Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

_____ _____

Samuel Hong Date

The recognition of DNA 5-methylcytosine: Studies on *Arabidopsis thaliana* DNA glycosylase ROS1 and basic leucine-zipper transcription factors in human and Epstein-Barr virus

By

Samuel Hong
Doctor of Philosophy

Graduate Division of Biological and Biomedical Science
Molecular and Systems Pharmacology

_____

Xiaodong Cheng, Ph.D.
Advisor

_____

Paul W. Doetsch, Ph.D.
Committee Member

_____

Eric A. Ortlund, Ph.D.
Committee Member

_____

Hyunsuk Shim, Ph.D.
Committee Member

Accepted:

_____

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

_____

Date

The recognition of DNA 5-methylcytosine: Studies on *Arabidopsis thaliana* DNA glycosylase ROS1 and basic leucine-zipper transcription factors in human and Epstein-Barr virus

By

Samuel Hong
B.S., Emory University, 2010

Advisor: Xiaodong Cheng, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Molecular and Systems Pharmacology
2016

Abstract

The recognition of DNA 5-methylcytosine: Studies on *Arabidopsis thaliana* DNA glycosylase ROS1 and basic leucine-zipper transcription factors in human and Epstein-Barr virus

By Samuel Hong

Eukaryotic DNA methylation, often a chemical modification of cytosine via methylation of the carbon-5, generates 5-methylcytosine (5mC) in genomes. This modified base serves as a critical epigenetic signal implicated in development, imprinting, immune responses, and various forms of diseases. Characterizing how DNA 5mC is recognized and regulated is critical to effectively understanding the function of DNA methylation. Previous investigations have shown that the base excision repair pathway can regulate active DNA demethylation—the enzyme-driven process of erasing and thus reversing the methyl modification signal. Particularly, Repressor of Silencing 1 (ROS1) and its paralogs in *Arabidopsis thaliana* can directly excise 5mC to reverse DNA methylation. A major portion of this dissertation describes the molecular mechanism of ROS1 activity. Specifically shown is the interaction between the C-terminal domain and the catalytic domain of ROS1, and the requirement of the C-terminal domain for the 5mC excision activity. This understanding expands the paradigm of DNA repair enzymes from their traditionally understood housekeeping roles to their extended roles in epigenetic regulations. In addition to the discoveries on how DNA 5mC is erased, understanding how proteins specifically recognize this modified base is also critical. It is widely generalized that 5mC is inhibitory for transcription factor binding. However, recent data show that certain transcription factors can preferentially recognize 5mC within specific sequences. As a major extension to this discovery, the other major portion of this dissertation describes the DNA sequence-specific recognition of methylated DNA by human AP-1 and Epstein-Barr virus AP-1-like transcription factors. The study provides the biochemical and structural basis of how DNA methylation can generate novel transcription factor binding sites to dynamically regulate transcription.

The recognition of DNA 5-methylcytosine: Studies on *Arabidopsis thaliana* DNA glycosylase ROS1 and basic leucine-zipper transcription factors in human and Epstein-Barr virus

By

Samuel Hong
B.S., Emory University, 2010

Advisor: Xiaodong Cheng, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Molecular and Systems Pharmacology
2016

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

12-O-Tetradecanoylphorbol-13-acetate (TPA)

12-O-Tetradecanoylphorbol-13-acetate response element (TRE)

5-carboxylcytosine (5caC)

5-formylcytosine (5fC)

5-hydroxymethylcytosine (5hmC)

5-hydroxymethyluracil (5hmU)

5-methylcytosine (5mC or M)

5-methylcytosine-phosphate-guanine dinucleotide (5mCpG)

5-methyluracil (5mU or T)

6-carboxy-fluorescein (FAM)

8-oxoguanine (8oxoG)

activation induced deaminase (AID)

Activator Protein 1 (AP-1)

adenine (Ade)

alpha-ketoglutarate (αKG)

amino acid carbon beta (Cβ)

AP endonuclease (APE)

apolipoprotein B mRNA-editing catalytc polypeptide (APOBEC)

apurinic/apyrimidinic (AP)

ATRX-DNMT3-DNMT3L (ADD)

base excision repair (BER)

basic helix-loop-helix (bHLH)

basic leucine-zipper (bZIP)

bovine serum albumin (BSA)

C-terminal domain (CTD)

cAMP response element (CRE)

cAMP response element-binding protein (CREB)

catalytic constant (K_{cat})

CCAAT-enhancer-binding protein (C/EBP)

CCAAT-enhancer-binding protein α (C/EBPα)

Chromomethylase (CMT)

CpG islands (CGI)

cytosine (Cyt)

cytosine carbon-5 (C5)

cytosine carbon-6 (C6)

cytosine-phosphate-guanine dinucleotide (CpG)

Demeter (DME)

deoxyribose-5-phosphate (dRP)

differentially methylated regions (DMR)

dissociation constant (K_D)

dithiothreitol (DTT)

DME-like 2 (DML2)

DME-like 3 (DML3)

DNA methyltransferase 1 (DNMT1)

DNA methyltransferase 3-like (DNMT3L)

DNA methyltransferase 3A (DNMT3A)

DNA methyltransferase 3B (DNMT3B)

domain rearranged methyltransferase (DRM)

embryonic stem (ES)

Epstein-Barr virus (EBV)

Escherichia coli (*E. coli*)

Flap endonuclease (Flap)

full-length (FL)

glutathione S-transferase (GST)

glycosylase domain (GD)

guanidine hydrochloride (Gua-HCl)

guanine (Gua)

Helix-hairpin-Helix (HhH)

hydrogen-deuterium exchange (HDX)

isopropyl β-D-1-thiogalactopyranoside (IPTG)

Kruppel-Like Factor 4 (Klf4)

methyl-CpG-binding domain (MBD)

methyl-CpG-binding domain protein 1 (MBD1)

methyl-CpG-binding domain protein 2 (MBD2)

methyl-CpG-binding domain protein 4 (MBD4)

methyl-CpG-binding protein 2 (MeCP2)

methylated TRE (meTRE)

methylated ZRE (meZRE)

Methyltransferase 1 (Met1)

Methyltransferase HhaI (M. HhaI)

MutY homolog (MYH)

oligonucleotides (oligo)

Polymerase β (Pol β)

polyunsaturated aldehyde (PUA)

Repressor of Silencing 1 (ROS1)

S-adenosyl-L-homocysteine (SAH)

S-adenosyl-L-methionine (SAM)

SET and RING finger associated (SRA)

signal transducer and activator of transcription 1 (STAT1)

tandem Tudor domain (TTD)

Ten-eleven translocation (Tet)

thymine (Thy)

thymine DNA glycosylase (TDG)

thymine-phosphate-guanine dinucleotide (TpG)

transcription activator-like effector (TALE)

Ubiquitin-like-specific protease 1 (ULP-1)

uracil DNA glycosylase (UDG or UNG)

Uracil glycosylase inhibitor protein (UGI)

Wilms Tumor 1 (WT1)

zinc finger (ZnF)

zinc finger and BTB domain containing 4 (ZBTB4)

zinc finger DNA 3'-phosphoesterase (ZDP)

zinc finger protein 57 (Zfp57)

ZRE (Zta response element)

LIST OF FIGURES

LIST OF TABLES

CHAPTER I.

General Introduction

DNA methylation

DNA modifications by enzymes have fundamental biological roles in many living organisms. In both prokaryotes and many eukaryotes, DNA cytosine can be methylated at the carbon-5 (C5) position by cytosine C5 methyltransferases that incorporate S-adenosyl-L-methionine (SAM) as a cofactor to generate 5-methylcytosine (5mC)[1, 2]. According to the reaction mechanism proposed by Wu and Santi[3, 4], the catalytic cysteine of a methyltransferase makes a nucleophilic attack on C6 of cytosine to form a covalent complex, followed by transferring of the methyl group from SAM to cytosine C5 (**Figure 1**). M.HhaI was the first DNA methyltransferase to be structurally characterized, and the crystal structure of M.HhaI-DNA-SAM ternary complex supported the proposed mechanism and demonstrated base flipping as a mode of accessing DNA base substrate[5]. Prokaryotic DNA methylation is often described in the context of bacteria-phage warfare, as extensively reflected in the restriction-modification systems[6]. In certain eukaryotes, however, DNA methylation is critically involved in transcriptional regulation of many biological processes.

Eukaryotic DNA methyltransferases are classified as maintenance methyltransferase or *de novo* methyltransferase[1, 7]. In mammals, the maintenance methyltransferase DNMT1 preferentially recognizes a hemi-methylated CpG dinucleotide over unmethylated DNA during DNA replication and methylates the daughter strand to maintain methylation patterns encoded in the mother strand[8-10]. UHRF1 is potentially engaged in the process of guiding DNMT1 to the hemi-methylated sites. The SRA domain of UHRF1 recognizes a hemi-methylated CpG[11-13] and is associated with guiding DNMT1 activities to the hemi-methylated

DNA[14, 15]. Aside from DNMT1, *de novo* DNA methyltransferases DNMT3A and DNMT3B can methylate both CpG and non-CpG sites[16-18]. Mammalian DNMT3A can directly associate with DNMT3L[19, 20], which contains an ADD domain that binds unmethylated lysine 4 of histone H3[21, 22]. In a similar way, UHRF1 has a TTD domain that recognizes tri-methylated lysine 9 of histone H3[23-25]. Thus, generation of DNA methylation is coordinated with relevant histone modifications in a larger chromatin context. In plants, Met1 acts as a maintenance methyltransferase, and other methyltransferases that belong to domain-rearranged methyltransferase (DRM) and chromo-methyltransferase (CMT) families act as *de novo* methyltransferases in both CpG and non-CpG contexts[26].

DNA methylome profiles in terms of distributions and patterns of 5mC within genomes provide a functional context of DNA methylation. Approximately 1% of a mammalian genome is methylated, primarily in CpG context[27, 28]. A plant genome can be methylated approximately 20-30% in both CpG and non-CpG context[29, 30]. Most transposons and repetitive regions in genomes are silenced by methylation[31, 32]. However, CpG-rich clusters of 500 to 2000 base pairs, known as CpG islands (CGI)[33], are found in gene promoter regions and remain largely unmethylated[34-36]. Approximately 50-70% of mammalian promoters contain CGI[35, 37]. Promoters with methylated CGI are associated with gene repression, while most promoters with unmethylated CGI are those of housekeeping genes with stable gene expression profiles[38, 39]. In contrast to promoter methylation in CGI, many transcriptionally active gene body regions are methylated with distinct enrichment patterns near exon-intron boundaries[30, 40, 41], indicating a potential role of DNA methylation for splicing. In both mammals and plants, germ cells undergo global genome-wide DNA demethylation[42-44]. Then, an embryo at the pluripotent stage contains the highest level of genome-wide methylation, including the methylation of CpA sites[40, 45]. Subsequent

differentiations are followed by a decreased amount of overall methylation and the establishment of differentially methylated regions (DMR) in both promoter regions and gene bodies with tissue-specific patterns[40, 46]. DMR patterns are thus associated with tissue-specific gene expression profiles. DMR can also be specifically established in paternal or maternal alleles as primarily shown in imprinting[47, 48]. In cancer cells, CGI methylation patterns can become aberrant such that tumor suppressor promoters are methylated, whereas proto-oncogene promoters are unmethylated[49, 50].

Proteins that specifically bind methylated CpG can mediate biological signals of DNA methylation. Certain proteins with the methyl-CpG-binding domain (MBD) are found in both mammals and plants and can preferentially bind a single, symmetrically methylated CpG compared to the unmethylated form[8, 51, 52]. Genetic evidence shows that MBD family proteins—such as MeCP2, MBD1, and MBD2—associate with repressive histone modifiers[53-55]. Particularly, MeCP2 is globally expressed in neurons and represses several genes as well as repetitive regions[56-60]. The lack of functional MeCP2 is linked to an intellectual disability known at Rett Syndrome[61]. Several genome-wide studies of MBD1 and MBD2 also show that they are involved in transcriptional regulation through gene repression[62-66]. Therefore, transcriptional inhibitory function of DNA methylation is partly mediated by readers of methylated DNA.

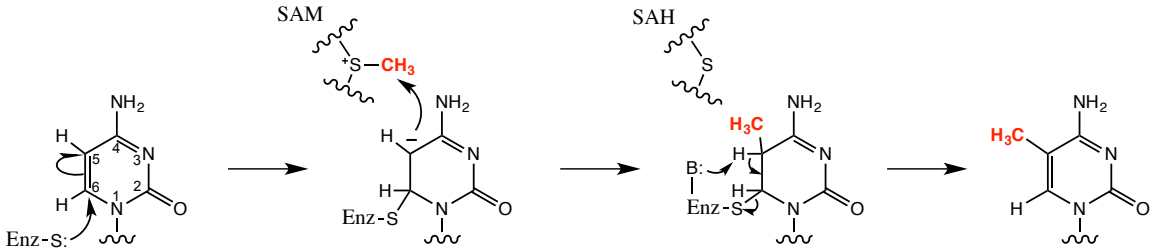


Figure 1. The reaction mechanism of cytosine C5 methylation (Adapted and modified from Wu and Santi 1987[3]).

SAM indicates S-adenosyl-L-methionine; SAH indicates S-adenosyl-L-homocysteine. The methyl group being transferred is shown in red.

Oxidative modifications of 5-methylcytosine and active DNA demethylation

In addition to 5mC, other chemical modifications of DNA are known. In prokaryotes, cytosine N4 and adenine N6 in genomes can also be methylated[67, 68], and DNA adenine N6 methylation is involved in bacterial host defense and gene regulation[69]. Also, bacteriophage have yet another form of DNA base known as 5-hydroxymethycytsine (5hmC), which is modified from 2'-deoxycytidine before its integration into the viral genome[70]. 5hmC in phage was initially discovered in 1953[71], but this particular base has garnered much attention recently due to the discovery of mammalian 5mC dioxygenase enzymes known as Ten-eleven translocation (Tet) proteins that oxidize 5mC to 5hmC by using α-ketoglutarate (αKG) and Fe(II) as cofactors[72, 73]. Subsequently, Tet dioxygenases were shown to further oxidize 5hmC to 5-formylcytosine (5fC) and then to 5-carboxylcytosine (5caC)[74, 75] (**Figure 2**). Genomic studies have revealed that 5hmC constitutes 5-10% of 5mC in mouse embryonic stem (ES) cells and approximately 40% of 5mC in mouse Purkinje neurons[72, 73]. 5hmC is more abundant in brain tissues compared to other tissues, and it can be enriched in promoters, gene bodies, and enhancer regions[72, 76, 77]. The level of 5fC and 5caC are substantially less than that of 5hmC—0.03% and 0.01% of 5mC respectively[75, 78]. While the function of modified bases generated by Tet activities is only beginning to be uncovered, each modified base may pose a different signal in cells. Particularly, a mass spectrometry study has revealed several proteins that specifically recognize 5mC as well as each of the oxidized bases in support of this idea[79].

Also, the discovery of Tet proteins has renewed interests in DNA demethylation pathways, as several mechanisms of DNA demethylation had been proposed[80]. The simplest mechanism that does not involve an enzyme would be a passive diffusion of 5mC during several rounds of DNA replication during which DNMT1 does not maintain the

methylation pattern[81, 82]. On the other hand, active DNA demethylation requires an enzyme-mediated activity without the need for DNA replication. There are records of activities whereby DNMT3A and DNMT3B directly remove the C5-methyl group of 5mC and/or C5-hydroxymethyl group of 5hmC[83, 84], though no *in vivo* data have yet to validate the activities. Also, it has been proposed that 5mC can be deaminated to thymine, which would be mismatched to guanine (G:T mismatch). The base excision repair (BER) pathway involving DNA glycosylases would then initiate a mismatch repair. In zebra fish, AID/APOBEC deaminases can generate thyimine and 5-hydroxymethyluracil (5hmU) mismatched to G by deamination of 5mC and 5hmC, after which monofunctional DNA glycosylases such as MBD4 and thymine DNA glycosylase (TDG) can excise the mismatched pyrimidine by hydrolyzing the glycosidic bond[85-88] (**Figure 3a**). The resulting apyrimidinic (AP) site would be subjected to downstream repair pathways, eventually involving DNA polymerase β and ligase activities to complete the repair processes (**Figure 4**). The Tet activities are also implicated in active DNA demethylation through BER, as TDG was discovered to excise 5fC and 5caC[74, 89, 90]. Indeed, a depletion of TDG in mouse ES cells was accompanied by increased levels of 5fC and 5caC[91, 92]. However, the increased amounts were still substantially low to adequately account for the full level of genome-wide demethylation observed in the mouse ES cells.

In addition, a direct removal of the modified base by 5mC DNA glycosylases has been proposed, and mammalian 5mC DNA glycosylase activities have previously been reported[93-95]. 5hmC DNA glycosylase activities have also been observed[96]. However, an enzyme responsible for any of such activity has not been identified. In *Arabidopsis thaliana*, on the other hand, *bone fide* 5mC DNA glycosylases have been clearly identified: ROS1, DME, DML2, and DML3[97-99]. They have a catalytic glycosylase domain homologous to *E. coli*

endonuclease III (Nth), a Helix-hairpin-Helix (HhH) fold DNA glycosylase/lyase known to contain an iron-sulfur cluster-binding site and excise damaged pyrimidines. Studies have shown that *Arabidopsis thaliana* 5mC DNA glycosylases are bifunctional glycosylase/lyase enzymes that both excise the base and cleaves the phosphate backbone via β-elimination or β,δ-elimination reaction[100-102] (**Figure 3b & Figure 4**). ROS1 shows overlapping substrate specificities partly shared by endonuclease III family enzymes[102]. After a base excision, the resulting single nucleotide gap with 3'- and 5'-phosphate termini after the elimination reaction is tailored by ZDP 3'-phosphatase to generate 3'-OH to initiate the downstream Pol β and Ligase activities to complete repair[103] (**Figure 4**).

Interestingly, ROS1 has been shown to excise 5mC and 5hmC but not 5fC and 5caC *in vitro*[85, 104, 105]. Thus, plant ROS1 and mammalian TDG have mutually exclusive substrate specificities for 5mC, 5hmC, 5fC, and 5caC: the first two specific for ROS1 and the latter two specific for mammalian TDG[85] (**Figure 2**). Particular residues within the catalytic glycosylase domain (GD) are involved in the specific recognition of substrate. In TDG, a single point mutation can alter the substrate specificity profile of the enzyme such that TDG becomes specific for 5caC in exclusion of other known substrates[106]. A catalytic mutation within ROS1 GD can abolish the glycosylase activity without abolishing the lyase activity[104]. However, an important observation in regards to the mechanism of 5mC and 5hmC excisions by ROS1 is the requirement of the enzyme's C-terminal domain (CTD) for the activity. ROS1 CTD is conserved among *Arabidopsis thaliana* 5mC DNA glycosylases. Studies on ROS1 CTD are covered in Chapter II.

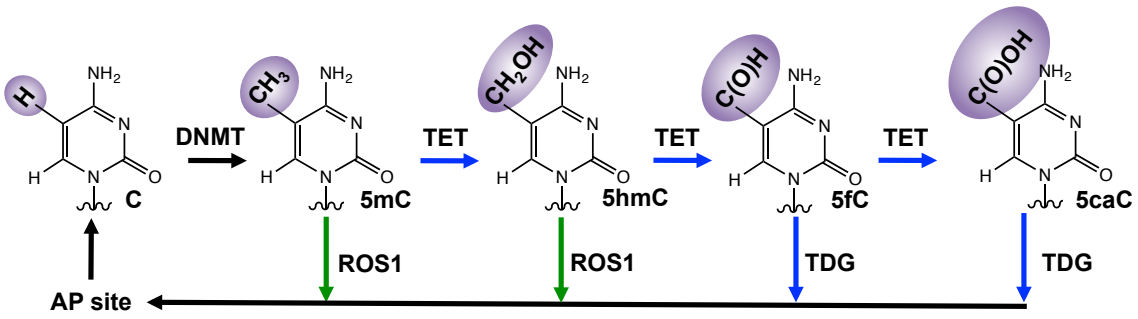


Figure 2. Generation and erasure of cytosine modifications.

C5 modifications are shown in purple; black arrow indicates both mammalian and plant systems; blue arrow indicates mammalian systems only; and green arrow indicates plant systems only. AP site indicates apyrimidinic site.

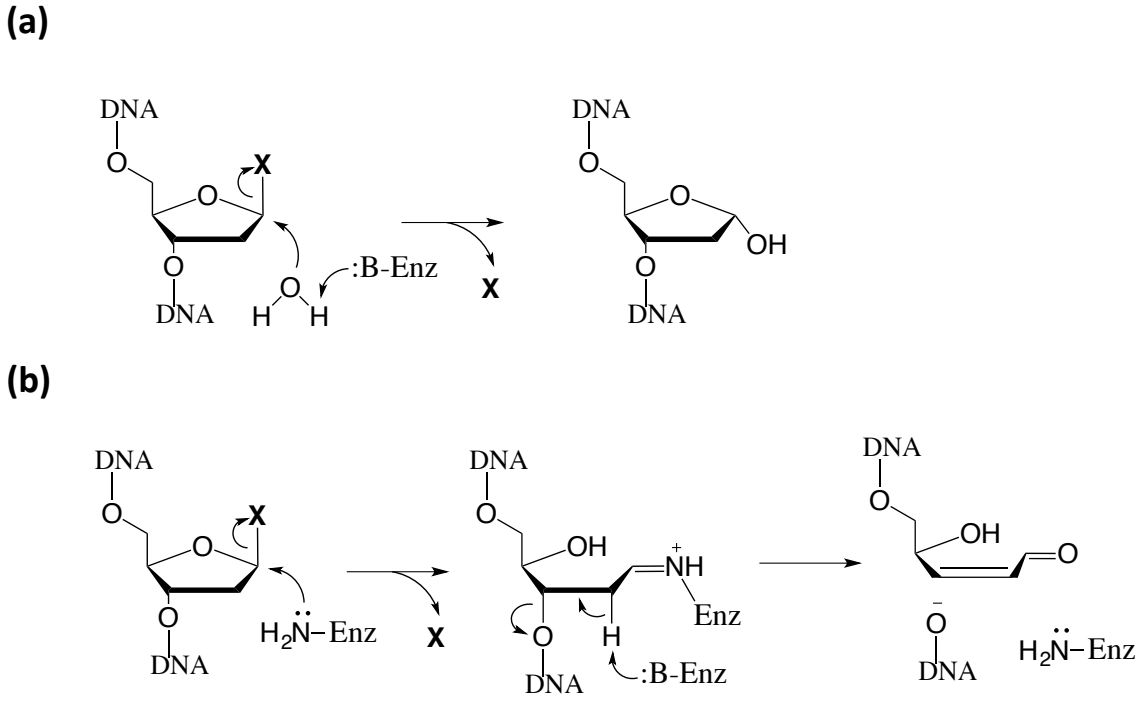


Figure 3. Reaction mechanisms of DNA glycosylases for generating AP sites (Adapted from Brooks 2013[107]).

(a) The reaction mechanism of monofunctional DNA glycosylases involving a hydrolysis of the glycosydic bond, leaving the AP site product. X indicates the substrate base. (b) The reaction mechanism of bifunctional DNA glycosylases involving a nucleophilic substitution of the substrate by a lysine side-chain, forming a Schiff base of a transient enzyme-DNA covalent complex. The following lyase reaction by β-elimination cleaves the C3'-phosphate bond, and the enzyme is released. The resulting AP site contains polyunsaturated aldehyde (PUA).

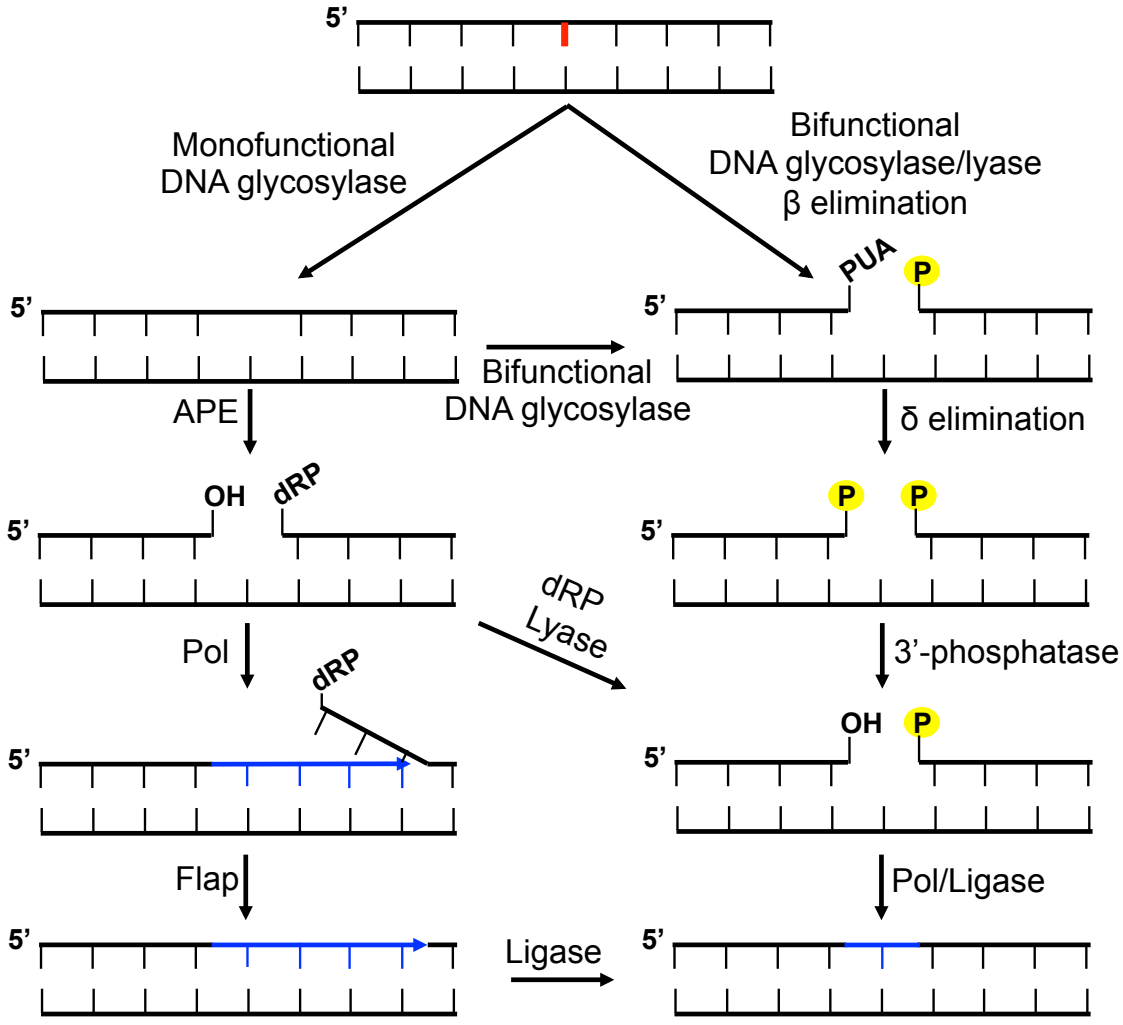


Figure 4. Overview of the base excision repair pathway.

The red line indicates the substrate base, and blue lines indicate newly incorporated nucleotides. APE: AP endonuclease; PUA: polyunsaturated aldehyde; dRP: deoxyribse-5-phosphate; Pol: Polymerase β; Flap: Flap endonuclease.

Sequence-specific recognition of 5-methylcytosine by transcription factors

The function of DNA methylation is perhaps most clearly elucidated by understanding how methylated DNA is specifically recognized during a given biological process. Studies on the recognition of 5mCpG by MBD proteins that associate with repressive chromatin modifiers have linked the function of DNA 5mC to gene silencing cascades as discussed previously. Studies have also shown that DNA methylation inhibits certain transcription factors from binding their response elements when the methylation occurs at the binding site[108]. Some of the transcription factors whose DNA binding is inhibited by DNA methylation within the cognate response element include Myc, STAT1, and CREB[109-111]. More recently, however, transcription factors with enhanced DNA binding capabilities upon CpG methylation within response elements have been discovered. Some zinc-finger (ZnF) family transcription factors, including Zfp57, ZBTB4, and Kaiso, have shown significant increases in DNA binding upon methylation within their binding sites[112-114]. In addition, two studies that utilized mass spectrometry and protein microarray have shown more than a dozen transcription factor candidates that may bind methylated DNA in a sequence-specific manner[79, 115].

Structural studies of ZnF 5mCpG-readers have shown that a single 5mCpG within the sequences are recognized by a conserved arginine involving a 5mC-Arg-G triad, which also recognizes TpG in an equivalent manner through its non-polar interaction with the C5-methyl group[116] (**Figure 5**). In the case of Zfp57, a glutamate, in addition to the arginine, is further involved in the recognition of the C5-methyl group of the same 5mC, and an ordered water network surrounds the symmetric 5mC in the opposite strand to further contribute to DNA binding[112]. Such arginine- and water-mediated recognition of 5mCpG have been observed in the crystal structure of MeCP2 in complex with methylated DNA[117], indicating a

common mode of 5mCpG recognition by distinct classes of proteins. Also, the function of 5mCpG-recognizing ZnF transcription factors has been associated with gene repression[118-121], suggesting that transcriptional inhibitory output of DNA methylation can be directed to specific sequences. Such a mode of repression would involve a more targeted gene inhibition than the inhibitory output by MBD proteins whose sequence specificity is confined to a single CpG dinucleotide.

In addition to ZnF transcription factors, some basic leucine-zipper (bZIP) family transcription factors have been shown to preferentially bind 5mCpG within specific sequences. The bZIP family is comprised of a large number transcription factors that function as homo- and/or –heterodimers that are known to bind several types of 7-bp to 14-bp consensus sequences containing inverted repeats of two identical half-sites, each bound by a monomer[122]. The consensus sequences can be categorized into three types, depending on the core 7- or 8-bp sequence. The first group contains a semi-palindromic sequence, 5'-TGAGTCA-3' (the middle base can be either G or C), also known as 12-O-Tetradecanoylphorbol-13-acetate (TPA) response element (TRE). AP-1 transcription factors such as Jun/Fos heterodimer and Jun/Jun homodimer are known to bind TRE. The second group contains 8-bp palindromic, TRE-like core sequence, 5'-TGACGTCA-3', known as cAMP response element (CRE) that contains a CpG in the middle of the sequence. CREB proteins are primarily known to bind CRE. The third group contains a distinct 8-bp palindromic sequence, 5'-TTGCGCAA-3', known as C/EBP consensus sequence that also contains a CpG and are primarily bound by C/EBP family proteins.

A study has shown that methylation of the CpG within CRE reduces CREB binding but enhances C/EBPσ binding to CRE and that methylated CRE-binding by C/EBPσ is associated with expression of tissue-specific genes in adipocytes[123]. The crystal structure of

C/EBPσ homodimer in complex with DNA containing C/EBP consensus sequences shows that a universally conserved arginine among bZIP family proteins is involved the recognition of the center CpG. The conformation of the arginine over the CpG is similar to the conformation seen in the 5mC-Arg-G triad[116, 124]. The structure of C/EBPσ in complex with DNA containing methylated CRE is not available. Yet, it is plausible that the observed conformation of the arginine of C/EBPσ over the CpG within the C/EBP consensus sequence may be conducive to recognize 5mCpG within the CRE sequence via a putative 5mC-Arg-G triad.

In two other studies, human Jun/Fos heterodimer and Jun/Jun homodimer, which binds TRE, were shown to preferentially bind 5mCpG in a TRE-like sequence, 5'-MGAGTCA-2' (where M is 5mC), termed meAP-1 (or meTRE)[125, 126]. A sequence comparison of TRE and meTRE shows that one thymine in TRE is switched to 5mC in meTRE, suggesting that this T-to-5mC switch is compatible for the protein-DNA interaction. Both thymine and 5mC are pyrimidines containing the C5-methyl group. Also, an AP-1-like Epstein-Barr virus (EBV) transcription factor Zta/Zta homodimer is known to bind TRE and other TRE-like methylated Zta response elements (meZREs) such as meZRE-2 (5'-TGAGMGA-3') in CpG methylation-dependent manner[127-129]. Unlike 5mCpG-binding ZnF transcription factors that repress genes, methylated DNA binding events involving human AP-1 and EBV Zta are associated with transcriptional activations[125, 127]. Therefore, the function of DNA methylation may include selective transcriptional activations. The recognition of 5mC by human AP-1 and EBV Zta in their methylated consensus sequences involves mechanisms distinguishable from the 5mC-Arg-G triad mechanism. The structural and biochemical studies of Jun/Jun and Zta/Zta for their recognition of methylated DNA are covered in Chapter III.

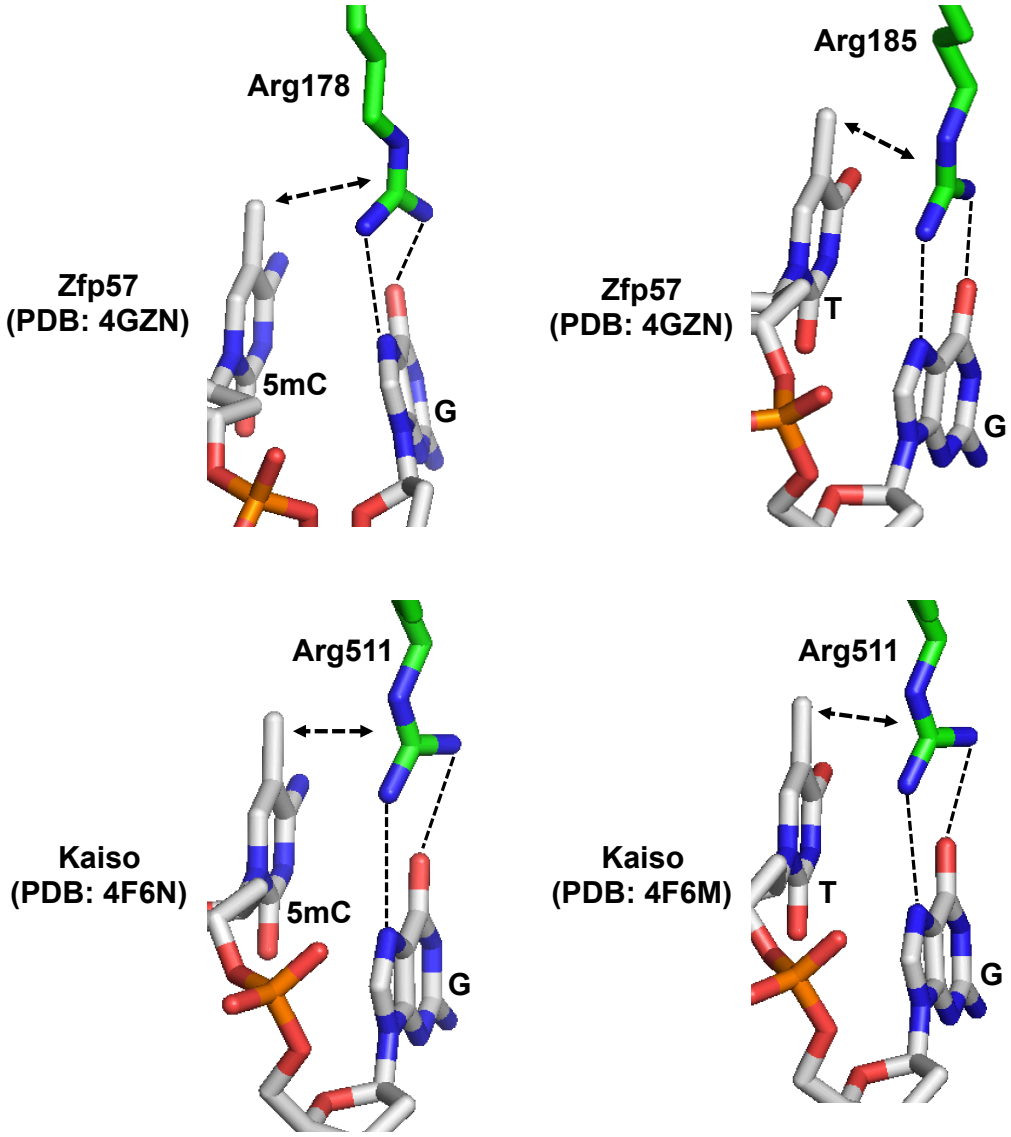


Figure 5. The recognition of 5mCpG and TpG by ZnF family Zfp57 and Kaiso.

The C5-methyl groups of 5mC and T are recognized by the arginine involving a non-polar interaction, while the polar ends (Arg-N^{η} atoms) are involved in the bifurcated recognition of 3'-Gua O6 and N7 atoms. The bidirectional arrow indicates a non-polar interaction.

CHAPTER II.

The carboxyl-terminal domain of ROS1 is essential for 5-methylcytosine DNA glycosylase activity*

Abstract

Arabidopsis thaliana Repressor of Silencing 1 (ROS1) is a multi-domain bifunctional DNA glycosylase/lyase, which excises 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC) as well as thymine and 5-hydroxymethyluracil (i.e., the deamination products of 5mC and 5hmC) when paired with a guanine, leaving an apyrimidinic (AP) site that is subsequently incised by the lyase activity. ROS1 is slow in base excision and fast in AP lyase activity, indicating that the recognition of pyrimidine modifications might be a rate-limiting step. In the C-terminal half, the enzyme harbors a helix-hairpin-helix DNA glycosylase domain followed by a unique C-terminal domain. We show that the isolated glycosylase domain is inactive for base excision but retains partial AP lyase activity. Addition of the C-terminal domain restores the base excision activity and increases the AP lyase activity as well. Furthermore, the two domains remain tightly associated and can be co-purified by chromatography. We suggest that the C-terminal domain of ROS1 is indispensable for the 5mC DNA glycosylase activity of ROS1.

** This chapter is adopted and modified from the following manuscript:*

Hong S, Hashimoto H, Kow YW, Zhang X, Cheng X. The carboxy-terminal domain of ROS1 is essential for 5-methylcytosine DNA glycosylase activity. J Mol Biol. 2014 Nov 11; 426 (22):3703-12.

** Author Contributions:*

S.H. performed all experiments, H.H. provided purified TDG and MBD4 enzymes, Y.W.K. assisted in data analysis, X.Z. and X.C. organized and designed the scope of the study. All authors were involved in analyzing data and preparing the manuscript.

Introduction

In eukaryotic genomes, DNA methyltransferases convert a proportion of cytosine into 5-methylcytosine (5mC)[1]. Mammalian ten-eleven-translocation (Tet) dioxygenases then convert a fraction of these to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) in consecutive oxidation reactions[73-75]. Mammalian TDG, named after thymine DNA glycosylase, excises the mismatched base from G:X mismatches, where X is uracil, thymine or 5-hydroxymethyluracil (5hmU). These are, respectively, the deamination products of cytosine, 5mC and 5hmC. In addition, TDG excises the Tet enzyme products 5fC and 5caC but not 5mC and 5hmC, when paired with a guanine[74, 85, 89, 90]. The resulting apurinic/apyrimidinic (AP) site is enzymatically converted to normal cytosine through the base excision repair pathway, altering DNA methylation patterns utilized for epigenetic controls. Mammalian DNA glycosylases that excise 5mC or 5hmC have not been identified but such activities have been reported[93-96].

In *Arabidopsis thaliana*, a family of 5mC DNA glycosylases has been identified: Repressor of Silencing 1 (ROS1)[99], Demeter (DME)[98], DME-like 2 (DML2) and DME-like 3 (DML3)[97]. ROS1 is a 1393-residue, multi-domain protein: the N-terminal domain containing a lysine-rich stretch involved in non-specific DNA binding and sliding along DNA[130, 131], followed by the central Helix-hairpin-Helix (HhH) DNA glycosylase domain containing an iron-sulfur (4Fe-4S) cluster[99], and a unique uncharacterized domain at the C-terminus. The central glycosylase domain (GD) has an atypical insertion of ~230 residues—whose sequence and length vary among the ROS1 family members—that is not found in other characterized HhH DNA glycosylases[100]. Like ROS1, mammalian Tet proteins have an atypical insertion into their catalytic domains, and the insertion is not required for the *in vitro* catalytic activity [132]. ROS1, and its family members, is a bifunctional DNA glycosylase/lyase

whose glycosylase activity excises a 5mC base from the DNA backbone and then its lyase activity cleaves the DNA backbone at the AP site[102, 133, 134].

The amino acids sequences within the C-terminal domain (CTD) are conserved among the ROS1 family members, but no homologous sequence has been found in other phyla. Introduction of random point mutations or deletions in the corresponding domain in DME resulted in abrogation of the 5mC excision activity[101]. Here we show that the isolated glycosylase domain of ROS1 does not possess the 5mC excision activity but partially retains the AP lyase activity. Addition of the CTD restores the 5mC excision activity. The two domains remain tightly associated and can be co-purified by chromatography.

Results

ROS1 glycosylase domain and the C-terminal domain

First, we constructed a deletion variant of ROS1, deleting the N-terminal 509 residues and replacing the internal insertion (residues 628-855) with a 5-residue linker, which we refer to as ROS1ΔN (**Figure 6a**). We measured the base excision and the AP lyase activities of the purified ROS1 full-length (FL), ROS1ΔN and its catalytic mutant D971N, using various 32-base pair (bp) DNA oligonucleotides (oligos), each containing a single variable base opposite a guanine (G:X pair), where X is C, 5mC, and 5hmC. These substrates bear the "natural" base pairs. Both FL and ΔN deletion excised 5mC and 5hmC but not C (**Figure 6b**). We further tested time course activities using the oligos with G:X, where X is C, 5mC, 5hmC, 5fC and 5caC as well as C, T and 5hmU that are deamination products of C, 5mC, and 5hmC respectively. 5hmC excision was weaker (by a factor of ~1.6) than 5mC excision for both ROS1 and ROS1ΔN, and no detectable activities were observed for 5fC and 5caC (**Figure 7a**). The *in vitro* excision activity on 5hmC has recently been reported for ROS1 and its family members (k_{cat} = 0.3-1 h^{-1} under single turnover conditions)[105, 135]. However, the significance of this activity is unclear, because no homologs of Tet dioxygenases have been identified in *Arabidopsis thaliana* and data on the existence of 5hmC in *Arabidopsis thaliana* are conflicting: one study detected no 5hmC[105], whereas another study found low levels of 5hmC in the DNA of leaves and flowers[136]. In addition to the base-paired substrates, ROS1ΔN is also active on G:T and G:5hmU mismatches, but no activity was observed on G:U mismatch (**Figure 7b**). The activity on G:T mismatch is comparable with that on G:5mC. This observation indicates that ROS1 is sensitive to pyrimidine modifications at the C5 position.

In the structurally characterized HhH DNA glycosylases, a conserved aspartate, Asp138 of *E. coli* endonuclease III[137], Asp138 of *E. coli* MutY[138], Asp238 of *E. coli* AlkA[139], Asp268 of human OGG1[140], and Asp534 of mouse MBD4[86, 141], has been suggested to activate a catalytic nucleophile (such as a water molecule or a nearby lysine residue) for the attack on the deoxyribose C1' carbon atom of the target nucleotide. The equivalent residue in ROS1 is Asp971[100], and the mutation of Asp971 to asparagine (D971N) abolished the base excision activity but not the AP lyase activity (**Figure 8**). One interesting observation is that the AP lyase activity of ROS1 is substantially faster than the base excision activity. Both ROS1 FL and ROS1ΔN showed ~90% cleavage of AP sites in 15 min compared to ~80% excision of 5mC over 20 h under the same conditions. ROS1 is known for slow turnover kinetics[102], and our observation of the fast AP lyase activity of ROS1 suggests that an initial stage of 5mC excision reaction, or probably the recognition of pyrimidine modifications, is a rate-limiting step.

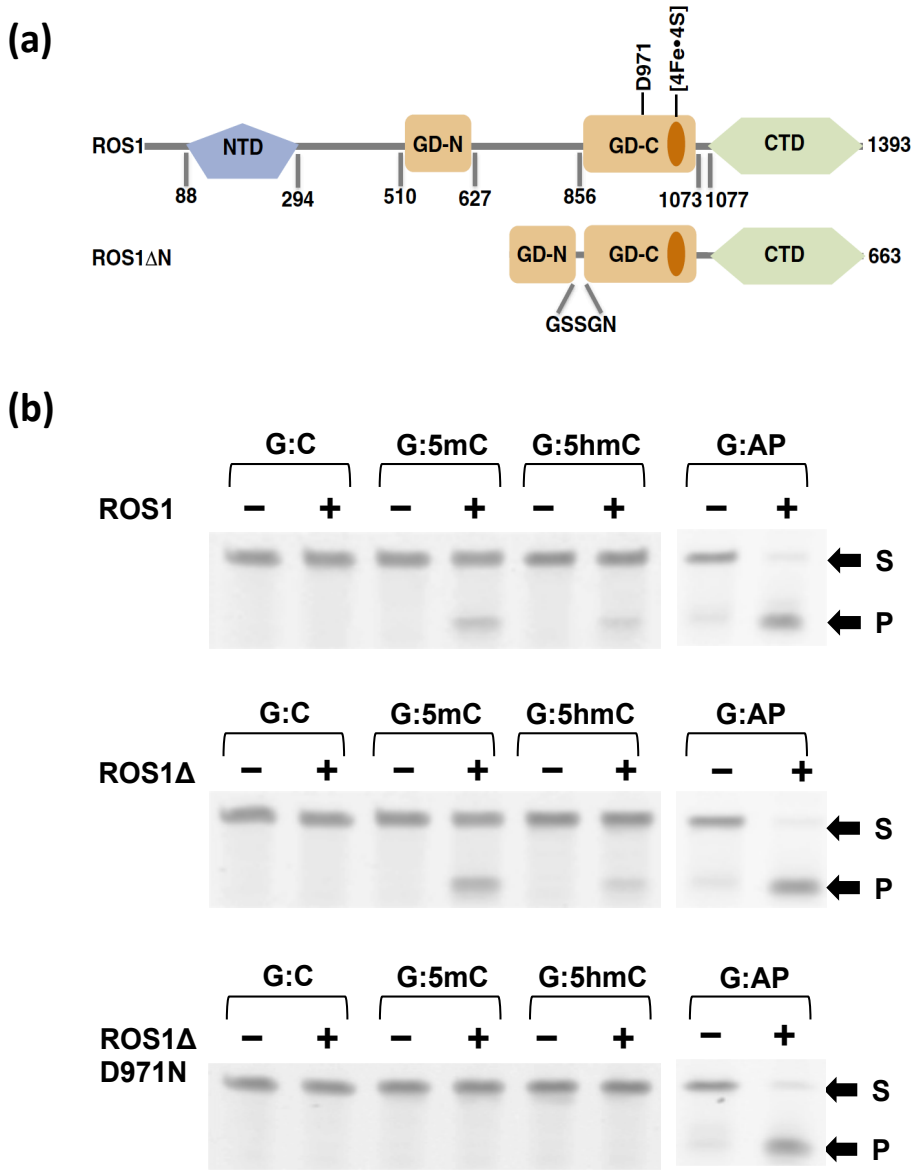


Figure 6. ROS1 glycosylase domain (GD) and the C-terminal domain (CTD).

(a) Domain organizations of ROS1 full-length (FL) and ROS1ΔN. (b) Activities of ROS1 FL (top panel), ROS1ΔN (middle panel), and ROS1ΔN D971N (bottom panel) on 32-bp oligos for indicated time under the single-turnover condition ($[S_{DNA}]$=50 nM and $[E_{FL}]$=100 nM or $[E_{\Delta N}]$=100 nM or $[E_{D971N}]$=500 nM). Labels S is for substrate and P is for product.

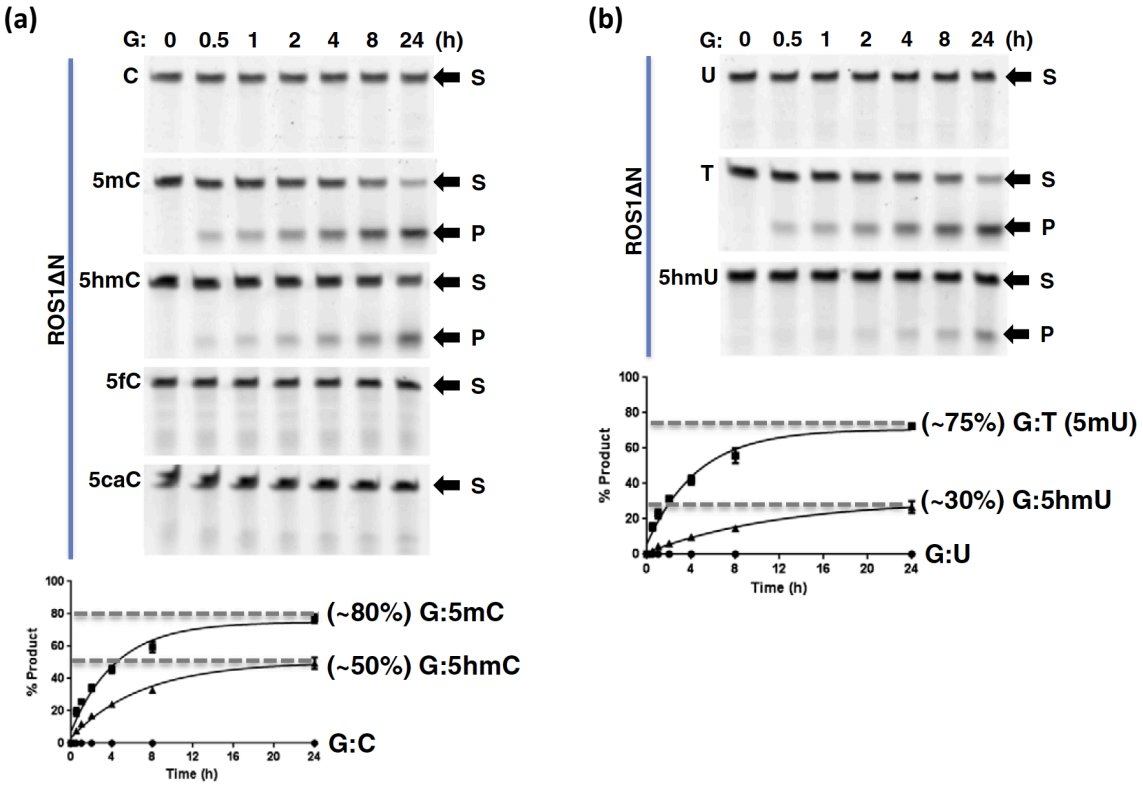


Figure 7. DNA glycosylase/lyase activities of ROS1ΔN.

(a) The time course (0–24 h) of ROS1ΔN reactions ($[E_{\Delta N}]$=500 nM) on five oligos ($[S_{DNA}]$=50 nM) with various modifications under the single-turnover condition. Data (± error bars) were averaged from three independent experiments (n=3). (b) The time course (0–24 h) of ROS1ΔN reactions ($[E_{\Delta N}]$=500 nM) on three oligos with G:X mismatches ($[S_{DNA}]$=50 nM). Data (± error bars) were averaged from three independent experiments (n=3).

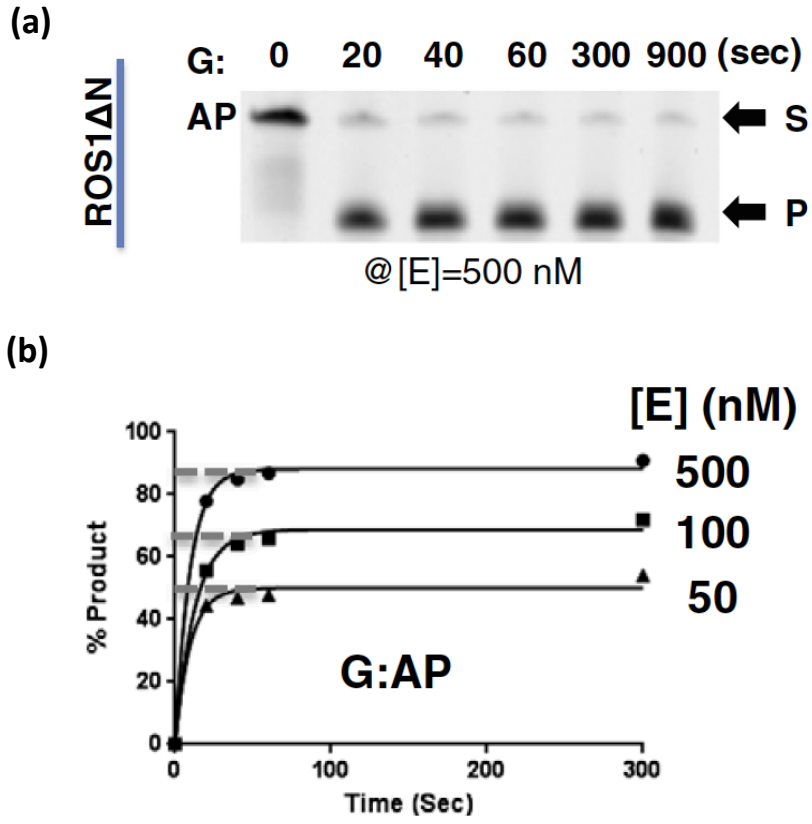


Figure 8. AP lyase activity of ROS1ΔN.

(a) The time course (0–15 min) of ROS1ΔN AP lyase reactions ([S_{DNA}]=50 nM). (b) The time-course of AP lyase reactions under three enzyme concentrations on oligo ([S_{DNA}]=50 nM).

ROS1 glycosylase domain and the C-terminal domain associate tightly

Most structurally characterized HhH DNA glycosylases, like endonuclease III[137], hOGG1[140], AlkA[139], and MBD4[86, 141], exist as or have an isolated glycosylase domain active on its own *in vitro*. We asked whether ROS1 glycosylase domain (GD) could function on its own and thus purified the isolated glycosylase domain (GD) and the C-terminal domain (CTD) individually. We note that the isolated domains, particularly CTD, were somewhat problematic during expression and/or purification with low yield, more impurity, and tendency to aggregate (see Materials and Methods). Nevertheless, GD is inactive on 5mC and 5hmC excisions while retaining residual AP lyase activity (**Figure 9a lane 2** & **Figure 9b**), whereas CTD alone did not show any activity (**Figure 9a lane 3**). Addition of the CTD (with estimated 3:1 molar ratio of CTD:GD) restored partial activity of base excision on 5mC and 5hmC and increased the AP lyase activity as well (**Figure 9a lane 4**).

We reasoned that, in order to restore the base excision activity of ROS1, the C-terminal domain must interact with the glycosylase domain, either directly or through DNA. To test this notion and to overcome the problems of the isolated GD and CTD, we engineered a new construct, termed as ROS1ΔN:P, in which the PreScission protease recognition sequence (LEVLFQGP) was inserted in the linker between GD and CTD (**Figure 10a**). The 8-residue insertion did not affect the 5mC and 5hmC excision activities (**Figure 10b lanes 2 & 3**) and AP lyase activities (**Figure 9b**). Approximately the same base excision and AP lyase activities were observed with and without the protease cleavage (**Figure 10b lanes 3 & 4**). Analytical size-exclusion chromatography measurements revealed that the two cleaved fragments of ROS1 associated together in presence of 500 mM NaCl (**Figure 10c-e**), suggesting that the interactions between the two domains are hydrophobic, a plausible reason that the isolated domains tend to aggregate in aqueous solution. Introducing

guanidine hydrochloride (0-2 M) showed a delayed peak for the cleaved fragments compared to the uncleaved form under the same conditions, indicating the dissociation of the two domains under denaturing conditions (**Figure 11**).

In a previous report by Mok and his colleagues[101], dozens of randomly generated point mutations mapped to the CTD of DME—a close paralog of ROS1—were shown to abolish 5mC DNA glycosylase activity by DME. Based on the identity of mutations in DME CTD, we likewise designed comparable mutations in the CTD of ROS1 in the background of ROS1ΔN:P— I1233M, W1234R, R1287Q, and D1309N. We used the following four principles to design the mutants: (1) the amino acid conserved among all four 5mC DNA glycosylases within *Arabidopsis thaliana* was given a priority for mutagenesis; (2) the amino acid that would likely cause mis-folding such as proline and glycine was not considered; (3) mutations that significantly alter chemical properties such exchanging hydrophobic residue with hydrophilic residue (e.g. valine to aspartate) or mutations that drastically alter the size of the side-chain (e.g. arginine to serine) were avoided; (4) mutations that resulted in a partial loss of the activity were considered. Out of the four mutants designed, only I1233M and R1287Q were purified comparably to the wild-type (WT). The mutants were subjected to DNA glycosylase and AP lyase activities as done previously, except that the AP lyase activity was performed in ~0 ºC. Compared to WT, I1233M did not show a significant change in the activities, whereas R1287Q showed significantly reduced overall base excision activities (**Figure 12a**) and the AP lyase activity (~4-fold reduction) (**Figure 12b**). Both I1233M and R1287Q were also subjected to size-exclusion chromatography with and without the protease cleavage. For both mutants, the peak elution volume before and after the cleavage remained the same as comparable to WT (**Figure 12c,d**), indicating that the mutations did not significantly affect the association between GD and CTD.

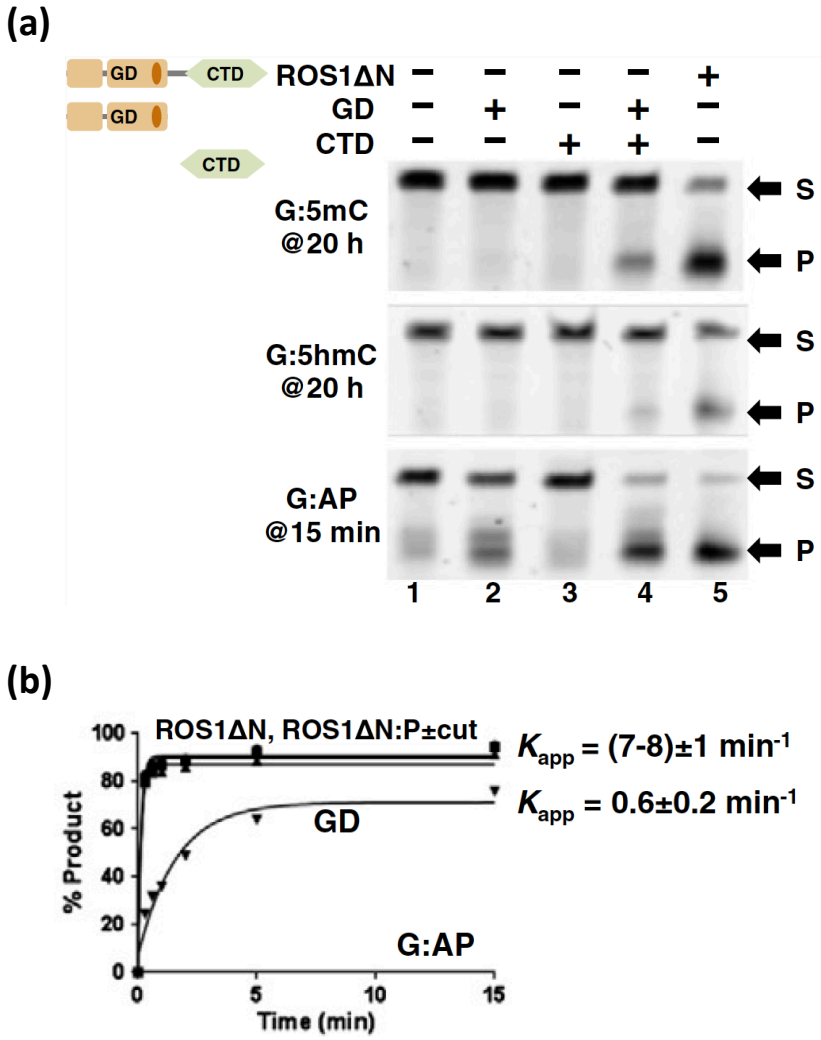


Figure 9. Effects of the C-terminal domain (CTD) on ROS1 glycosylase domain (GD) activity.

(a) Activities of ROS1ΔN ($[E_{\Delta N}]$=0.5 μM), the glycosylase domain ($[E_{GD}]$≈0.5 μM), and the C-terminal domain ($[E_{CTD}]$≈1.5 μM) on 32-bp oligos ($[S_{DNA}]$ = 50 nM) at 20 h (G:5mC and G:5hmC) or 15 min (G:AP) reactions. (b) The time course of AP lyase activities of ROS1ΔN, ROS1ΔN:P (with and without the protease cleavage) and GD.

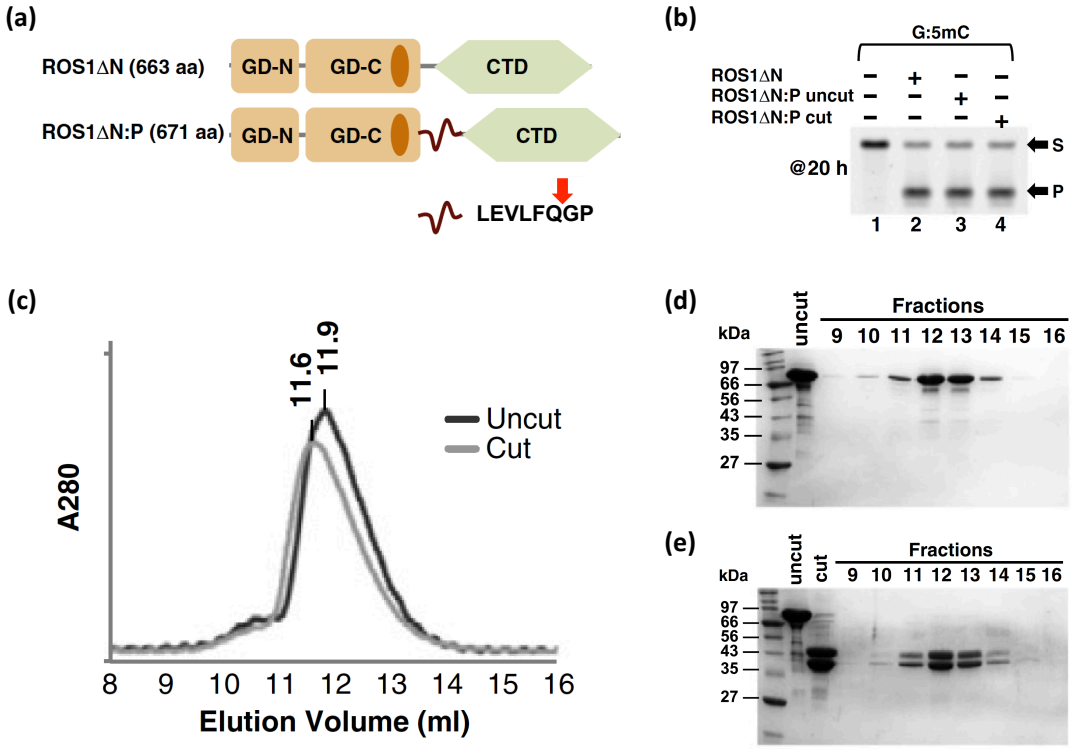


Figure 10. ROS1 glycosylase domain and the C-terminal domain associate tightly.

(a) The Prescission protease recognition sequence (LEVLFQGP) was inserted between ROS1 GD and CTD. (b) Activities of ROS1ΔN and ROS1ΔN:P before and after the protease cleavage, on 32-bp oligos (G:5mC) at 20 h reaction in room temperature ([E]=500 nM and [S_{DNA}]=50 nM). (c) Elution profiles of ROS1ΔN:P in two consecutive runs on a Superdex 200 (10/300 GL) column (GE Healthcare) before and after the protease cleavage, in 20 mM Tris-HCl (pH 8.0), 5% glycerol, 1 mM dithiothreitol, and 500 mM NaCl. Peak heights reflected relative OD280 absorbance and the retention volume shown as fractions. (d and e) SDS-PAGE (15%) analyses of S200 fractions containing ROS1ΔN:P, before and after the protease cleavage.

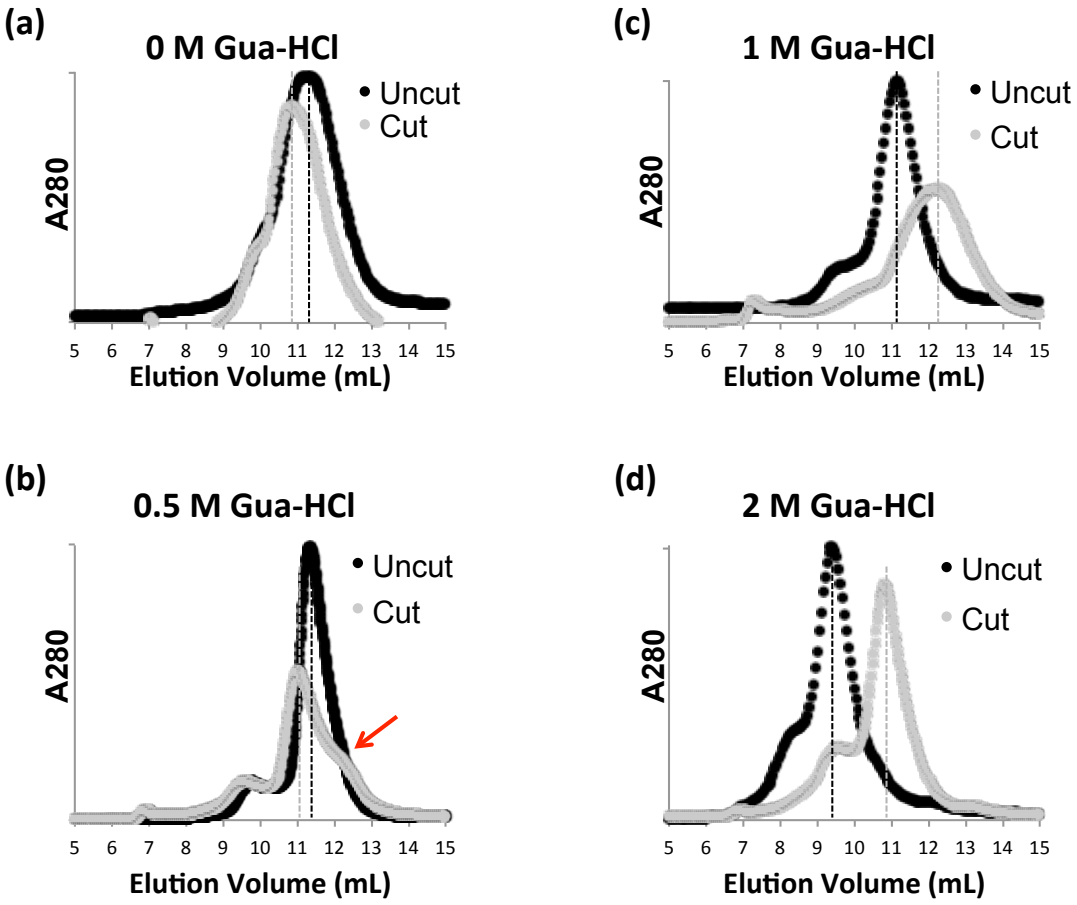


Figure 11. ROS1 glycosylase domain and the C-terminal domain dissociate in the presence of guanidine hydrochloride (Gua-HCl).

(a-d) Elution profiles of ROS1ΔN:P in consecutive runs on a Superdex 200 (10/300 GL) column (GE Healthcare) before and after the protease cleavage, in 20 mM Tris-HCl (pH 8.0), 5% glycerol, 1 mM dithiothreitol, 500 mM NaCl, and Gua-HCl in the concentration of 0, 0.5, 1, and 2 M. The peak height (y-axis) reflects relative OD280 absorbance as function of the retention volume in the x-axis. Red arrow in the panel (b) indicates a small amount of dissociated fragments shown as a delayed peak.

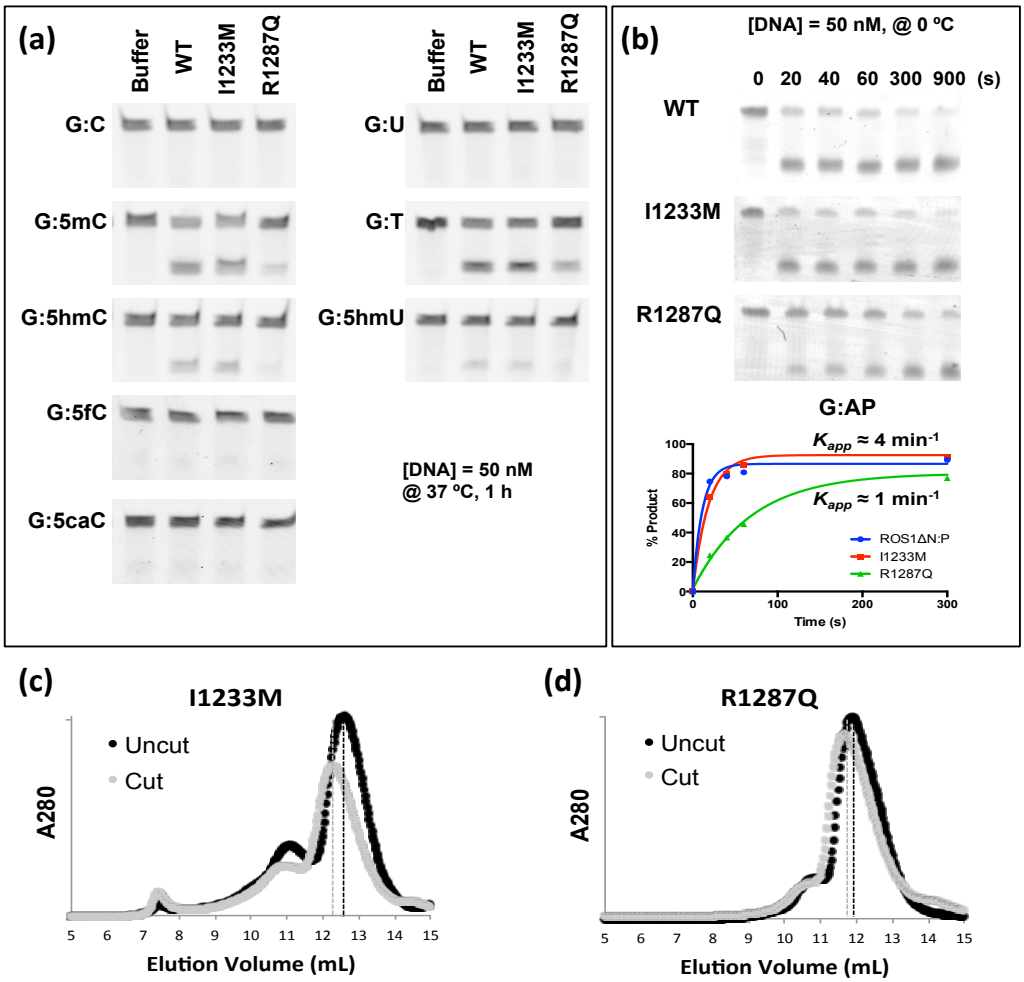


Figure 12. Effects of selected CTD mutagenesis on DNA glycosylase and lyase activities. (a) Activities of ROS1ΔN:P WT, I1233M, and R1287Q reactions ([$E_{\Delta N}$]=500 nM) on five oligos ([S_{DNA}]=50 nM) with various modifications and three oligos with G:X mismatches ([S_{DNA}]=50 nM) under the single-turnover condition. (b) The time course (0-15 min) of AP lyase activities of ROS1ΔN:P WT, I1233M, and R1287Q ([E]=0.5 μM) on 32-bp oligos ([S_{DNA}] = 50 nM) at 0 °C. (c and d) Elution profiles of ROS1ΔN:P I1233M and R1287Q, each in two consecutive runs on a Superdex 200 (10/300 GL) column (GE Healthcare) before and after the protease cleavage, in 20 mM Tris-HCl (pH 8.0), 5% glycerol, 1 mM dithiothreitol, and 500 mM NaCl. Peak heights reflected relative OD280 absorbance.

Mouse MYH does not possess 5-methylcytosine DNA glycosylase activity

We were intrigued by the observation that adding ROS1 CTD could restore the base excision activity by ROS1 GD. We asked whether ROS1 CTD could also allow other glycosylases to be active on 5mC and 5hmC for four reasons: (1) ROS1 shares some common substrates (such as 5-hydroxyuracil) with several DNA glycosylases (Nth1, Neil1)[102], (2) some of them are oxidized pyrimidine-specific DNA glycosylases that have been characterized in mammalian cells (Nth1, Neil1/2, TDG)[107], (3) several mouse DNA glycosylases (Neil1/2, Nth1 and Ogg1) were identified to bind 5mC- or 5hmC-containing oligos in a DNA pull-down experiment combined with quantitative mass spectrometry[79], and (4) mouse MutY homolog (mMYH) has recently been suggested to possess 5mC DNA glycosylase activity[142]. However, none of the mammalian enzymes we examined (**Figure 13**), including mMYH (**Figure 14**), showed 5mC or 5hmC DNA glycosylase activity with and without the addition of ROS1 CTD, whereas they were active on their respective substrates. Furthermore, addition of ROS1 CTD had no effect on the activities of mMBD4 and mMYH on their cognate substrates (**Figure 13b and 14a**). We were unable to observe the suggested 5mC activity for mMYH using either the 32-bp oligos (**Figure 14a**) or the 71-bp sequence from the mouse IL-2 promoter used by Wu and Zheng[142] (**Figure 14b**).

Among the HhH DNA glycosylases, mammalian MYH (MutY homolog) shares a similar domain organization as that of ROS1ΔN. MutY cleaves the adenine opposite 8-oxoguanine (8oxoG), which arises from unrepaired oxoG after DNA replication. In the structure of *Bacillus stearothermophilus* MutY in complex with DNA, the C-terminal domain recognizes 8oxoG and the opposite Ade flips out into the active site of the glycosylase domain where the excision occurs[143]. Mouse MYH (mMYH) is also known to excise Ade opposite Gua with comparable efficiencies as that of A:8oxoG[144]. For comparison with

ROS1, we generated the analogous mMYH:P construct, in which the PreScission protease recognition sequence was inserted in the linker connecting the mMYH glycosylase domain and its C-terminal domain (**Figure 15a**). The separated glycosylase domain of mMYH has much reduced activity on adenine excision (**Figure 15b lanes 3 & 4**), similar to *E. coli* MutY where the glycosylase domain alone has reduced activity[138, 145]. Unlike ROS1, the protease-cleaved mMYH fragments eluted as two separate and delayed peaks in the size exclusion chromatography (**Figure 15c-e**), clearly showing that the two domains dissociated after the cleavage.

We attempted to test whether ROS1 CTD could allow mMYH glycosylase domain to be active on 5mC by generating a hybrid enzyme (**Figure 16a**). The fusion enzyme has reduced activity on G:A mismatch (**Figure 16b lanes 5 & 6**), similar to that of cleaved mMYH glycosylase domain, but is not active on a G:5mC pair (**Figure 16b land 3**).

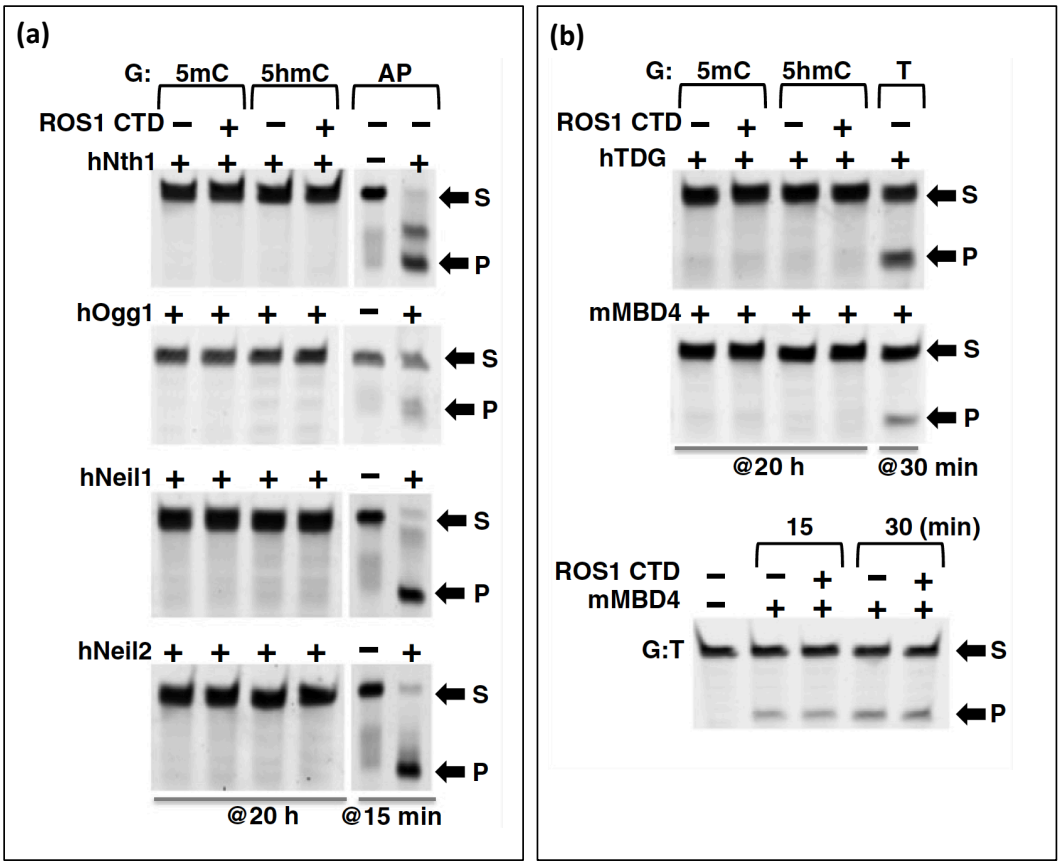


Figure 13. ROS1 CTD and other DNA glycosylases.

(a) Bifunctional DNA glycosylases, with and without ROS1 CTD ($[E_{CTD}]\approx1.5$ μM), on G:5mC and G:5hmC 32-bp oligos ($[S_{DNA}]=50$ nM) at 20 h reactions. AP lyase activities (15 min; 1 h for hOgg1) under the same condition are shown as positive controls. The enzyme concentrations used were 0.1 μg μl^{-1} of hNth1[146], 1.6 U of hOGG1 (catalog #M0241; New England Biolabs), and 50 ng μl^{-1} of hNeil1[147] and hNeil2[148]. (b) The glycosylase domains of hTDG[85] and mMBD4[86] ([E]=500 nM), incubated with and without ROS1 CTD ($[E_{CTD}]\approx1.5$ μM), on G:5mC and G:5hmC 32-bp oligos ($[S_{DNA}]=50$ nM) at 20 h reactions. Activities on G:T mismatch (30 min) under the same condition are shown as positive controls. Bottom panel: the activity of mMBD4 on G:T substrate is unaffected by the addition of ROS1 CTD.

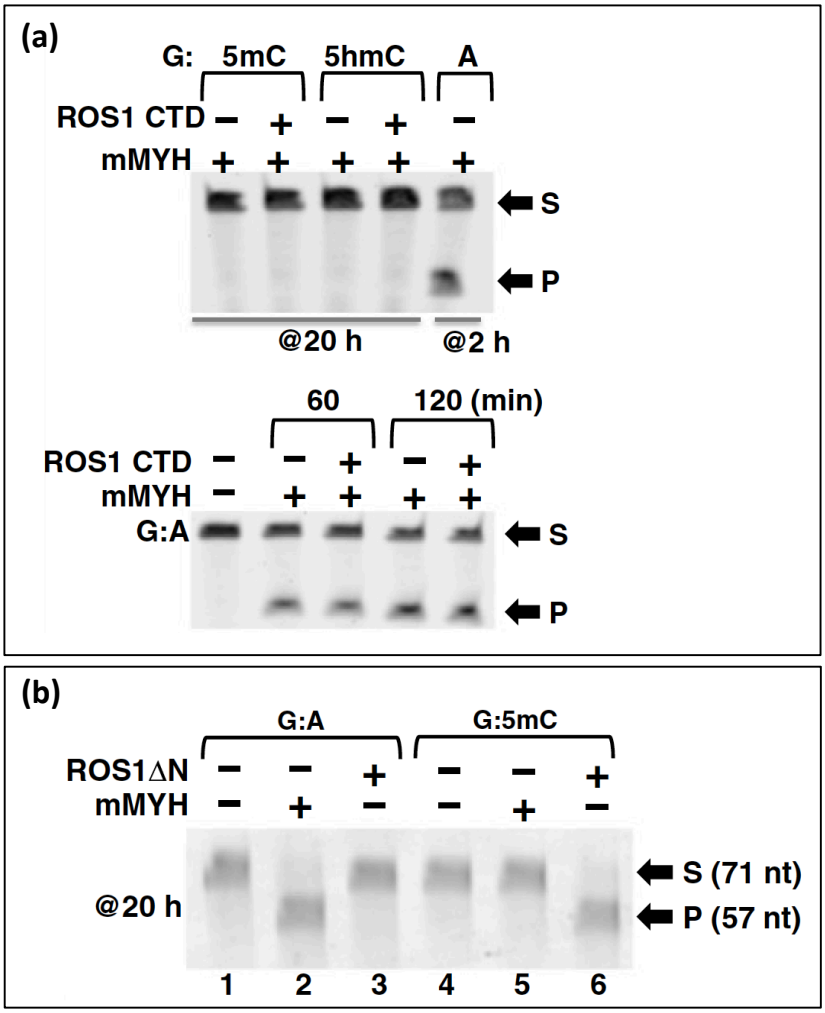


Figure 14. ROS1 CTD and mouse MutY homolog (mMYH).

(a) mMYH ([E]=500 nM), with and without ROS1 CTD ([E_{CTD}]≈1.5 μM), on G:5mC and G:5hmC 32-bp oligos ([S_{DNA}]=50 nM) at 20 h reactions. Activities on G:A mismatch (2 h) under the same condition is shown as a positive control. Bottom panel: the activity of mMYH on G:A substrate is unaffected by the addition of ROS1 CTD. (b) mMYH or ROS1ΔN ([E]=500 nM) on 71-bp oligos from the mouse IL-2 promoter [142] ([S_{DNA}]=50 nM) at 20 h reactions. mMYH is only active on G:A mismatch (lane 2), while ROS1ΔN is active on G:5mC (lane 6) under the same condition.

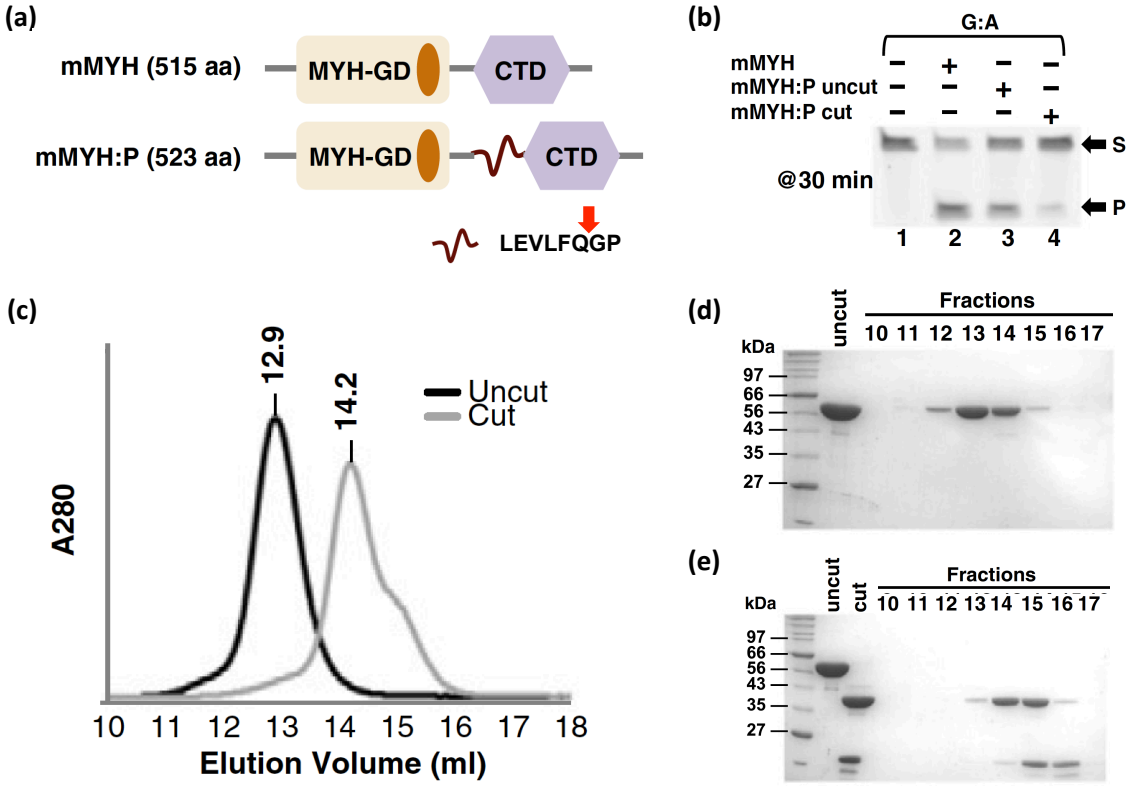


Figure 15. mMYH glycosylase domain and the C-terminal domain do not associate.

(a) The Prescission protease recognition sequence (LEVLFQGP) was inserted between mMYH GD and CTD. (b) Activities of mMYH and mMYH:P before and after the protease cleavage, on 32-bp oligos (G:A) at 30 min reaction in 37 ºC ([E]=500 nM and $[S_{DNA}]$=50 nM). (c) Elution profiles of mMYH:P in two consecutive runs on a Superdex 200 (10/300 GL) column, before and after the protease cleavage, in 20 mM Tris-HCl (pH 8.0), 5% glycerol, 1 mM dithiothreitol, and 500 mM NaCl. Peak heights reflected relative OD280 absorbance. (d and e) SDS-PAGE (15%) analyses of S200 fractions containing mMYH:P, before and after the protease cleavage.

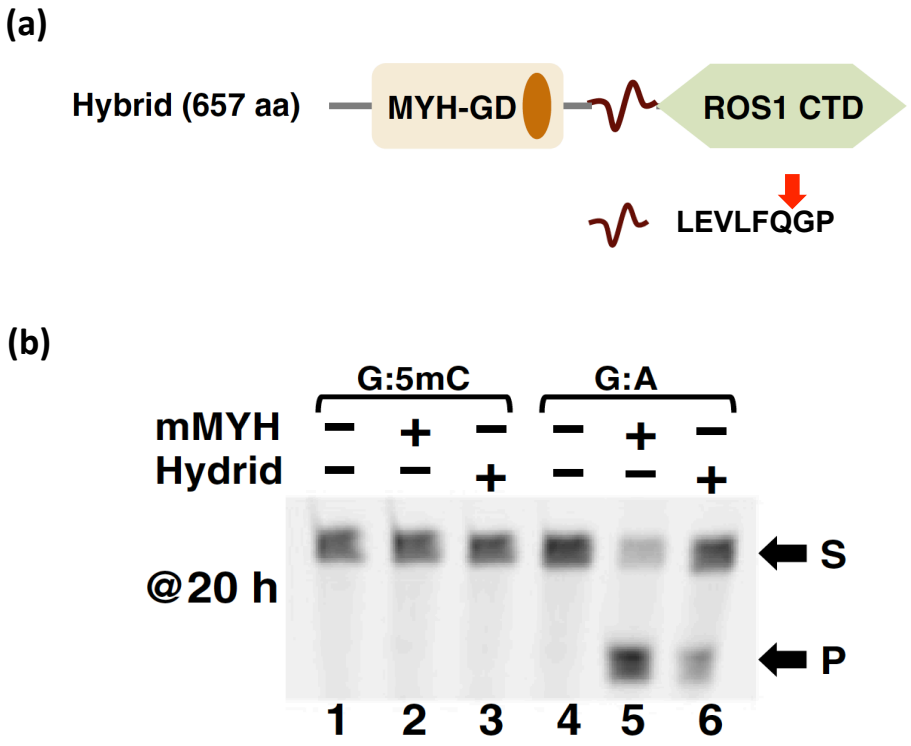


Figure 16. mMYH:GD-ROS1:CTD hybrid.

(a) The hybrid enzyme generated by fusing the N-terminal mMYH glycosylase domain to the ROS1 C-terminal domain. (b) Activities of mMYH and the Hybrid at 20 h reaction ([E]=500 nM and [S_{DNA}]=50 nM) on G:5mC and G:A 32-bp oligos.

Discussion

Extensive studies on DNA repair glycosylase enzymes, such as human uracil DNA glycosylase (hUNG) and 8-oxoguanine DNA glycosylases (hOGG and bacterial MutM), showed that they recognize damaged bases through a multi-step interrogation process [reviewed in [149] and [150]]. Allowing only a true substrate to reach the active site, these enzymes distort DNA by bending it followed by intrahelical interrogation to detect a lesion, flipping potential substrate nucleotides to varying degrees and rejecting non-substrate nucleotide back to DNA helices. The two-domain structure of MutY in complex with DNA containing an 8oxoG:A mismatch[143] revealed that the C-terminal domain contributes specific contacts to the intra-helically stacked 8oxoG lesion, which are functionally important for the lesion recognition and thus enzymatic excision of the extra-helical adenine opposite 8oxoG by the catalytic glycosylase domain. In other words, the two domains of MutY are primarily responsible, respectively, for essential interaction with the bases on opposite DNA strands; as changing 8oxoG:A to C:A significantly reduces the activity of MYH from calf thymus[151]. In the case of ROS1, the two domains, the glycosylase domain and the C-terminal domain, strongly associate with each other and seem to be insensitive to the base identity paired with the modified cytosine, as ROS1ΔN is active on all four pairs of 5mC:X or 5hmC:X (X=G, A, T, or C) (**Figure 17**). However, somehow the target 5mC or 5hmC must be recognized intrahelically, flipped out and delivered to the active site of the glycosylase domain to allow excision. The precise way in which the two interacting domains of ROS1 mediate specific DNA recognition and excision awaits the solution of a protein-DNA complex structure.

One possibility is that the C-terminal domain stabilizes the glycosylase domain and stimulates its intrinsic excision and lyase activities. A precedent is mammalian DNA methyltransferase 3A (DNMT3A) and its interaction with DNMT3-like protein (DNMT3L).

DNMT3A has a low activity on its own and forms oligomers. Interaction with DNMT3L disrupts the DNMT3A oligomer, forming a DNMT3L-3A tetramer via the catalytic domain of DNMT3A, and stimulates the DNMT3A activity[19, 152-154]. Previous published works by Ariza and colleagues showed, based on structural homology modeling and site-directed mutagenesis, that the ROS1 glycosylase domain interacts with both strands of DNA[100, 155]. It was suggested that residues Phe589 and Tyr1028 in the glycosylase domain are involved in the recognition of flipped-out 5mC in the cleavage center[100], and residues Arg903 and Met905 interact with the orphan guanine in the complementary strand[155]. However, these suggested interactions would be post-base flipping and would not account for the steps of the intrahelical modification interrogation that precedes the specific extrahelical base recognition. It is conceivable that the C-terminal domain could have a nonspecific DNA binding activity and thus stimulates the modification interrogation process by the glycosylase domain. Alternatively, we speculate that the C-terminal domain is a novel DNA substrate recognition domain responsive to pyrimidine modifications at the C5 position. The C-terminal domain might function in the early steps of intrahelical interrogation to detect the C5 modification and facilitate base flipping by the glycosylase domain for specific binding in the active site.

Arabidopsis thaliana ROS1 and mammalian TDG are the two DNA glycosylases currently implicated in so-called active DNA demethylation pathways by removing a modified cytosine base[156]: ROS1 excises 5mC and 5hmC but not 5fC and 5caC, whereas TDG removes 5fC and 5caC but not 5mC and 5hmC[85, 90]. It is worthy to note that ROS1 is inactive on G:U[102, 134], whereas TDG, even named after thymine DNA glycosylase, has much faster activity on G:U mismatch[106, 157]. The four chemically modified forms of cytosine might not be equivalent in terms of base pairing. A strong intramolecular hydrogen bond has been

observed between the exocyclic N4 amino group (NH_2) and the carbonyl oxygen (O=C) at ring carbon-5 position of 5fC, in the free nucleoside form[158, 159], and the carboxyl moiety (COO^-) of 5caC in the protein bound form[160]. It was hypothesized that the existence of such an intra-base hydrogen bond would shift the amino-imino equilibrium[106, 161, 162], which would enable 5fC and 5caC to form two, instead of three, hydrogen bonds with an opposite guanine, equivalent to a G:T or G:U 'wobble' pair. Previously observed mutagenic potential of 5fC and 5caC *in vivo* and *in vitro*[158, 161-164] suggested the possible existence of the imino tautomeric form. TDG might take advantage of the tendency of G:5fC and G:5caC to form a mismatch-like wobble hydrogen bonding pattern and turn them into substrates, whereas ROS1 is insensitive to mismatches.

Besides ROS1, which recognizes and excises 5mC from the 'natural' G:5mC base pair, another enzyme, PabI in *Pyrococcus abyssi*, initially identified as a restriction enzyme, actually is a sequence-specific adenine DNA glycosylase[165]. The dimeric PabI recognizes a palindromic 5'-GTAC-3', hydrolyses the *N*-glycosidic bond between the adenine base and the sugar, and produces two opposing AP sites that are subsequently cleaved by AP endonucleases to introduce a double-strand break. Thus, not every DNA glycosylase is involved in DNA repair, and some may generate damage.

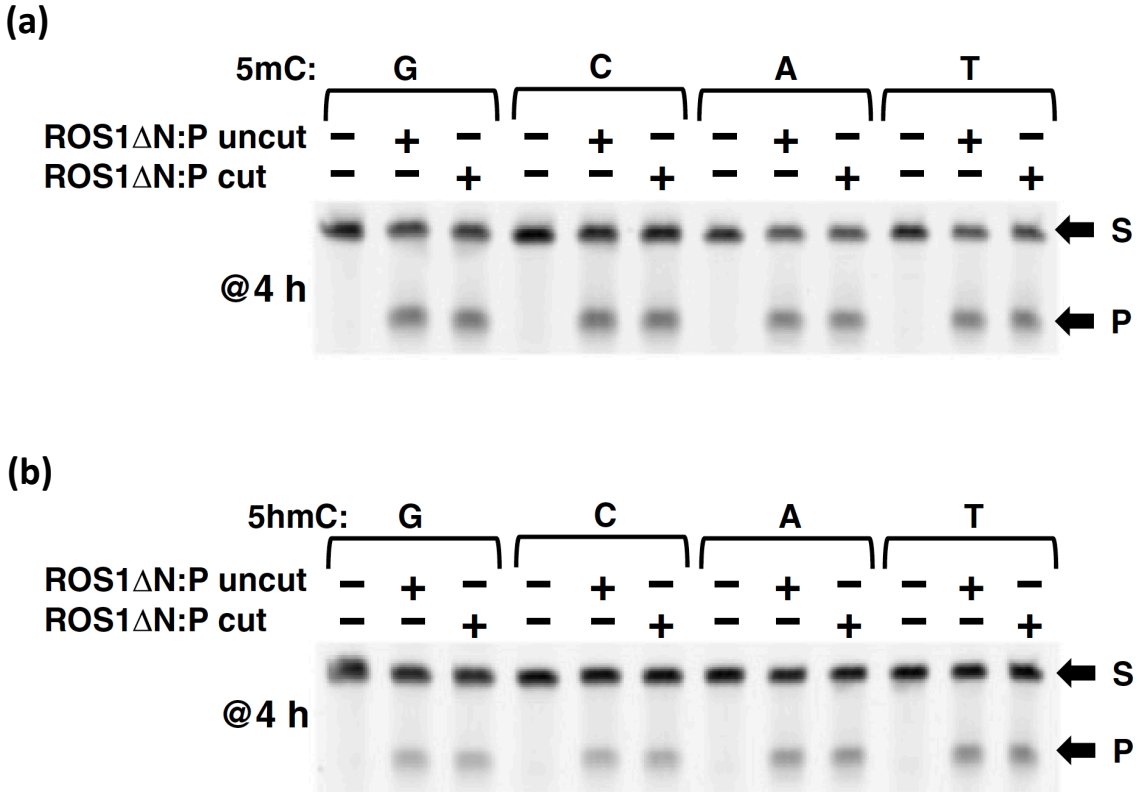


Figure 17. The effect of mismatching 5mC and 5hmC for excision activities.

(a and b) The effect of the opposing base on 5mC and 5hmC excision by ROS1ΔN:P before and after the PreScission protease cleavage. Reactions were performed with [E]=500 nM and [S]=50 nM for 4 h.

Materials and Methods

Protein Expression and Purification

The full-length (FL) ROS1 (pXC1135) was expressed in *Escherichia coli* dcm⁻ BL21-CodonPlus(DE3)-RIL (Stratagene) as a 6xHis fusion in a pET28a vector (Novagen). ROS1ΔN (pXC1256), ROS1ΔN—D971N (pXC1273), ROS1ΔN:P (pXC1327), ROS1ΔN:P—I1233M (pXC1375), ROS1ΔN:P—R1287Q (pXC1391), ROS1:GD (pXC1278), ROS1:CTD (pXC1297), mMYH (pXC1321), mMYH:P (pXC1332), and mMYH-ROS1 hybrid (pXC1338) were ligated into a modified pET28b vector (Novagen) as an N-terminal 6xHis-SUMO fusion and expressed in *E. coli dcm⁻* BL21-CodonPlus(DE3)-RIL (Stratagene). Bacterial cells were cultured in LB media at 37 ºC, and protein expression was induced at 16 ºC overnight or at 23ºC for 2 h (ROS1 FL). Cells were harvested and stored in -80 ºC. Cell pellet was thawed and lysed by sonication in lysis buffer [20 mM Tris-HCl (pH 7.5), 500 mM NaCl, 5 % glycerol and 1 mM dithiothreitol]. Lysate was clarified by centrifugation at 18,000 rpm for 1 h. The fusion protein was isolated on a Nickel-charged HisTrap affinity column (GE Healthcare).

For ROS1:FL, eluted fractions from the nickel column were further purified on tandem HiTrap Q and HiTrap SP ion exchange columns (GE Healthcare). For ROS1ΔN, ROS1ΔN—D971N, ROS1Δ:P, ROS1ΔN:P—I1233M, ROS1ΔN:P—R1287Q, and mMYH-ROS1 hybrid, eluted fractions from the nickel column were purified on Heparin affinity column (GE Healthcare) followed by cleavage of the 6xHis-SUMO tag via yeast ubiquitin-like-specific protease 1 (ULP-1; purified in-house), and then purified on a HiTrap Q column. For ROS1ΔN, 6xHis-SUMO tag-cleaved sample was passed through Glutathione Sepharose 4B (GE Healthcare) pre-bound with GST (glutathione S-transferase)-tagged Uracil glycosylase inhibitor protein (GST-UGI; purified in-house)[166] to remove any residual *E. coli*

uracil DNA glycosylase activity. ROS1ΔN, ROS1ΔN—D971N, ROS1Δ:P, ROS1ΔN:P—I1233M, and ROS1ΔN:P—R1287Q were further purified by Superdex 200 16/60 size exclusion column (GE Healthcare).

For ROS1:GD, ROS1:CTD, mMYH, and mMYH:P, eluted fractions from the nickel column were cleaved of their 6xHis-SUMO tag and further purified on an Heparin column. For ROS1:GD, mMYH, and mMYH:P, eluted fractions from the Heparin column were further purified on Superdex 200 (16/60 or 10/300 GL) size-exclusion column (GE healthcare). Final protein concentrations were estimated by absorbance at 280 nm for ROS1ΔN (absorbance coefficient ε=1.084), ROS1ΔN:P (ε=1.071), ROS1 GD (ε=1.450), mMYH (ε=1.472), mMYH:P (ε=1.456) or by Coomassie Blue staining using Bovine Serum Albumin as a standard for ROS1:FL and CTD. Compared with ROS1ΔN, the isolated ROS1:GD had lower yield and a broader peak on size exclusion column with some of the fractions overlapping with a major *E. coli* contaminant. ROS1:CTD had lower yield and higher impurity. In addition, ROS1:CTD was aggregated as it eluted in void volume after loaded on size-exclusion chromatography (data now shown).

DNA glycosylase activity assay

Activities of ROS1 and its variants, and other DNA glycosylases, on various DNA oligos labeled with 6-carboxy-fluorescein (FAM) were performed in reaction buffer [50 mM Tris-HCl pH 8.0, 1 mM ethylenediamine-tetraacetic acid (EDTA), 1 mM DTT and 0.1% bovine serum albumin] for the indicated time in room temperature (~23 ºC or otherwise indicated). For reactions with ROS1ΔN:P and mMYH:P after being cleaved by the Prescission protease (purified in-house), the protease (0.1 μg μl^{-1}) was present in the reaction mixture. Reactions were stopped by adding 2 μl Proteinase K (1 mg ml^{-1}) and incubating at 50 °C for 15 min.

For substrates except G:AP, reactions were stopped by adding 0.1 M NaOH and incubating at 95ºC on a heat block for 10 min. An aliquot (20 μl) of loading buffer (98% formamide, 1 mM EDTA, and trace amount of bromophenol blue and xylene cyanole) was added, and samples were heated in 95ºC on a heat block for 10 min. Samples were then immediately transferred to ice water to cool and loaded on a 10 cm x 10 cm denaturing PAGE gel containing 15% acrylamide, 7 M urea and 24% formamide in 1x Tris-Borate-EDTA (TBE). For G:AP substrates, reactions were stopped by adding 20 μl of loading buffer and samples were loaded on the gel without heating. The gel was run in 1x TBE at 200 V for 75 min. Typhoon Trio+ (GE Healthcare) was used to visualize the intensities from FAM-labeled DNA. The image-processing program ImageJ was used to quantify the intensities and data points were fit to a curve using Prism 6.0 (GraphPad).

Various 32-bp oligos labeled with FAM (synthesized by New England Biolabs) were used as substrates: (FAM)-5′-TCG GAT GTT GTG GGT CAG **X**GC ATG ATA GTG TA-3′ (where X = C, 5mC, 5hmC, 5fC, 5caC, U, T, or 5hmU) and its complementary strand 5′-TAC ACT ATC ATG C**Y**C TGA CCC ACA ACA TCC GA-3′ (where Y = G, A, T, or C). Oligo containing G:AP was generated by incubating G:U oligo with 1 Unit of *E. coli* uracil DNA glycosylase (catalog #M0280; New England Biolabs) for 30 min in room temperature (~23 °C). In addition, 71-bp oligos (synthesized by Sigma) were used for testing mMYH activity: (FAM)-5'-CAT GAG TTA CTT TTG TGT CTC CAC CCC AAA GAG GAA AAT TTG TTT CAT ACA GAA GG**X** GTT CAT TGT ATG AA-3′ (where X = A or 5mC) and its complementary strand 5'-TTC ATA CAA TGA ACG CCT TCT GTA TGA AAC AAA TTT TCC TCT TTG GGG TGG AGA CAC AAA AGT AAC TCA TG-3'.

Acknowledgements

We thank J.K. Zhu (Purdue University) for providing the ROS1 FL construct, J.I. Cohen (NIH) for the GST-UGI construct, S. Mitra (University of Texas Medical Branch at Galveston) for purified hNth1, hNeil1, and hNeil2, and J.R. Horton for comments. National Institutes of Health grant GM049245-21 supported this work to X.C. (who is a Georgia Research Alliance Eminent Scholar).

CHAPTER III.

Structural basis of methylated DNA recognition by human AP-1 and Epstein-Barr virus Zta transcription factors

Abstract

AP-1 is a classic basic leucine-zipper (bZIP) family transcription factor that binds the TPA response element (TRE; TGAGTCA). AP-1 can also bind a methylated response element (meTRE; MGAGTCA where M = 5mC). Homologous to AP-1, Epstein-Barr virus Zta/Zta homodimer also binds TRE and recognizes several methylated Zta response elements (meZRE) in a CpG methylation-dependent manner, one of which is meZRE-2 (TGAGMGA where M = 5mC). In this study, we have solved the crystal structures of Jun/Jun homodimer in complex with oligonucleotides containing meTRE and Zta/Zta homodimer in complex with oligonucleotides containing meZRE-2. The two structures reveal that Jun Ala265 and Zta Ser186 are involved in the specific recognition of T by one monomer and 5mC by the other monomer in the cognate methylated sequences. In addition, the highly conserved asparagine residues in both Jun and Zta show alternative conformations by each monomer for the recognition of different half-site sequences within meTRE and meZRE-2. Fluorescence polarization-based DNA binding analysis supports the observations in the structures. Our results demonstrate novel modes of 5mC recognition and explain the mechanism of DNA methylation-dependent, sequence-specific transcription factor binding by bZIP family proteins.

Introduction

DNA methylation is generally thought to inhibit transcription factor binding events, but mounting evidence shows that several families of transcription factors can preferentially bind 5mCpG within specific sequences[79, 112, 113, 115, 119, 123, 167-169]. The classic basic leucine zipper (bZIP) transcription factor family of AP-1 (e.g. Jun/Jun homodimer and Jun/Fos heterodimer) is critically involved in various regulations including oncogenesis, proliferations, and apoptosis[122, 170]. AP-1 activates a set of genes by binding a 7-bp 12-O-Tetradecanoylphorbol-13-acetate (TPA)-response element (TRE; 5'-TGAGTCA-3') as well as a methylated response element known as meAP-1 site (termed as meTRE; 5'-MGAGTCA-3' where M = 5mC)[125, 126, 171]. The sequence of meTRE is reminiscent of a distinct set of 5mCpG-containing 7-bp DNA sequences known as methylated Z response elements (meZREs), which are bound by AP-1-like Epstein-Barr virus (EBV) Zta/Zta homodimer (BZLF-1, Zebra, or Z) in a CpG methylation-dependent manner[127-129]. EBV is a human B cell-infecting gamma-herpesvirus, and its genome is heavily methylated during the latent stage of the virus cycle[172, 173]. Early lytic cycle activation is related to events in which EBV Zta/Zta binds TRE as well as preferentially recognizing methylated DNA containing meZREs, a notable example of which is meZRE-2 (5'-TGAGMGA-3')[129, 174].

Both human AP-1 and Zta are thus considered bZIP family transcription factors that bind the classic TRE, yet recognizing 5mCpG within different sequence contexts: AP-1 recognizes meTRE, and Zta/Zta recognizes meZRE-2. An alignment of amino acids sequences of AP-1 transcription factors and Zta shows that four DNA base-contacting amino acids are highly conserved except for Zta Ser186, which is equivalent to Jun Ala266 in AP-1. Zta Ser186 has been shown critical for methyl-dependent meZRE-2 binding[127, 175, 176]. Although a crystal structure of Zta bZIP homodimer-DNA complex was previously solved,

the reported structure has S186A mutation (AP-1 mimicry) and oligonucleotides containing TRE, not meZRE-2[177]. Comparing meZRE-2 to meTRE in DNA sequences shows that the relative position of 5mCpG within each sequence is different, indicating that the recognition of 5mC within meTRE would involve a distinct base-contacting amino acid other than the one equivalent to Zta Ser186. To date, how AP-1 and Zta/Zta recognize 5mC within their cognate methylated response elements is unknown.

Using isolated Jun and Zta bZIP domains, each forming a homodimer, we here report two high-resolution crystal structures of Jun/Jun-DNA complex containing meTRE and Zta/Zta-DNA complex containing meZRE-2. Both Jun/Jun and Zta/Zta recognize DNA by phosphate backbone contacts and base-specific contacts contributing to overall DNA binding affinities. The structures show that the C5-methyl groups of 5mC within meTRE and meZRE-2 are specifically recognized. Double-stranded DNA sequence comparison of TRE, meTRE, and meZRE-2 reveals that the position of the specifically recognized 5mC within meTRE and meZRE-2 aligns with the position of one of four T bases in TRE that are recognized by the equivalent amino acids in Jun (Ala265) and Zta (Ser186). We have generalized this observation as "T-to-5mC switch" to represent the model in which DNA methylation can uncover hidden binding sites for AP-1 and Zta/Zta by allowing 5mC to replace and mimic T for specific protein-DNA interactions.

Results

Overall structures

The crystal structure of Jun bZIP homodimer (Jun^A/Jun^B)-DNA complex containing a hemi-methylated meTRE was solved in the space group of C2 and refined to 1.89 Å (**Figure 18a**). The oligonucleotides used contained a 5'-terminal Ade extension on one strand and 5'-terminal Thy extension on the other strand, and both ends of the oligonucleotides were involved in protein-to-DNA contacts via symmetry related molecules for crystal packing. The overall structure in an asymmetric unit resembles the classic Jun/Fos-DNA complex structure characterized by two long α-helical monomers docking on the major groove of DNA via the basic region and forming a leucine-zipper dimer via the C-terminus[171]. The high-resolution data allowed 120 water molecules to be positioned, which are mostly involved in coating both major and minor groove of DNA as well as mediating various electrostatic protein-DNA interactions.

Next, the crystal structure of EBV Zta bZIP homodimer (Zta^A/Zta^B)-DNA complex containing a fully methylated meZRE-2 sequence, or 5'-TGAGCGA-3' (underlined CpG and the complementary CpG methylated), was solved in the space group of C2 and refined to 2.25 Å (**Figure 18b**). The oligonucleotides used contained 5'-terminal Ade or Thy in each end. Crystal packing shows one end of the oligonucleotides involved in a protein-to-DNA contact and the other involved in a DNA-to-DNA contact in a head-to-head fashion. The 5'-terminal Thy and the adjacent C:G base pair are partly overlapped with the equivalent bases in the adjacent symmetry-related molecule. The overall structure in an asymmetric unit reveals the similar N-terminal DNA binding regions and the C-terminal dimerization features as shown in the structure of Zta/Zta-DNA complex containing the TRE sequence[177].

The two structures—human Jun/Jun-DNA and EBV Zta/Zta-DNA complexes containing methylated DNA—are remarkably similar in that both Jun and Zta form homodimers via the C-terminal leucine-zipper regions of the domains and recognize the same number of bases in the major groove via the basic N-terminal regions of the α-helices. However, the two structures significantly differ in that the C-terminal tails of Zta/Zta homodimer present disordered loops, each spanning approximately 12 amino acids. In addition to the leucine zipper-like dimerization formed between two helices, the disordered C-terminal tails are engaged in salt bridge-mediated dimerization as previously reported[177].

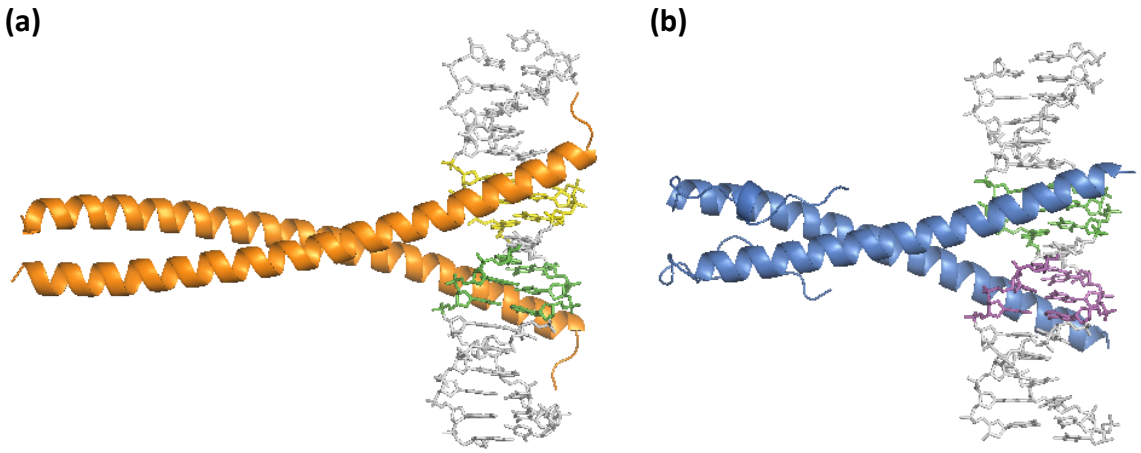


Figure 18. Overall Structures of Jun/Jun-DNA and Zta/Zta-DNA complexes.

(a) Jun^{A}/Jun^{B} (orange) binds meTRE through the major groove, and the identical monomers recognize non-identical half-sites: 5' half-(MGA) is indicated by yellow and 5' half-(TGA) is indicated by green. (b) Zta^{A}/Zta^{B} (skyblue) recognizes non-identical half-sites of meZRE-2: 5' half-(TGA) in green and 5' half-(TMG) in purple are indicated.

Response elements containing asymmetric half-sites

Three 7-bp response elements—TRE, meTRE, and meZRE—are relevant to this study. Each sequence contains two half-sites, and each half-site is recognized by a single monomer of Jun or Zta (**Figure 19a**). Within the classic TRE (TGAGTCA) in the double-stranded context, the middle G:C base pair is flanked by two symmetric sites of 5' half-(TGA), or equivalently 3' half-(TCA) in the complementary strand. In contrast to TRE, meTRE and meZRE-2 contain asymmetric half-sites. In meTRE, one of the half-sites is identical to 5' half-(TGA) of TRE, and the other half-site is 5' half-(MGA). meZRE-2 also has 5' half-(TGA) and a distinctive 5' half-(TMG). In the N-terminal basic region of Jun and Zta, four core amino acids in conserved positions are involved in the recognition of DNA bases within a single half-site plus the middle G or C (**Figure 19b**). The four core amino acids of Jun and Zta differ at a single position corresponding to Jun Ala266 and Zta Ser186. This Ala-to-Ser difference in the position partly distinguishes how Jun^A/Jun^B and Zta^A/Zta^B interact with the cognate sequences.

In the structure of Jun^A/Jun^B-DNA complex containing meTRE (**Figure 19c**), Jun^A recognizes 5' half-(MGA) from -3 to -1 positions via Asn262, Ala265, and Ala266. Jun^B engages the equivalent amino acids to recognizes 5' half-(TGA) from +3 to +1 positions. The middle G:C base pair is recognized by both monomers via conserved Arg270: $Arg270^A$ recognizing C (0) via a water molecule and $Arg270^B$ recognizing G (0). In the structure of Zta^A/Zta^B-DNA complex containing meZRE-2 (**Figure 19d**), Zta^A recognizes 5' half-(TGA) from -3 to -1 positions via Asn182, Ala185, and Ser186; and Zta^B recognizes 5' half-(TMG) from +3 to +1 positions via the equivalent amino acids. The middle G:C base pair is recognized likewise by both monomers with $Arg190^A$ recognizing C (0) via a water molecule and $Arg190^B$ recognizing G (0). Therefore, Jun^B and Zta^A monomers can both recognize 5'

half-(TGA) via the four core amino acids despite the Ala-to-Ser difference (**Figure 20a**). Conversely, Jun^A and Zta^B engage the same core amino acids to recognize distinct half-sites containing 5mCpG: Jun^A recognizing 5' half-(MGA) (**Figure 20b**) and Zta^B recognizing 5' half-(TMG) (**Figure 20c**). The Ala-to-Ser difference in this case is critical for the recognition of different methylate sequences.

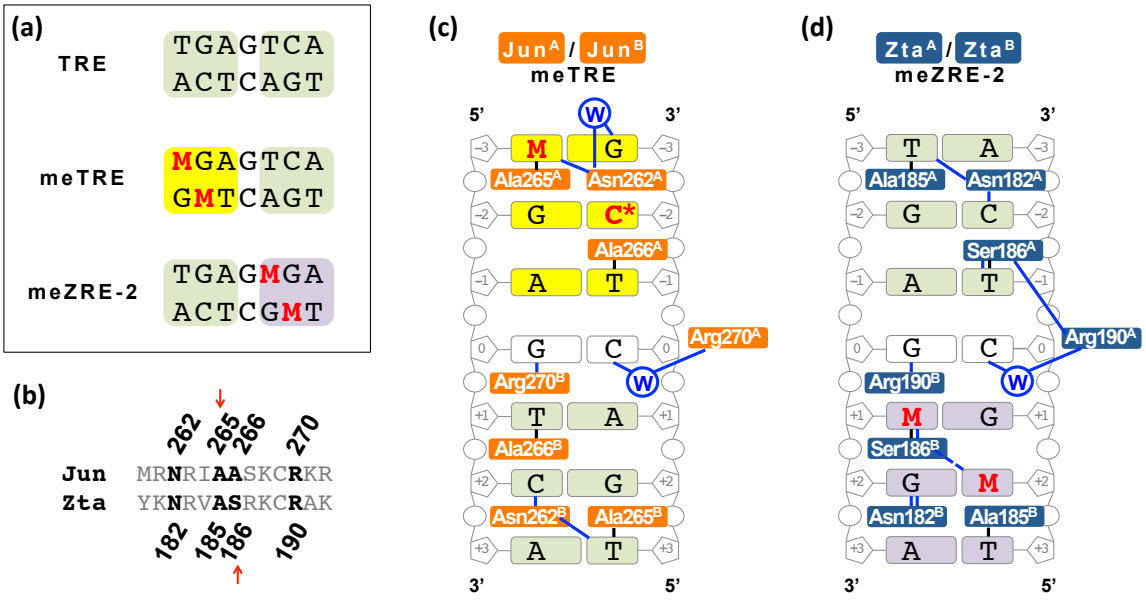


Figure 19. Summary of Jun/Jun-DNA and Zta/Zta-DNA base-specific interactions.

(a) The DNA sequences of the three response elements are aligned. Each half-sites in double strands is shown in corresponding colors. 5mC is indicated by M in red. (b) The basic regions of human Jun and EBV Zta are aligned. Bolded letters indicate the four core residues involved in base-specific contacts. The red arrow indicates the residue that directly recognizes 5mC. (c and d) Schematic representations of base-specific interactions of Jun^{A}/Jun^{B}-DNA confined to meTRE-2 and Zta^{A}/Zta^{B}-DNA confined to meZRE-2 are shown. Van der Waals contacts are indicated by black lines, and electrostatic and H bond interactions are indicated by blue lines ("W" indicates a water molecule). Each residue is indicated by the corresponding identity of the monomer A or B. The Jun^{A}/Jun^{B}-DNA complex structure contains hemi-methylated meTRE sequence in which the asterisk (*) indicates the unmethylated Cyt.

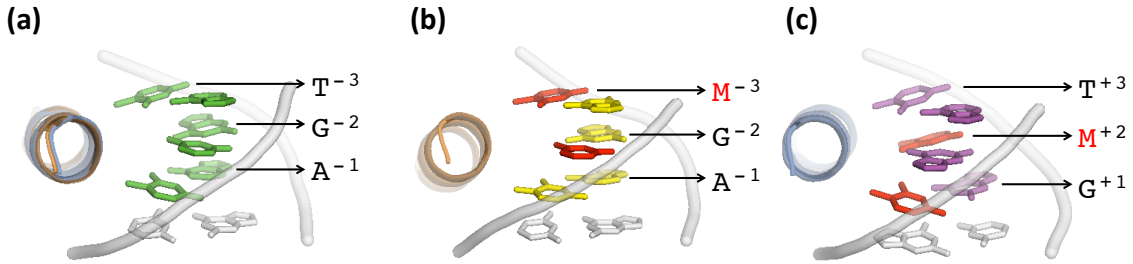


Figure 20. Comparison of 5' half-(TGA), 5' half-(MGA), and 5' half-(TMG).

(a) A structural view of Zta^{A} over 5' half-(TGA) in green with an overlay of Jun^{B} is shown. (b) Jun^{A} over 5' half-(MGA) in yellow is shown. 5mC (M in red) bases of the double-stranded 5mCpG within the half-site are indicated in red. (c) Zta^{B} over 5' half-(TMG) in purple is shown.

Position-specific pyrimidine C5-methyl group recognitions: "T-to-5mC switch"

The chemical structures of T and 5mC are very similar, as both bases are pyrimidines containing the C5-methyl group. T is otherwise known as 5-methyluracil (5mU). In light of this similarity, a comparison of the three response elements—TRE, meTRE, and meZRE-2—reveals that the positions of carbon-5 (C5)-methyl-presenting bases (T or 5mC) within the sequences are conserved at -3, -1, +1, and +3 positions where 5mC replaces T in one of the positions within meTRE and meZRE-2 in respect to TRE (**Figure 21a**). The C5-methyl groups of either 5mC or T bases at -3 and -1 positions are symmetrically related to those at +3 and +1 respectively (indicated by ± sign). In TRE, four T bases are found in all four positions. In the structure of Jun/Fos-DNA complex (PDB: 1FOS) containing TRE, Jun Ala265 and Ala266 over one 5' half-(TGA) have van der Waals contacts to the C5-methyl group of T (-3)and T (-1) respectively via the Ala-C^{β} with the interatomic distance of 3.7-3.9 Å[171] (**Figure 21b,c**). Over the other 5' half-(TGA), Fos Ala150 and Ala151, which are equivalent to the two alanine residues from Jun, have the equivalent van der Waals contacts to T (+3) and T (+1) respectively (**Figure 21d,e**).

In comparing meTRE to TRE sequences(**Figure 21a top and middle panels**), M (-3) replaces T (-3) in one of the conserved positions ("T-to-5mC switch" at -3 position), forming 5mCpG from -3 to -2 positions. The T bases in the other conserved positions (-1, +1, and +3) remain the same as in TRE. In the structure of Jun^{A}/Jun^{B}-DNA complex containing meTRE, Jun^{A} Ala265 over 5' half-(MGA) has a van der Waals contact to the C5-methyl group of M (-3) (**Figure 21f**). Jun^{A} Ala266 likewise recognizes T (-1) of 5' half-(MGA) (**Figure 21g**). Therefore, Jun Ala265 can effectively recognize both T and 5mC via the van der Waals contact between Ala-C^{β} and the C5-methyl group. Jun^{B} Ala265 and

Ala266 over 5' half-(TGA) recognize T (+3) and T (+1) respectively, involving the similar van der Waals contact (**Figure 21h-i**).

An equivalent pattern is observed in the structure of Zta^A/Zta^B-DNA complex containing meZRE-2, though a "T-to-5mC switch" is found at a different position. In comparing meZRE-2 to TRE sequences (**Figure 21a top and bottom panels**), 5mC replaces T at the conserved +1 position, along with a C-to-G change at +2 position, forming 5mCpG from +1 to +2 positions. Zta^A Ala185 over 5' half-(TGA) recognizes T (-3) via the typical van der Waals contact (**Figure 21j**), and Zta^B Ala185 over 5' half-(TMG) recognizes T (+3) in much the same way (**Figure 21m**). On the other hand, Zta Ser186 is equivalent to Jun Ala266, and thus involve a serine side-chain instead of alanine to contact the base in -1 and +1 positions. Over 5' half-(TGA), Zta^A Ser186 recognizes T (-1) via the H bond-donating Ser-O^{γ} that contacts T (-1) O4 with an interatomic distance of 2.9 Å, and the Ser-C^{β} contacts the C5-methyl group via a van der Waals contact (**Figure 21k**). The van der Waals contact between the Ser-C^{β} and the C5-methyl group is similar to that between an Ala-C^{β} and a C5-methyl group. Over the other half-site of meZRE-2, or 5' half-(TMG), Zta^B Ser186 coordinates M (+1) in much the similar way as the T (-1) recognition (**Figure 21l**). Yet, Ser-O^{γ} would now accept an H bond from M (+1) N4, while Ser-C^{β} maintains the typical van der Waals contact to the C5-methyl group of M (+1). Zta^A Ser186 and Zta^B Ser186 are thus distinguishable in that Zta^A Ser186 over 5' half-(TGA) donates an H bond to T (-1) O4, whereas Zta^B Ser186 over 5' half-(TMG) accepts an H bond from M (+1) N4. Nevertheless, both Zta^A Ser186 and Zta^B Ser186 recognize the C5-methyl group of T (-1) and M (+1), involving the van der Waals contact by Ser-C^{β}.

Here, it is important to note that the C5-methyl group of 5mC is biologically different from that of T, as only 5mC is generated by DNA methyltransferases whose

activities are differently regulated under various biological cues. Therefore, T is permanently "methylated" at the C5 position, while the C5 methyl group of 5mC is dynamically regulated. The preservation of the key interactions by Jun Ala265 and Zta Ser186 for the recognition of both T and 5mC in different half-site contexts indicates that the C5-methyl group of 5mC can effectively mimic that of T for a protein-DNA interaction. In other words, 5mC can functionally replace T for specific DNA binding. Our observations directly present the mechanism by which AP-1 and Zta/Zta recognize their cognate methylated response elements whose sequences differ from TRE in that the most significant change in the sequence involves a "T-to-5mC switch".

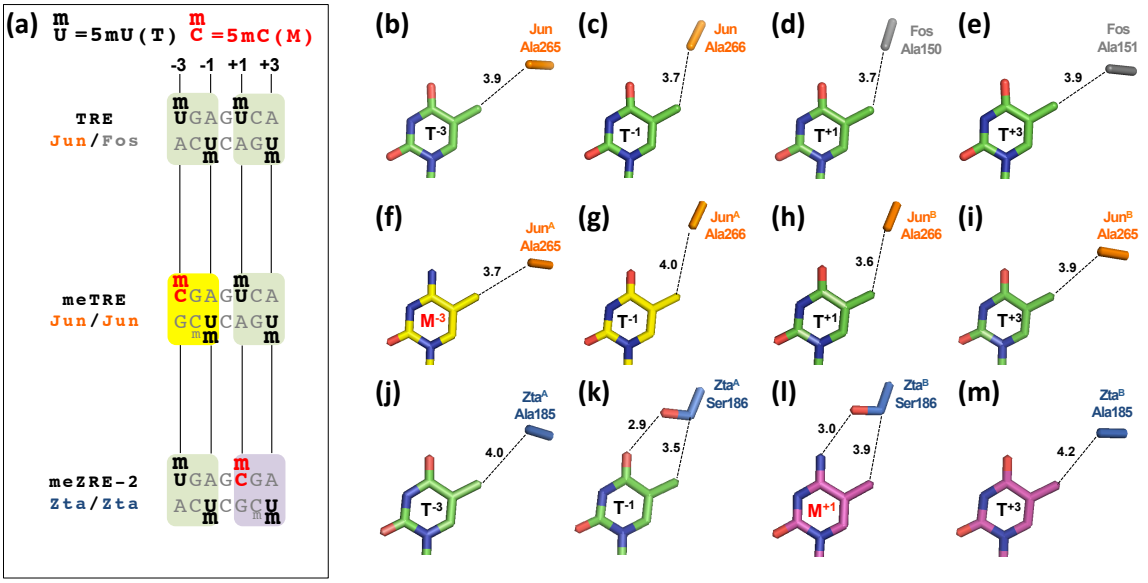


Figure 21. Position-specific C5-methyl group recognitions by Jun/Jun and Zta/Zta.

(a) Conserved positions (-3, -1, +1, and +3) of four C5-methyl groups (indicated by "m") of 5mU (T) bases in black and 5mC (M) bases in red within the three response elements are shown. Double-stranded half-sites within the sequences are shown in colors: 5' half-(TGA) in green, 5' half-(MGA) in yellow, and 5' half-(TMG) in purple. (b-e) The recognition of the C5-methyl groups of T (-3), T (-1), T (+1), and T (+3) by conserved alanine residues in Jun and Fos (PDB: 1FOS). (f-i) The recognition of the C5-methyl groups of M (-3), T (-1), T (+1), and (T+3) by conserved alanine residues in Jun. (j-m) The recognition of T (-3), T (-1), M (+1), and (T+3) by the conserved alanine and serine residue in Zta.

Methyl-dependent binding in solution

In our structures, we have observed that 5mC bases in the conserved positions—M (-3) of meTRE and M (+1) of meZRE-2—are specifically recognized. However, we have not been able to directly address the specific recognition of 5mC bases at the "non-conserved" positions (-2 in meTRE and +2 in meZRE-2) in the complementary strands within the CpG dinucleotide context. For the crystallization of Jun^A/Jun^B-DNA complex, we used double-stranded oligonucleotides containing meTRE sequence with a hemi-methylated CpG: M (-3) and C (-2). We did not observe any amino acid side-chain that may account for the recognition of the methyl group at -2 position if the base in the position was methylated. For the crystallization of Zta^A/Zta^B-DNA complex, we used double-stranded oligonucleotides containing meZRE-2 sequence with a fully methylated CpG: M (+1) and M (+2). Still, we did not observe any direct recognition of the C5-methyl group of M (+2) in the structure. We were then motivated to better understand the effect of methylation in each strand for DNA binding and to quantitatively demonstrate the "T-to-5mC switch" model in solution to examine that the C5-methyl groups in the conserved positions determine specific DNA binding. Using fluorescence polarization analysis, we measured the dissociation constant (K_D) for meTRE DNA binding by Jun^A/Jun^B and meZRE-2 DNA binding by Zta^A/Zta^B as a function of different methylation conditions: no methylation (C/C), methylation of CpG in one strand (M/C or C/M), and methylation of CpG in both strands (full methylation; M/M).

In the analysis of Jun/Jun DNA binding in the meTRE sequence context under the different methylation conditions (**Figure 22a**), K_D for full methylation (M/M) was determined to 107 nM and for no methylation (C/C) determined to 472 nM (4- to 5-fold weaker binding). K_D for methylation in the conserved position-only (M/C) was determined

to 96 nM, essentially equivalent to K_D for full methylation. On the other hand, K_D for methylation in the "non-conserved" position in the other strand (C/M) was determined to 396 nM, approaching K_D for no methylation. Therefore, methylation in the conserved -3 position significantly contributed to specific DNA binding, whereas methylation in the "non-conserved" position was insensitive to specific DNA binding. Our analysis agrees with results from previous studies measuring the effect of methylation in meTRE sequence context for Jun/Fos or Jun/Jun binding, performed using different methods[125, 126, 178]. The reported data as well as our analysis clearly support that methyl-specific DNA binding by Jun/Jun for meTRE is mediated by methylation in the conserved position where the "T-to-5mC switch" is found.

For Zta/Zta-DNA binding in meZRE-2 sequence context, several studies have reported that full methylation (M/M) results in significantly stronger binding compared to no methylation (C/C), primarily shown via electronic mobility shift assays (EMSA) [127, 129]. However, to our knowledge, the effect of methylation in each strand on DNA binding affinities had not been accounted for. In our analysis of Zta/Zta-DNA binding under the different methylation conditions (**Figure 22b**), K_D for full methylation (M/M) was determined to 6 nM, and for no methylation (C/C) it was determined to 122 nM (20-fold weaker binding). K_D for methylation in the conserved position (M/C) was determined to 12 nM, which is only 2-fold weaker than K_D for full methylation and 10-fold stronger than K_D for no methylation. K_D for methylation in the other strand (C/M) was determined to 54 nM, approximately 2-fold stronger than K_D for no methylation. The observation that the effect of methylation in one strand is stronger than that in the other resonates with the previous studies on methyl-dependent DNA binding by proteins such as several MBD proteins, Zfp57, and Klf4, all showing the similar 2-fold increase in DNA binding[8, 112, 168]. For some of

these proteins, a network of water molecules in the vicinity of the C5-methyl group is considered to contribute to DNA binding[112, 117]. However, we have not observed a comparable network of water molecules over the related M^{+2} position in our Zta/Zta-DNA complex structure, likely due to the limitation in the resolution of our structural data. Nevertheless, the effect of the methylation in the conserved position at M^{+1} regardless of the methylation status in the other strand is clear and supports the "T-to-5mC switch" model.

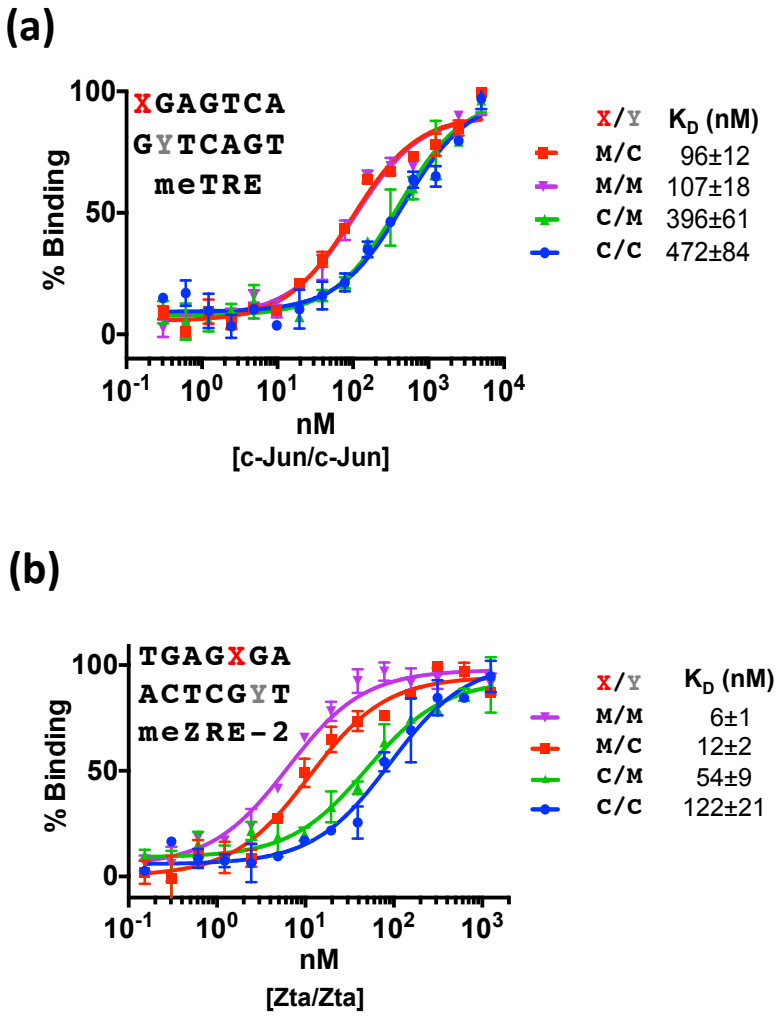


Figure 22. C5-methyl-dependnet DNA binding by Jun/Jun and Zta/Zta in solution.

(a) Effects of different methylation status in the conserved position (X in red) in one strand and/or non-conserved position in the other strand (Y in gray) on meTRE DNA binding by Jun/Jun in solution are shown compared to no methylation (X/Y = M/M, M/C, C/M, and C/C). Binding affinities was measured by fluorescence polarization analysis. (b) The effects of methylation status on meZRE-2 DNA binding by Zta/Zta in solution are shown.

Effects of oxidative modifications on DNA binding

Mammalian genomes have three other forms of modified cytosine in addition to 5mC: 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC)[73-75]. Mammalian Ten-eleven translocation (Tet) dioxygenases can oxidize the C5-methyl group of 5mC to generate 5hmC, and further oxidations of 5hmC can produce 5fC and then 5caC[132, 179-181]. Understanding how the five forms of cytosine (C, M, 5hmC, 5fC, and 5caC) are specifically recognized within the context of various protein-DNA interactions can shed light on fundamental mechanisms of epigenetic regulations by DNA modifications. Particularly, both AP-1 and Tet dioxygenases are implicated in several types of malignancies[122, 182]. Also, Zta/Zta can bind the host genome in the EBV-infected cells via meZRE-2 or meZRE-2-like sequence context[183, 184] and may respond to Tet dioxygenase activities *in vivo*[185]. We therefore tested the effect of additional cytosine C5 modifications for Jun/Jun-DNA binding (meTRE sequence background) and Zta/Zta-DNA binding (meZRE sequence background). Because the effect of methylation in the conserved position was critical for methyl-specific binding, we introduced the five forms of cytosine at the conserved position in the background of methylation in the non-conserved position (X/M; where X = C, M, 5hmC, 5fC, and 5caC).

In our analysis, both Jun/Jun and Zta/Zta showed decreased DNA binding affinities with the introduction of the oxidative modifications. For meTRE DNA binding by Jun/Jun in reference to full methylation (M/M), 5hmC (5hmC/M) showed less than 2-fold weaker binding, followed by both 5fC and 5caC (5fC/M and 5caC/M) showing ~3-fold weaker binding (**Figure 23a**). However, all modifications presented stronger binding than no modification at the conserved position (C/M). For Zta/Zta meZRE-2 DNA binding in reference to full methylation (M/M), 5hmC (5hmC/M) showed ~4-fold weaker binding, and

5fC (5fC/M) and 5caC (5caC/M) showed ~10-fold and ~22-fold weaker binding respectively (**Figure 23b**). 5hmC and 5fC, but not 5caC, presented stronger binding than no modification at the conserved position (C/M).

The effect of weak DNA binding with the introduction of the oxidative modifications can be effectively explained by our structures, as the van der Waals contact between an amino acid side-chain C^{β} and the C5-methyl group of 5mC is critical for methylated DNA binding by both Jun/Jun and Zta/Zta. Each successive oxidative modification accompanied by increasing bulkiness would progressively disrupt this key interaction. Such progressive reduction in specific DNA binding by the modifications has also been observed in Zfp57 and Klf4 DNA binding studies[112, 168]. 5hmC in both strands of CpG within meZRE-2 was also shown to reduce meZRE-2 DNA binding by Zta/Zta in a recent report[185], correlating with our observation. These results suggest that, oxidative modifications may serve as graduated signals to progressively reduce the activity of certain 5mC-binding transcription factors during the event of active DNA demethylation. On the other hand, proteins with 5caC-reading capabilities, such as WT1 transcription factor[186], Tet3 CXXC domain[187], and RNA Pol II elongation complex[188], may be distinctively signaled by the oxidative modifications.

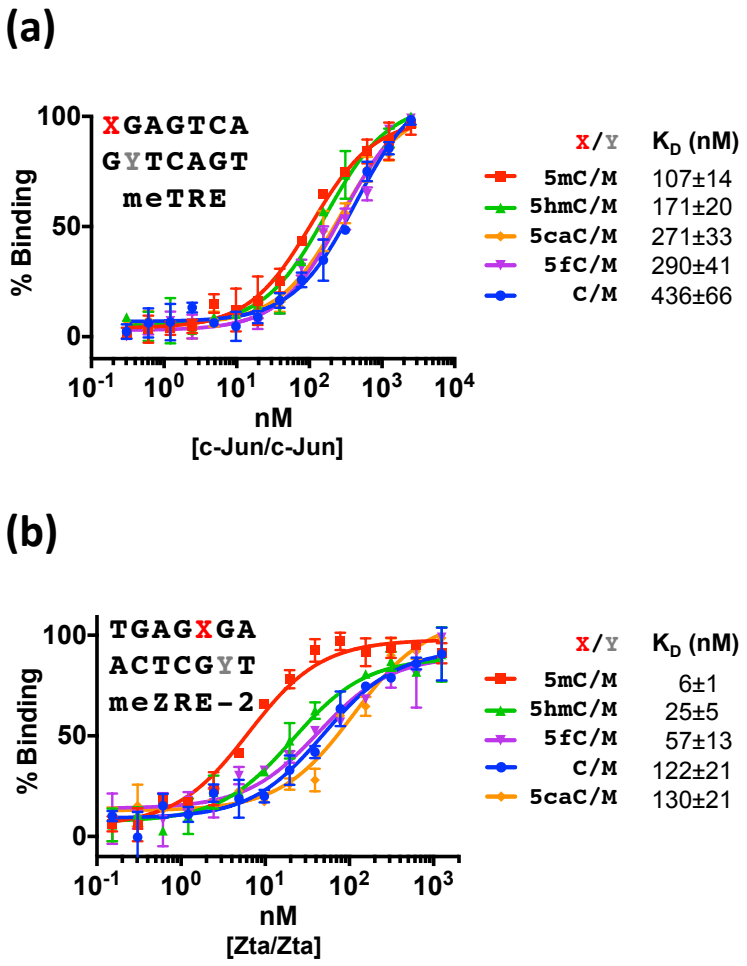


Figure 23. Effects of oxidative modifications on DNA binding by Jun/Jun and Zta/Zta. (a) Oxidative modifications in the conserved position (X) in the background of methylation in the other strand (Y = M) on meTRE DNA binding by Jun/Jun in solution are shown compared to no methylation or methylation in the corresponding position (X = C, 5mC, 5hmC, 5fC, and 5caC). (b) The effects of oxidative modifications on meZRE-2 DNA binding by Zta/Zta in solution are shown.

Resolving the difference of asymmetric half-sites

While the "T-to-5mC switch" preserves the recognition of the C5 methyl groups by Jun Ala265 and Zta Ser186 upon DNA methylation, methylation-dependent DNA binding would not be possible without the adaptation of other base-contacting core amino acids to accommodate different half-site sequences. Indeed, we have observed that Jun Asn262 and Zta Asn182, which are highly conserved asparagine residues based on the alignment of bZIP family proteins, undergo the most significant changes in the rotamer χ angles for alternative sequence recognitions. Both Jun^{B} Asn262 and Zta^{A} Asn182 over 5' half-(TGA) have the same orientation, recognizing T (-3) O4 via the H bond-donating Asn-N$^{\delta}$ and recognizing C (-2) N4 via the H bond-accepting Asn-O$^{\delta}$ (**Figure 24a**). This asparagine orientation is conventional to other bZIP family transcription factors that recognize 5' half-(TGA)[171, 189, 190]. Jun^{A} Asn262 over 5' half-(MGA), however, has the χ_1 and χ_2 angles significantly changed in reference to its conventional conformation, to coordinate M (-3) N4 via Asn-O$^{\delta}$ and G^{-3} O6 via Asn-N$^{\delta}$ involving a water molecule (**Figure 24b**).

Zta^{B} Asn182 over 5' half-(TMG) shows another distinct alternative conformation in which the χ_1 angle shows a minor shift, but the χ_2 angle is rotated nearly 180° to coordinate Ade (-3) N6 via the Asn-O$^{\delta}$, and to coordinate the O6 and N7 atoms of Gua^{-2} via the Asn-N$^{\delta}$ (**Figure 24c**). Previous studies have shown that the corresponding asparagine residues in other bZIP proteins that recognize 5' half-(TTA), such as yeast PapI and human C/EBP subfamily, can likewise adapt to a different sequence context compared to the conventional 5' half-(TGA)[124, 191]. Therefore, the flexibility of the conserved asparagine and its ability to form both the H-bond acceptor and donor account for how Jun/Jun and Zta/Zta recognize asymmetric half-sites, provided that any change in DNA sequence preserves other key interactions critical for specific DNA binding.

In addition to the conserved asparagine, Zta Ser186 is also distinctively engaged in H-bond interactions in each half-site of meZRE-2, apart from its role in the recognition of T and 5mC in the conserved ±1 positions. In our Zta^A/Zta^B-DNA complex structure, Zta^A Ser186 and Zta^B Ser186 present different networks of interactions over 5' half-(TMG) and 5' half-(TGA) within meZRE-2. The "T-to-5mC switch" within meZRE-2 engages an H bond donator-acceptor alteration for Ser186 as described previously: Zta^A Ser186 Oγ donates an H bond to T (-1) O4, and Zta^B Ser186 Oγ accepts an H bond from M (+1) N4 (**Figure 25a,b**). Also, DNA sequences adjacent to T (-1) and M (+1) are not symmetric, as they are part of distinct half-sites. Consequently, neighboring atoms near each Ser186 within Zta^A/Zta^B are involved in a distinct network of interactions. Over 5' half-(TGA), Zta^A Ser186 recognizes T (-1) as well as the Zta^A Arg190, as the Ser-O^γ accepts an H bond from Arg-N^η that also has an H bond with a water molecule to coordinate C (0) N4 (**Figure 25a**). In contrast, Zta^B Ser186 over 5' half-(TMG) recognizes M (+1) as well as M (+2) by accepting an H bond from M (+2) N4 (**Figure 25b**). Also, Zta^B Arg190 does not engage Zta^B Ser186 as in the Zta^A Ser186-Arg190 interaction but is involved in the bifurcated recognition of G (0) (**Figure 25b**).

Interestingly, the orientations involved in the recognition of middle G:C at position 0 by Zta^A Arg190 and Zta^B Arg190 are conventional to most bZIP proteins recognizing the "pseudo-palindromic" response elements such as TRE or TRE-containing sites[171, 177, 192]. This middle G:C base pair can be switched (C:G) in TRE for AP-1 binding, as the conserved arginine from each monomer would switch the orientation. However, such switching may not be allowed for meZRE-2 DNA binding by Zta^A/Zta^B, as the neighboring Zta^A Ser186 and Zta^B Ser186, having their specific orientations in relation to the arginine residues, may prevent such flexibility. The consequence would be that meZRE-2 binding by Zta^A/Zta^B

would require the middle G:C to be fixed in such that G (0) is always 3' to the 5' half-(TGA) and 5' to the 3' half-(MGA). Alternatively, C (0) would be fixedly 3' to 5' half-(TMG) and 5' to 3' half-(TCA). This prediction is supported by ChIP-seq data from other studies, showing that middle G or C can be varied in TRE for AP-1 (Jun) but fixed in meZRE-2 for Zta[183, 193].

The recognition of M (+2) N4 by Zta^{B} Ser186 (**Figure 25b**) points out a critical aspect of how Zta/Zta recognizes meZRE-2, as Jun and other AP-1 transcription factors presenting alanine (Jun Ala266) in the corresponding position would lack this particular interaction. Previous studies showed that AP-1 does not activate promoters via meZRE-2 binding and that Zta S186A mutant has a reduced meZRE-2 binding capability compared to the wild type (WT)[127, 175, 176]. Particularly, Yu et al. showed that Zta S186A mutant rendered an inability to activate promoters via meZRE-2 binding, whereas Jun A266S mutant led to a gain-of-function resembling Zta WT to activate promoters via meZRE-2[127].

We were therefore motivated to quantitatively measure the effect of Zta S186A and Jun A266S mutants for DNA binding in the meZRE-2 sequence context under the background of no methylation (C/C) and full methylation (M/M) of the CpG within the sequence. Subsequent results showed that meZRE-2 DNA binding by Zta S186A for full methylation (M/M) with K_D of 201 nM was ~13-fold weaker compared to Zta WT, suggesting the loss of M^{+2} N4 recognition by Zta^{B} Ser186-O^{γ} (**Figure 25c**). Interestingly, Zta S186A for full methylation (M/M) still showed approximately 6-fold stronger binding compared to no methylation (C/C). This difference is likely due to that the mutant Zta Ala186 still recognizes the C5-methyl group of M (+1) within meZRE-2 in the same manner as shown in the interaction between Jun Ala266 and the C5-methyl group of T within meTRE. Comparatively, Jun A266S for full methylation (M/M) showed approximately 4-fold stronger meZRE-2 DNA binding compared to Jun WT in the same background,

suggesting that Jun A266S may now engage in the recognition of M (+2) N4 within meZRE-2 and become capable of specific binding (**Figure 25d**). Also, Jun A266S for full methylation (M/M) showed 6-fold stronger meZRE-2 DNA binding compared to no methylation (C/C), while Jun WT did not present such binding affinity difference in response to changing the methylation status. These results suggest that meZRE-2 DNA binding by Jun WT is non-specific without the Ala-to-Ser mutation. Therefore, our structures and DNA binding assay results effectively present Zta Ser186 as a critical factor for both methyl-specific and meZRE-2 sequence-specific DNA binding by Zta/Zta.

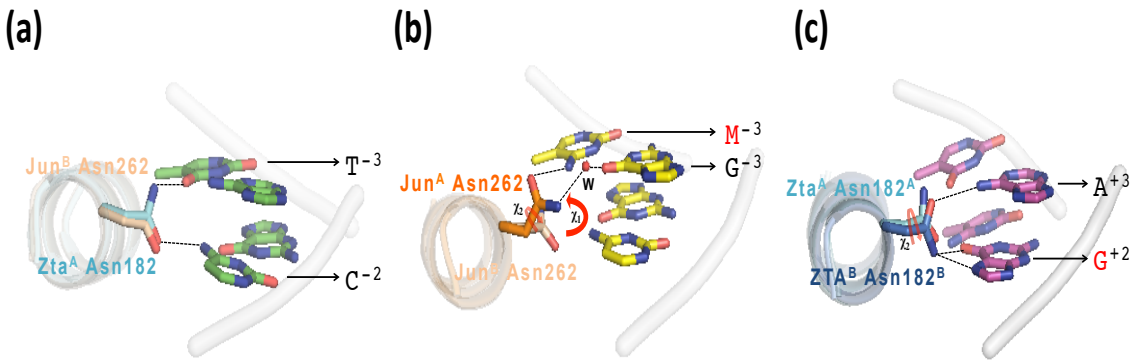


Figure 24. Alternative conformations adapted by the conserved asparagine for engaging asymmetric half-sites.

(a) A structural view of Zta^{A} Asn182 (light blue) coordinating T (-3) O4 and C (-2) N4 within 5' half-(TGA) of meZRE-2 is shown. The basic region of Jun^{B}, which also recognizes 5' half-(TGA) of meTRE, is structurally aligned to that of Zta^{A}. Jun^{B} Asn262 (light orange) shows a conformational equivalence to Zta^{A} Asn182. (b) Jun^{A} Asn262 (darker orange) coordinating M (-3) N4 and G (-3) O6 via a water molecule (W) over 5' half-(MGA) of meTRE. Compared to Jun^{B} Asn262 (light orange) over the other half-site of 5' half-(TGA), Jun^{A} Asn262 shows a different conformation with the χ_1 angle swung by nearly 90 ° and the χ_2 angle also rotated. (c) Zta^{B} Asn182 (darker blue) coordinating O6 and N7 atoms of G (+2) via Asn-N$^{\delta}$ and Ade (+3) N6 via Asn-O$^{\delta}$ over 5' half-(TMG) of meZRE-2. Compared to Zta^{A} Asn182 (light blue) over the other half-site of 5' half-(TGA), the χ_2 angle is rotated approximately 180°.

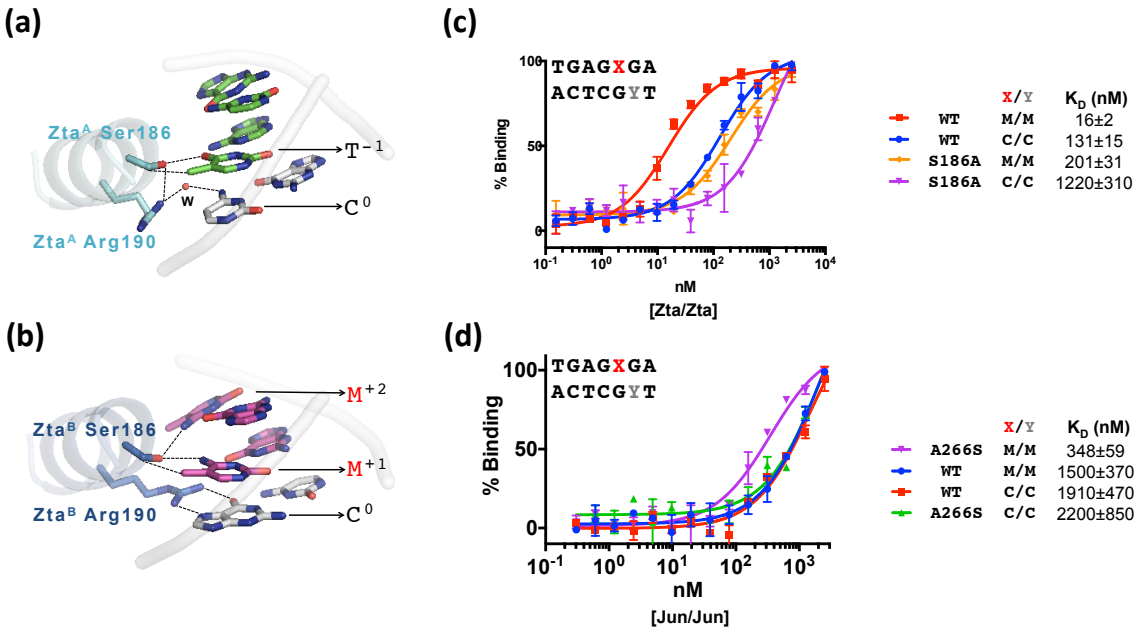


Figure 25. The recognition of T and 5mC by Zta Ser186.

(a) A structural view of Zta^{A} Ser186 and Zta^{A} Arg190 over 5' half-(TGA) of meZRE-2 is shown. Zta^{A} Ser186 recognizes T (-1) as described (Fig. 2K). In addition, Ser-O$^{\gamma}$ H bonds with Zta^{A} Arg190-N$^{\eta}$ that coordinates C (0) N4 via a water molecule (W). (b) Over the other half-site of 5' half-(TMG), Zta^{B} Ser186 recognizes M (+1). In addition, the Ser-O$^{\gamma}$ accepts an H bond from M (+2). The adjacent Zta^{B} Arg190 has a bifurcated coordination to G (0) via Arg-N$^{\eta}$ atoms. (c) Fluorescence analysis of the effects of full (M/M) or no methylation (C/C) on meZRE-2 DNA binding by Zta/Zta was shown in the wild-type (WT) background and the S186A mutant background. Binding affinities (K_D) measured are indicated. S186A reduces meZRE-2 binding but still retains methyl-specific binding. (d) Fluorescence analysis of the effects of full (M/M) or no methylation (C/C) on meZRE-2 DNA binding by Jun/Jun was shown in the wild-type (WT) background and A266S mutant background. Binding affinities (K_D) measured are indicated. Jun WT shows non-specific binding regardless of methylation status, whereas A266S increases methyl-specific binding.

Discussion

Historically, prokaryotic and eukaryotic 5mC recognition can be categorized into two structurally distinct modes of interactions. In one mode, 5mC is flipped out of DNA helix and thus extrahelically recognized, as most extensively shown in SET and RING finger-associated (SRA) domains[13, 194, 195]. In the other mode, 5mC within a CpG dinucleotide is recognized via a non-base flipping mechanism involving the 5mC-Arg-G triad[116]. The examples include methyl-CpG binding domain (MBD)-containing MeCP2 as well as C2H2 zinc finger (ZnF) family transcription factors such as Kaiso, Zfp57, and Klf4 that bind 5mCpG within specific sequences[112, 117, 168, 169]. Recent studies have shown that there are other families of transcription factors including the bZIP family that can recognize 5mCpG within specific sequences besides the ZnF family[79, 167]. As the first structural demonstration of such, our Jun^A/Jun^B-DNA Zta^A/Zta^B-DNA complex structures reveal distinct modes of the recognition of 5mCpG compared to the 5mC-Arg-G triad mode.

In the 5mC-Arg-G triad, the arginine side-chain has a non-polar interaction with the C5-methyl group of 5mC, and two Arg-N^{η} atoms are engaged in bifurcated interactions with the 3'-Gua for the recognition of 5mCpG in one strand[116]. The same interaction may be adopted for the recognition of TpG. For meTRE binding by Jun/Jun, however, Ala265 side-chain recognizes 5mCpG by contacting the C5-methyl group via Ala-C^{β}. Asn262 nearby then engages in an alternative conformation to adopt the CpG dinucleotide context, as previously described. Therefore both Asn262 and Ala265 from one Jun monomer over 5' half-(MGA) recognize the double-stranded 5mCpG in which only one C5-methyl group is recognized. The same Asn262 and Ala265 can recognize the TpG dinucleotide, involving a different conformation of Asn262 over 5' half-(TGA).

meZRE DNA binding by Zta/Zta shows yet another distinct mechanism of 5mCpG recognition. In the structure of Zta/Zta-DNA containing meZRE-2, Ser186 recognizes 5mCpG by contacting the C5 methyl group of 5mC by Ser-C^{β} and the N4 atom by Ser-O^{γ}. The same Ser-O^{γ} then engages in the recognition of 5mC N4 in the other strand. Zta Ser186 thus primarily recognizes double-stranded 5mCpG by recognizing both 5mC bases by the N4 atoms but recognizes only one C5-methyl group. In addition, Zta Asn182 recognizes Gua O6 and N7 within 5mCpG. Despite such different modes of 5mCpG recognitions, both of our structures and other structures showing 5mC-Arg-G triads point to the principle of "T-to-5mC switch" in such that the C5-methyl group of 5mC can effectively equate the C5-methyl group of T for transcription factor binding.

Further studies call for a systematic understanding in which various sequence-specific 5mC readers control gene regulations in response to extra-cellular cues and in relation to intra-cellular chromatin states. Studies have shown that bZIP family proteins may bind methylated CpG in distal promoter regions for gene activations, whereas proximal promoter regions of transcriptionally active genes are primarily unmethylated[123, 125, 183]. Particularly interesting for future directions would be to broadly understand how DNA methylation and demethylation events at genomic regions bound by such 5mC-binding transcription factors are regulated.

Materials and Methods

Protein Expression and Purification

Human Jun bZIP (residues 254-315 containing C269S mutation) with wild-type Ala266 (Jun WT, pXC1398) or A266S mutant (Jun A266S, pXC1440) was expressed as an N-terminal 6xHis-SUMO (HisSUMO) fusion via modified pET28b vector (Novagen) in *Escherichia coli* BL21-CodonPlus(DE3)-RIL (Stratagene). EBV Zta bZIP (residues 175-236 containing C189S mutation) with wild-type Ser186 (Zta WT, pXC1416) or S186A mutant (Zta S186A, pXC1455) was expressed under the same background as Jun bZIP. Bacterial cells were cultured in LB at 37 ºC, and protein expression was induced at 16 ºC overnight by adding 0.5 mM isopropyl-β-d-1-thiogalactopyranoside (IPTG). Cells were harvested and stored in -80 ºC. Cell pellets were thawed and lysed by sonication in 20 mM sodium phosphate pH 7.4, 500 mM NaCl, 25 mM imidazole, 5% (v/v) glycerol, and 1 mM DTT. Lysate was clarified by centrifugation at 18,000 rpm for 1 h, and the fusion protein was isolated on a Nickel-charged HisTrap affinity column (GE Healthcare). Eluted fractions from the nickel column were pooled.

For the purification of Jun bZIP, ubiquitin-like-specific protease 1 (ULP-1; purified in-house) was added to the pooled nickel column fractions, followed by overnight incubation at 16 ºC to completely cleave the HisSUMO tag. The tag-cleaved sample was then loaded to tandem HiTrap-Q/HiTrap-Heparin column (GE Healthcare), followed by elution from the Heparin column using a linear gradient of NaCl (500 mM to 2 M). The eluted fractions were loaded onto Superdex 200 16/60 size exclusion column (GE Healthcare) in buffer containing 20 mM Tris-HCl pH 7.5, 500 mM NaCl, 5% (v/v) glycerol, and 1 mM DTT. The final concentration of the purified homodimer was estimated by Bradford protein assay (Bio-Rad no. 500-0205).

For Zta bZIP purification, the pooled Ni column fractions were loaded to tandem HiTrap-Q/HiTrap-Heparin column (GE Healthcare) in Zta buffer (Tris-HCl pH 7.5, 150 mM ammonium acetate, 150 mM NaCl, and 1 mM DTT). The Heparin column was then eluted using a linear gradient of NaCl (150 mM to 2 M). The eluted fractions were pooled and dialyzed against the Zta buffer in presence of ULP-1 in 4 ºC to cleave the HisSUMO tag. The dialyzed sample was then loaded to Heparin column followed by elution as before. The eluted fractions were then dialyzed against the Zta buffer again, concentrated, and then loaded to Superdex 200 16/60 size exclusion column in the Zta buffer. Elution from the column showed a single peak corresponding to the expected Zta bZIP homodimer size. The final concentration of the purified homodimer was estimated by measuring absorbance at 280 nm.

Crystallography

For Jun bZIP homodimer (Jun/Jun)-DNA complex, purified Jun/Jun was mixed with annealed oligonucleotides containing methylated meTRE sequence (hemi-methylated CpG, See Table 1) in a molar ratio of ~ 1:1. The final complex was concentrated to ~1 mM in 20 mM Tris-HCl (pH 7.5), 100 mM NaCl, and 5 % v/v glycerol. Initial screening was performed by the sitting-drop method, and select conditions were optimized by the hanging-drop method. The final rod-shaped crystals appeared at 16 ºC within 3 days in mother liquor containing 0.05 M Citric Acid, 0.05 M Bis-Tris-Propane, and 16% w/v polyethylene glycol 3350 at pH 5.0.

For Zta bZIP homodimer (Zta/Zta)-DNA complex crystallization, purified Zta/Zta was mixed with annealed oligonucleotides containing methylated meZRE-2 sequence (fully methylated CpG) in a molar ratio of ~1:1. The final complex was concentrated to ~1 mM in

20 mM Tris-HCl (pH 7.5), 150 mM ammonium acetate, 150 mM NaCl, and 1 mM DTT. A wide range of screening resulted in the formation of a well-diffracting crystal at 16 ºC within 2 months in mother liquor containing 0.2 M sodium phosphate monobasic monohydrate and 20% w/v polyethylene glycol 3,350.

Crystals were diffracted at the SER-CAT 22ID beamline at the Advanced Photon Source, Argonne National Laboratory, and the diffraction data were processed using HKL2000[196]. Crystallographic phase for Jun/Jun-DNA and Zta/Zta-DNA complexes were determined by molecular replacement using the coordinates from previous structures (PDB 1FOS for Jun and PDB 2C9L for Zta). Model refinements were performed using PHENIX[197]. Graphics for the figures were generated using PyMol (DeLano Scientific, LLC). Detail X-ray data collection results are summarized in **Table 1**.

Fluorescence-based DNA binding Assay

Fluorescence polarization assay was performed using Synergy 4 microplate reader (BioTek) to measure DNA binding by Jun/Jun and Zta/Zta. For DNA binding assay, Jun bZIP (WT and A266S) and Zta bZIP (WT and S186A) were purified as HisSUMO tag-uncleaved forms by following the same purification procedure used for tag-cleaved Jun bZIP, except for the addition of ULP-1. 6-caboxy-fluorescein (FAM)-labeled dsDNA probe (5 nM) was incubated with increasing concentration of proteins in 20 mM Tris-HCl pH 7.5, 5% glycerol, and 185 mM NaCl (for Jun/Jun) or 225 mM NaCl (for Zta/Zta). The sequences of the probe for Jun/ Jun were 5'-GGAXGAGTCATAG-3' and FAM-5'-CTATGACTYGTCC-3' (where X and Y are C, M, 5hmC, 5fC, or 5caC); and the sequences for the probe for Zta/Zta were 5'-CTATGAGXGATCC-3' and FAM-5'-GGATYGCTCATAG-3' (where X and Y are C, M, 5hmC, 5fC, or 5caC). K_D values were calculated as [mP] = [maximum mP] x

[C]/(K_D + [C]) + [baseline mP], and % binding was calculated as ([mP] – [baseline mP])/([maximum mP] – [baseline mP]) (where mP is milipolarization and [C] is protein concentration). Average K_D values and standard errors are indicated.

Table 1. Jun/Jun-DNA and Zta/Zta-DNA crystals data collection and refinement.

Protein	Human Jun DBD homodimer	EBV Zta DBD homodimer
DNA (M = 5mC)	5′ AATGGAMGAGTCATAGGAG 3′ 3′ TACCTGCTCAGTATCCTCT 5′	5′ AAGCACTGAGMGATGAAG 3′ 3′ TCGTGACTCGMTACTTCT 5′
Beamline	SER-CAT AP 22ID	SER-CAT AP 22ID
Wavelength (Å)	1.0	1.0
Space group	C2	C2
Cell dimensions		
a, b, c (Å)	158.87, 42.49, 45.17	95.549, 26.732, 99.673
α, β, γ (°)	90, 98.01, 90	90 97.248 90
Resolution (Å)*	35.00-1.89 (1.96-1.89)	35.00-2.25 (2.33-2.25)
Rmerge*	0.034 (0.405)	0.086 (0.885)
$<I/\sigma I>$*	134.05 (2.63)	16.8 (1.09)
Completeness (%)*	98.5 (90.1)	88.7 (48.9)
Redundancy*	6.9 (4.2)	7.2 (2.8)
Observed reflections	164,675	77,911
Unique reflections*	23,735 (2181)	10789 (579)
Refinement		
Resolution (Å)	1.89	2.25
Number of reflections	23,712	10,726
Rwork / Rfree	20.1/23.7	25.3/29.2
Number of atoms	1890	1765
Average B-factors (Å^2)	48.0	85.0
Wilson B-factors (Å^2)	33.9	54.8
RMS deviations		
Bond lengths (Å)	0.011	0.005
Bond angle (°)	1.325	0.679
All atom clash score	2.11	1.85
Ramachandran Favored (%)	99.2	96.6
Additional allowed	0	0
Cβ deviation	0	0

* Data for the highest-resolution shell are in parentheses.

CHAPTER IV.

Discussions and Future Directions

Comparison of 5-methylcytosine and thymine

5mC and thymine as pyrimidines within DNA share a common feature of possessing the C5-methyl group (**Figure 26**). The methyl group of 5mC is regulated by DNA methyltransferases, whereas the methyl group of thymine is not. An important observation from the ROS1 substrate specificity studies (See Chapter II) is that ROS1 is comparatively active for the excision of both 5mC base-paired with guanine and thymine mismatched to guanine, whereas it is not active on uracil mismatched to guanine[102, 104]. Therefore, ROS1 does not share TDG's characteristic mismatch repair, which is substantially active on uracil as well as thymine mismatched to guanine. Such a distinctive substrate specificity profile by ROS1 clearly suggests that the substrate base recognition is responsive to the presence of the C5-methyl group. It is not clear how ROS1 distinguishes between the thymine mismatched to guanine as opposed to the one base-paired to adenine. Nevertheless, the recognition of both 5mC and thymine by ROS1 is reminiscent of transcription factors that recognize 5mC and T equivalently. As shown in Chapter III, our structures of bZIP transcription factors in complex with 5mC-containing DNA show that the C5-methyl groups of 5mC and thymine could be equivalent for protein-DNA interactions, involving an alanine or serine side-chain to form van der Waals contacts. Some ZnF family transcription factors also equivalently recognize 5mC and T via a different mode of interaction involving the 5mC-Arg-G triad[116].

The ability of ROS1 to recognize and excise thymine mismatched to guanine may be a feature of protection from 5mC deamination. It is widely recognized that many 5mCpG sites in the genomes are prone to deamination during which the generation of thymine from

5mC would initially generate a T:G mismatch[198, 199]. Subsequent rounds of DNA replication without repairing the mismatch would establish a C:G to T:A transition mutation. As such, some deamination events in plants may be disruptive to genetic regulation. For instance, *Arabidopsis thaliana* 5mC DNA glycosylases can activate silenced genes by promoter demethylation[99]. If the methylated promoter sequence becomes deaminated and results in transition mutations, the promoter function would be compromised. Also, methylated rice retrotransposon Tos17 is activated by 5mC DNA glycosylases in response to environmental stress elements[200]. Deamination of transposable elements that results in deactivation would interfere with the organism's adaptive responses to changing environments. Therefore, the correction of T:G mismatches through thymine excision by 5mC DNA glycosylases may be a DNA repair housekeeping feature in addition to their role in DNA demethylation.

A similar mechanism of a protection from deamination is also present in mammalian systems. Mammalian MBD4 has 5mCpG-binding MBD domain as well as a DNA glycosylase domain that excises thymine mismatched to guanine[201]. Mammalian MBD4 can thus bind methylated CpG islands and allow the glycosylase domain to excise thymine mismatched to guanine in the vicinity. Interestingly, the plant MBD4 homolog lacks the MBD domain while preserving the glycosylase domain, and thus the 5mCpG-binding function and the mismatched thymine repair function appear to be unlinked in plants[202]. Thymine excision activity by plant 5mC DNA glycosylases and mammalian MBD4 may thus be functionally equivalent in preventing deamination of 5mC in 5mCpG-rich regions of plant genomes.

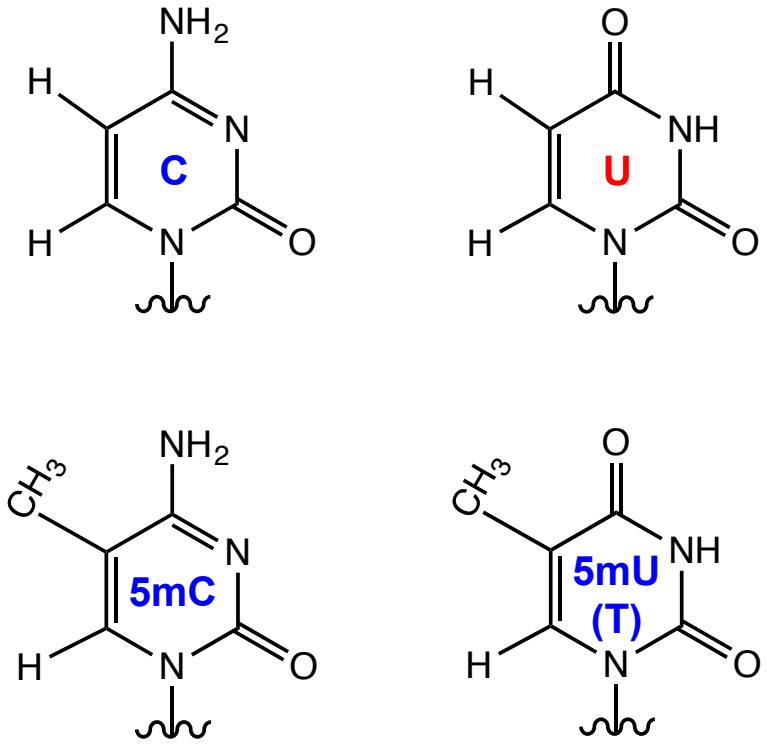


Figure 26. Pyrimidines of nucleic acids.

The methyl group of 5mC is in dynamic equilibrium within the genome. On the other hand, the methyl group of thymine (5mU) is fixed. The unmethylated form of thymine—uracil—is not a building block of genomic DNA (blue) but of RNA (red). For protein-DNA interactions in which the C5-methyl group of thymine is specifically recognized, the C5-methyl group of 5mC may substitute for thymine to enable DNA methylation-dependent binding events.

Role of 5-methylcytosine-binding transcription factors

While 5mC deamination may be mutagenic, many of the eukaryotic transcription factor-binding sites near and upstream of transcription start sites contain TpGs that are thought to have originated from deamination of 5mCpGs[178]. Therefore, 5mC deamination presents an evolutionary process that has shaped the function of many gene promoters. In light of such a perspective, it is noteworthy that the binding sites of AP-1 transcription factors as well as other bZIP family proteins recognize TpGs within the response element sequences. The ability of AP-1 to bind 5mCpG in lieu of TpG within the sequences may thus be an aspect of evolutionary memory.

Alternatively, the above dual binding ability may be biologically driven for specific functions. Previous evidence shows that TRE binding by AP-1 occurs near transcription start sites, whereas meTRE binding predominantly occurs more than 5 kb away from transcription start sites[125]. Further, methylated CRE binding by bZIP family C/EBPα in adipocytes for tissue-specific gene expression occurs at an enhancer region[123]. Another study shows that many active enhancers are methylated[203]. Interestingly, EBV Zta was also shown to bind the genome of EBV-infected host cells in distal regulatory regions[183]. Although DNA methylation status of the regions was not directly addressed in the study, EBV-infected cells can have significantly elevated levels of DNA methylation throughout the host genome, as shown in other studies[204, 205]. Collectively, these data suggest that DNA methylation may control transcription factor-binding events in distant regulatory regions in genomes. Therefore, future studies investigating the role of methylated DNA binding by certain transcription factors during well-defined biological processes can expand our fundamental understanding of the function of DNA methylation.

The recognition of oxidative modifications

The existence of the oxidative derivatives of DNA 5mC generated by Tet dioxygenases poses the question of how the oxidized bases differently influence protein-DNA interactions. Several 5mC-binding transcription factors such as human AP-1 and EBV Zta showed reduced binding affinities for the oxidized bases (See Chapter III). Also, many MBD family proteins showed reduced binding affinities for oxidized bases[8]. ROS1 also showed reduced activities towards 5hmC compared to 5mC, followed by even further decrease in the activity for 5fC and 5caC (See Chapter II). On the other hand, ZnF family WT1 and basic helix-loop-helix (bHLH) family Tcf3-Ascl1 heterodimer can have significantly increased binding affinities for 5caC within specific sequences compared to the unmodified base or other modified bases in the same sequence background[186, 206]. Specifically, the crystal structure of WT1 in complex with oligos containing 5caC displays the specific recognition of the C5-carboxyl group of 5caC[186]. Further, a study utilizing a mass spectrometry pull-down experiment with oligos containing 5mC, 5hmC, 5fC, or 5caC has revealed several proteins that may preferentially bind a particular modified base[79]. Therefore, each form of cytosine modification by methylation and iterative oxidations can serve as a distinct epigenetic signal. Identifying additional readers that specifically recognize a particular oxidative derivative of 5mC would be critical to support this hypothesis.

One of clearly demonstrated ways of recognizing a base by a reader domain involves base flipping, which is a mode of protein-DNA interaction that different classes of proteins have adapted. As previously mentioned, the SRA domain of UHRF1 recognizes 5mC by base flipping[11-13, 195]. While structurally distinct from the SRA domain, DNA glycosylases also flip bases for the extra-helical recognition in the active site. The crystal structure of the TDG catalytic domain in complex with oligos containing 5caC shows that the C5-carboxyl group

of 5caC is specifically recognized in the active site[89]. TDG not only recognizes 5caC, but also recognizes 5fC, thymine, uracil, and 5-hydroxymethyluracil[74, 85, 90]. Introducing a point mutation to the binding pocket can allow the enzyme to be specific for 5caC[106, 207]. Also, some of mammalian 5mC- or 5hmC-binding proteins discovered from the mass spectrometry pull-down experiment are DNA glycosylases such as Ogg1, Nth1, Neil1-2[79]. *In vitro* DNA glycosylase assays of these enzymes revealed a lack of specific activities towards 5mC or 5hmC (See Chapter II). However, they may able to remove the bases in concert with other proteins, as exemplified by the C-terminal domain of ROS1, which is required together with the catalytic glycosylase domain for the base excision activities.

Future directions for ROS1

The study on the C-terminal domain of ROS1 5mC DNA glycosylases from Chapter II has clearly demonstrated that the domain is essential for the enzyme's activity. The C-terminal domain may stabilize the glycosylase domain for reaction and/or engage in DNA recognition to convey the substrate base to the glycosylase domain (**Figure 27**). Isolating individual domains resulted in unstable aggregates that compromised further experiments for characterizing DNA binding or protein-protein interactions. Insertion of a protease recognition sequence between the domains and introducing cleavage by protease unlinked the two domains, though they still tightly associated afterwards.

Introducing an optimal amount of denaturing agent such as guanidine hydrochloride can disrupt the protein-protein interaction between the C-terminal domain and the glycosylase domain while minimally affecting overall folding of each individual domain. Comparing hydrogen-deuterium exchange (HDX) mass spectrometry[208] of the two domains before and after introducing the denaturing agent may reveal the regions involved in the domain-domain interaction. The hypothesis regarding whether the C-terminal domain recognizes DNA can also be tested in a similar fashion whereby HDX mass spectrometry analysis of ROS1 before and after the addition of substrate DNA can be compared to reveal the region in the C-terminal domain engaged in DNA binding. Attempts to crystallize ROS1 with or without substrate DNA have failed so far, however, continued efforts to eventually solve the structure of ROS1 can be pursued. The structure of ROS1-DNA complex would clearly reveal how the two domains are engaged and involved in the recognition of the base substrate.

In addition to the study of understanding the mechanism of ROS1, a separate study for applying ROS1 5mC DNA glycosylase activity for epigenomic editing may be

informative. Because ROS1 and its family of 5mC DNA glycosylases are the only enzymes known to directly remove 5mC for active DNA demethylation, the enzyme can be targeted to a specific locus within a mammalian genome as a fused component of engineered modular proteins such as transcription activator-like effectors (TALE) and ZnF proteins that can bind specific DNA sequences as designed[209, 210]. The idea of epigenomic editing by delivering enzymes to a specific genomic locus to locally alter the chromatin state has been discussed[211]. ROS1 could be ideally applied as a DNA methylation eraser tool for various studies investigating stem cell functions, immune responses, and cancer epigenetics.

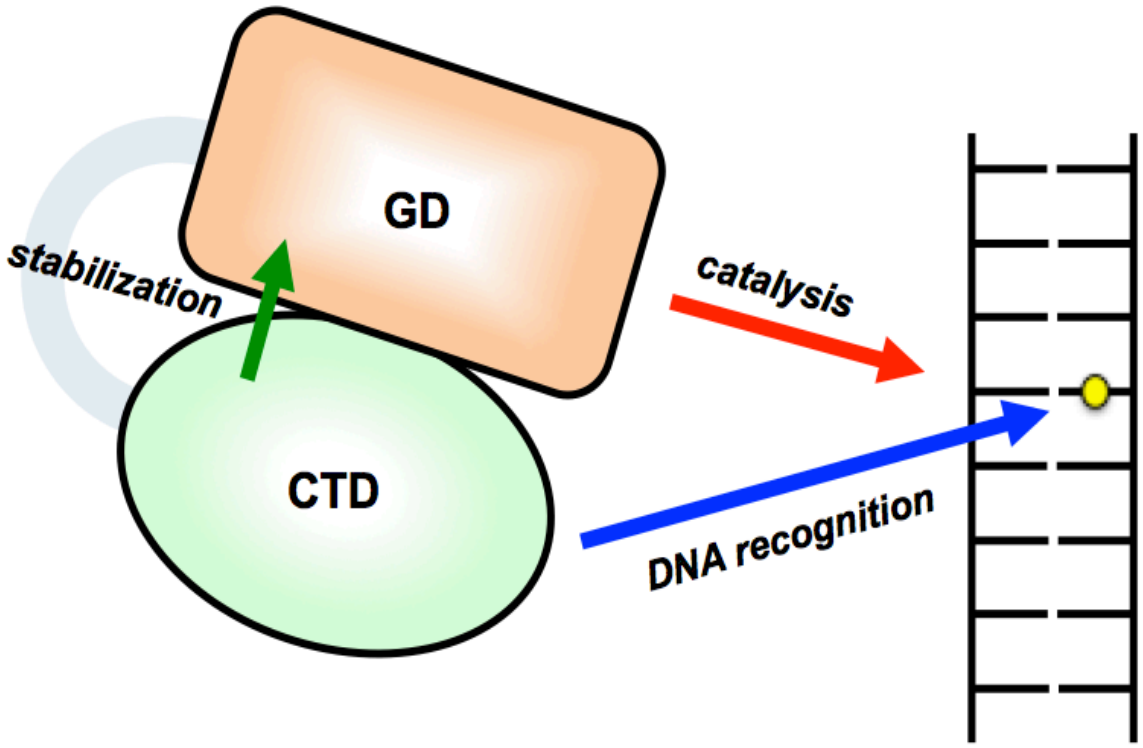


Figure 27. Model for the reaction mechanism of ROS1.

ROS1 CTD stabilizes GD and/or recognizes DNA, specifically or non-specifically, to facilitate the catalytic activity of the substrate base excision. Yellow circle indicates the C5-methyl group of 5mC or T opposite G.

APPENDIX

Table 2. Constructs generated.

pXC #	Protein	Name	Vector	Comments
1135	Arabidopsis thaliana ROS1 (Full-length; 1-1393) (received from Jian-Kang Zhu, PhD)	ROS1 FL	pET28b	Expresses; purifies as crude aggregates with *E. coli* proteins with **overall yield of 0.03 mg per 6 L**; possessed DNA glycosylase activities
1203	*Arabidopsis thaliana* ROS1 N-terminal deletion of residues 1-455		His-Sumo	Expresses well; completes His-Sumo cut; purifies as crude aggregates with *E. coli* proteins with **overall yield of 0.1 mg per 1 L**; possessed DNA glycosylase activities
1214	*Arabidopsis thaliana* ROS1 N-terminal deletion of residues 1-465 and internal deletion of 628-855 replaced by five linker residues (GSSGN)		His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak as a monomer (on 4th prep, **overall yield of 0.2 mg per 1 L**); prep-to-prep variations in purifications with some degradations observed; possessed activities; time-course reactions with different substrates performed successfully for the first time
1233	*Arabidopsis thaliana* ROS1 N-terminal deletion of residues 1-465 and internal deletion of 628-855 replaced by five linker residues (GSSGN) with D971N mutation		His-Sumo	Expresses well; completes His-Sumo cut; did not purify well as pXC1214 (WT); some degradations observed; possessed the AP lyase but not DNA glycosylase activities; **average yield of 0.3 mg per 6 L**
1245	*Arabidopsis thaliana* ROS1 N-terminal deletion of residues 1-465 and internal deletion of 628-855 replaced by five linker residues (GSSGN)		pGEX 6P-1	No expression

1256	*Arabidopsis thaliana* ROS1 N-terminal deletion of residues 1-509 and internal deletion of 628-855 replaced by five linker residues (GSSGN)	ROS1ΔN	His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak as a monomer with an early bump; purifications duplicated with consistency; **average overall yield of 10 mg per 6 L**; time-course reactions with different substrates performed; did not crystallize
1348	*Arabidopsis thaliana* ROS1 N-terminal deletion of residues 1-471 and internal deletion of 628-855 replaced by five linker residues (GSSGN)		His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak as a monomer with a **low overall yield of 0.3 mg per 6 L** compared to pXC1256; possessed activities
1349	*Arabidopsis thaliana* ROS1 N-terminal deletion of residues 1-566 and internal deletion of 628-855 replaced by five linker residues (GSSGN)		His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak as a monomer with a **very low overall yield of 0.03 mg per 6 L** compared to pXC1256; possessed activities
1273	*Arabidopsis thaliana* ROS1 N-terminal deletion of residues 1-509 and internal deletion of 628-855 replaced by five linker residues (GSSGN) with D971N mutation	ROS1ΔN D971N	His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak as a monomer; purifications duplicated with consistency; **average overall yield of 6 mg per 6 L**; refined S200 column purification with **overall yield of 2 mg per 6 L** (for crystallization trials); possessed AP lyase activities; did not crystallize

1274	*Arabidopsis thaliana* ROS1 N-terminal deletion of residues 1-509 and internal deletion of 627-883		His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak as a monomer; purifications duplicated with consistency; **average overall yield of 1 mg per 6 L**; possessed activities
1301	*Arabidopsis thaliana* ROS1 N-terminal deletion of residues 1-509 and internal deletion of 627-883 with D971N mutation		His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak as a monomer; purifications duplicated with consistency; **average overall yield of 1 mg per 6 L**; possessed AP lyase activity; did not crystallize
1276	*Arabidopsis thaliana* ROS1 N-terminal deletion of residues 1-509 and internal deletion of 628-870		His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak as a monomer; **average overall yield of 1 mg per 6 L**; possessed activities
1271	ROS1ΔN with the C-terminal deletion of residues 1171-1393		His-Sumo	Expresses well; completes His-Sumo cut; purifies as crude aggregates with *E. coli* proteins with degradations
1278	ROS1ΔN with the C-terminal deletion of residues 1074-1393	ROS1 GD	His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column with a peak dragging from void volume range and difficult to duplicate; tendency to degrade; **initial yield in the 1st prep of 0.5 mg per 6 L**; possessed AP lyase activity

1297	ROS1 FL N-terminal deletion of residues 1-1080	ROS1 CTD	His-Sumo	Expresses well; completes His-Sumo cut; purifies as aggregate with degradations and elutes in void volume range in the S200 column; **overall yield of 1 mg per 6 L** activates ROS1 GD for 5mC excision
1327	*Arabidopsis thaliana* ROS1 N-terminal deletion of residues 1-509 and internal deletion of residues 628-855 replaced by five linker residues (GSSGN) plus the PreScission recognition sequence (LEVLFQGP) inserted between residues 1076 and 1077	ROS1ΔN:P	His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak as a monomer; purifications duplicated with consistency; **average overall yield of 5 mg per 6 L**; possessed activities before and after the PreScission protease cut; did not crystallize
1333	ROS1ΔN:P with D971N mutation	ROS1ΔN:P D971N	His-Sumo	Expresses well; completes His-Sumo and PreScission cut; purifies through the S200 column and elutes as a single peak as a monomer; purifications duplicated with consistency; **average overall yield of 2 mg per 6 L**; possessed AP lyase activity before and after the PreScission protease cut; did not crystallize
1375	ROS1ΔN:P with I1233M mutation	ROS1ΔN:P I1233M	His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak as a monomer; purifications duplicated with consistency; **average overall yield of 0.5 mg per 6 L**; possessed activities

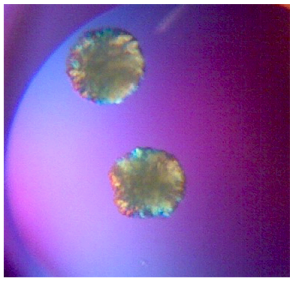
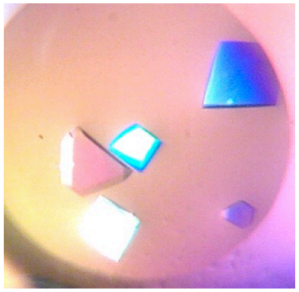
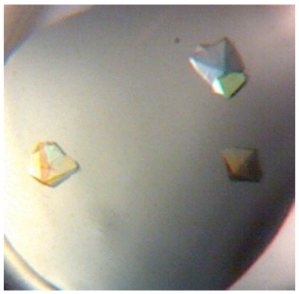
1376	ROS1ΔN:P with W1234R mutation	ROS1ΔN:P W1234R	His-Sumo	Expresses well; completes His-Sumo cut; purifies as aggregates and degradations
1391	ROS1ΔN:P with R1287Q mutation	ROS1ΔN:P R1287Q	His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak as a monomer; purifications duplicated with consistency; **average overall yield of 3.5 mg per 6L**; possessed reduced activities compared to ROS1ΔN:P wild-type
1392	ROS1ΔN:P with D1309N mutation	ROS1ΔN:P D1309N	His-Sumo	Expresses well; completes His-Sumo cut; purifies as aggregates and degradations
1316	human MutY homolog (Full-length 1-535)		His-Sumo	Expresses well; totally insoluble
1318	human MutY homolog N-terminal deletion of residues 1-64		His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak as a monomer with some degradations; overall yield of 0.5 mg per 6 L
1321	mouse MutY homolog (Full-length 1-515)	mMYH	His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak as a monomer; over 90% protein lost during purification due to human error with **resulting yield of 0.7 mg per 6 L**; possessed activities

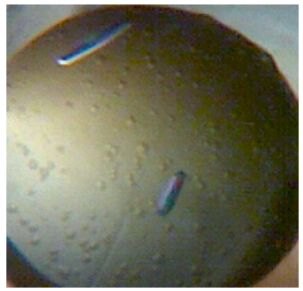
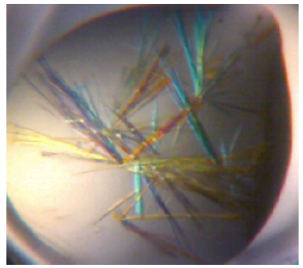
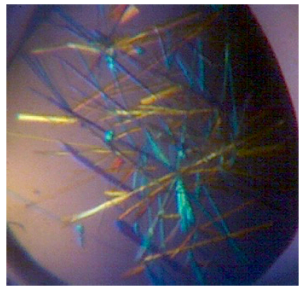
1331	mMYH with the PreScission recognition sequence insert between 322 and 323		His-Sumo	Expresses well; completes His-Sumo and PreScission cut; purifies through the S200 column and elutes as a single peak as a monomer; degradations after the S200 column;
1332	mMYH with the PreScission recognition sequence (LEVLFQGP) inserted between residue 330 and 331	mMYH:P	His-Sumo	Expresses well; completes His-Sumo and PreScission cut; purifies through the S200 column and elutes as a single peak as a monomer; **overall yield of 10 mg per 6 L**; possessed activities
1338	Hybrid of N-terminal mMYH GD (residues 1-330) followed by the PreScission recognition sequence (LEVLFQGP) and C-terminal ROS1 CTD (residues 1081-1393)	mMYH-ROS1 hybrid	His-Sumo	Expresses well; completes His-Sumo cut and PreScission cut; purifies through the S200 column and elutes as a single peak as a monomer with nicks; possessed mMYH GD activities
1201	*Arabidopsis thaliana* DML3 Full-length		pET28a	Expresses moderately; low solubility; minimally purifies with degradations
1354	*Arabidopsis thaliana* DML3 residues 395-1105 (DML3ΔN)	DML3ΔN	His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak as a monomer with minor degradations; **overall yield of 2 mg per 6 L**; possessed activities
1355	Arabidopsis thaliana DML3 residues 395-775 (DML3 GD)	DML3 GD	His-Sumo	Expresses well; completes His-Sumo cut; purifies as aggregate with degradations

1367	Arabidopsis thaliana DML3 residues 395-1105 with the PreScission recognition sequence (LEVLFQGP) inserted between residues 775 and 776 (DML3ΔN:P)	DML3ΔN:P	His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak as a monomer with minor degradations; **overall yield of 3 mg per 6 L**; possessed activities
1270	pGEX-2T Uracil DNA glycosylase inhibitor protein (received from Jeffrey Cohen, MD at NIH)	GST-Ugi	pGEX-2T	Expresses and purifies well on the GST column; estimated yield to be **more than 10 mg per 2 L**
1394	human c-FOS bZIP domain (residues 135-200)		His-Sumo	No expression
1395	human c-JUN bZIP domain (residues 254-315)		His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak; formed complex with pXC1396 purified product with **overall yield of 4 mg per 6 L**
1396	human c-FOS bZIP domain (residues 135-200)		pGEX 6P-1	Expresses well; completes PreScission cut; formed complex with pXC1395 purified product with **overall yield of 4 mg per 6 L**
1397	Epstein-Barr virus Zta bZIP domain (residues 175-236)		His-Sumo	Expresses well; largely precipitates with His-Sumo cut without 150 mM ammonium acetate; purifications scheme optimized with pXC1416

1398	human c-JUN bZIP domain (254-315) (pXC1395) C269S	Jun	His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak; **overall yield of 30 mg per 6 L**; crystal structure in DNA-bound form solved
1399	human c-FOS bZIP domain (135-200) (pXC1399) C154S	Fos	pGEX 6P-1	Expresses well; completes PreScission cut; purifies well and formed complex with pXC1396 purified product with **overall yield of 17 mg per 6 L**
1416	Epstein-Barr virus Zta bZIP domain (residues 175-236) C189S	Zta	His-Sumo	Expresses well; completes His-Sumo cut without precipitation if digested in presence of 150 mM ammonium acetate; purifies through the S200 column and elutes as a single peak; **overall yield of 10 mg per 6 L**; crystal structure in DNA-bound form solved
1440	Jun with A266S mutation	Jun A266S	His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak; **overall yield of 35 mg per 6 L**
1455	Zta with S186A mutation	Zta S186A	His-Sumo	Expresses well; purified as His-Sumo tagged form; purifies through the S200 column and elutes as a single peak; **overall yield of 30 mg per 6 L**
1456	Jun with A265S mutation	Jun A265S	His-Sumo	Expresses well; completes His-Sumo cut; purifies through the S200 column and elutes as a single peak; **overall yield of 30 mg per 6 L**

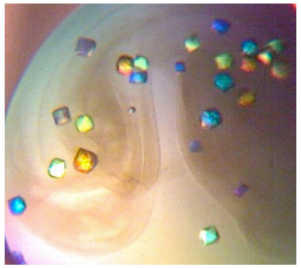
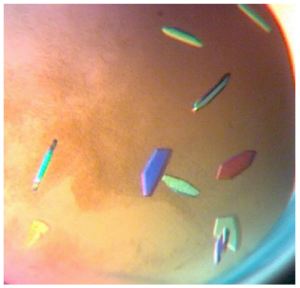
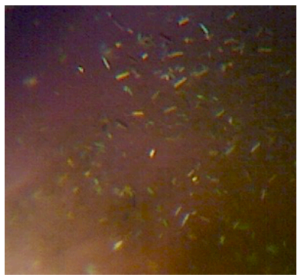
Table 3. Oligonucleotides used for Fos/Jun and Jun/Jun crystallization trials.

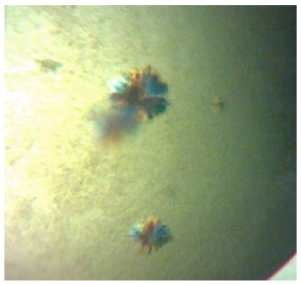
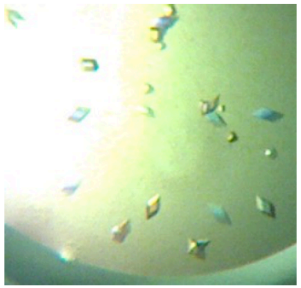
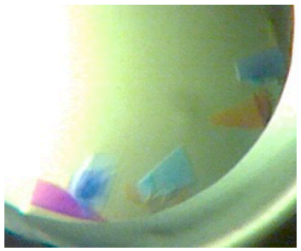
Name	Sequence*	Comments	
19+1	TTCTCCTA**TGACTCG**TCCAT AGAGGAT**ACTGAGM**AGGTAA		Protein: Fos/Jun (pXC1398/pXC1399) Condition: 50 mM Bis-Tris (pH 6.4), 10% PEG400, 55 mM $MgCl_2$, 2.0 mM Spermine, 5 mM DTT, and 300 mM NaCl Conditions and oligonucleotides designed based on the reported Fos/Jun structure (PDB 1FOS) and pertinent information; did not obtain singly separated crystals
			Protein: Fos/Jun (pXC1398/pXC1399) Condition: No mother liquor, Protein/DNA complex solution containing 10 mM Tris (pH 7.5-8.0), 50-100 mM NaCl Crystals appeared as the drop underwent a minor evaporation in the well without mother liquor; diffracted to 5 Å by home X-ray; data not collected
			Protein: Fos/Jun (pXC1398/pXC1399) Condition: 0.1 M Hepes (pH 7.5), 3.0 M NaCl Crystals in similar shape appeared in other similar conditions in buffer with neutral pH and high concentration of salt; diffracted to 3 Å with C2 space group (97 x 83 x 69, $\beta = 131°$); molecular replacement by PHENIX only found protein dimers without DNA

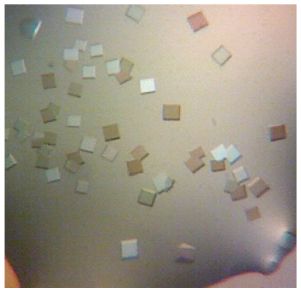
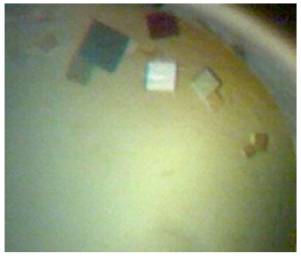
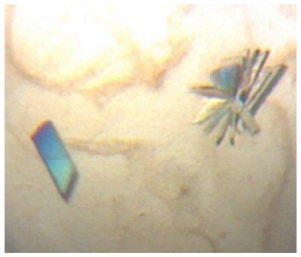
19	TCTCCTA**TGACTCG**TCCAT AGAGGAT**ACTGAGM**AGGTA		Protein: Fos/Jun (pXC1398/pXC1399) Condition: 0.2 M KI, 20% PEG3350 Did not diffract; no significant crystals formed throughout different screening conditions
18+1**	TCTCCTA**TGACTCG**TCCAT GAGGAT**ACTGAGM**AGGTAA		Protein: Jun/Jun (pXC1398) Condition: 0.2 M $NaH_2PO4•H2O$, 20% PEG3350 Crystals appeared in 1- to 2-day period; diffracted to 3-4 Å by home X-ray; crystals with similar shape appeared in similar conditions characterized by buffer with pH near 5 with 20% PEG3350
			Protein: Jun/Jun (pXC1398) Condition: 0.05 M Citric acid, 0.05 M Bis-Tris propane (pH 5.0), 16% PEG3350 Crystals appeared in 1-day period; diffracted to 3 Å by home X-ray; the condition duplicated in hanging drop and crystals appeared in an optimized condition diffracted to 1.9 Å by APS ID-22 X-ray

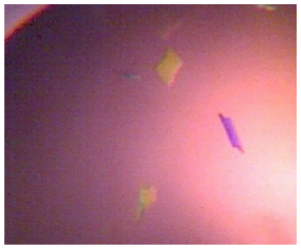
* The top sequences are from 5' to 3' and the bottom sequences are from 3' to 5'; M indicates 5mC; meTRE sequence is in bold.
** Formed a crystal that diffracted to 1.9 Å.

Table 4. Summary of oligonucleotides used for Zta/Zta crystallization trials.

Name	Sequence*	Comments	
18+1	TACTTCA**TCGCTCA**GTGCT TGAAGT**AGMGAGT**CACGAA		Condition: 20 mM Tris (pH 7.5), 0.2 M CaCl2, 16% PEG3350 Crystals appeared within 3-5 days; diffracted to 10 Å with APS ID-22 X-ray; attempt to optimize conditions did not yield better-diffracting crystals
			Condition: 0.2 M KH_2PO_4, 20% PEG3350 Crystals appeared after several weeks; diffracted 4 Å in home X-ray; mounted crystals lost due to machine error
			Condition: 0.2 M $NaH_2PO4•H_2O$, 20% PEG3350 Tiny crystals appeared after several weeks; diffracted to 4 Å in APS ID-22 X-ray; space group of P1 (37 x 46 x 57, 92° 108° 105 °); molecular replacement found Protein-DNA complex

18+1T	**TACTTCATCGCTCA**GTGCT TGAAGT**AGMGAGT**CACGA	Small irregularly shaped crystals appeared in 0.2 M NH_4SO_4, 0.1 M Sodium cacodylate trihydrate (pH 6.5), 30% PEG8000, Sodium citrate tribasic dehydrate (pH 6.5), 20% 2-Propanol, 20% PEG4000; and 0.2 M NH_4H_2PO4, 20% PEG3350	
18+1B	ACTTCA**TCGCTCA**GTGCT TGAAGT**AGMGAGT**CACGAA		Condition: 0.1 M Succinic acid (pH 7.0), 12% PEG3350 Irregularly shaped crystals appeared in similar conditions with buffer typically containing 20% PEG3350; diffracted weak
18	ACTTCA**TCGCTCA**GTGCT TGAAGT**AGMGAGT**CACGAA		Condition: 0.1 M Imidazol pH7.0, 2% PEG400, 24% PEG-MME5000 Singly shaped crystals appeared within 3 days; initially diffracted to 6 Å in APS ID-22 X-ray and later to 3.3 Å after optimization of the condition; space group of P 3_2 2 1 (70 x 70 x 170); molecular replacement by PHENIX found two protein-DNA complexes in an asymmetric unit
			Condition: 0.15 M Li_2SO_4• H_2O, 0.1 M Citric acid (pH 3.5), 18% PEG6000 Plate-like crystals appeared within 3-5 days; diffracted to 6 Å in APS ID-22 X-ray

17+1**	TCTTCA**TMGCTCA**GTGCT GAAGT**AGMGAGT**CACGAA		Condition: Condition: 0.1 M Na_2HPO_4:Citric acid (pH 4.2), 30% PEG300 Singly separated crystals appeared after 3-5 days; diffracted to 6 Å in APS ID-22 X-ray
			Condition: 0.2 M CaCl2, 20% 2-propanol Singly separated crystals appeared after 3-5 days; diffracted to 8 Å in APS ID-22 X-ray
			Condition: 0.2 M $NaH_2PO4•H_2O$, 20% PEG3350 A single rod-shaped crystal on the left that diffracted to 2.4 Å in APS ID-22 X-ray appeared after several weeks; other cluttered crystals appeared around a similar time and diffracted to 5 Å

17+1T	TCTTCA**TMGCTCA**GTGCT GAAGT**AGMGAGT**CACGA	No crystal formed	
17+1B	CTTCA**TMGCTCA**GTGCT GAAGT**AGMGAGT**CACGAAA	Small irregularly shaped crystals appeared in 4% Tacsimate™ (pH 4.0), 23% PEG3350; 0.1 M Imidazol pH7.0, 2% PEG400, 24% PEG-MME5000; 4% 2-Methyl-2,4-pentanediol, 0.1 M Citric acid (pH 3.5), 20% PEG1500	
17	CTTCA**TMGCTCA**GTGCT GAAGT**AGMGAGT**CACGAA		Condition: 0.02 M $ZnCl_2$, 20% PEG3350 Singly separated crystals appeared after 3-5 days; diffracted to 7 Å in APS ID-22 X-ray
16+1	TTTCA**TMGCTCA**GTGCT AAGT**AGMGAGT**CACGAA	Crystals appeared in a number of conditions containing 20%PEG3350 and buffer with pH range of 5-6 after using crystals seeds from the Zta-DNA complex with 17+1 in the condition containing 0.2 M $NaH_2PO4•H_2O$, 20% PEG3350; crystals diffracted to 7 Å in APS ID-22 X-ray	

* The top sequences are from 5' to 3' and the bottom sequences are from 3' to 5'; M indicates 5mC; meZRE-2 sequence is in bold.
** Formed a crystal that diffracted to 2.4 Å.

REFERENCES

1. Goll, M. G., and Bestor, T. H. (2005) Eukaryotic cytosine methyltransferases, *Annual review of biochemistry 74*, 481-514.

2. Kumar, S., Cheng, X., Klimasauskas, S., Mi, S., Posfai, J., Roberts, R. J., and Wilson, G. G. (1994) The DNA (cytosine-5) methyltransferases, *Nucleic acids research 22*, 1-10.

3. Wu, J. C., and Santi, D. V. (1987) Kinetic and catalytic mechanism of HhaI methyltransferase, *The Journal of biological chemistry 262*, 4778-4786.

4. Wu, J. C., and Santi, D. V. (1985) On the mechanism and inhibition of DNA cytosine methyltransferases, *Progress in clinical and biological research 198*, 119-129.

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

6. Roberts, R. J., Vincze, T., Posfai, J., and Macelis, D. (2015) REBASE--a database for DNA restriction and modification: enzymes, genes and genomes, *Nucleic acids research 43*, D298-299.

7. Law, J. A., and Jacobsen, S. E. (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals, *Nature reviews. Genetics 11*, 204-220.

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

9. Song, J., Rechkoblit, O., Bestor, T. H., and Patel, D. J. (2011) Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation, *Science 331*, 1036-1040.

10. Hermann, A., Goyal, R., and Jeltsch, A. (2004) The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites, *The Journal of biological chemistry 279*, 48350-48359.

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

12. Avvakumov, G. V., Walker, J. R., Xue, S., Li, Y., Duan, S., Bronner, C., Arrowsmith, C. H., and Dhe-Paganon, S. (2008) Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1, *Nature 455*, 822-825.

13. Hashimoto, H., Horton, J. R., Zhang, X., Bostick, M., Jacobsen, S. E., and Cheng, X. (2008) The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix, *Nature 455*, 826-829.

14. Sharif, J., Muto, M., Takebayashi, S., Suetake, I., Iwamatsu, A., Endo, T. A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., Okamura, K., Tajima, S., Mitsuya, K., Okano, M., and Koseki, H. (2007) The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA, *Nature 450*, 908-912.

15. Bostick, M., Kim, J. K., Esteve, P. O., Clark, A., Pradhan, S., and Jacobsen, S. E. (2007) UHRF1 plays a role in maintaining DNA methylation in mammalian cells, *Science 317*, 1760-1764.

16. Suetake, I., Miyazaki, J., Murakami, C., Takeshima, H., and Tajima, S. (2003) Distinct enzymatic properties of recombinant mouse DNA methyltransferases Dnmt3a and Dnmt3b, *Journal of biochemistry 133*, 737-744.

17. Gowher, H., and Jeltsch, A. (2001) Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non-processive

manner and also methylates non-CpG [correction of non-CpA] sites, *Journal of molecular biology 309*, 1201-1208.

18. Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P., and Jaenisch, R. (2000) Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a, *Proceedings of the National Academy of Sciences of the United States of America 97*, 5237-5242.
19. Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A., and Cheng, X. (2007) Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation, *Nature 449*, 248-251.
20. Gowher, H., Liebert, K., Hermann, A., Xu, G., and Jeltsch, A. (2005) Mechanism of stimulation of catalytic activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L, *The Journal of biological chemistry 280*, 13341-13348.
21. Zhang, Y., Jurkowska, R., Soeroes, S., Rajavelu, A., Dhayalan, A., Bock, I., Rathert, P., Brandt, O., Reinhardt, R., Fischle, W., and Jeltsch, A. (2010) Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail, *Nucleic acids research 38*, 4246-4253.
22. Ooi, S. K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S. P., Allis, C. D., Cheng, X., and Bestor, T. H. (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA, *Nature 448*, 714-717.
23. Cheng, J., Yang, Y., Fang, J., Xiao, J., Zhu, T., Chen, F., Wang, P., Li, Z., Yang, H., and Xu, Y. (2013) Structural insight into coordinated recognition of trimethylated histone H3 lysine 9 (H3K9me3) by the plant homeodomain (PHD) and tandem tudor

domain (TTD) of UHRF1 (ubiquitin-like, containing PHD and RING finger domains, 1) protein, *The Journal of biological chemistry 288*, 1329-1339.

24. Rothbart, S. B., Krajewski, K., Nady, N., Tempel, W., Xue, S., Badeaux, A. I., Barsyte-Lovejoy, D., Martinez, J. Y., Bedford, M. T., Fuchs, S. M., Arrowsmith, C. H., and Strahl, B. D. (2012) Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation, *Nature structural & molecular biology 19*, 1155-1160.
25. Nady, N., Lemak, A., Walker, J. R., Avvakumov, G. V., Kareta, M. S., Achour, M., Xue, S., Duan, S., Allali-Hassani, A., Zuo, X., Wang, Y. X., Bronner, C., Chedin, F., Arrowsmith, C. H., and Dhe-Paganon, S. (2011) Recognition of multivalent histone states associated with heterochromatin by UHRF1 protein, *The Journal of biological chemistry 286*, 24300-24311.
26. Chan, S. W., Henderson, I. R., and Jacobsen, S. E. (2005) Gardening the genome: DNA methylation in Arabidopsis thaliana, *Nature reviews. Genetics 6*, 351-360.
27. Weber, M., Davies, J. J., Wittig, D., Oakeley, E. J., Haase, M., Lam, W. L., and Schubeler, D. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells, *Nature genetics 37*, 853-862.
28. Jones, P. A. (1999) The DNA methylation paradox, *Trends in genetics : TIG 15*, 34-37.
29. Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning, *Nature 452*, 215-219.
30. Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W., Chen, H., Henderson, I. R., Shinn, P., Pellegrini, M., Jacobsen, S. E., and Ecker, J. R. (2006) Genome-wide high-

resolution mapping and functional analysis of DNA methylation in arabidopsis, *Cell 126*, 1189-1201.

31. Slotkin, R. K., and Martienssen, R. (2007) Transposable elements and the epigenetic regulation of the genome, *Nature reviews. Genetics 8*, 272-285.

32. Yoder, J. A., Walsh, C. P., and Bestor, T. H. (1997) Cytosine methylation and the ecology of intragenomic parasites, *Trends in genetics : TIG 13*, 335-340.

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

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

35. Saxonov, S., Berg, P., and Brutlag, D. L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters, *Proceedings of the National Academy of Sciences of the United States of America 103*, 1412-1417.

36. Rollins, R. A., Haghighi, F., Edwards, J. R., Das, R., Zhang, M. Q., Ju, J., and Bestor, T. H. (2006) Large-scale structure of genomic methylation patterns, *Genome research 16*, 157-163.

37. Illingworth, R. S., and Bird, A. P. (2009) CpG islands--'a rough guide', *FEBS letters 583*, 1713-1720.

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

39. Zhu, J., He, F., Hu, S., and Yu, J. (2008) On the nature of human housekeeping genes, *Trends in genetics : TIG 24*, 481-484.

40. Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T., Low, H. M., Kin Sung, K. W., Rigoutsos, I., Loring, J., and Wei, C. L. (2010) Dynamic changes in the human methylome during differentiation, *Genome research 20*, 320-331.

41. Ball, M. P., Li, J. B., Gao, Y., Lee, J. H., LeProust, E. M., Park, I. H., Xie, B., Daley, G. Q., and Church, G. M. (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells, *Nature biotechnology 27*, 361-368.

42. Hsieh, T. F., Ibarra, C. A., Silva, P., Zemach, A., Eshed-Williams, L., Fischer, R. L., and Zilberman, D. (2009) Genome-wide demethylation of Arabidopsis endosperm, *Science 324*, 1451-1454.

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

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

45. Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q. M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences, *Nature 462*, 315-322.

46. Yagi, S., Hirabayashi, K., Sato, S., Li, W., Takahashi, Y., Hirakawa, T., Wu, G., Hattori, N., Hattori, N., Ohgane, J., Tanaka, S., Liu, X. S., and Shiota, K. (2008) DNA methylation profile of tissue-dependent and differentially methylated regions (T-

DMRs) in mouse promoter regions demonstrating tissue-specific gene expression, *Genome research 18*, 1969-1978.

47. Voon, H. P., and Gibbons, R. J. (2016) Maintaining memory of silencing at imprinted differentially methylated regions, *Cellular and molecular life sciences : CMLS 73*, 1871-1879.

48. Nordin, M., Bergman, D., Halje, M., Engstrom, W., and Ward, A. (2014) Epigenetic regulation of the Igf2/H19 gene cluster, *Cell proliferation 47*, 189-199.

49. Jones, P. A., and Baylin, S. B. (2007) The epigenomics of cancer, *Cell 128*, 683-692.

50. Laird, P. W., and Jaenisch, R. (1996) The role of DNA methylation in cancer genetic and epigenetics, *Annual review of genetics 30*, 441-464.

51. Zemach, A., and Grafi, G. (2007) Methyl-CpG-binding domain proteins in plants: interpreters of DNA methylation, *Trends in plant science 12*, 80-85.

52. Hendrich, B., and Bird, A. (1998) Identification and characterization of a family of mammalian methyl-CpG binding proteins, *Molecular and cellular biology 18*, 6538-6547.

53. Sarraf, S. A., and Stancheva, I. (2004) Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly, *Molecular cell 15*, 595-605.

54. Ng, H. H., Zhang, Y., Hendrich, B., Johnson, C. A., Turner, B. M., Erdjument-Bromage, H., Tempst, P., Reinberg, D., and Bird, A. (1999) MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex, *Nature genetics 23*, 58-61.

55. Nan, X., Ng, H. H., Johnson, C. A., Laherty, C. D., Turner, B. M., Eisenman, R. N., and Bird, A. (1998) Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex, *Nature 393*, 386-389.

56. Muotri, A. R., Marchetto, M. C., Coufal, N. G., Oefner, R., Yeo, G., Nakashima, K., and Gage, F. H. (2010) L1 retrotransposition in neurons is modulated by MeCP2, *Nature 468*, 443-446.

57. Carouge, D., Host, L., Aunis, D., Zwiller, J., and Anglard, P. (2010) CDKL5 is a brain MeCP2 target gene regulated by DNA methylation, *Neurobiology of disease 38*, 414-424.

58. Deng, V., Matagne, V., Banine, F., Frerking, M., Ohliger, P., Budden, S., Pevsner, J., Dissen, G. A., Sherman, L. S., and Ojeda, S. R. (2007) FXYD1 is an MeCP2 target gene overexpressed in the brains of Rett syndrome patients and Mecp2-null mice, *Human molecular genetics 16*, 640-650.

59. Zhou, Z., Hong, E. J., Cohen, S., Zhao, W. N., Ho, H. Y., Schmidt, L., Chen, W. G., Lin, Y., Savner, E., Griffith, E. C., Hu, L., Steen, J. A., Weitz, C. J., and Greenberg, M. E. (2006) Brain-specific phosphorylation of MeCP2 regulates activity-dependent Bdnf transcription, dendritic growth, and spine maturation, *Neuron 52*, 255-269.

60. Yu, F., Zingler, N., Schumann, G., and Stratling, W. H. (2001) Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription, *Nucleic acids research 29*, 4493-4501.

61. Amir, R. E., Van den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., and Zoghbi, H. Y. (1999) Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2, *Nature genetics 23*, 185-188.

62. Liu, C., Teng, Z. Q., Santistevan, N. J., Szulwach, K. E., Guo, W., Jin, P., and Zhao, X. (2010) Epigenetic regulation of miR-184 by MBD1 governs neural stem cell proliferation and differentiation, *Cell stem cell 6*, 433-444.

63. Clouaire, T., de Las Heras, J. I., Merusi, C., and Stancheva, I. (2010) Recruitment of MBD1 to target genes requires sequence-specific interaction of the MBD domain with methylated DNA, *Nucleic acids research 38*, 4620-4634.

64. Li, X., Barkho, B. Z., Luo, Y., Smrt, R. D., Santistevan, N. J., Liu, C., Kuwabara, T., Gage, F. H., and Zhao, X. (2008) Epigenetic regulation of the stem cell mitogen Fgf-2 by Mbd1 in adult neural stem/progenitor cells, *The Journal of biological chemistry 283*, 27644-27652.

65. Barr, H., Hermann, A., Berger, J., Tsai, H. H., Adie, K., Prokhortchouk, A., Hendrich, B., and Bird, A. (2007) Mbd2 contributes to DNA methylation-directed repression of the Xist gene, *Molecular and cellular biology 27*, 3750-3757.

66. Kransdorf, E. P., Wang, S. Z., Zhu, S. Z., Langston, T. B., Rupon, J. W., and Ginder, G. D. (2006) MBD2 is a critical component of a methyl cytosine-binding protein complex isolated from primary erythroid cells, *Blood 108*, 2836-2845.

67. Goedecke, K., Pignot, M., Goody, R. S., Scheidig, A. J., and Weinhold, E. (2001) Structure of the N6-adenine DNA methyltransferase M.TaqI in complex with DNA and a cofactor analog, *Nature structural biology 8*, 121-125.

68. Blumenthal, R. M., Gregory, S. A., and Cooperider, J. S. (1985) Cloning of a restriction-modification system from Proteus vulgaris and its use in analyzing a methylase-sensitive phenotype in Escherichia coli, *Journal of bacteriology 164*, 501-509.

69. Low, D. A., Weyand, N. J., and Mahan, M. J. (2001) Roles of DNA adenine methylation in regulating bacterial gene expression and virulence, *Infection and immunity 69*, 7197-7204.

70. Warren, R. A. (1980) Modified bases in bacteriophage DNAs, *Annual review of microbiology 34*, 137-158.

71. Wyatt, G. R., and Cohen, S. S. (1953) The bases of the nucleic acids of some bacterial and animal viruses: the occurrence of 5-hydroxymethylcytosine, *The Biochemical journal 55*, 774-782.

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

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

74. He, Y. F., Li, B. Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., Sun, Y., Li, X., Dai, Q., Song, C. X., Zhang, K., He, C., and Xu, G. L. (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA, *Science 333*, 1303-1307.

75. Ito, S., Shen, L., Dai, Q., Wu, S. C., Collins, L. B., Swenberg, J. A., He, C., and Zhang, Y. (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine, *Science 333*, 1300-1303.

76. Globisch, D., Munzel, M., Muller, M., Michalakis, S., Wagner, M., Koch, S., Bruckl, T., Biel, M., and Carell, T. (2010) Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates, *PloS one 5*, e15367.

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

78. Pfaffeneder, T., Hackner, B., Truss, M., Munzel, M., Muller, M., Deiml, C. A., Hagemeier, C., and Carell, T. (2011) The discovery of 5-formylcytosine in embryonic stem cell DNA, *Angewandte Chemie 50*, 7008-7012.

79. Spruijt, C. G., Gnerlich, F., Smits, A. H., Pfaffeneder, T., Jansen, P. W., Bauer, C., Munzel, M., Wagner, M., Muller, M., Khan, F., Eberl, H. C., Mensinga, A., Brinkman, A. B., Lephikov, K., Muller, U., Walter, J., Boelens, R., van Ingen, H., Leonhardt, H., Carell, T., and Vermeulen, M. (2013) Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives, *Cell 152*, 1146-1159.

80. Zhu, J. K. (2009) Active DNA demethylation mediated by DNA glycosylases, *Annual review of genetics 43*, 143-166.

81. Rougier, N., Bourc'his, D., Gomes, D. M., Niveleau, A., Plachot, M., Paldi, A., and Viegas-Pequignot, E. (1998) Chromosome methylation patterns during mammalian preimplantation development, *Genes & development 12*, 2108-2113.

82. Monk, M., Adams, R. L., and Rinaldi, A. (1991) Decrease in DNA methylase activity during preimplantation development in the mouse, *Development 112*, 189-192.

83. Chen, C. C., Wang, K. Y., and Shen, C. K. (2013) DNA 5-methylcytosine demethylation activities of the mammalian DNA methyltransferases, *The Journal of biological chemistry 288*, 9084-9091.

84. Chen, C. C., Wang, K. Y., and Shen, C. K. (2012) The mammalian de novo DNA methyltransferases DNMT3A and DNMT3B are also DNA 5-hydroxymethylcytosine dehydroxymethylases, *The Journal of biological chemistry 287*, 33116-33121.

85. Hashimoto, H., Hong, S., Bhagwat, A. S., Zhang, X., and Cheng, X. (2012) Excision of 5-hydroxymethyluracil and 5-carboxylcytosine by the thymine DNA glycosylase

domain: its structural basis and implications for active DNA demethylation, *Nucleic acids research 40*, 10203-10214.

86. Hashimoto, H., Zhang, X., and Cheng, X. (2012) Excision of thymine and 5-hydroxymethyluracil by the MBD4 DNA glycosylase domain: structural basis and implications for active DNA demethylation, *Nucleic acids research 40*, 8276-8284.

87. Cortellino, S., Xu, J., Sannai, M., Moore, R., Caretti, E., Cigliano, A., Le Coz, M., Devarajan, K., Wessels, A., Soprano, D., Abramowitz, L. K., Bartolomei, M. S., Rambow, F., Bassi, M. R., Bruno, T., Fanciulli, M., Renner, C., Klein-Szanto, A. J., Matsumoto, Y., Kobi, D., Davidson, I., Alberti, C., Larue, L., and Bellacosa, A. (2011) Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair, *Cell 146*, 67-79.

88. Rai, K., Huggins, I. J., James, S. R., Karpf, A. R., Jones, D. A., and Cairns, B. R. (2008) DNA demethylation in zebrafish involves the coupling of a deaminase, a glycosylase, and gadd45, *Cell 135*, 1201-1212.

89. Zhang, L., Lu, X., Lu, J., Liang, H., Dai, Q., Xu, G. L., Luo, C., Jiang, H., and He, C. (2012) Thymine DNA glycosylase specifically recognizes 5-carboxylcytosine-modified DNA, *Nature chemical biology 8*, 328-330.

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

91. Song, C. X., Szulwach, K. E., Dai, Q., Fu, Y., Mao, S. Q., Lin, L., Street, C., Li, Y., Poidevin, M., Wu, H., Gao, J., Liu, P., Li, L., Xu, G. L., Jin, P., and He, C. (2013) Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming, *Cell 153*, 678-691.

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

93. Vairapandi, M. (2004) Characterization of DNA demethylation in normal and cancerous cell lines and the regulatory role of cell cycle proteins in human DNA demethylase activity, *Journal of cellular biochemistry 91*, 572-583.

94. Vairapandi, M., Liebermann, D. A., Hoffman, B., and Duker, N. J. (2000) Human DNA-demethylating activity: a glycosylase associated with RNA and PCNA, *Journal of cellular biochemistry 79*, 249-260.

95. Vairapandi, M., and Duker, N. J. (1993) Enzymic removal of 5-methylcytosine from DNA by a human DNA-glycosylase, *Nucleic acids research 21*, 5323-5327.

96. Cannon, S. V., Cummings, A., and Teebor, G. W. (1988) 5-Hydroxymethylcytosine DNA glycosylase activity in mammalian tissue, *Biochemical and biophysical research communications 151*, 1173-1179.

97. Ortega-Galisteo, A. P., Morales-Ruiz, T., Ariza, R. R., and Roldan-Arjona, T. (2008) Arabidopsis DEMETER-LIKE proteins DML2 and DML3 are required for appropriate distribution of DNA methylation marks, *Plant molecular biology 67*, 671-681.

98. Gehring, M., Huh, J. H., Hsieh, T. F., Penterman, J., Choi, Y., Harada, J. J., Goldberg, R. B., and Fischer, R. L. (2006) DEMETER DNA glycosylase establishes MEDEA polycomb gene self-imprinting by allele-specific demethylation, *Cell 124*, 495-506.

99. Gong, Z., Morales-Ruiz, T., Ariza, R. R., Roldan-Arjona, T., David, L., and Zhu, J. K. (2002) ROS1, a repressor of transcriptional gene silencing in Arabidopsis, encodes a DNA glycosylase/lyase, *Cell 111*, 803-814.

100. Ponferrada-Marin, M. I., Parrilla-Doblas, J. T., Roldan-Arjona, T., and Ariza, R. R. (2011) A discontinuous DNA glycosylase domain in a family of enzymes that excise 5-methylcytosine, *Nucleic acids research 39*, 1473-1484.

101. Mok, Y. G., Uzawa, R., Lee, J., Weiner, G. M., Eichman, B. F., Fischer, R. L., and Huh, J. H. (2010) Domain structure of the DEMETER 5-methylcytosine DNA glycosylase, *Proceedings of the National Academy of Sciences of the United States of America 107*, 19225-19230.

102. Ponferrada-Marin, M. I., Roldan-Arjona, T., and Ariza, R. R. (2009) ROS1 5-methylcytosine DNA glycosylase is a slow-turnover catalyst that initiates DNA demethylation in a distributive fashion, *Nucleic acids research 37*, 4264-4274.

103. Martinez-Macias, M. I., Qian, W., Miki, D., Pontes, O., Liu, Y., Tang, K., Liu, R., Morales-Ruiz, T., Ariza, R. R., Roldan-Arjona, T., and Zhu, J. K. (2012) A DNA 3' phosphatase functions in active DNA demethylation in Arabidopsis, *Molecular cell 45*, 357-370.

104. Hong, S., Hashimoto, H., Kow, Y. W., Zhang, X., and Cheng, X. (2014) The carboxy-terminal domain of ROS1 is essential for 5-methylcytosine DNA glycosylase activity, *Journal of molecular biology 426*, 3703-3712.

105. Jang, H., Shin, H., Eichman, B. F., and Huh, J. H. (2014) Excision of 5-hydroxymethylcytosine by DEMETER family DNA glycosylases, *Biochemical and biophysical research communications 446*, 1067-1072.

106. Hashimoto, H., Zhang, X., and Cheng, X. (2013) Selective excision of 5-carboxylcytosine by a thymine DNA glycosylase mutant, *Journal of molecular biology 425*, 971-976.

107. Brooks, S. C., Adhikary, S., Rubinson, E. H., and Eichman, B. F. (2013) Recent advances in the structural mechanisms of DNA glycosylases, *Biochimica et biophysica acta 1834*, 247-271.

108. Boyes, J., and Bird, A. (1991) DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein, *Cell 64*, 1123-1134.

109. Perini, G., Diolaiti, D., Porro, A., and Della Valle, G. (2005) In vivo transcriptional regulation of N-Myc target genes is controlled by E-box methylation, *Proceedings of the National Academy of Sciences of the United States of America 102*, 12117-12122.

110. Chen, B., He, L., Savell, V. H., Jenkins, J. J., and Parham, D. M. (2000) Inhibition of the interferon-gamma/signal transducers and activators of transcription (STAT) pathway by hypermethylation at a STAT-binding site in the p21WAF1 promoter region, *Cancer research 60*, 3290-3298.

111. Iguchi-Ariga, S. M., and Schaffner, W. (1989) CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation, *Genes & development 3*, 612-619.

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

113. Sasai, N., Nakao, M., and Defossez, P. A. (2010) Sequence-specific recognition of methylated DNA by human zinc-finger proteins, *Nucleic acids research 38*, 5015-5022.

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

115. Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H. N., Shin, J., Cox, E., Rho, H. S., Woodard, C., Xia, S., Liu, S., Lyu, H., Ming, G. L., Wade, H., Song, H., Qian, J., and Zhu, H. (2013) DNA methylation presents distinct binding sites for human transcription factors, *eLife 2*, e00726.

116. Liu, Y., Zhang, X., Blumenthal, R. M., and Cheng, X. (2013) A common mode of recognition for methylated CpG, *Trends in biochemical sciences 38*, 177-183.

117. Ho, K. L., McNae, I. W., Schmiedeberg, L., Klose, R. J., Bird, A. P., and Walkinshaw, M. D. (2008) MeCP2 binding to DNA depends upon hydration at methyl-CpG, *Molecular cell 29*, 525-531.

118. Anvar, Z., Cammisa, M., Riso, V., Baglivo, I., Kukreja, H., Sparago, A., Girardot, M., Lad, S., De Feis, I., Cerrato, F., Angelini, C., Feil, R., Pedone, P. V., Grimaldi, G., and Riccio, A. (2016) ZFP57 recognizes multiple and closely spaced sequence motif variants to maintain repressive epigenetic marks in mouse embryonic stem cells, *Nucleic acids research 44*, 1118-1132.

119. Quenneville, S., Verde, G., Corsinotti, A., Kapopoulou, A., Jakobsson, J., Offner, S., Baglivo, I., Pedone, P. V., Grimaldi, G., Riccio, A., and Trono, D. (2011) In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions, *Molecular cell 44*, 361-372.

120. Lopes, E. C., Valls, E., Figueroa, M. E., Mazur, A., Meng, F. G., Chiosis, G., Laird, P. W., Schreiber-Agus, N., Greally, J. M., Prokhortchouk, E., and Melnick, A. (2008) Kaiso contributes to DNA methylation-dependent silencing of tumor suppressor genes in colon cancer cell lines, *Cancer research 68*, 7258-7263.

121. Weber, A., Marquardt, J., Elzi, D., Forster, N., Starke, S., Glaum, A., Yamada, D., Defossez, P. A., Delrow, J., Eisenman, R. N., Christiansen, H., and Eilers, M. (2008) Zbtb4 represses transcription of P21CIP1 and controls the cellular response to p53 activation, *The EMBO journal 27*, 1563-1574.

122. Eferl, R., and Wagner, E. F. (2003) AP-1: a double-edged sword in tumorigenesis, *Nature reviews. Cancer 3*, 859-868.

123. Rishi, V., Bhattacharya, P., Chatterjee, R., Rozenberg, J., Zhao, J., Glass, K., Fitzgerald, P., and Vinson, C. (2010) CpG methylation of half-CRE sequences creates C/EBPalpha binding sites that activate some tissue-specific genes, *Proceedings of the National Academy of Sciences of the United States of America 107*, 20311-20316.

124. Miller, M., Shuman, J. D., Sebastian, T., Dauter, Z., and Johnson, P. F. (2003) Structural basis for DNA recognition by the basic region leucine zipper transcription factor CCAAT/enhancer-binding protein alpha, *The Journal of biological chemistry 278*, 15178-15184.

125. Gustems, M., Woellmer, A., Rothbauer, U., Eck, S. H., Wieland, T., Lutter, D., and Hammerschmidt, W. (2014) c-Jun/c-Fos heterodimers regulate cellular genes via a newly identified class of methylated DNA sequence motifs, *Nucleic acids research 42*, 3059-3072.

126. Tulchinsky, E. M., Georgiev, G. P., and Lukanidin, E. M. (1996) Novel AP-1 binding site created by DNA-methylation, *Oncogene 12*, 1737-1745.

127. Yu, K. P., Heston, L., Park, R., Ding, Z., Wang'ondu, R., Delecluse, H. J., and Miller, G. (2013) Latency of Epstein-Barr virus is disrupted by gain-of-function mutant cellular AP-1 proteins that preferentially bind methylated DNA, *Proceedings of the National Academy of Sciences of the United States of America 110*, 8176-8181.

128. Bergbauer, M., Kalla, M., Schmeinck, A., Gobel, C., Rothbauer, U., Eck, S., Benet-Pages, A., Strom, T. M., and Hammerschmidt, W. (2010) CpG-methylation regulates a class of Epstein-Barr virus promoters, *PLoS pathogens 6*, e1001114.

129. Bhende, P. M., Seaman, W. T., Delecluse, H. J., and Kenney, S. C. (2004) The EBV lytic switch protein, Z, preferentially binds to and activates the methylated viral genome, *Nature genetics 36*, 1099-1104.

130. Ponferrada-Marin, M. I., Roldan-Arjona, T., and Ariza, R. R. (2012) Demethylation initiated by ROS1 glycosylase involves random sliding along DNA, *Nucleic acids research 40*, 11554-11562.

131. Ponferrada-Marin, M. I., Martinez-Macias, M. I., Morales-Ruiz, T., Roldan-Arjona, T., and Ariza, R. R. (2010) Methylation-independent DNA binding modulates specificity of Repressor of Silencing 1 (ROS1) and facilitates demethylation in long substrates, *The Journal of biological chemistry 285*, 23032-23039.

132. Hashimoto, H., Pais, J. E., Zhang, X., Saleh, L., Fu, Z. Q., Dai, N., Correa, I. R., Jr., Zheng, Y., and Cheng, X. (2014) Structure of a Naegleria Tet-like dioxygenase in complex with 5-methylcytosine DNA, *Nature 506*, 391-395.

133. Agius, F., Kapoor, A., and Zhu, J. K. (2006) Role of the Arabidopsis DNA glycosylase/lyase ROS1 in active DNA demethylation, *Proceedings of the National Academy of Sciences of the United States of America 103*, 11796-11801.

134. Morales-Ruiz, T., Ortega-Galisteo, A. P., Ponferrada-Marin, M. I., Martinez-Macias, M. I., Ariza, R. R., and Roldan-Arjona, T. (2006) DEMETER and REPRESSOR OF SILENCING 1 encode 5-methylcytosine DNA glycosylases, *Proceedings of the National Academy of Sciences of the United States of America 103*, 6853-6858.

135. Brooks, S. C., Fischer, R. L., Huh, J. H., and Eichman, B. F. (2014) 5-methylcytosine recognition by Arabidopsis thaliana DNA glycosylases DEMETER and DML3, *Biochemistry 53*, 2525-2532.

136. Yao, Q., Song, C. X., He, C., Kumaran, D., and Dunn, J. J. (2012) Heterologous expression and purification of Arabidopsis thaliana VIM1 protein: in vitro evidence for its inability to recognize hydroxymethylcytosine, a rare base in Arabidopsis DNA, *Protein expression and purification 83*, 104-111.

137. Thayer, M. M., Ahern, H., Xing, D., Cunningham, R. P., and Tainer, J. A. (1995) Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure, *The EMBO journal 14*, 4108-4120.

138. Guan, Y., Manuel, R. C., Arvai, A. S., Parikh, S. S., Mol, C. D., Miller, J. H., Lloyd, S., and Tainer, J. A. (1998) MutY catalytic core, mutant and bound adenine structures define specificity for DNA repair enzyme superfamily, *Nature structural biology 5*, 1058-1064.

139. Hollis, T., Ichikawa, Y., and Ellenberger, T. (2000) DNA bending and a flip-out mechanism for base excision by the helix-hairpin-helix DNA glycosylase, Escherichia coli AlkA, *The EMBO journal 19*, 758-766.

140. Bruner, S. D., Norman, D. P., and Verdine, G. L. (2000) Structural basis for recognition and repair of the endogenous mutagen 8-oxoguanine in DNA, *Nature 403*, 859-866.

141. Wu, P., Qiu, C., Sohail, A., Zhang, X., Bhagwat, A. S., and Cheng, X. (2003) Mismatch repair in methylated DNA. Structure and activity of the mismatch-specific thymine glycosylase domain of methyl-CpG-binding protein MBD4, *The Journal of biological chemistry 278*, 5285-5291.

142. Wu, L., and Zheng, Q. (2014) Active demethylation of the IL-2 Promoter in CD4+ T cells is mediated by an inducible DNA glycosylase, Myh, *Molecular immunology 58*, 38-49.

143. Fromme, J. C., Banerjee, A., Huang, S. J., and Verdine, G. L. (2004) Structural basis for removal of adenine mispaired with 8-oxoguanine by MutY adenine DNA glycosylase, *Nature 427*, 652-656.

144. Ushijima, Y., Tominaga, Y., Miura, T., Tsuchimoto, D., Sakumi, K., and Nakabeppu, Y. (2005) A functional analysis of the DNA glycosylase activity of mouse MUTYH protein excising 2-hydroxyadenine opposite guanine in DNA, *Nucleic acids research 33*, 672-682.

145. Chmiel, N. H., Golinelli, M. P., Francis, A. W., and David, S. S. (2001) Efficient recognition of substrates and substrate analogs by the adenine glycosylase MutY requires the C-terminal domain, *Nucleic acids research 29*, 553-564.

146. Ikeda, S., Biswas, T., Roy, R., Izumi, T., Boldogh, I., Kurosky, A., Sarker, A. H., Seki, S., and Mitra, S. (1998) Purification and characterization of human NTH1, a homolog of Escherichia coli endonuclease III. Direct identification of Lys-212 as the active nucleophilic residue, *The Journal of biological chemistry 273*, 21585-21593.

147. Hazra, T. K., Izumi, T., Boldogh, I., Imhoff, B., Kow, Y. W., Jaruga, P., Dizdaroglu, M., and Mitra, S. (2002) Identification and characterization of a human DNA glycosylase for repair of modified bases in oxidatively damaged DNA, *Proceedings of the National Academy of Sciences of the United States of America 99*, 3523-3528.

148. Hazra, T. K., Kow, Y. W., Hatahet, Z., Imhoff, B., Boldogh, I., Mokkapati, S. K., Mitra, S., and Izumi, T. (2002) Identification and characterization of a novel human DNA

glycosylase for repair of cytosine-derived lesions, *The Journal of biological chemistry 277*, 30417-30420.

149. Parikh, S. S., Putnam, C. D., and Tainer, J. A. (2000) Lessons learned from structural results on uracil-DNA glycosylase, *Mutation research 460*, 183-199.

150. David, S. S., O'Shea, V. L., and Kundu, S. (2007) Base-excision repair of oxidative DNA damage, *Nature 447*, 941-950.

151. McGoldrick, J. P., Yeh, Y. C., Solomon, M., Essigmann, J. M., and Lu, A. L. (1995) Characterization of a mammalian homolog of the Escherichia coli MutY mismatch repair protein, *Molecular and cellular biology 15*, 989-996.

152. Jurkowska, R. Z., Anspach, N., Urbanke, C., Jia, D., Reinhardt, R., Nellen, W., Cheng, X., and Jeltsch, A. (2008) Formation of nucleoprotein filaments by mammalian DNA methyltransferase Dnmt3a in complex with regulator Dnmt3L, *Nucleic acids research 36*, 6656-6663.

153. Suetake, I., Shinozaki, F., Miyagawa, J., Takeshima, H., and Tajima, S. (2004) DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction, *The Journal of biological chemistry 279*, 27816-27823.

154. Chedin, F., Lieber, M. R., and Hsieh, C. L. (2002) The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a, *Proceedings of the National Academy of Sciences of the United States of America 99*, 16916-16921.

155. Parrilla-Doblas, J. T., Ponferrada-Marin, M. I., Roldan-Arjona, T., and Ariza, R. R. (2013) Early steps of active DNA demethylation initiated by ROS1 glycosylase require three putative helix-invading residues, *Nucleic acids research 41*, 8654-8664.

156. Kohli, R. M., and Zhang, Y. (2013) TET enzymes, TDG and the dynamics of DNA demethylation, *Nature 502*, 472-479.

157. Bennett, M. T., Rodgers, M. T., Hebert, A. S., Ruslander, L. E., Eisele, L., and Drohat, A. C. (2006) Specificity of human thymine DNA glycosylase depends on N-glycosidic bond stability, *Journal of the American Chemical Society 128*, 12510-12519.

158. Munzel, M., Lischke, U., Stathis, D., Pfaffeneder, T., Gnerlich, F. A., Deiml, C. A., Koch, S. C., Karaghiosoff, K., and Carell, T. (2011) Improved synthesis and mutagenicity of oligonucleotides containing 5-hydroxymethylcytosine, 5-formylcytosine and 5-carboxylcytosine, *Chemistry 17*, 13782-13788.

159. Burdzy, A., Noyes, K. T., Valinluck, V., and Sowers, L. C. (2002) Synthesis of stable-isotope enriched 5-methylpyrimidines and their use as probes of base reactivity in DNA, *Nucleic acids research 30*, 4068-4074.

160. Liu, Y., Olanrewaju, Y. O., Zhang, X., and Cheng, X. (2013) DNA recognition of 5-carboxylcytosine by a Zfp57 mutant at an atomic resolution of 0.97 A, *Biochemistry 52*, 9310-9317.

161. Kamiya, H., Murata-Kamiya, N., Karino, N., Ueno, Y., Matsuda, A., and Kasai, H. (2000) Mutagenicity of 5-formyluracil in mammalian cells, *Nucleic acids symposium series*, 81-82.

162. Kamiya, H., Tsuchiya, H., Karino, N., Ueno, Y., Matsuda, A., and Harashima, H. (2002) Mutagenicity of 5-formylcytosine, an oxidation product of 5-methylcytosine, in DNA in mammalian cells, *Journal of biochemistry 132*, 551-555.

163. Kellinger, M. W., Song, C. X., Chong, J., Lu, X. Y., He, C., and Wang, D. (2012) 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription, *Nature structural & molecular biology 19*, 831-833.

164. Karino, N., Ueno, Y., and Matsuda, A. (2001) Synthesis and properties of oligonucleotides containing 5-formyl-2'-deoxycytidine: in vitro DNA polymerase

reactions on DNA templates containing 5-formyl-2'-deoxycytidine, *Nucleic acids research 29*, 2456-2463.

165. Miyazono, K., Furuta, Y., Watanabe-Matsui, M., Miyakawa, T., Ito, T., Kobayashi, I., and Tanokura, M. (2014) A sequence-specific DNA glycosylase mediates restriction-modification in Pyrococcus abyssi, *Nature communications 5*, 3178.

166. Reddy, S. M., Williams, M., and Cohen, J. I. (1998) Expression of a uracil DNA glycosylase (UNG) inhibitor in mammalian cells: varicella-zoster virus can replicate in vitro in the absence of detectable UNG activity, *Virology 251*, 393-401.

167. Spruijt, C. G., and Vermeulen, M. (2014) DNA methylation: old dog, new tricks?, *Nature structural & molecular biology 21*, 949-954.

168. Liu, Y., Olanrewaju, Y. O., Zheng, Y., Hashimoto, H., Blumenthal, R. M., Zhang, X., and Cheng, X. (2014) Structural basis for Klf4 recognition of methylated DNA, *Nucleic acids research 42*, 4859-4867.

169. Buck-Koehntop, B. A., Stanfield, R. L., Ekiert, D. C., Martinez-Yamout, M. A., Dyson, H. J., Wilson, I. A., and Wright, P. E. (2012) Molecular basis for recognition of methylated and specific DNA sequences by the zinc finger protein Kaiso, *Proceedings of the National Academy of Sciences of the United States of America 109*, 15229-15234.

170. Karin, M., Liu, Z., and Zandi, E. (1997) AP-1 function and regulation, *Current opinion in cell biology 9*, 240-246.

171. Glover, J. N., and Harrison, S. C. (1995) Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA, *Nature 373*, 257-261.

172. Kenney, S. C., and Mertz, J. E. (2014) Regulation of the latent-lytic switch in Epstein-Barr virus, *Seminars in cancer biology 26*, 60-68.

173. Fernandez, A. F., Rosales, C., Lopez-Nieva, P., Grana, O., Ballestar, E., Ropero, S., Espada, J., Melo, S. A., Lujambio, A., Fraga, M. F., Pino, I., Javierre, B., Carmona, F. J., Acquadro, F., Steenbergen, R. D., Snijders, P. J., Meijer, C. J., Pineau, P., Dejean, A., Lloveras, B., Capella, G., Quer, J., Buti, M., Esteban, J. I., Allende, H., Rodriguez-Frias, F., Castellsague, X., Minarovits, J., Ponce, J., Capello, D., Gaidano, G., Cigudosa, J. C., Gomez-Lopez, G., Pisano, D. G., Valencia, A., Piris, M. A., Bosch, F. X., Cahir-McFarland, E., Kieff, E., and Esteller, M. (2009) The dynamic DNA methylomes of double-stranded DNA viruses associated with human cancer, *Genome research 19*, 438-451.

174. Farrell, P. J., Rowe, D. T., Rooney, C. M., and Kouzarides, T. (1989) Epstein-Barr virus BZLF1 trans-activator specifically binds to a consensus AP-1 site and is related to c-fos, *The EMBO journal 8*, 127-132.

175. Bhende, P. M., Seaman, W. T., Delecluse, H. J., and Kenney, S. C. (2005) BZLF1 activation of the methylated form of the BRLF1 immediate-early promoter is regulated by BZLF1 residue 186, *Journal of virology 79*, 7338-7348.

176. Francis, A., Ragoczy, T., Gradoville, L., Heston, L., El-Guindy, A., Endo, Y., and Miller, G. (1999) Amino acid substitutions reveal distinct functions of serine 186 of the ZEBRA protein in activation of early lytic cycle genes and synergy with the Epstein-Barr virus R transactivator, *Journal of virology 73*, 4543-4551.

177. Petosa, C., Morand, P., Baudin, F., Moulin, M., Artero, J. B., and Muller, C. W. (2006) Structural basis of lytic cycle activation by the Epstein-Barr virus ZEBRA protein, *Molecular cell 21*, 565-572.

178. He, X., Tillo, D., Vierstra, J., Syed, K. S., Deng, C., Ray, G. J., Stamatoyannopoulos, J., FitzGerald, P. C., and Vinson, C. (2015) Methylated Cytosines Mutate to

Transcription Factor Binding Sites that Drive Tetrapod Evolution, *Genome biology and evolution 7*, 3155-3169.

179. Hu, L., Lu, J., Cheng, J., Rao, Q., Li, Z., Hou, H., Lou, Z., Zhang, L., Li, W., Gong, W., Liu, M., Sun, C., Yin, X., Li, J., Tan, X., Wang, P., Wang, Y., Fang, D., Cui, Q., Yang, P., He, C., Jiang, H., Luo, C., and Xu, Y. (2015) Structural insight into substrate preference for TET-mediated oxidation, *Nature 527*, 118-122.

180. Hashimoto, H., Pais, J. E., Dai, N., Correa, I. R., Jr., Zhang, X., Zheng, Y., and Cheng, X. (2015) Structure of Naegleria Tet-like dioxygenase (NgTet1) in complexes with a reaction intermediate 5-hydroxymethylcytosine DNA, *Nucleic acids research 43*, 10713-10721.

181. Hu, L., Li, Z., Cheng, J., Rao, Q., Gong, W., Liu, M., Shi, Y. G., Zhu, J., Wang, P., and Xu, Y. (2013) Crystal structure of TET2-DNA complex: insight into TET-mediated 5mC oxidation, *Cell 155*, 1545-1555.

182. Cimmino, L., Dawlaty, M. M., Ndiaye-Lobry, D., Yap, Y. S., Bakogianni, S., Yu, Y., Bhattacharyya, S., Shaknovich, R., Geng, H., Lobry, C., Mullenders, J., King, B., Trimarchi, T., Aranda-Orgilles, B., Liu, C., Shen, S., Verma, A. K., Jaenisch, R., and Aifantis, I. (2015) TET1 is a tumor suppressor of hematopoietic malignancy, *Nature immunology 16*, 653-662.

183. Ramasubramanyan, S., Osborn, K., Al-Mohammad, R., Naranjo Perez-Fernandez, I. B., Zuo, J., Balan, N., Godfrey, A., Patel, H., Peters, G., Rowe, M., Jenner, R. G., and Sinclair, A. J. (2015) Epstein-Barr virus transcription factor Zta acts through distal regulatory elements to directly control cellular gene expression, *Nucleic acids research 43*, 3563-3577.

184. Heather, J., Flower, K., Isaac, S., and Sinclair, A. J. (2009) The Epstein-Barr virus lytic cycle activator Zta interacts with methylated ZRE in the promoter of host target gene egr1, *The Journal of general virology 90*, 1450-1454.

185. Wille, C. K., Nawandar, D. M., Henning, A. N., Ma, S., Oetting, K. M., Lee, D., Lambert, P., Johannsen, E. C., and Kenney, S. C. (2015) 5-hydroxymethylation of the EBV genome regulates the latent to lytic switch, *Proceedings of the National Academy of Sciences of the United States of America 112*, E7257-7265.

186. Hashimoto, H., Olanrewaju, Y. O., Zheng, Y., Wilson, G. G., Zhang, X., and Cheng, X. (2014) Wilms tumor protein recognizes 5-carboxylcytosine within a specific DNA sequence, *Genes & development 28*, 2304-2313.

187. Jin, S. G., Zhang, Z. M., Dunwell, T. L., Harter, M. R., Wu, X., Johnson, J., Li, Z., Liu, J., Szabo, P. E., Lu, Q., Xu, G. L., Song, J., and Pfeifer, G. P. (2016) Tet3 Reads 5-Carboxylcytosine through Its CXXC Domain and Is a Potential Guardian against Neurodegeneration, *Cell reports 14*, 493-505.

188. Wang, L., Zhou, Y., Xu, L., Xiao, R., Lu, X., Chen, L., Chong, J., Li, H., He, C., Fu, X. D., and Wang, D. (2015) Molecular basis for 5-carboxycytosine recognition by RNA polymerase II elongation complex, *Nature 523*, 621-625.

189. Pogenberg, V., Consani Textor, L., Vanhille, L., Holton, S. J., Sieweke, M. H., and Wilmanns, M. (2014) Design of a bZip transcription factor with homo/heterodimer-induced DNA-binding preference, *Structure 22*, 466-477.

190. Schumacher, M. A., Goodman, R. H., and Brennan, R. G. (2000) The structure of a CREB bZIP.somatostatin CRE complex reveals the basis for selective dimerization and divalent cation-enhanced DNA binding, *The Journal of biological chemistry 275*, 35242-35247.

191. Fujii, Y., Shimizu, T., Toda, T., Yanagida, M., and Hakoshima, T. (2000) Structural basis for the diversity of DNA recognition by bZIP transcription factors, *Nature structural biology 7*, 889-893.

192. Kurokawa, H., Motohashi, H., Sueno, S., Kimura, M., Takagawa, H., Kanno, Y., Yamamoto, M., and Tanaka, T. (2009) Structural basis of alternative DNA recognition by Maf transcription factors, *Molecular and cellular biology 29*, 6232-6244.

193. Qiao, Y., He, H., Jonsson, P., Sinha, I., Zhao, C., and Dahlman-Wright, K. (2016) AP-1 Is a Key Regulator of Proinflammatory Cytokine TNFalpha-mediated Triple-negative Breast Cancer Progression, *The Journal of biological chemistry 291*, 5068-5079.

194. Horton, J. R., Wang, H., Mabuchi, M. Y., Zhang, X., Roberts, R. J., Zheng, Y., Wilson, G. G., and Cheng, X. (2014) Modification-dependent restriction endonuclease, MspJI, flips 5-methylcytosine out of the DNA helix, *Nucleic acids research 42*, 12092-12101.

195. Rajakumara, E., Law, J. A., Simanshu, D. K., Voigt, P., Johnson, L. M., Reinberg, D., Patel, D. J., and Jacobsen, S. E. (2011) A dual flip-out mechanism for 5mC recognition by the Arabidopsis SUVH5 SRA domain and its impact on DNA methylation and H3K9 dimethylation in vivo, *Genes & development 25*, 137-152.

196. Otwinowski, Z., Borek, D., Majewski, W., and Minor, W. (2003) Multiparametric scaling of diffraction intensities, *Acta crystallographica. Section A, Foundations of crystallography 59*, 228-234.

197. Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: a comprehensive Python-based

system for macromolecular structure solution, *Acta crystallographica. Section D, Biological crystallography 66*, 213-221.

198. Sved, J., and Bird, A. (1990) The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model, *Proceedings of the National Academy of Sciences of the United States of America 87*, 4692-4696.

199. Duncan, B. K., and Miller, J. H. (1980) Mutagenic deamination of cytosine residues in DNA, *Nature 287*, 560-561.

200. La, H., Ding, B., Mishra, G. P., Zhou, B., Yang, H., Bellizzi Mdel, R., Chen, S., Meyers, B. C., Peng, Z., Zhu, J. K., and Wang, G. L. (2011) A 5-methylcytosine DNA glycosylase/lyase demethylates the retrotransposon Tos17 and promotes its transposition in rice, *Proceedings of the National Academy of Sciences of the United States of America 108*, 15498-15503.

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

202. Ramiro-Merina, A., Ariza, R. R., and Roldan-Arjona, T. (2013) Molecular characterization of a putative plant homolog of MBD4 DNA glycosylase, *DNA repair 12*, 890-898.

203. Charlet, J., Duymich, C. E., Lay, F. D., Mundbjerg, K., Dalsgaard Sorensen, K., Liang, G., and Jones, P. A. (2016) Bivalent Regions of Cytosine Methylation and H3K27 Acetylation Suggest an Active Role for DNA Methylation at Enhancers, *Molecular cell 62*, 422-431.

204. Niller, H. H., Banati, F., Salamon, D., and Minarovits, J. (2016) Epigenetic Alterations in Epstein-Barr Virus-Associated Diseases, *Advances in experimental medicine and biology 879*, 39-69.

205. Birdwell, C. E., Queen, K. J., Kilgore, P. C., Rollyson, P., Trutschl, M., Cvek, U., and Scott, R. S. (2014) Genome-wide DNA methylation as an epigenetic consequence of Epstein-Barr virus infection of immortalized keratinocytes, *Journal of virology 88*, 11442-11458.

206. Golla, J. P., Zhao, J., Mann, I. K., Sayeed, S. K., Mandal, A., Rose, R. B., and Vinson, C. (2014) Carboxylation of cytosine (5caC) in the CG dinucleotide in the E-box motif (CGCAG|GTG) increases binding of the Tcf3|Ascl1 helix-loop-helix heterodimer 10-fold, *Biochemical and biophysical research communications 449*, 248-255.

207. Hashimoto, H., Zhang, X., and Cheng, X. (2013) Activity and crystal structure of human thymine DNA glycosylase mutant N140A with 5-carboxylcytosine DNA at low pH, *DNA repair 12*, 535-540.

208. Wales, T. E., and Engen, J. R. (2006) Hydrogen exchange mass spectrometry for the analysis of protein dynamics, *Mass spectrometry reviews 25*, 158-170.

209. Mussolino, C., and Cathomen, T. (2012) TALE nucleases: tailored genome engineering made easy, *Current opinion in biotechnology 23*, 644-650.

210. Dreyer, A. K., and Cathomen, T. (2012) Zinc-finger nucleases-based genome engineering to generate isogenic human cell lines, *Methods in molecular biology 813*, 145-156.

211. de Groote, M. L., Verschure, P. J., and Rots, M. G. (2012) Epigenetic Editing: targeted rewriting of epigenetic marks to modulate expression of selected target genes, *Nucleic acids research 40*, 10596-10613.