments using isothermal titration calorimetry showed that MBD_{MBD4} has similar affinities for the $^{5m}CG/^{5m}CG$ sequence and the $^{5m}CG/TG$ mismatch unlike other MBD domain proteins. The diverse DNA binding specificity of MBD_{MBD4} seems to underlie the multi-functionality of MBD4 protein. Our crystal structures of MBD_{MBD4} complexed with a series of mismatch and methylated CG sequences provide the molecular basis of the broad substrate specificity of MBD_{MBD4}. MBD_{MBD4} shares overall DNA recognition mode with MBD_{MeCP2}, MBD_{MBD1} and $MBD_{MBD2,}$. Nevertheless, the multi-specificity of MBD_{MBD4} is provided by the cavity filled with ordered water molecules on the DNA binding surface in addition to flexibility of Arg finger-2. The indirect contacts mediated by ordered water molecules are important as well as the direct contacts by MBD_{MBD4}, which was also observed in the crystal structure of the MBD_{MeCP2}-DNA complex[10]. Interestingly, more ordered water molecules are identified in the DNA interface of MBD_{MBD4} than those ovserved in the MBD_{MeCP2}-DNA complex structure as a consequence of the wider space generated by the flipping of Tyr96 of the MBD_{MBD4}. The recognition scheme of the ^{5m}C bases through a hydration water on the major groove side is almost completely conserved in both MBD_{MBD4}-DNA and MBD_{MeCP2}-DNA complexes. The conserved water molecule locate within a hydrogen bonding distance from the amino group at the 4^{th} position and within a van der Waals distance from the 5-methyl group of the ^{5m}C base (Figures S5 & S6). The water molecule interacting with the upper strand ^{5m}C, named W2 in each of the MBD_{MBD4}-DNA complexes (Figures

S7b-e), forms hydrogen bonds with Asp94 of MBD_{MBD4}, which counterpart, W183 in the MBD_{MeCP2}-DNA complex interacts with Asp121. The water molecule bound to the lower strand ^{5m}C, W32 in Figure S6a, makes hydrogen bonds with Tyr123 and Arg133 in the MBD_{MeCP2}-DNA complex. However, in the DNA interface of MBD_{MBD4},-$^{5m}CG/^{5m}CG$ complex, the corresponding water molecule, W1 forms the hydrogen bonding network with other water molecules instead of Tyr94 directing the opposite site (Figure S6b). Of these water molecules, W28 and W46, in addition to W1, are within a distance of weak interaction with the 5-methyl group of ^{5m}C and the water mediated interaction appears to compensate for the lack of the Tyr-^{5m}C interaction of MBD_{MeCP2}-$^{5m}CG/^{5m}CG$ complex (Figure S6b). In the MBD_{MBD4}-^{5m}CG/TG complex, W1, W25 and W42 make analogous contacts with the 5-methyl group of the thymine base (Figure S6c). In the MBD_{MBD4}-$^{5m}CG/^{hm}CG$ complex, the hydroxyl group of the ^{hm}C base is involved in the hydrogen bonding network (Figure S6e). The hydrogen bonding network within the cavity of MBD_{MBD4} involves the phosphate groups of the DNA backbone and Asp94 and Lys104 side chains (Figures S6b-e). Although the width of the space on the protein-DNA interface is different depending on the DNA sequence, the change in the position and the number of ordered water molecules absorb the difference. The water-mediated hydrogen bonding network of MBD_{MBD4} seems to make it possible to bind to the diverse substrates specifically.

MBD_{MBD4} binds to the major groove side of undisturbed canonical B-form DNA in the common manner to MBD_{MeCP2}, MBD_{MBD1} and MBD_{MBD2}. On the other hand, a model building study using structures of homologous glycosylases suggests that the glycosylase domain of MBD4 binds to the minor groove side accompanied by ~70 °

bent of DNA[24]. Docking of MBD_{MBD4} to the model of the glycosylase domain bound to DNA clearly shows that these two domains can not bind to a target site simultaneously owing to a significant steric clash between the DNA phosphate backbone and the MBD domain (Figure S13). The isolated MBD domain of MBD4 has been also shown to inhibit the catalytic activity of the glycosylase domain *in vitro*[25], indicating that the DNA substrate should be handed over from MBD_{MBD4} to the glycosylase domain in the full-length context. In addition, MBD_{MBD4} does not make any contact with bases other than the CpG sequence, thereby it is allowed to bind to the symmetrical $^{5m}CG/^{5m}CG$ site in both directions equally as observed in the flipping motion of MBD_{MBD1} on its target DNA[26]. In contrast, the asymmetrical structural features of the $^{5m}CG/TG$ mismatch prevent the flipping motion of the MBD_{MBD4}. In the crystal structure, Arg106 adopts the conformation suitable for recognition of the mismatch pair. However, a model MBD_{MBD4} bound to the mismatch site in the opposite direction, in which the guanidino group of Arg84 is supposed to recognize the target site, shows an obvious steric clash between the carbonyl oxygen at the 4^{th} position of thymine ring and the side chain of Asp94 (Figure S14). MBD_{MBD4} presumably has a role to guide the glycosylase domain to the target mismatch base during the hand-over process in addition to the role to tether its glycosylase domain to the ^{5m}C-rich region. Taken together with non-specific DNA binding mode of MBD_{MBD4}, the MBD_{MBD4} is assumed to be a domain required for its proper localization to the ^{5m}C rich regions where the ^{5m}C to T conversion frequently occurs and for the rapid target search. Further structural and biochemical analysis of the full length MBD4 are essential to fully understand its molecular mechanism.

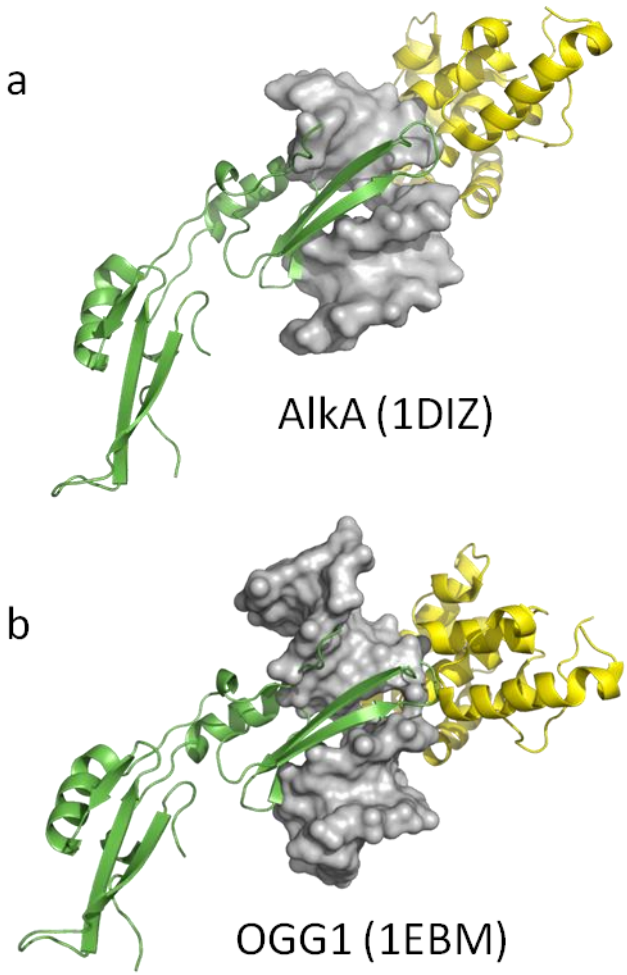


Figure S13. The ternary complex model based on the crystal structures of AlkA complexed with product DNA (a, PDB ID; 1DIZ) and OGG1 complexed with the substrate DNA (b, PDB ID; 1EBM). There are apparent steric clashes between the bent DNA molecules and MBD domain.

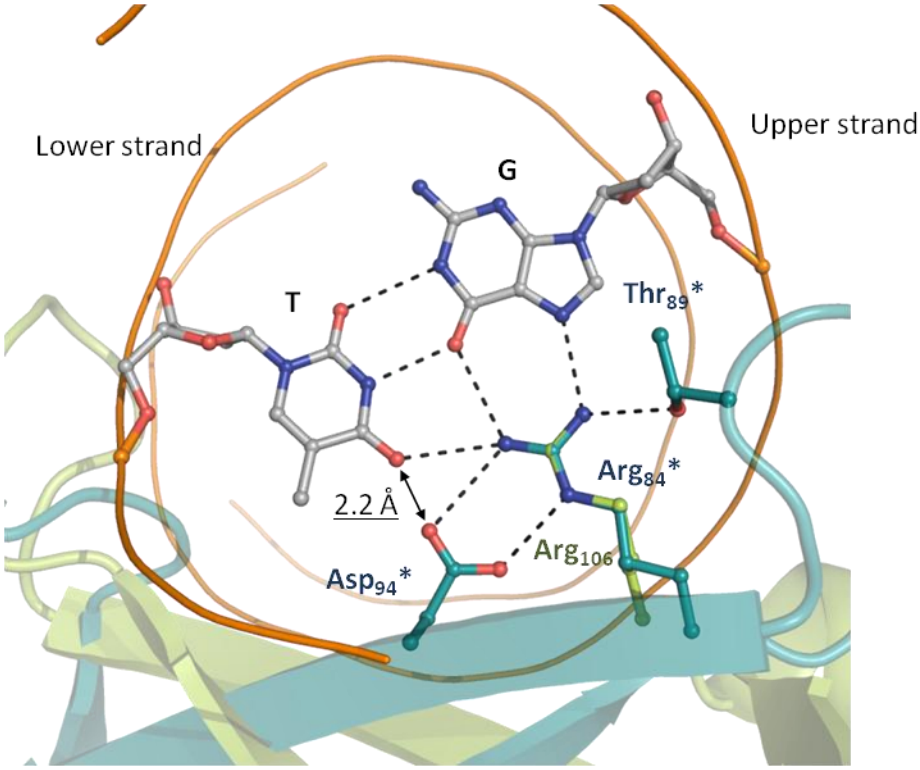


Figure S14. The structural model of the MBD domain of MBD4 bound to the ^{5m}CG/TG sequence in the direction opposite to that observed in the crystal. The guanidino group goup of Arg84 of the model molecule (colored in blue) is overlaid onto that of Arg106 in the MBD_{MBD4}-^{5m}CG/TG complex structure (colored in green). The steric clash between the carboxyl groups of the thymine base and the Asp94 is expected.

Our biochemical and structural studies demonstrated that the DNA binding surface of MBD_{MBD4} tolerates various modified pyrimidine rings including the deamination or and oxidation products of the ^{5m}C base in a methylated CpG site. The formyl group of ^{fo}C is allowed to fit in the cavity without any steric clash with the protein molecule as well as the hydroxyl group of ^{hm}C or ^{hm}U. The relatively low affinity of MBD_{MBD4} for $^{5m}CG/^{ca}CG$ could be attributed to the electrostatic repulsion between the carboxyl group of the ^{ca}C base and the side chain carboxyl of Asp94. The symmetric oxidative modification on both ^{5m}C bases in the CpG sequence perturbs the MBD_{MBD4} binding because of a lack of space enough for accepting the further modification of the 5^{th} position of the upper strand ^{5m}C base (Figure S7). The interface with the upper strand ^{5m}C base remains narrower space in contrast to that with the lower strand ^{hm}C base (Figure S7). This is consistent with the previous report that MBD4 or any of other MBD proteins does not bind to the totally hydroxymethylated DNA fragment[27,28].

In spite of broad spectra of the binding target of MBD_{MBD4}, full length MBD4 exhibits the glycosylase activity toward only the mismatch thymine and ^{hm}U bases. The oxidative products of ^{5m}C such as ^{hm}C, ^{fo}C and ^{ca}C are unsusceptible for the digestion by MBD4. These findings suggest that MBD4 is involved in the DNA demethylation pathway through the deamination of cytosine or ^{hm}C base mediated by AID/APOBEC family deaminases[14,15,29]. Thus, MBD4 is supposed not to be involved in the DNA demethylation pathway through the direct excision of oxidized cytosine bases[16,19]. Considering the recent findings that the ^{hm}C, ^{fo}C and ^{ca}C bases have long lifetimes during preimplantation development[30,31], these oxidative cytosine bases might function as regulatory markers antagonistic to ^{5m}C *in vivo.* MBD4 therefore could be a candidate to recognize these bases, although further investigations are required for

understanding the role of MBD4 in the biology of oxidative cytosine bases.

No glycosylase activity of MBD4 for the ^{hm}C base was detected in our glycosylation assays. Both MBD4 and TDG however exhibited the activity toward the non-mismatch thymine base at the 5' neighboring position of the ^{hm}C base (Figure S15). The non-mismatch base excisions were observed only for the thymine adjacent to ^{hm}C in a sequence T^{hm}CGA/T^{5m}CGA, but not for that next to ^{5m}C. These activities were relatively low compared with the mismatch base excision activities. The ^{hm}C bases have been reported to lower the annealing temperature and increase the flexibility of DNA duplex[32]. Therefore we speculate that perturbation of the neighboring base pair stability by ^{hm}C reinforce the glycosylase activity of MBD4. We would like to address the possibility that the ^{hm}C dependent removal of the unmismatch thymine base is relevant to the DNA demethylation involving the long patch base excision repair system which displaces the ^{hm}C base with unmodified cytosine base. Furthermore, we found that MBD4 and TDG cleaves off the ^{5m}C and ^{hm}C bases flanked by three sequential thymine bases at their 5' side (Figure S15). The ^{hm}C was more efficiently excised than the ^{5m}C (Figure S15). These activities might be induced by perturbation of the target base pairing and structural flexibility of DNA by the T-tract-like DNA sequence. Thus glycosylation activities of MBD4 seem somehow modulated by flexibility and instability of the target site influenced by flanked sequences. The biological relevance of such ^{hm}C dependent un-mismatch base excision and T-rich sequence dependent ^{5m}C or ^{hm}C removal should be further investigated.

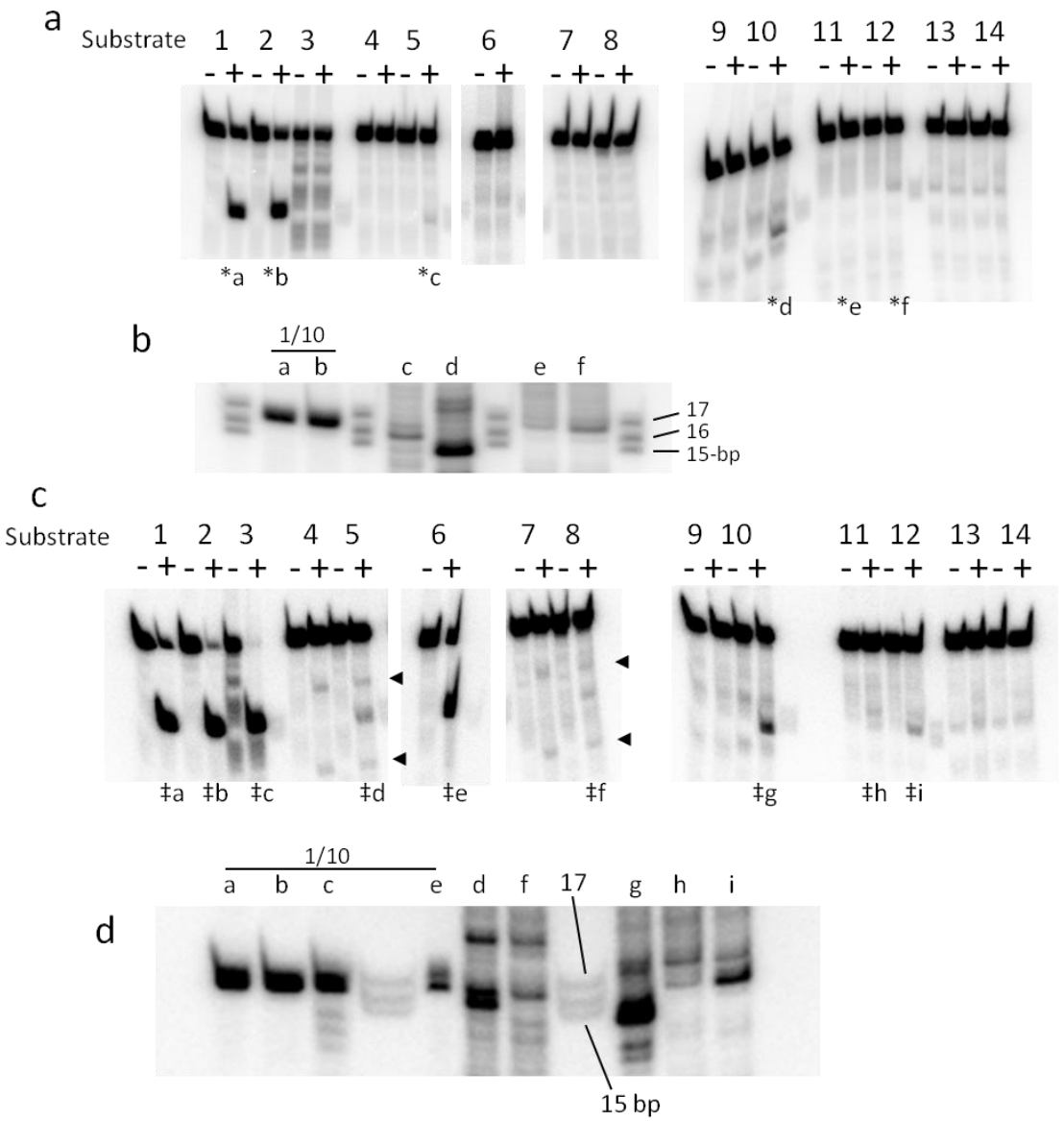


Figure S15. (**a**) The glycosylase assays toward various DNA sequences (listed in Table S1) using MBD4 as an enzyme. The DNA substrate 1 and 2 yielded significant digestion bands (*a and *b) and 5, 10, 11 and 12 yielded slight digestion bands (*c-*f). (**b**) Samples *c-*f and 1/10 amounts of samples *a and *b were re-run on the same gel to compare the size of product bands. The 15-, 16- and 17-bp oligo nucleotide fragments were used as DNA size marker. The result implies that MBD4 has significant activity to excise thymine (*a) and ^{hm}U (*b) paired with guanine and weak activity toward the

thymine paired with adenine at 5' side of the ^{hm}C (*c and *d) and the ^{5m}C (*e) and ^{hm}C (*f) bases flanked by three sequential thymine bases at their 5' side. (**c**) The glycosylase assay using TDG as an enzyme. The DNA substrate 1, 2, 3 and 6 yielded significant digestion bands (‡a-‡d) and 5, 8, 10, 11 and 12 yielded slight digestion bands (‡e-‡i). Relatively high non-specific thymine glycosylase activity of TDG was observed (indicated with arrowheads). (**d**) Samples ‡d, ‡f, ‡g, ‡h and ‡i and 1/10 amounts of samples ‡a, ‡b, ‡c and ‡e were re-run on the same gel to compare the size of digested band. The results imply that TDG has significant activity to excise thymine (‡a), ^{hm}U (‡b), ^{ca}C (‡c) and ^{fo}C (‡d) and weak activity toward the thymine paired with adenine at 5' side of the ^{hm}C (‡d and ‡g) and the ^{5m}C (‡h) and ^{hm}C (‡i) bases flanked by three sequential thymine bases at their 5' side.

Table S1. The DNA sequences used in the glycosylase assay shown in figure S15.

	32P-labeled strand	un-labeled strand
1	GGCTAAATACCTGGGC T**TG** AAGTGAACTGATTGCC	GGCAATCAGTTCACTT ^{5m}**CG**A GCCCAGGTATTTAGCC
2	GGCTAAATACCTGGGC T^{hm}**UG** AAGTGAACTGATTGCC	GGCAATCAGTTCACTT ^{5m}**CG**A GCCCAGGTATTTAGCC
3	GGCTAAATACCTGGGC T^{ca}**CG** AAGTGAACTGATTGCC	GGCAATCAGTTCACTT ^{5m}**CG**A GCCCAGGTATTTAGCC
4	GGCTAAATACCTGGGC T^{5m}**CG** AAGTGAACTGATTGCC	GGCAATCAGTTCACTT ^{5m}**CG**A GCCCAGGTATTTAGCC
5	GGCTAAATACCTGGGC T^{hm}**CG** AAGTGAACTGATTGCC	GGCAATCAGTTCACTT ^{5m}**CG**A GCCCAGGTATTTAGCC
6	GGCTAAATACCTGGGC T^{fo}**CG** AAGTGAACTGATTGCC	GGCAATCAGTTCACTT ^{5m}**CG**A GCCCAGGTATTTAGCC
7	GGCTAAATACCTGGGC A^{5m}**CG**T AGTGAACTGATTGCC	GGCAATCAGTTCACT A^{5m}**CG**T GCCCAGGTATTTAGCC
8	GGCTAAATACCTGGGC A^{hm}**CG**T AGTGAACTGATTGCC	GGCAATCAGTTCACT A^{5m}**CG**T GCCCAGGTATTTAGCC
9	GGCAATCAGTTCAC TT^{5m}**CG** AGCCCAGGTATTTAGCC	GGCTAAATACCTGGGCT ^{5m}**CG**AA GTGAACTGATTGCC
10	GGCAATCAGTTCAC TT^{hm}**CG** AGCCCAGGTATTTAGCC	GGCTAAATACCTGGGCT ^{5m}**CG**AA GTGAACTGATTGCC
11	GGAGAAGCGCTTTCTTT^{5m}**CG** CGAGCACCCTGAACCA	TGGTTCAGGGTGCTCG ^{5m}**CG**AAA GAAAGCGCTTCTCC
12	GGAGAAGCGCTTTCTTT^{hm}**CG** CGAGCACCCTGAACCA	TGGTTCAGGGTGCTCG ^{5m}**CG**AAA GAAAGCGCTTCTCC
13	GGCAATCAGTTCACT T^{5m}**CG**T GCCCAGGTATTTAGCC	GGCTAAATACCTGGGC A^{5m}**CG**A AGTGAACTGATTGCC
14	GGCAATCAGTTCACT T^{5m}**CG**T GCCCAGGTATTTAGCC	GGCTAAATACCTGGGC A^{hm}**CG**A AGTGAACTGATTGCC

We have shown that the MBD domain of MBD4 has the wider binding substrate specificity compared to the MBD domain of MBD1. Our crystal structures of the MBD_{MBD4}-DNA complexes provided the structural basis of the binding. The wide DNA binding specificity of the MBD domain seems to underlie the multi-functionality of MBD4 and appears to be important for the role as a glycosylase to search for DNA damage. Our glycosylase assays indicated that MBD4 can cut thymine and ^{hm}U bases mispaired with guanine but is not active for oxidized cytosine bases, ^{fo}C and ^{ca}C which have been shown recently to be excised by TDG. Although the detailed mechanism of the DNA demethylation pathway is controversial, our glycosylase experiments indicate that MBD4 can be involved in the proposed pathway which includes the deamination of cytosine or hydroxymethylcytosine bases.

Methods

Protein expression and purification.

A DNA fragment encoding MBD_{MBD4} (residues 69-136) was amplified by PCR and cloned into a bacterial expression vector pGEX4T-3 (GE Healthcare Biosciences) engineered for protein expression with an N-terminal Glutathione-S-transferase (GST) tag and small ubiquitin-like modifier-1 (SUMO-1) fusion tag. The GST-SUMO-1 fusion MBD_{MBD4} was overexpressed in *E. coli* strain BL21(DE3). Cells were grown at 37 °C in Luria–Bertani medium (LB) containing 50 μg/ml ampicillin to an optical density of 0.5–0.6 at 660 nm, and then induced with 0.2 mM isopropyl β-d-thiogalactoside (IPTG) for 15 hours at 18 °C. The following steps were carried out at 4°C. Cells were harvested by centrifugation, and lysed by sonication in 50 mM Tris-HCl pH 8.0 buffer containing 300 mM NaCl, 1 mM dithiothreitol (DTT), 5% glycerol, 0.1% Triton X-100 and 1 mM phenylmethylsulfonyl fluoride (PMSF). The clarified lysate was loaded onto Glutathione Sepharose 4 Fast Flow beads (GE Healthcare) after the debris was removed by centrifugation. MBD_{MBD4} was eluted from the beads with elution buffer containing 10 mM glutathione, and the tag was removed by SENP2 protease treatment. The tag-free MBD_{MBD4} was further purified by sequential column chromatography steps using HiTrap Heparine HP and HiLoad 16/60 Superdex 75 columns (GE Healthcare Biosciences). Purified protein in the final elution buffer containing 10 mM Hepes–NaOH pH 7.4, 150 mM NaCl and 2 mM DTT was concentrated using an Amicon Ultra 3,000 cut-off membrane concentrator (Millipore).

To introduce the selenomethionine, L116 of MBD4 was substituted with methionine. Se-Met containing MBD_{MBD4} was expressed in modified M9 medium supplemented with 60 mg/L of selenomethionine , 100mg/L of lysine, phenylalanine and threonine

and 50 mg/L of isoleucine and valine[33]. The expression and purification of the selenomethionine-labeled protein were carried out by the same procedure as that for the native protein.

Crystallization of the MBD_{MBD4}–DNA complexes

For crystallization of the complex, MBD_{MBD4} at the concentration of 200-800 μM was mixed with annealed each DNA fragment at a 1:1 molar ratio.

Crystals of MBD_{MBD4} in the complex with the 14-bp oligomer containing ^{5m}CG/TG were obtained in a hanging drop consisting of 1.0 μl complex solution mixed with 1.0 μl precipitant solution containing 10% PEG 10K, 100 mM Na acetate pH 4.4 and 200 mM NaCl. Selenium derivative crystals were prepared under the same conditions using the Se-Met induced L116M mutant. Native crystals belongs to space group C2221 with cell constants a = 89.074 Å, b = 94.989 Å, c = 54.738 Å, α = 90°, β = 90°, γ = 90°. Isomorphic crystals of the complexes with the 14-bp ^{5m}CG/^{5m}CG and ^{5m}CG/^{hm}CG fragments were obtained from the similar conditions using PEG 1500 as a precipitant. Crystallization conditions and cell constants of all the MBD_{MBD4}-DNA complex crystals are listed in TableS1.

Data collection and structure determination.

All crystals were flash-frozen at 100 K in cryoprotectant containing 20% ethylene glycol. X-ray diffraction data sets were collected at a wavelength of 1.0000 Å on beamline BL-5A and BL-17A at Photon Factory, and were processed with the program HKL2000[34].

Phases of Se-Met labeled MBD_{MBD4} L116M in complex with 14-bp ^{5m}CG/TG were

obtained by the single wavelength anomalous dispersion method using the programs SOLVE and RESOLVE[35,36]. The initial model was built using the program COOT[37] and used as a search model to solve the crystal structure of the MBD_{MBD4} in complex with 14-bp ^{5m}CG/TG by the molecular replacement method with the program molrep from the CCP4 suit[38]. The structural model was further refined using Phenix suite[39], yielding a crystallographic R factor of 18.81% and a free R factor of 21.53% to 2.0 Å. Each structure of the MBD_{MBD4} - ^{5m}CG/^{5m}CG, MBD_{MBD4} - ^{5m}CG/^{hm}CG or MBD_{MBD4} - 11-bp ^{5m}CG/TG complex was solved by a molecular replacement method, using the structure of MBD_{MBD4}-14mer ^{5m}CG/TG as a search model. The crystallographic, diffraction and refinement statistics are summarized in supplementary Tables 1 and 2. The stereochemical quality of the final models was assessed using MolProbity[40]. All figures of protein and DNA molecules were produced using PyMOL (W. L. DeLano; http://www.pymol.org). The protein-DNA interactions revealed by our crystal structures are summarized in supplementary figures S16-S19.

The model of the MBD_{MBD4}-^{5m}CG/^{hm}UG or MBD_{MBD4}-^{5m}CG/^{fo}CG complex was built based on the structure of the MBD_{MBD4}-^{5m}CG/TG or MBD_{MBD4}-^{5m}CG/^{5m}CG complex. The coordinate of the ^{hm}C base from the crystal structure of MBD_{MBD4}-^{5m}CG/^{hm}CG complex or the small molecule crystal structure of the ^{fo}C base (Cambridge Crystallographic Data Centre code: 843055) was overlaid onto the thymine or ^{5m}C base in the lower strand to predict the position of the hydroxyl oxygen atom[41].

DNA binding assay.

Isothermal titration calorimetry measurements were performed on a MicroCal itc200 instrument at 25 °C. Protein solutions were exchanged into the ITC measurement buffer of 25 mM HEPES-NaOH buffer (pH 7.4) containing 100 mM NaCl and 0.1 mM TCEP

by gel-filtration chromatography or dialysis. Each 14-bp synthesized oligonucleotide (Gene Design Inc.) dissolved in 25 mM ammonium acetate solution. Complementary strands were mixed at an equimolar ratio, and annealed by decreasing temperature from 95°C down to 4 °C for 12 h. Each annealed DNA duplex was dried and dissolved in the ITC buffer. The DNA solution at 10-20 μM in the calorimetric cell was titrated with protein solution at 100-400 μM. Binding constants were calculated by fitting the data using the ITC data analysis module of Origin 7.0 (OriginLab Corporation).

Gel mobility shift assays and competitive binding assays were also performed in the ITC buffer. The upper strand of 14mer ^{5m}CG/^{5m}CG DNA fragment was radioisotope-labeled at the 5′ end with T4 polynucleotide kinase (TOYOBO) and ^{32}P-γ-ATP (Muromachi Kagaku, Tokyo). The labeled strand was then mixed with 1.2-fold amount of the complementary strand and annealed. Samples containing 3 μM of the MBD protein, 1 μM of radioisotope-labeled ^{5m}CG/^{5m}CG and 0, 1, 2, 4 μM non-label competitor DNA fragment were incubated for 30 min at 4 °C and then separated by a native gel electrophoresis and visualized with a Fuji BAS-2000 phosphor imager. Each of the bound to total ratios is estimated by quantification of the band on the native gel. Three data sets by independent experiments are fitted simultaneously to the Morrison's equation by a non-linear, least square method, to draw a fitting curve[42].

Glycosylase assay.

32P radioisotope-labelling was introduced to each 35mer oligonucleotide by kinase reaction described above. The labeled strand was then annealed with 1.2-fold amount of the complementary strand. Glycosylase assays were performed in a 10 μl reaction mixture containing 10 mM Tris-HCl, pH8.0, 0.1 mM EDTA and 0.1 mg/ml BSA. 40 or

400 nM of the indicated DNA duplexes were incubated with 200 or 2000 nM of human TDG or mouse MBD4 at 37°C for 1 hour. Reactions were terminated by the addition of 10 μl of a reaction stop solution containing 0.2 M NaOH and 20 mM EDTA followed by incubation at 95°C for 10 min. After addition of 60 μl of 10 M Urea, 20 μl of each sample was subjected to electrophoresis in a 9 M Urea/20% PAGE and visualized with a Fuji BAS-2000 phosphor imager. The ^{fo}C containing oligonucleotide was digested by the incubation at 95°C for 10 min under alkaline pH condition regardless of the enzymatic activity. The reactions for the ^{fo}C containing oligo nucleotide were therefore stopped by the addition of 10 μl of the stop solution followed by incubation at 70°C for 5 min.

Table S2. Crystallographic data and data collection statistics

Crystal	14mer ^{5m}CG/TG (SeMet)	14mer ^{5m}CG/TG	14mer ^{5m}CG/^{5m}CG	14mer ^{5m}CG/^{hm}CG	11mer ^{5m}CT/TG
Crystallization condition	12% PEG10,000 0.1 M Na acetate pH=4.4 0.2 M NaCl	12% PEG10,000 0.1 M Na acetate pH=4.4 0.2 M NaCl	7% PEG10,000 0.1 M Na acetate pH=4.4 0.2 M NaCl	8% PEG 1,500 0.1 M Na acetate pH=3.9 0.1 M Na cacodylate pH=5.4	2 mM Zn acetate 0.1 M Na Cacodylate
X-ray source	PF-NE3A	PF BL-5	PF-NW12	PF-BL17	Spring8-BL38
wave-length (Å)	0.97923 (peak)	1.0000	1.0000	1.0000	1.0000
Space group	C2221	$C222_1$	$C222_1$	$C222_1$	P1
Unit cell parameters (Å, °)	a=88.752 b=97.588 c=54.959 α=β=γ=90	a=89.074 b=94.989 c=54.738 α=β=γ=90	a=89.182 b=93.829 c=55.357 α=β=γ=90	a=88.693 b=97.758 c=55.725 α=β=γ=90	a=30.324 b=33.781 c=60.035 α=75.317 β=77.743 γ=87.352
Resolution range (Å)[1]	50-2.7 (2.8-2.7)	50-2.0 (2.07-2.0)	50-2.2 (2.28-2.2)	50-2.19 (2.27-2.19)	50-2.5 (2.59-2.5)
Total observations	38796	114920	86212	74818	20848
Unique reflections[1]	6051 (363)	16062 (1591)	12048 (1192)	11517 (701)	7225 (531)
Multiplicity[1]	6.4 (5.6)	7.2 (7.3)	7.2 (7.3)	6.5 (4.1)	2.9 (1.7)
R_{merge}[1,2]	0.097 (0.308)	0.031 (0.365)	0.078(0.477)	0.052 (0.317)	0.058 (0.125)
Completeness (%)[1]	88.6 (54.0)	99.7 (100)	99.6 (99.9)	89.6 (54.7)	92.8 (67.4)
$\langle I/\sigma(I)\rangle$[1]	12.7 (5.2)	18.2 (4.3)	12.4 (4.5)	14.7 (4.1)	16.3 (5.4)

[1] Numbers in parentheses are the values for the highest resolution shell of each data set.

[2] Rmerge = $\Sigma h\, \Sigma i \mid I(h)i - \langle I(h)\rangle \mid / \Sigma h\, \Sigma i\, I(h)i$,where $I(h)$ is the intensity of reflection h, Σh is the

sum of all measured reflections and Σi is the sum of i measurements of reflection.

Table S3. Refinement statistics

Crystal	14mer ^{5m}CG/TG	14mer ^{5m}CG/^{5m}CG	14mer ^{5m}CG/^{hm}CG	11mer ^{5m}CT/TG
Resolution range (Å)	32.5-2.00	28.3-2.20	34.7-2.40	33.0-2.53
R_{work} (%)[1,2]	18.81 (22.61)	18.92 (23.32)	19.06 (28.38)	19.34 (24.85)
R_{free} (%) [1,2]	21.53 (25.81)	21.06 (29.61)	21.99 (32.24)	23.65 (33.47)
RMS Deviations				
bond length (Å)	0.005	0.003	0.004	0.007
bond angle (°)	1.148	1.045	1.008	1.161
Ramachandran plot				
favored (%)	98.41	100	98.36	99.18
allowed (%)	100	100	100	100
Number of atoms				
Protein	525	520	511	1030
DNA	570	570	571	429
Water	152	120	71	55
Ligand	48	37	8	7

[1] R_{work} and R_{free} = (Σhkl ||*F*o| - |*F*c||) / Σhkl |*F*o|, where the free reflections (5% of the total

used) were held aside for R_{free} throughout refinement.

[2] Numbers in parentheses are the values for the highest resolution shell of each data set.

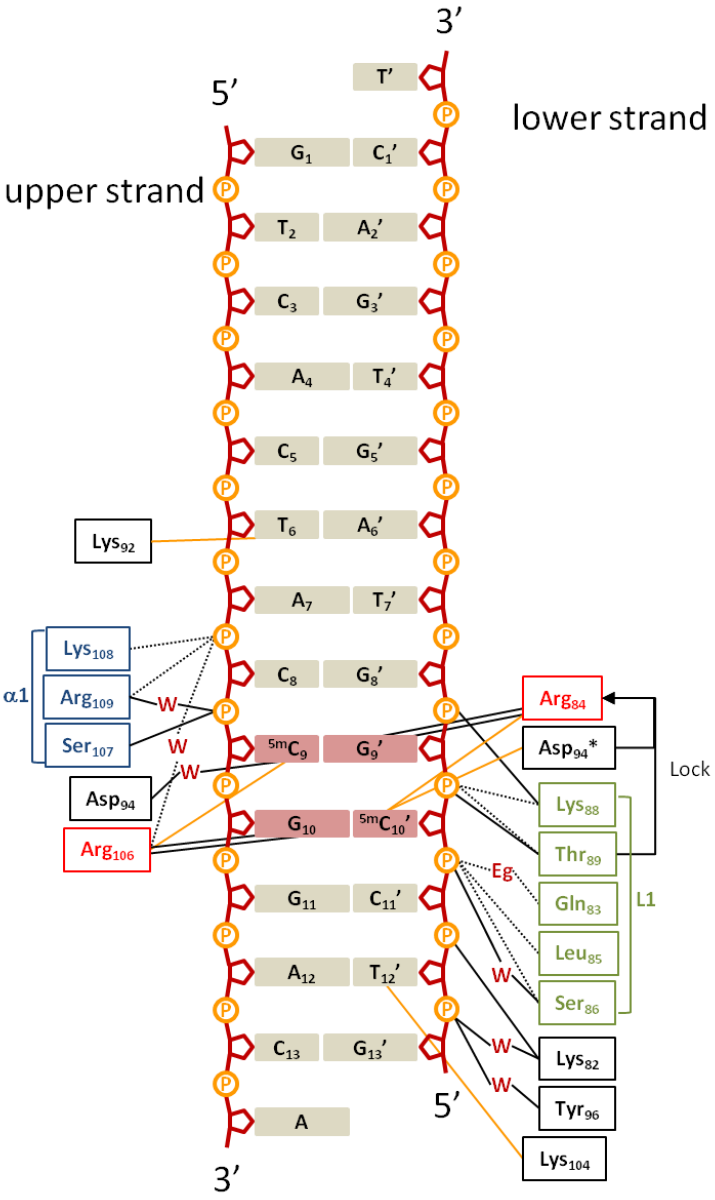


Figure S16. The summary of the interaction between MBD_{MBD4} and the 14-bp $^{5m}CG/^{5m}CG$ fragment. The black solid and dotted lines indicate the hydrogen bonds associated with main chains and side chains of MBD_{MBD4}, respectively. The orange solid lines indicate van der Waals contacts to the DNA bases. The amino acid residues within the L1 loop and the a1 helix are colored in green and blue, respectively. The asterisks denote that the residues are duplicated on this figure. "W" and "Eg" indicate an ordered water molecule and an ethylene glycol molecule, respectively.

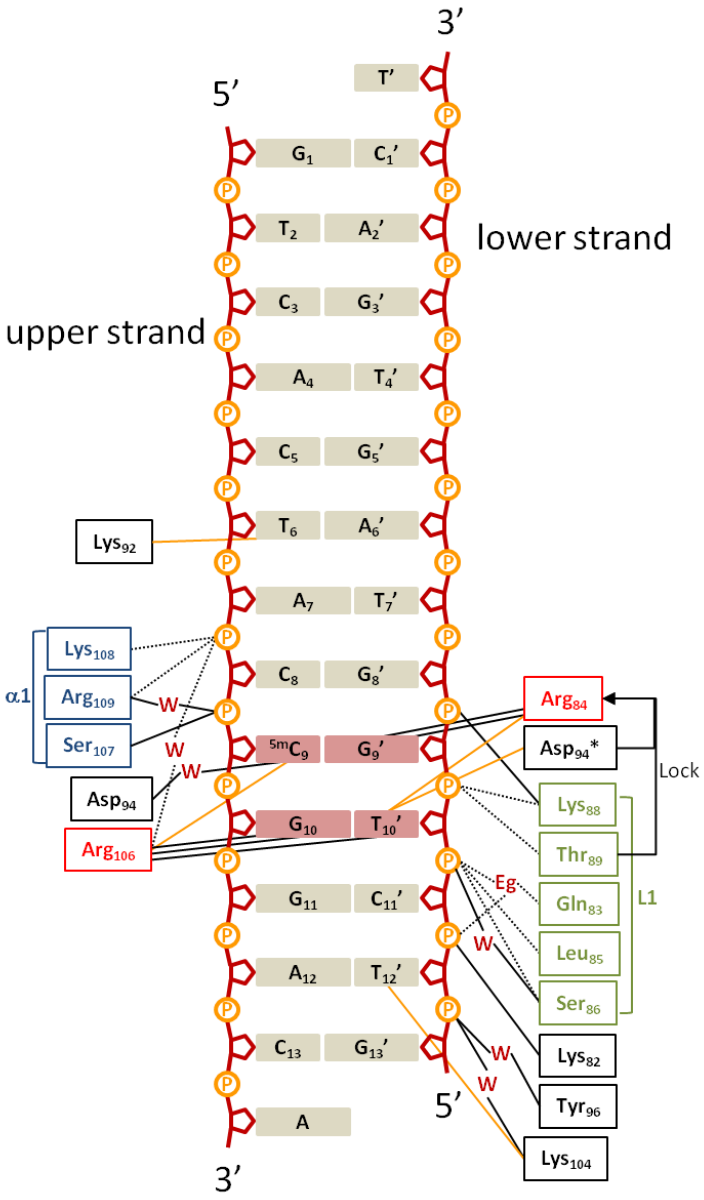


Figure S17. The summary of the interaction between MBD_{MBD4} and the 14-bp $^{5m}CG/TG$ fragment.

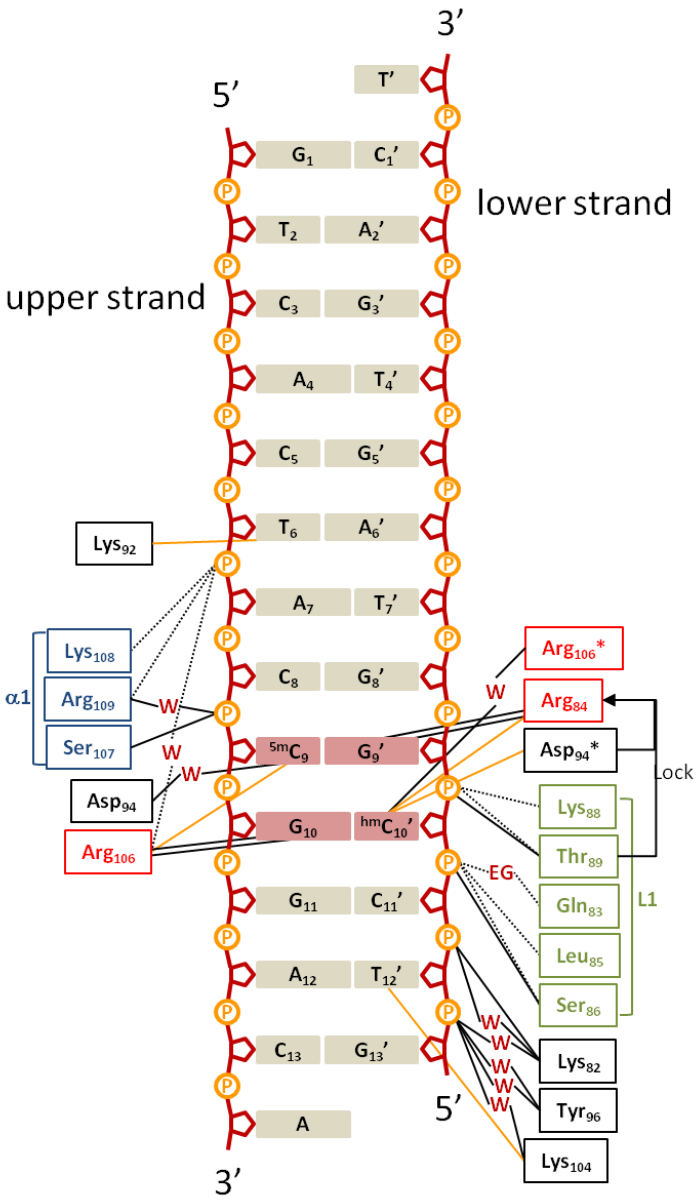


Figure S18. The summary of the interaction between MBD_{MBD4} and the 14-bp $^{5m}CG/^{hm}CG$ fragment.

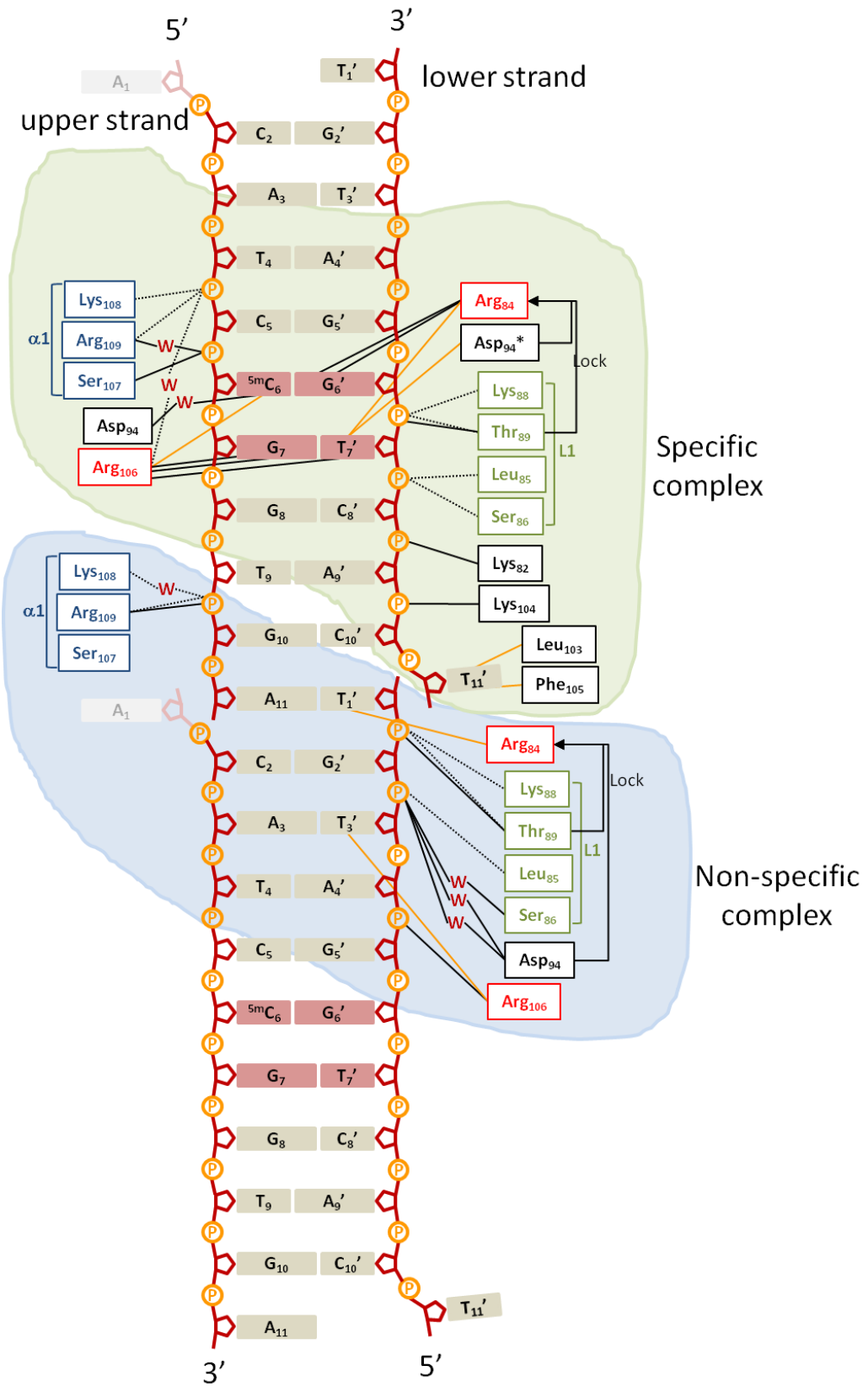

5′
3′
upper strand
lower strand
A1
T1′
C2
G2′
A3
T3′
T4
A4′
C5
G5′
5mC6
G6′
G7
T7′
G8
C8′
T9
A9′
G10
C10′
A11
T11′
α1
Lys108
Arg109
Ser107
Asp94
Arg106
W
Arg84
Asp94*
Lock
Lys88
Thr89
Leu85
Ser86
L1
Lys82
Lys104
Leu103
Phe105
Specific complex
Non-specific complex
3′
5′

Figure S19. The summary of the interaction between MBD_{MBD4} and the 11-bp $^{5m}CG/TG$ DNA fragment. The symmetry related molecules of DNA molecules are stacked with each other. The translucent DNA base is invisible in this crystal. The amino acid residues of MBD_{MBD4} bound to the specific and non-specific site are listed on the green and blue back ground, respectively.

References

1. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev* **16**, 6-21 (2002).
2. Li, E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet* **3**, 662-73 (2002).
3. Ehrlich, M. et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res* **10**, 2709-21 (1982).
4. Robertson, K.D. & Wolffe, A.P. DNA methylation in health and disease. *Nat Rev Genet* **1**, 11-9 (2000).
5. Bogdanović, O. & Veenstra, G.J. DNA methylation and methyl-CpG binding proteins: developmental requirements and function. *Chromosoma* **118**, 549-65 (2009).
6. Riccio, A. et al. The DNA repair gene MBD4 (MED1) is mutated in human carcinomas with microsatellite instability. *Nat Genet* **23**, 266-8 (1999).
7. Millar, C.B. et al. Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. *Science* **297**, 403-5 (2002).
8. 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**, 301-4 (1999).
9. Ohki, I. et al. Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. *Cell* **105**, 487-97 (2001).
10. Ho, K.L. et al. MeCP2 binding to DNA depends upon hydration at methyl-CpG. *Mol Cell* **29**, 525-31 (2008).

11. Scarsdale, J.N., Webb, H.D., Ginder, G.D. & Williams, D.C. Solution structure and dynamic analysis of chicken MBD2 methyl binding domain bound to a target-methylated DNA sequence. *Nucleic Acids Res* **39**, 6741-52 (2011).

12. Kangaspeska, S. et al. Transient cyclical methylation of promoter DNA. *Nature* **452**, 112-5 (2008).

13. Métivier, R. et al. Cyclical DNA methylation of a transcriptionally active promoter. *Nature* **452**, 45-50 (2008).

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

15. Rai, K. et al. DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45. *Cell* **135**, 1201-12 (2008).

16. He, Y.F. et al. Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science* (2011).

17. Kim, M.S. et al. DNA demethylation in hormone-induced transcriptional derepression. *Nature* **461**, 1007-12 (2009).

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

19. Maiti, A. & Drohat, A.C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: Potential implications for active demethylation of CpG sites. *J Biol Chem* (2011).

20. Hunter, W.N. et al. The structure of guanosine-thymidine mismatches in B-DNA at 2.5-A resolution. *J Biol Chem* **262**, 9962-70 (1987).

21. Olson, W.K. et al. A standard reference frame for the description of nucleic acid

base-pair geometry. *J Mol Biol* **313**, 229-37 (2001).

22. Liu, P., Burdzy, A. & Sowers, L.C. Repair of the mutagenic DNA oxidation product, 5-formyluracil. *DNA Repair (Amst)* **2**, 199-210 (2003).

23. Kondo, E., Gu, Z., Horii, A. & Fukushige, S. The thymine DNA glycosylase MBD4 represses transcription and is associated with methylated p16(INK4a) and hMLH1 genes. *Mol Cell Biol* **25**, 4388-96 (2005).

24. Wu, P. et al. Mismatch repair in methylated DNA. Structure and activity of the mismatch-specific thymine glycosylase domain of methyl-CpG-binding protein MBD4. *J Biol Chem* **278**, 5285-91 (2003).

25. Aziz, M.A., Schupp, J.E. & Kinsella, T.J. Modulation of the activity of methyl binding domain protein 4 (MBD4/MED1) while processing iododeoxyuridine generated DNA mispairs. *Cancer Biol Ther* **8**, 1156-63 (2009).

26. Inomata, K. et al. Kinetic and thermodynamic evidence for flipping of a methyl-CpG binding domain on methylated DNA. *Biochemistry* **47**, 3266-71 (2008).

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

28. Valinluck, V. et al. Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2). *Nucleic Acids Res* **32**, 4100-8 (2004).

29. Guo, J.U., Su, Y., Zhong, C., Ming, G.L. & Song, H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell* **145**, 423-34 (2011).

30. Inoue, A., Shen, L., Dai, Q., He, C. & Zhang, Y. Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. *Cell Res* **21**, 1670-6 (2011).

31. Inoue, A. & Zhang, Y. Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science* **334**, 194 (2011).

32. Wanunu, M. et al. Discrimination of Methylcytosine from Hydroxymethylcytosine in DNA Molecules. *J Am Chem Soc* (2010).

33. Doublié, S. Preparation of selenomethionyl proteins for phase determination. *Methods Enzymol* **276**, 523-30 (1997).

34. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Macromolecular Crystallography, Pt A* **276**, 307-326 (1997).

35. Terwilliger, T. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr D Biol Crystallogr* **55**, 849-61 (1999).

36. Terwilliger, T. Maximum-likelihood density modification. *Acta Crystallogr D Biol Crystallogr* **56**, 965-72 (2000).

37. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**, 2126-32 (2004).

38. Vagin, A. & Teplyakov, A. MOLREP: an Automated Program for Molecular Replacement. *Journal of Applied Crystallography* **30**, 1022-1025 (1997).

39. Adams, P.D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-21 (2010).

40. Lovell, S.C. et al. Structure validation by Calpha geometry: phi,psi and Cbeta deviation. *Proteins* **50**, 437-50 (2003).

41. Münzel, M. et al. Improved synthesis and mutagenicity of oligonucleotides containing 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine. *Chemistry* **17**, 13782-8 (2011).

42. Morrison, J.F. Kinetics of the reversible inhibition of enzyme-catalysed reactions by tight-binding inhibitors. *Biochim Biophys Acta* **185**, 269-86 (1969).

CHAPTER 3

De novo DNA methylation is balanced by 5-hydroxymethylcytosine-mediated passive DNA demethylation in mouse embryonic stem cells

Abstract

Methylation at the C5 position of cytosine in DNA has important roles in genome function and is dynamically reprogrammed during early embryonic and germ cell development. Recently, 5-hydroxymethylcytosin (hmC) is found to be generated by oxidation of 5-methylcytosine (5mC) by the TET family enzymes that are highly expressed in embryonic stem (ES) cells. Cytosine hydroxymethylation has become the center of attention because it may be involved in the DNA demethylation pathways. In this study, we provide insights into the nature of hmC bases which is generated mainly from 5mC bases deposited by "*de novo*" DNA methyltransferases and is progressively diluted by cell division, resulting in passive DNA demethylation.

Introduction

DNA methylation is one of the principle regulators of the epigenetic landscape that defines gene expression programs[1]. In mammals, there are two classes of DNA methyltransferases, maintenance DNA methyltransferase, Dnmt1 which copies parental DNA methylation pattern to daughter cells with the help of ubiquitin-like with PHD and RING finger domains 1 (Uhrf1)[2-6] and *de novo* DNA methyltransferases, Dnmt3a and Dnmt3b, which are responsible for establishing DNA methylation patterns during embryogenesis and gametogenesis[7]. While mechanisms of establishment and maintenance of DNA methylation by DNA methyltransferases are well characterized[8,9], it is less clear which enzymatic machinery is responsible for DNA demethylation or even which pathways lead to DNA demethylation[10,11]. Recently, it was discovered that the TET family of Fe(II) and 2-oxoglutarate dependent enzymes, TET1, 2 and 3, oxidize 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (hmC)[12-14]. 5mC hydroxylation by TET enzymes has become the center of attention because it is involved in DNA demethylation.

TET1 and TET2 are highly expressed in mouse embryonic stem (ES) cells and are rapidly down-regulated upon differentiation while TET3 is especially prevalent in oocytes and zygotes[15,16]. TET proteins have been shown to have important roles in cancer and stem cell biology[12,14,17,18]. In ES cells lacking all three DNA methyltransferases, hmC and 5mC bases are lost completely implicating that hmC bases are produced exclusively by 5mC hydroxylation in ES cells[19,20]. In rapidly dividing ES cells, hmC bases seem to be progressively diluted by cell division. hmC bases are lost also by further conversion into 5-formylcytosine and 5-carboxylcytosine by TET enzymes or into 5-hydroxymethyluracil by AID/APOBEC family enzymes, followed by

DNA repair initiated by glycosylase reaction[21-25]. Thus the amount of hmC bases in ES cells seems to be produced by sequential reactions of DNA methyltransferases and TET enzymes and decreased through cell division mediated and independent mechanisms.

Here, we have quantified 5mC and hmC bases within a series of DNMT enzyme knock out ES cell lines and cells treated with cell cycle inhibitors to move the equilibrium of the amount of hmC bases and gain insights into the molecular details behind the equilibrium. Our results imply that Dnmt3a and Dnmt3b, rather than Dnmt1, supply most of the 5mC bases for subsequent hydroxylation by TET enzymes and that a large portion of hmC is stable throughout the cell cycle of ES cells and replication dependent dilution is the major pathway to reduce hmC content. Our results highlight the dynamic nature of hmC bases, which are progressively produced and diluted by cell division, working antagonistically to *de novo* DNA methylation by inducing passive DNA demethylation.

Results and discussion

DNA hydroxymethylation is dependent on de novo DNA methyltransferases.

The hmC base is generated from pre-existing 5mC by the enzymatic activities of Tet family enzymes. To examine the relationship between DNA methylation and hydroxymethylation, we quantified the amount of occupied CpG sites and hmC bases of a series of J1 ES cell lines depleted of Dnmt enzymes. We have conducted the biochemical assays using the bacterial DNA methyltransferase, SssI and the β-glucosyltransferase (βGT) of T4 phage, to quantify the occupied CpG sites and hmC bases essentially as described before[26]. SssI transfers methyl group from S-adenosylmethionine to the cytosine bases within the context of CpG, then the amount of vacant CpG sites and, consequently, the amount of occupied CpG sites can be estimated from the SssI reactivity. βGT is a glucosyltransferase that transfers glucose from UDP-glucose to the hmC base[26]. The linear calibration curves for these experiments were constructed using the control DNA fragments, each of which contains only cytosine, 5mC or hmC in the context of CpG sequence (Figure 1).

Surprisingly, the ratios of the occupied amount of the CpG sites and hmC bases were not constant among these Dnmt knock out cell lines, suggesting that the amount of hmC is not just a reflection of the amount of 5mC. In the triple DNMT knock out ES cell line, TKO, which lacks all the Dnmt enzymes, the 5mC and hmC bases were undetectable in our system, consistent with the notion that 5mC is prerequisite for hmC existence (Figure 2). The amount of occupied CpG sites in Dnmt1 knock out cells (1KO) was about 25% of that in J1 cells whereas the amount of hmC in 1KO is about 60 % of that in wild-type J1 cells (Figure 2). Thus, the amount of hmC is not severely affected by Dnmt1 knouck out. In contrast, the ES cells lacking both of *de novo* DNA

methyltransferases, Dnmt3a and Dnmt3b (7aa/bb) looses almost all of the hmC bases while the amount of the occupied CpG sites remains at significant level, about 50% of that in J1 cells, suggesting that the production of hmC bases is largely dependent on *de novo* DNA methylation activities of Dnm3a and Dnmt3b (Figure 2). Thus, hmC production largely depends on *de novo* DNA methyltransferases, Dnmt3a and Dnm3b, and less on maintenance DNA methyltransferase, Dnmt1.

The Dnmt3a or Dnmt3b single knock out cell line (3aKO or 3bKO) has similar levels of occupied CpG sites and hmC bases compared with those in J1 cells, indicating that these enzymes work redundantly with regard to the production of 5mC and hmC bases (Figure 2). The low level of hmC bases in 7aa/bb cells could be recovered by stable introduction of Dnmt3a1 or its splicing variant, Dnmt3a2 which lacks amino-terminal 219 amino acid residues (3aTAP or 3a2TAP), indicating these variants can contribute equivalently in hmC production pathway (Figure 2).

The mRNA expression levels of Tet proteins were not decreased in 7aa/bb and all the remaining Dnmt knock out cell lines examined, suggesting that the change in transcription of TET enzymes cannot explain the low level of hmC bases in 7aa/bb cells (Figure 3).

The large difference in the amount of hmC between 1KO and 7aabb cell lines clearly shows that *de novo* DNA methyltransferases, Dnmt3a and Dnmt3b supply most of the 5mC bases for subsequent hydroxylation by TET enzymes. Considering that the hydroxylation of 5mC might result in DNA de-methylation as recent reports have implied[21-25,27], the large decrease in the amount of hmC bases within 7aa/bb cells can be explained by the starvation of 5mC at the hmC positive sites because of the lack of *de novo* DNA methylation. The retained 5mC level in 7aa/bb cells may reflect the

restricted distribution of TET enzymes as revealed by recent genome-wide hmC localization analyses, which show that hmC bases are associated with CpG islands, gene bodies and bivalent domains and relatively depleted from heterochromatic region such as satellite repeats in ES cells[19,20,28].

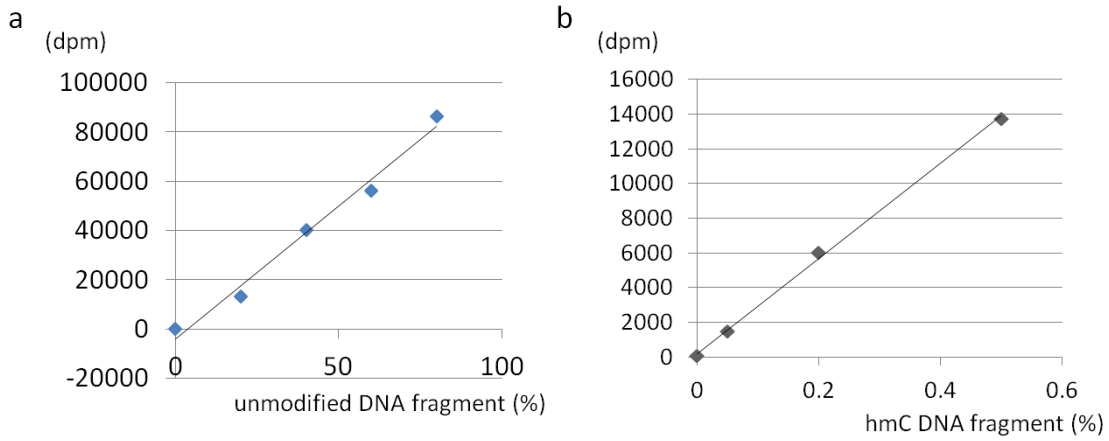


Figure 1. The representative linear calibration curves for the 5mC and hmC quantification assays. (**a**) The scintillation counts after SssI reaction for 200 ng of standard DNA samples, which are the mixtures of C and 5mC fragments at ratios of 0:1, 1:4, 2:3, 3:2, and 4:1. (**b**) The radio-activities of the filter papers after βGT assay for 200 ng of the C standard DNA spiked with 0, 0.1, 0.4 and 1 ng of the hmC standard fragment.

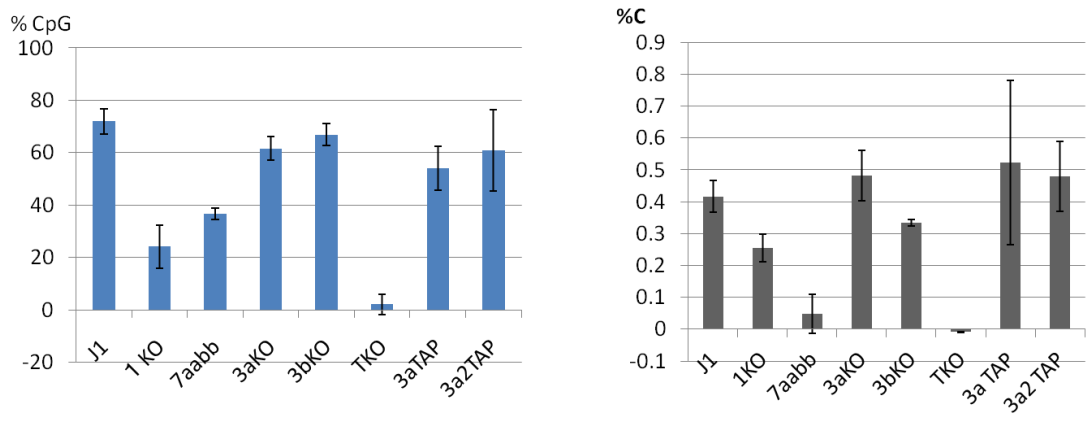


Figure 2. The results of quantification assay of occupied CpG sites determined by the SssI enzymatic assay (a) and hmC bases by βGT (b). The values represent the average of three experiments with independent genomic DNA samples. Black bars represent standard deviations.

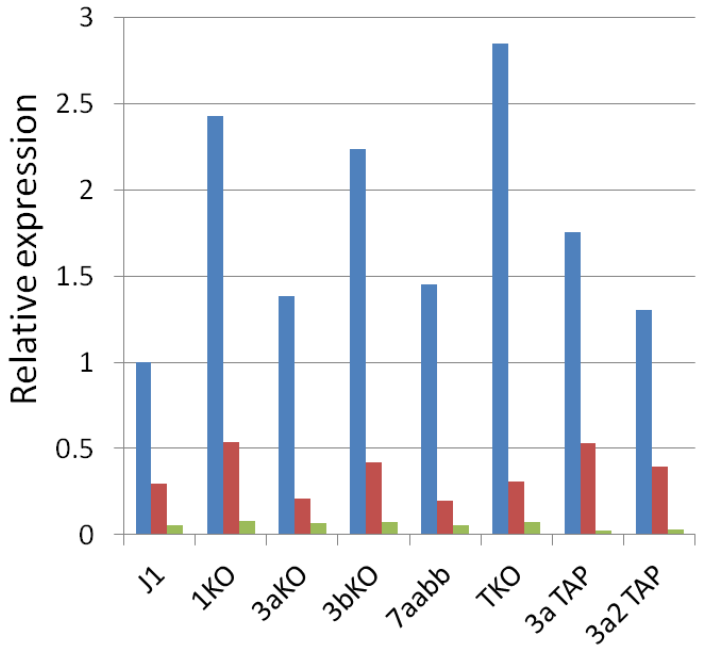


Figure 3. Relative mRNA expression of TET1, TET2 and TET3 in the series of Dnmt knock out cell lines are shown as bars in blue, red and green, respectively. The Tet1 mRNA expression level in the wild-type J1 cells has been set to 1.

Cell cycle arrest results in increased global hmC content.

Next, we examined the mechanism in which hmC bases disappear in ES cells. There are two possible pathways to decrease the global hmC content, one is the cell division dependent dilution where hmC bases of a parent cell will be equally divided into two daughter cells, and the other is replication-independent active demethylation where enzymatic activities remove hmC bases. To gain insight into this process, we have used DNA replication inhibitor, aphidicolin to stop the cell cycle at the S-phase. The hmC content of the aphidicolin treated cells double by 24 hours (Figure 4). The experiment was repeated with the other replication inhibitor, hydroxyurea to exclude the possibility of the side effect of aphidicolin and obtained the similar result. Time course analysis demonstrated progressive increment in hmC content in the presence of aphidicolin and reduction after washing out (Figure 5). These results clearly indicated the significant contribution of replication dependent dilution to reduce hmC bases. Contrary to the replication coupled DNA methylation by Dnmt1 and Uhrf1, 5mC hydroxylation by TET enzymes seemed not to depend on DNA replication as hmC production was not inhibited by DNA replication inhibitors.

The hmC content was also increased by G0/G1 arrest by serum depletion in culture medium but not changed in the presence of Nocodazole, a microtubule destabilizing agent, indicating that TET enzymes cannot access condensed chromosome of M-phase cells and/or TET enzymes are inactivated or degraded at M phase (Figure 4).

Although recent reports suggest that the hmC is an intermediate of active DNA demethylation pathway where hmC bases are further converted to hydroxymethyluracil, formylcytosine or carboxycytosine bases followed by base excision repair[21-25], our

results indicate that the majority of hmC bases are stable throughout the cell cycle and lost via replication dependent dilution in ES cells.

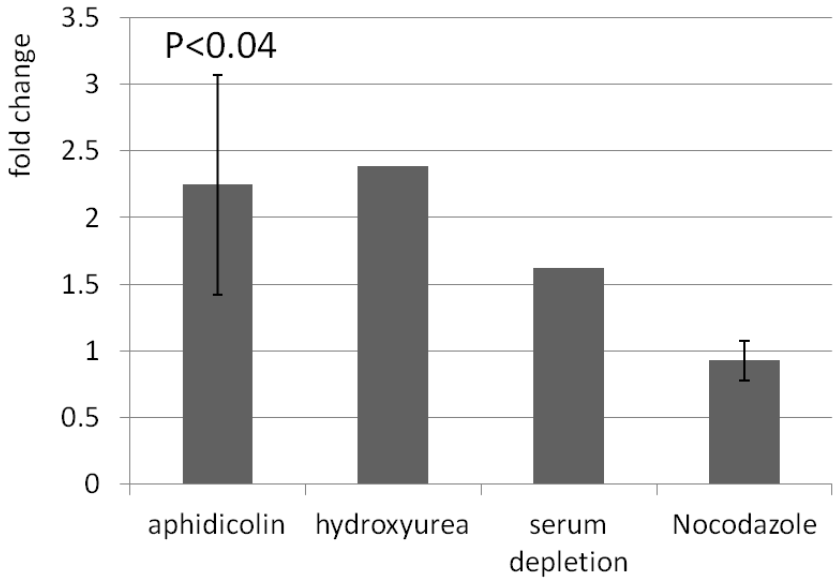


Figure 4. Changes in hmC content of J1 cells after cell growth inhibition by aphidicolin, hydroxyurea, serum depletion and Nocodazole treatments. The values represent the fold changes from the control cells which were proliferating exponentially. The values for aphidicolin and Nocodazole treatments represent averages of three independent experiments with standard deviations (for aphidicolin paired t test, $P < 0.04$).

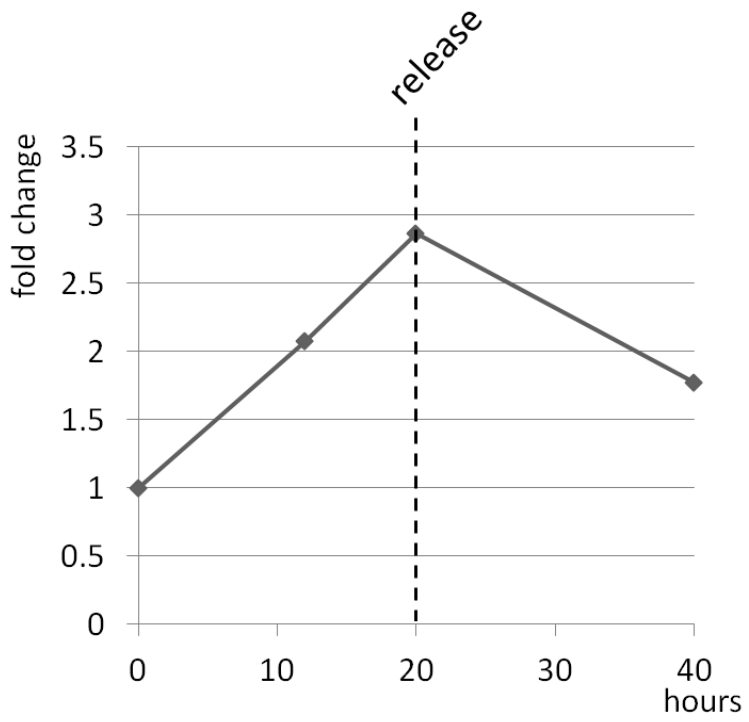


Figure 5. Time course dependent accumulation of hmC bases in the presence of aphidicolin. Fold changes from the control cells (0-hour sample) were plotted as in figure 4. After aphidicolin treatment for 20 hours, the cells were washed and incubated in the normal ES medium for another 20 hours.

The hydroxylation of 5mC inhibits maintainance DNA methylation system.

The molecular details of the DNA demethylation is largely unknown. However, previous report shows that the hydroxylation of 5mC disturbs the maintenance DNA methyltransferase activity of Dnmt1[27].

Since we have a quantitative system for measuring DNA methyltransferase activity, we first examined the maintenance DNA methyltransferase activity of Dnmt1. Consistent with previous report, hydroxylation at hemimethyl-CpG (5mCG/CG) site has largely diminished the catalytic activity of Dnmt1 (Figures 6 & 7). The Dnmt1 activity toward the hemihydroxymethyl-CpG (hmCG/CG) site is below 10% of the activity toward 5mCG/CG site. In contrast, *de novo* DNA methylases, Dnmt3a and Dnmt3b are active for all of the substrate DNA sequences, non-methyl, hemi-methyl and hemi-hydroxymethyl CpG sequences consistent with the lack of recognition of complementary strand by *de novo* DNA methyltransferases (Figures 8 & 9).

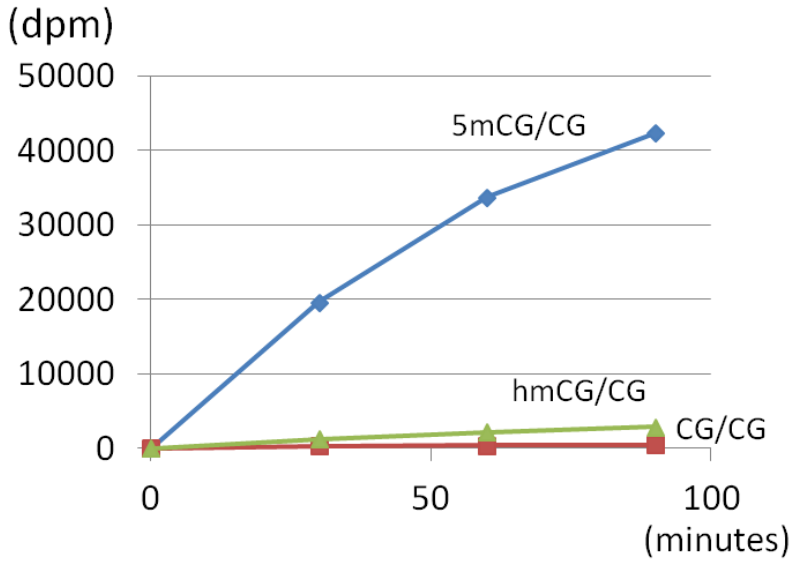


Figure 6. The DNA substrate specificity of maintenance DNA methyltransferase, Dnmt1. Radio-activities incorporated into the substrate DNA fragments after incubation for 30, 60 and 90 minutes were analyzed. The ratio of reaction rate for the 5mCG/CG and hmCG/CG substrates was estimated to be about 10:1 from the slopes of the linear fitting curves.

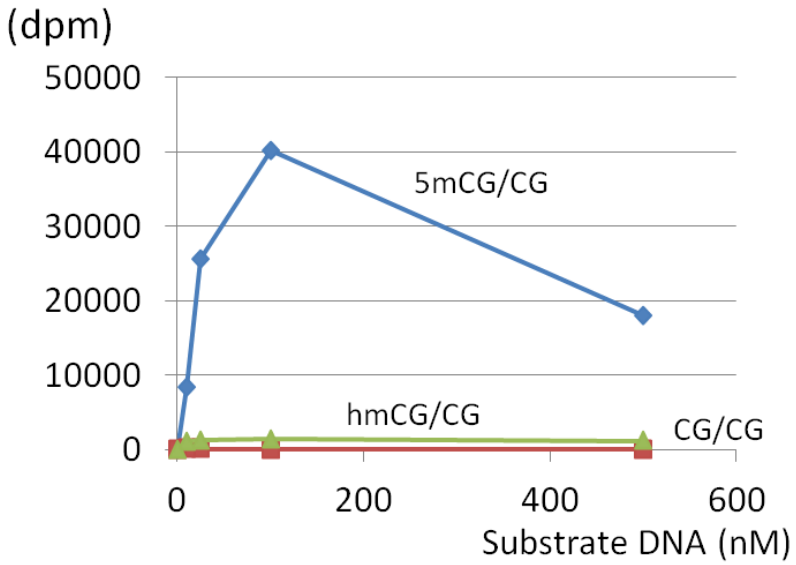


Figure 7. The substrate concentration dependence for Dnmt1 methyltransferase reactions with increasing amount of substrate DNA fragments at the concentration of 0, 10, 25, 100 or 500 nM.

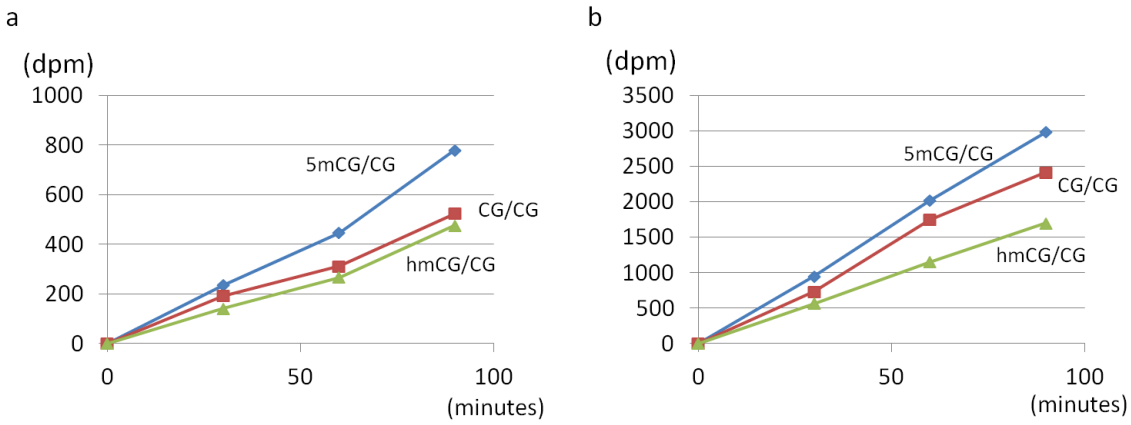


Figure 8. The DNA substrate specificity of *de novo* DNA methyltransferases, Dnmt3a (a) and Dnmt3b (b). Radio-activities incorporated into the substrate DNA fragments after incubation for 30, 60 and 90 minutes were plotted.

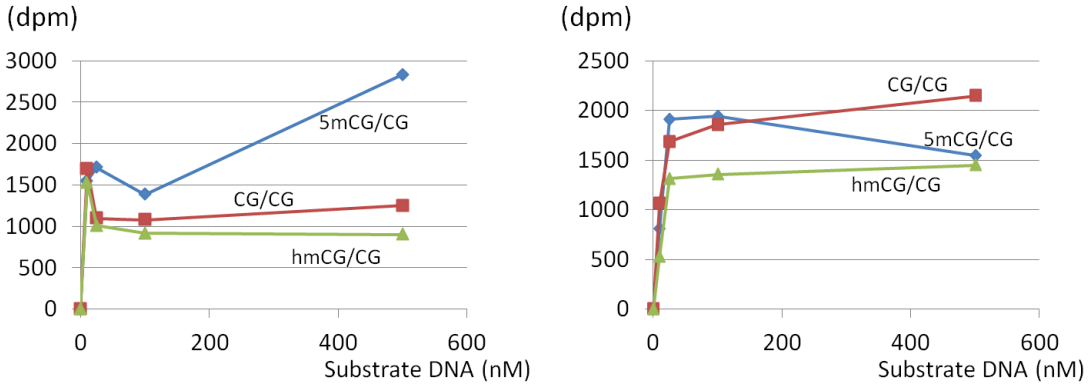


Figure 9. DNA-substrate dose dependency in the reaction of *de novo* DNA methyltransferases. The concentration of substrate DNA fragments were 0, 10, 25, 100 or 500 nM.

Uhrf1 is also an essential protein for maintenance DNA methylation and the Uhrf1 knock out ES cells exhibit global DNA hypomethylation[2,3]. The SRA (SET and RING associated) domain of Uhrf1 flips out the 5mC base from the double stranded DNA to recognize the hemi-methylated CpG site[4-6]. The methyl group of the 5mC base is recognized via extensive contacts within the pocket on the DNA binding surface of the SRA domain[4-6]. The tight recognition pocket seems to reject the hmC base to be accommodated (Figure 10a).

Thus we examined the binding between the SRA domain and the 12-bp DNA fragment containing a 5mCG/CG, hmCG/CG or CG/CG site. We observed that the SRA domain prefers the 5mCG/CG over the hmCG/CG although the hmCG/CG fragment is preferred over the CG/CG (Figure 10b). This result is contrary to the recent report which shows that the SRA domain can bind to both of the 5mCG/CG and hmCG/CG sites with similar affinity[29]. This discrepancy can be caused by the strong nonspecific DNA binding of the SRA domain. We have had trouble showing the selectivity of the SRA domain in the gelshift experiments using 42-bp DNA fragments (data not shown). With our experimental conditions using 12-bp DNA fragments, we could show the binding preference of the SRA domain for the 5mCG/CG site over the hmCG/CG and CG/CG sites.

To show the binding specificity of the SRA domain more clearly, we performed the competitive DNA binding assay. The 32P-labeled 12-bp duplex containing a 5mCG/CG site bound by the SRA domain is competed off by the increasing amount of unlabeled 12-bp duplexes containing a 5mCG/CG, hmCG/CG or CG/CG site. This experiment showed the significantly stronger competition effect of the 5mCG/CG fragment than that of the hmCG/CG fragment indicating the tighter binding between the SRA and

5mCG/CG (Figure 11, compare lanes 3 and 7). These results show that the hydroxylation of 5mC will disturb the maintenance DNA methylation system not only by inhibiting the catalytic activity of Dnmt1 but also by degrading recognition by the SRA domain of Uhrf1.

Our cellular experiments indicated that hmC production depends largely on *de novo* DNA methyltransferases and most part of hmC reduction is mediated by cell division and our biochemical analyses showed that 5mC hydroxylation results in passive DNA demethylation. Combined with the limited hmC distribution revealed by recent genome-wide analyses[19], it could be said that cytosine hydroxymethylation by TET enzymes is a mechanism to protect some genomic regions from *de novo* DNA methyltransferase activity by inducing passive DNA demethylation in ES cells. Recent reports have shown that hmC, 5-formylcytosine and 5-caroxylcytosine bases are produced by TET3 in the paternal pronuclei and are diluted progressively by cell division through early embryonic development, indicating that the zygotic paternal DNA demethylation is initiated by conversion of 5mC to hmC followed by cell division dependent dilution[30,31].

It seems appropriate to consider that the majority of DNA demethylation is accomplished via 5mC hydroxylation followed by passive dilution in dividing cells because the passive demethylation pathway appears to be safer than other proposed DNA demethylation mechanisms which rely on DNA repair machinery[11]. It should be noted that our results do not exclude the partial contribution of active DNA demethylation in ES cells.

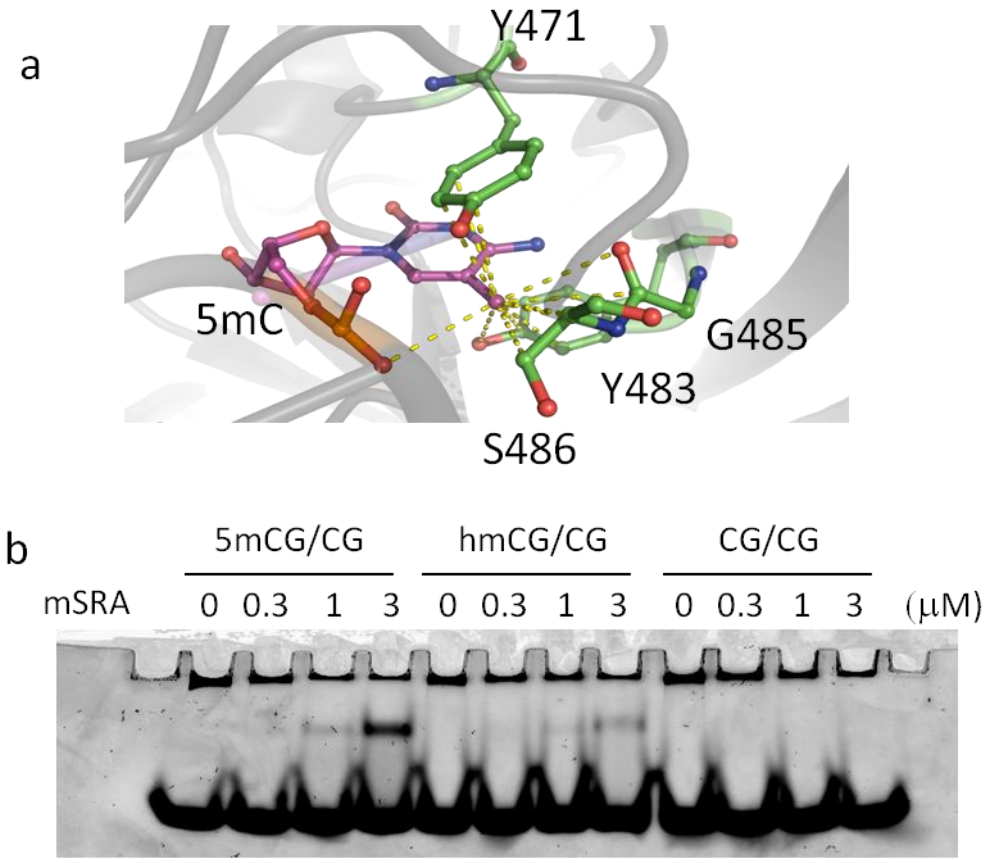


Figure 10. (**a**) The tight recognition of 5mC by Uhrf1 revealed by the crystal structure of the SRA domain of Uhrf1 in complex with 5mCG/CG (PDB code; 2ZKD). The flipped 5mC base and the protein side chains which are critical for 5mC recognition are shown as stick models in purple and green, respectively. Yellow dotted lines represent van der Waals contacts (3.5 - 4.2 Å). (**b**) The DNA binding specificity of the SRA domain was assessed by electrophoretic mobility shift assay stained with GelGreen. Each of the sequence of substrate DNA fragments is indicated above.

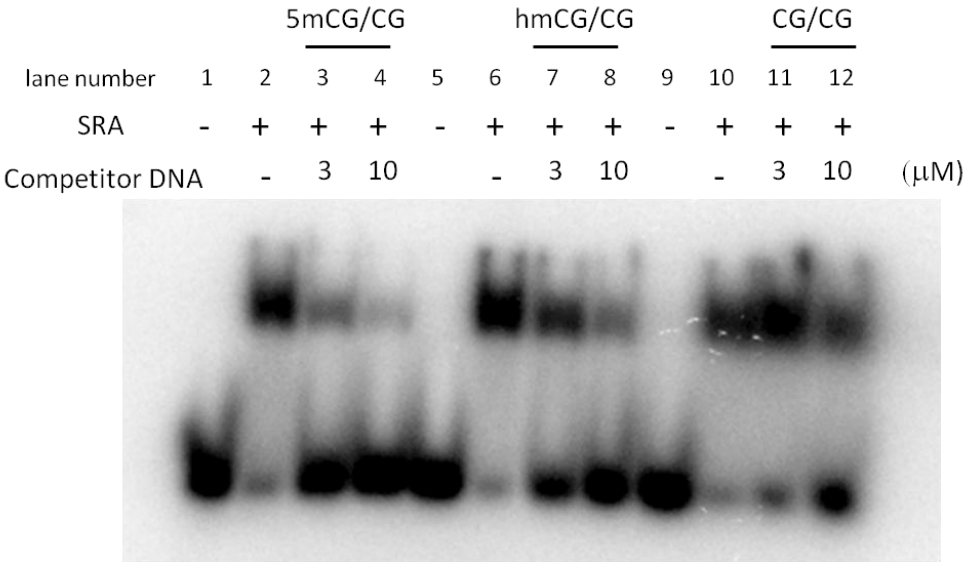


Figure 11. The complex between the SRA and 32P-labeled 5mCG/CG was competed off by non-labeled 5mCG/CG, hmCG/CG or CG/CG fragment. The sequence and the amout of the competitor DNA were listed above the autoradiograph.

Methods

Enzymatic assay using βGT assay and SssI for 5mC and hmC quantification.

The DNA fragment coding β- glucosyltransferase (βGT) was amplified by PCR using T4 phage genomic DNA as a template. βGT enzyme is expressed by E. coli expression system and purified by affinity chromatography. The βGT assay was performed essentially as described[26] with modifications. Briefly, 200 ng of genomic DNA was incubated with 10 pmol of βGT and 1.91 kBq of 3H-labeled UDP-glucose at 25 ˚C for a hour in the 25 μl reaction buffer: 50 mM potassium acetate, 20 mM Tris-acetate, 10 mM Magnesium Acetate and 1 mM Dithiothreitol. The reaction was stopped by incubation at 55 ˚C in the presence of 20 μg of proteinase K (Takara), 1% sodium dodecil sulphate and loaded onto the DE81 filter paper (GE healthcare) followed by liquid scintillation counting.

Similarly, 200 ng of genomic DNA was treated by 2U of SssI (New England Biolabs) and 5.3 μM S-[methyl-3H]adenosyl-L-methionine (15 Ci/mmol; Perkin Elmer) at 37 ˚C in the 10 μl reaction buffer: 50 mM NaCl, 10 mM Tris-HCl, 10 mM MgCl2, 1 mM Dithiothreitol. After 1 hour of incubation, the radioactivity was determined using a scintillation counter.

The C and hmC standard DNA fragments were prepared by PCR with dCTP and dhmCTP, respectively. The C standard fragment was further treated with SssI and non-labeled SAM to prepare the 5mC standard DNA fragment. The 5mC labeled ratio was estimated to be over 95%.

Cell culture and cell growth inhibitor treatments.

J1 murine embryonic stem cells were cultured in Dulbecco's modified Eagle's

medium supplemented with 0.1 mM 2-mercaptoethanol, 1,000 U/ml leukemia inhibitory factor (ESGRO, Gibco) and 20% KSR (Gibco). Dnmt knock out J1 cell lins are kind gift from Dr. Okano. Cells are treated with 5 μM of aphidicolin or 1 mM of hydroxyurea, to arrest at the S phase. Serum depleted medium, 20% of KSR is replaced by 1% of Fetal bovin serum (Gibco) is used to arrest the cells at the G0/G1 phase as described before[32]. 200 ng/ml of Nocodazole (Sigma) is used to arrest cells at the M phase.

RT-qPCR.

Total RNA was purified with Trizol (Invitrogen). Template cDNA synthesized using Superscript II reverse transcriptase (Invitrogen) with random hexamers was amplified with the specific primer sets for Tet1, Tet2, Tet3 and Gapdh. The mRNA transcripts of TET enzymes were quantified relative to that of Gapdh.

Kinetic assays of Dnmt enzymes.

Recombinant mouse Dnmt enzymes, Dnmt1 (291-1620), Dnmt3a and Dnmt3b were prepared as described before[33,34]. DNA methylation activities were determined essentially as described previously[34]. Briefly, the methylation reaction mixture contained 80 ng of Dnmt enzyme, 30 nM hemimethylated oligonucleotide substrates (35-bp containing a 5mCG/CG, hmCG/CG or CG/CG site; purchased from Genedesign. Inc.) and 5.3 μM S-[methyl-3H]adenosyl-L-methionine (15 Ci/mmol; Perkin Elmer) in 25 μl of reaction buffer comprising 5% glycerol, 0.5 mM EDTA, 0.2 mM dithiothreitol, 0.1 mg/ml BSA and 20 mM Tris/HCl (pH 7.4). After incubation at 37 °C, the radioactivity was determined using a scintillation counter.

DNA binding assays.

Recombinant protein of the SRA domain from mouse Uhrf1 was prepared as described befor[4] and the oligo nucleotide fragments were purchased from Genedesign. The 12-bp oligo nucleotides, 5'-CTACCGGATTGC-3' and 5'-GCAATCXGGTAG-3', where X is C, 5mC or hmC, was annealed to prepare 12-bp CG/CG, 5mCG/CG and hmCG/CG duplexes. To prepare 32P-labeled DNA for competitive DNA binding assay, the first strand was radioisotope-labeled at the 5′ end with T4 polynucleotide kinase (TOYOBO) and ^{32}P-γ-ATP (Muromachi Kagaku, Tokyo). The labeled strand was then mixed with 1.2-fold amount of the complementary strand and annealed.

Gel mobility shift and competitive binding assays were performed in the binding buffer; 25 mM HEPES-NaOH (pH 7.4), 100 mM NaCl and 0.1 mM TCEP. Samples containing 0, 0.3, 1 or 3 μM of the SRA protein, 1 μM of non-labeled DNA and 250 ng of nonspecific DNA competitor poly (dI-dC)(dI-dC) duplex (Sigma) were incubated for 30 min at 4 °C and then separated by a native gel electrophoresis followed by GelGreen staining (Biotium, Inc.). Samples containing 5 μM of the SRA protein, 1 μM of radioisotope-labeled 5mCG/CG, 0, 3, 10 μM of non-label competitor DNA fragment and 250 ng poly (dI-dC)(dI-dC) duplex were incubated for 30 min at 4 °C and then separated by a native gel electrophoresis and visualized with a Fuji BAS-2000 phosphor imager.

References

1. Suzuki, M.M. & Bird, A. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**, 465-76 (2008).
2. Sharif, J. et al. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature* **450**, 908-12 (2007).
3. Bostick, M. et al. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science* **317**, 1760-4 (2007).
4. Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y. & Shirakawa, M. Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. *Nature* **455**, 818-21 (2008).
5. Avvakumov, G.V. et al. Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. *Nature* **455**, 822-5 (2008).
6. Hashimoto, H. et al. The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. *Nature* **455**, 826-9 (2008).
7. Goll, M. & Bestor, T. Eukaryotic cytosine methyltransferases. *Annu Rev Biochem* **74**, 481-514 (2005).
8. Li, E., Bestor, T.H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915-26 (1992).
9. Okano, M., Bell, D., Haber, D. & Li, E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247-57 (1999).
10. Ooi, S.K. & Bestor, T.H. The colorful history of active DNA demethylation. *Cell* **133**, 1145-8 (2008).

11. Wu, S.C. & Zhang, Y. Active DNA demethylation: many roads lead to Rome. *Nat Rev Mol Cell Biol* **11**, 607-20 (2010).

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

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

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

15. Wossidlo, M. et al. 5-Hydroxymethylcytosine in the mammalian zygote is linked with epigenetic reprogramming. *Nat Commun* **2**, 241 (2011).

16. Gu, T.P. et al. The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. *Nature* **477**, 606-10 (2011).

17. Delhommeau, F. et al. Mutation in TET2 in myeloid cancers. *N Engl J Med* **360**, 2289-301 (2009).

18. Ko, M. et al. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **468**, 839-43 (2010).

19. Ficz, G. et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398-402 (2011).

20. Williams, K. et al. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343-8 (2011).

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

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

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

24. Guo, J.U., Su, Y., Zhong, C., Ming, G.L. & Song, H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell* **145**, 423-34 (2011).

25. Maiti, A. & Drohat, A.C. Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: Potential implications for active demethylation of CpG sites. *J Biol Chem* (2011).

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

27. Valinluck, V. & Sowers, L.C. Endogenous cytosine damage products alter the site selectivity of human DNA maintenance methyltransferase DNMT1. *Cancer Res* **67**, 946-50 (2007).

28. Pastor, W.A. et al. Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. *Nature* **473**, 394-7 (2011).

29. Frauer, C. et al. Recognition of 5-hydroxymethylcytosine by the Uhrf1 SRA domain. *PLoS One* **6**, e21306 (2011).

30. Inoue, A., Shen, L., Dai, Q., He, C. & Zhang, Y. Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. *Cell Res* **21**, 1670-6 (2011).

31. Inoue, A. & Zhang, Y. Replication-dependent loss of 5-hydroxymethylcytosine in mouse preimplantation embryos. *Science* **334**, 194 (2011).

32. Zhang, E., Li, X., Zhang, S., Chen, L. & Zheng, X. Cell cycle synchronization of embryonic stem cells: effect of serum deprivation on the differentiation of embryonic bodies in vitro. *Biochem Biophys Res Commun* **333**, 1171-7 (2005).

33. Takeshita, K. et al. Structural insight into maintenance methylation by mouse DNA methyltransferase 1 (Dnmt1). *Proc Natl Acad Sci U S A* **108**, 9055-9 (2011).

34. Suetake, I., Miyazaki, J., Murakami, C., Takeshima, H. & Tajima, S. Distinct enzymatic properties of recombinant mouse DNA methyltransferases Dnmt3a and Dnmt3b. *J Biochem* **133**, 737-44 (2003).

PART III

SUMMARY AND GENERAL CONCLUSION

Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain.

In chapter 1, I have determined the crystal structures of the ADD domain of DNMT3A in an unliganded form and in a complex with the N-terminal tail of histone H3. Combined with the results of biochemical analysis, the complex structure indicates that DNMT3A recognizes the unmethylated state of Lys4 in histone H3. This finding suggests that the recruitment of DNMT3A onto chromatin, and thereby *de novo* DNA methylation, is mediated by recognition of the histone modification state by its ADD domain. Furthermore, biochemical and NMR analyses demonstrate a mutually exclusive binding of the ADD domain of DNMT3A and the chromo domain of heterochromatin protein 1α to the H3 tail. These results imply that *de novo* DNA methylation by DNMT3A requires alteration of chromatin structure. Although not much attention was paid to the relationship between histone modification readers which recognize neighboring modifications on a histone peptide chain, these relationship should be important for unravelling the "histone code" composed of chemical modifications of histone proteins.

The structural basis of the versatile DNA recognition by the Methyl CpG binding domain of MBD4.

In chapter 2, I have determined the crystal structures of MBD domain of MBD4 bound to various sequences on double stranded DNA. The structural and biochemical analyses show that the MBD domain of MBD4 can bind to the fully methylated CpG sequence and its deamination product with similar affinity and provided the structural basis for the broad substrate specificity of the MBD domain of MBD4. The flexible DNA binding surface of MBD4 appeared to be critical for the versatile DNA

recognition. The cavity observed in the protein-DNA interface drove me to examine the binding between the MBD domain and the DNA fragment carrying the oxidatively modified cytosine bases and found that 5-hydroxymethylcytosine, 5-formylcytosine and 5-hydroxymethyluracil can be accommodated in the cavity without any steric clash. Although further study is needed to clarify the biological relevance of the interaction between MBD4 and the oxidatively modified cytosine bases, it is fascinating to speculate that the interaction is involved in the mechanism for DNA de-methylation.

The metabolism of 5-methyl- and 5-hydroxymethylcytosine bases.

In chapter 3, I have analyzed the steady state equilibrium of the production and the reduction of 5-hydroxymethylcytosine bases in mouse embryonic stem cells. I have used a series of Dnmt knock out cell lines and cell growth inhibitors to move the equilibrium, resulting in decreased and increased hmC content in the cells, respectively. Almost all the 5-hydroxymethylcytosine bases were lost when the *de novo* DNA methyltransferases were absent, implying the antagonistic relationship between the *de novo* DNA methylation and hydroxylation of 5-methylcytosine. The large increase in 5-hydroxymethylcytosine content of the cells treated with cell growth inhibitors suggests that the replication dependent dilution contributes significantly to reduce 5-hydroxymethylcytosine content in the equilibrium. Combined with the *in vitro* observations that 5-hydroxymethylcytosine bases are not recognized by Dnmt1 and Uhrf1, TET-dependent hydroxylation of 5-methylcytosine seems to contribute significantly to the passive DNA demethylation pathway in ES cells. It may be said that the hydroxylation of 5-methylcytosine is a machinery to protect genomes from *de novo* DNA methylation by inducing passive DNA demethylation.

Concluding remarks

Epigenetic marks such as DNA methylation and histone methylation alter protein-protein, protein-DNA or protein-RNA interactions, although the marks are usually small in size. In this thesis, I used X-ray crystallography, in addition to biochemical assays, as a tool to study the epigenetic regulation at atomic resolution. In part II chapter 1, I have shown that DNMT3A recognizes histone H3 N-terminal tail depending on the modification state of H3 Lys4. DNMT3A is a reader and a writer of epigenetic marks. This kind of the network between epigenetic marks appears to participate in the sophisticated biological system which is noise-resistant and flexible to environmental stimuli. In part II chapter 2, I have demonstrated the broad binding specificity of MBD4 by the structural study of the MBD domain of MBD4. The broad binding specificity appears to be another layer of complexity of biological system. The structural study led me to pay attention to the 5-hydroxymethylcytosine, newly discovered DNA modification in mammals. In part II chapter 3, I presented the study about the 5-hydroxymethylcytosine production. Although the whole picture cannot be held at this time, the hydroxylation of 5-methylcytosine may be the mechanism to reverse 5-methylcytosine and is becoming increasingly important in the field of DNA methylation.