*8276–8284* *Nucleic Acids Research, 2012, Vol. 40, No. 17* *Published online 27 June 2012*
*doi:10.1093/nar/gks628*

# Excision of thymine and 5-hydroxymethyluracil by the MBD4 DNA glycosylase domain: structural basis and implications for active DNA demethylation

**Hideharu Hashimoto, Xing Zhang and Xiaodong Cheng***

Department of Biochemistry, Emory University School of Medicine, 1510 Clifton Road, Atlanta, GA 30322, USA

Received April 26, 2012; Revised May 24, 2012; Accepted June 4, 2012

## ABSTRACT

**The mammalian DNA glycosylase—methyl-CpG binding domain protein 4 (MBD4)—is involved in active DNA demethylation via the base excision repair pathway. MBD4 contains an N-terminal MBD and a C-terminal DNA glycosylase domain. MBD4 can excise the mismatched base paired with a guanine (G:X), where X is uracil, thymine or 5-hydroxy methyluracil (5hmU). These are, respectively, the deamination products of cytosine, 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC). Here, we present three structures of the MBD4 C-terminal glycosylase domain (wild-type and its catalytic mutant D534N), in complex with DNA containing a G:T or G:5hmU mismatch. MBD4 flips the target nucleotide from the double-stranded DNA. The catalytic mutant D534N captures the intact target nucleotide in the active site binding pocket. MBD4 specifically recognizes the Watson–Crick polar edge of thymine or 5hmU via the $O_2$, $N_3$ and $O_4$ atoms, thus restricting its activity to thymine/uracil-based modifications while excluding cytosine and its derivatives. The wild-type enzyme cleaves the *N*-glycosidic bond, leaving the ribose ring in the flipped state, while the cleaved base is released. Unexpectedly, the $C_1'$ of the sugar has yet to be hydrolyzed and appears to form a stable intermediate with one of the side chain carboxyl oxygen atoms of D534, via either electrostatic or covalent interaction, suggesting a different catalytic mechanism from those of other DNA glycosylases.**

## INTRODUCTION

Mammalian DNA glycosylases have been proposed to be involved in active DNA demethylation via the base excision repair pathway (1–4). MBD4 contains both an N-terminal methyl-CpG binding domain (MBD) and a C-terminal DNA glycosylase domain (Supplementary Figure S1a) that acts on G:T and G:U mismatches (5). In zebrafish, the activation-induced cytidine deaminase (AID) and MBD4 cooperate to demethylate DNA (1). Consistent with a role in DNA demethylation in mammals, AID is required to demethylate pluripotency genes during reprogramming of the somatic genome in embryonic stem cell fusions (6), and AID-deficient animals are less efficient in erasure of DNA methylation in primordial germ cells (7). It is noteworthy that AID promotes 5-methylcytosine (5mC or M) deamination, resulting in thymine (1,8), as well as 5-hydroxymethylcytosine (5hmC or H) deamination, which would produce 5-hydroxymethyluracil (5hmU) (3) (Figure 1). There are three mammalian ten eleven translocation (Tet) proteins that convert 5-methylcytosine (5mC or M) to 5hmC (9). 5hmC is a constituent of nuclear DNA, present in many tissues and cell types (9–11). The genomic content of 5hmU (<3.5 pmol) is orders of magnitude lower than that of 5hmC (hundreds of pmols) (10), suggesting that modification products of 5hmC are probably short lived and possibly removed by subsequent enzymatic reactions.

The importance of MBD4 for mutation avoidance in mammals and in maintaining genome stability is confirmed by an increase in 5mC to T mutations and increased occurrence of colon carcinoma in $Mbd4^{-/-}$ mice (13). Consistent with this observation, MBD4 is mutated in 26–43% of human colorectal tumors that show microsatellite instability (14). Here we show, by means of X-ray crystallography, that the mouse MBD4 glycosylase domain (residues 411–554; Supplementary Figure S1b) binds to a G:T or G:5hmU mismatch in the context of a CpG dinucleotide. The mismatched nucleotide (T or 5hmU) is flipped completely out of the DNA helix and is positioned in a binding pocket with Watson–Crick polar hydrogen bonds specific for thymine/uracil-based modifications.

## MATERIALS AND METHODS

### Expression and purification of MBD4 glycosylase domain

Hexahistidine-SUMO tagged mouse MBD4 residues 411–554 (pXC1064) and its mutants D534N (pXC1088), K536A (pXC1160) and Y514F (pXC1162) were expressed

*To whom correspondence should be addressed. Tel: +1 404 727 8491; Fax: +1 404 727 3746; Email: xcheng@emory.edu

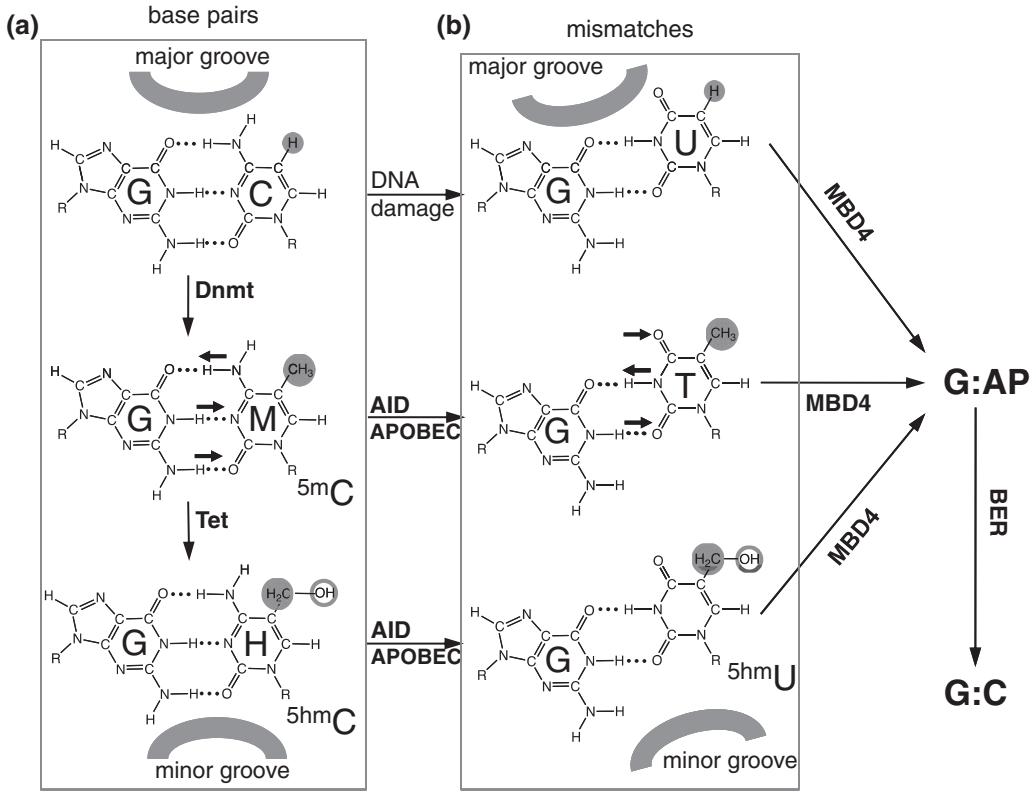


**Figure 1.** A putative pathway of DNA demethylation involving DNA methylation by DNMTs, hydroxylation by Tet proteins, deamination by AID [or members of APOBEC superfamily (12)] and base excision by MBD4 linked to base excision repair (BER). DNA major groove and minor groove sides are indicated. Small arrows in G:M pair and G:T mismatch indicate the hydrogen bond donors and acceptors for 5mC and thymine bases. (**a**) C, 5mC (M) and its oxidized derivative 5hmC (H) form base pairs with an opposite G, (**b**) deamination-linked mismatches.

in *Escherichia coli* BL21(DE3)-Gold cells from the RIL-Codon plus plasmid (Stratagene). Expression cultures were grown at 37°C until $OD_{600}$ reached 0.5, shifted to 16°C and then induced by adding 0.4 mM isopropyl β-D-1-thiogalactopyranoside. Cell pellets were suspended with 4× volume of Ni buffer (300 mM NaCl, 20 mM sodium phosphate pH 7.4, 20 mM imidazole, 1 mM dithiothreitol and 0.25 mM phenylmethylsulfonyl fluoride) and sonicated for 5 min (1 s on and 2 s off). The lysate was clarified by centrifugation twice at 38 000*g* for 30 min. His6-SUMO fusion proteins were isolated on a nickel-charged chelating column (GE Healthcare). The His6–SUMO tag was cleaved by ULP-1 protease (16 h at 4°C), leaving two extraneous N-terminal amino acids (HisMet). The cleaved protein was loaded onto tandem ion exchange HiTrap-Q and HiTrap-SP columns (GE-Healthcare), eluted from the SP column and further purified by Superdex 75 (16/60) in the presence of 500 mM NaCl.

### Crystallography of MBD4 and DNA complexes

For co-crystallization, the MBD4 proteins (WT or D534N mutant) and annealed G:X mismatch oligonucleotide (5′-TC AGC**G**CATGG-3′ and 5′-CCATG**X**GCTGA-3′; where X = T or 5hmU, synthesized by Sigma or New England Biolabs, respectively) were mixed at a 1:1 ratio and diluted 5× with buffer (20 mM HEPES pH 7.0, 1 mM dithiothreitol) to reduce the salt concentration to ~100 mM, and then concentrated to ~0.4 mM at ~4°C. The complex crystals appeared after 1–7 days at 16°C under the condition 25% polyethylene glycol 3350, 200 mM NaCl, 100 mM Bis–Tris–HCl pH 5.6. Crystals were cyroprotected by soaking in mother liquor with 20% ethylene glycol. The X-ray diffraction data sets for wild-type and D534N mutant co-crystals with DNA were collected at the SER-CAT beamlines 22BM-E and 22ID-D, respectively, and processed using HKL2000 (15). The structures were solved by molecular replacement using PHENIX (16), using the mouse MBD4 glycosylase domain apo structure (PDB 1NGN) (17) as the searching model. Electron density for DNA was easily interpretable, using the model-building program Coot (18). PHENIX/Refinement scripts were used for refinement, and the statistics shown in Supplementary Table S1 were calculated for the entire resolution range. The $R_{free}$ and $R_{work}$ values were calculated for 5% (randomly selected) and 95%, respectively, of observed reflections. The structures were solved, built and refined independently.

### DNA glycosylase activity assay

MBD4 glycosylase activity assays were performed using various oligonucleotides labeled with 6-carboxy-fluor escein (FAM), and the excision of the target base was monitored by denaturing gel electrophoresis following NaOH hydrolysis (Supplementary Figure S1c). In Figure 2a, MBD4 protein (0.5 μM) and equal amount of

double-stranded FAM labeled 32-bp duplexes were mixed in 20 µl nicking buffer (10 mM Tris–HCl, pH 8.0, 1 mM EDTA, 0.1% BSA) and incubated at 37°C for 30 min. Reactions were stopped by adding 2 µl of 1 N NaOH, boiled for 10 min, and 20 µl of loading buffer (98% formamide, 1 mM EDTA and 1 mg/ml of Bromophenol Blue and Xylene Cyanole) was added. The samples were boiled for another 10 min and then immediately cooled in ice water and loaded onto a 10 cm × 10 cm 15% denaturing gel containing 7 M urea, 24% formamide, 15% acrylamide and 1× TBE. The gels were run in 1× TBE buffer for 60 min at 200 V. FAM-labeled single-stranded DNA was visualized under UV exposure. The following 32-bp oligonucleotides were synthesized at the New England Biolabs:

(FAM)-5′-TCGGATGTTGTGGGTCAG**X**GCATGA TAGTGTA-3′ (where **X** = C, 5mC, 5hmC, U, T, 5hmU or 5caC) and 5′-TACACTATCATGCGCTGACCCACA ACATCCGA-3′

For enzymatic reactions under single turnover conditions (Figure 2b), the FAM-labeled 32-bp duplexes (0.25 µM) and 10-fold excess of MBD4 catalytic domain (2.5 µM) were incubated for 0–10 min at room temperature (~22°C) and processed as described earlier. The intensities of the FAM-labeled DNA were measured by Typhoon Trio+ (GE Healthcare) and quantified by the image-processing program ImageJ (NIH). The data were fitted by non-linear regression using software GraphPad PRISM 5.0 d (GraphPad Software Inc.): [Product] = $P_{max}(1 - e^{-kt})$, where $P_{max}$ is the product plateau level, $k$ is the observed rate constant, and t is the reaction time.

### DNA binding assay

As shown in Figure 2c, MBD4 protein (1.0 µM) and 0.5 µM of the 32-bp FAM-labeled DNA were mixed in 20 µl nicking buffer and incubated at 37°C for 15 min. As shown in Figure 2d, the 32-bp FAM-labeled DNA

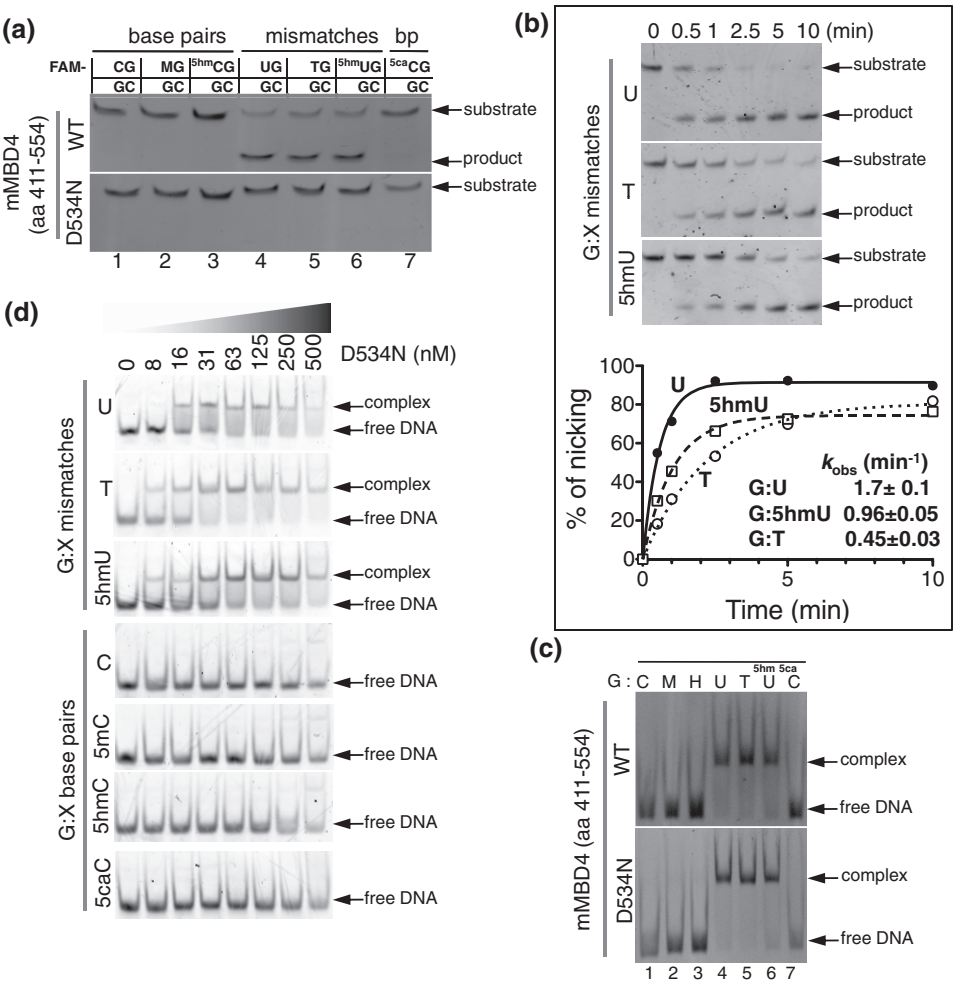


**Figure 2.** Base excision and binding activities of MBD4 glycosylase domain on double-stranded DNA containing various forms of CpG dinucleotide. (**a**) Double-stranded 32-bp oligonucleotides bearing a single CpG dinucleotide were incubated with equal amount of MBD4 at 37°C for 30 min. The oligonucleotide was labeled with FAM on the top strand and the modification status is indicated (M = 5mC). The products of the reactions were separated on a denaturing polyacrylamide gel, and the FAM-labeled strand was excited by UV and photographed. (**b**) The activity of MBD4 catalytic domain on G:U (top panel), G:T (middle panel) and G:5hmU (bottom panel) substrates under single turnover conditions ($[E_{MBD4}]$ = 2.5 µM and $[S_{DNA}]$ = 0.25 µM) at pH 8.0 and room temperature (~22°C). (**c**) DNA binding assays were performed by incubating 0.5 µM FAM-labeled oligonucleotides with 1 µM of MBD4 at 37°C for 15 min. (**d**) DNA binding assays were performed by incubating 20 nM FAM labeled oligonucleotides with an increased amount of D534N mutant at ~22°C for 30 min.

(20 nM) and MBD4 D534N mutant (with an indicated amount) were incubated at ~22°C for 30 min. Samples were loaded onto a 10 cm × 10 cm 10% native acrylamide gel in 0.5× or 1× TBE buffer and run 25–40 min at 100 V.

## RESULTS

### MBD4 specifically binds and cleaves thymine/uracil-based mismatches

We analyzed the specificity of the glycosylase and binding activities of MBD4 glycosylase domain using various 32-bp DNA oligonucleotides, with each containing a single G:X modification within the CpG sequence, where X = C, M, H, U, T, 5hmU or 5-carboxylcytosine (5caC). As expected, substrates bearing G:U, G:T and G:5hmU mismatches were efficiently cleaved (Figure 2a), but glycosylase activity was not detected with MBD4 on oligonucleotides bearing the 'natural' G:C and G:M base pairs, or G:H and G:5caC, which presumably preserve Watson–Crick base pair hydrogen bonds (Supplementary Figure S2a). Therefore, MBD4 is capable of acting on deamination-linked mismatches, with the observed rate constants ($k_{obs}$) of 1.7, 1.0 or 0.45 $min^{-1}$ for G:U, G:5hmU and G:T substrates, respectively, under single turnover conditions ($[E_{MBD4}] >> [S_{DNA}]$) (Figure 2b). The differences in the MBD4 glycosylase activity on G:U and G:T substrates are in agreement with previous findings (5,19). A non-catalytic mutant (D534N of MBD4) is inactive on all substrates (Figure 2a).

The glycosylase activity correlates well with the ability to form a specific complex in electrophoresis mobility-shift assays (Figure 2c), under the condition of 2:1 molar ratio of enzyme to DNA. Substrate oligonucleotides bind to WT and D534N mutants, while no binding is observed for the cytosine-based non-substrate oligonucleotides. The estimated dissociation constants ($K_D$) are <30 nM for all three mismatched substrates (Figure 2d).

We performed crystallographic analyses on complexes of D534N mutant MBD4 glycosylase domain with G:T or G:5hmU-containing 11-bp DNA as well as the complex of the wild-type enzyme with a G:T mismatch, which resulting in a product complex with a ribose sugar site. We determined the structures to the resolutions of 2.4 Å (D534N) and 2.8 Å (WT), respectively (Supplementary Table S1).

### The structures of D534N MBD4 glycosylase domain bound to G:T or G:5hmU DNA

MBD4 belongs to the family of Helix-hairpin-Helix (HhH) DNA glycosylases including hOGG1, MutY, EndoIII and AlkA (17). The protein component of MBD4, an all-helical structure containing 11 helices, is highly similar to that of MBD4 in the absence of DNA (17), with a root mean squared (rms) deviation of ~0.7 Å when comparing 144 pairs of Cα atoms. On the other hand, the 11-bp DNA duplex undergoes substantial protein-induced distortions. The phosphate backbone at the central G:T mismatch and C:G pair is bent ~65°, due to, in part, intercalation of the Leu482 side chain into the minor groove between the Cyt and Gua bases of the unmodified strand (Figure 3a and b). Concurrently, the thymine nucleotide flips out, and an arginine finger (Arg442) penetrates into the space left by the flipped thymine and interacts with two phosphate groups (3′ −1 phosphate and 5′ +2 phosphate) (Figure 3c and d). The intrahelical orphaned guanine hydrogen bonds with the main chain carbonyl oxygen atoms of Arg442 and Leu480 (Figure 3e). Asn441 and Leu485 sit in the minor groove, that is markedly widened by the severe bending of DNA, and interact weakly with the neighboring G:C pair (Figure 3f). No interaction to the neighboring G:C pair in the major groove is observed, suggesting that the modification status (methyl, hydroxymethyl or carboxyl) of the cytosine C5 position would have no impact on MBD4 activity (20).

The thymine is flipped out of DNA helix and inserted into the open active site cleft (Figure 3g) sandwiched between Leu440 at the bottom and Leu421 and Lys536 at the top (Figure 3h), as predicted by our previous modeling study (17). Although these residues are not conserved in HhH DNA glycosylases, similar stacking appears to be conserved in hOGG1 (22) and in MutY (23). The polar groups of the thymine base along the Watson–Crick edge are all hydrogen bonded with the protein (Figure 3i). The $O_2$ oxygen atom accepts hydrogen bonds from the side chains of Tyr514 (hydroxyl group) and Gln423 (amino group), the $N_3$ nitrogen atom donates a hydrogen bond to the carbonyl oxygen of Gln423 side chain and the $O_4$ oxygen atom accepts a hydrogen bond from the main-chain amide nitrogen of Val422. It is noteworthy that a thymine base has the opposite hydrogen bond potential at the ring positions of 3 and 4 compared to a cytosine base (or its $C_5$ derivatives), of which the ring $N_3$ nitrogen can only be a hydrogen-bond acceptor and the exocyclic $N_4$ amino group can only donate a hydrogen bond (Figure 1). Therefore, the network of hydrogen bonds in the MBD4 active site will not accommodate the Watson–Crick edge of a cytosine and exclusion of cytosine base by the active site allows MBD4 to discriminate against all cytosine derivatives, including 5caC, which is a substrate for mammalian thymine DNA glycosylase (TDG) (24). We do note, however, that the discrimination of a mismatched thymine/uracil from that of G:C pair could also occur prior to base flipping.

The mutated catalytic residue D534N is roughly perpendicular to the sugar ring of the flipped thymine with the side chain nitrogen atom forming a hydrogen bond with the $O_3'$ atom (2.7 Å), while the side chain oxygen atom is ~3.7 Å from the $C_1'$ atom and 4.6 Å from the $N_1$ atom—the two atoms that form the *N*-glycosidic bond (Figure 3i). No water molecule is found nearby. Superimposition with DNA containing a normal intrahelical thymine reveals that the everted yet uncleaved thymine ring rotates around the glycosidic bond as well as bends slightly relative to the sugar ring (Figure 3j). This is reminiscent of the structures of uracil DNA glycosylase–DNA complex with a flipped uracil (25) and *Methanobacterium thermoformicicum* mismatch glycosylase MIG–DNA complex with an extrahelical thymine (26), where the enzyme-induced distortion was suggested to facilitate the cleavage of the glycosidic bond.

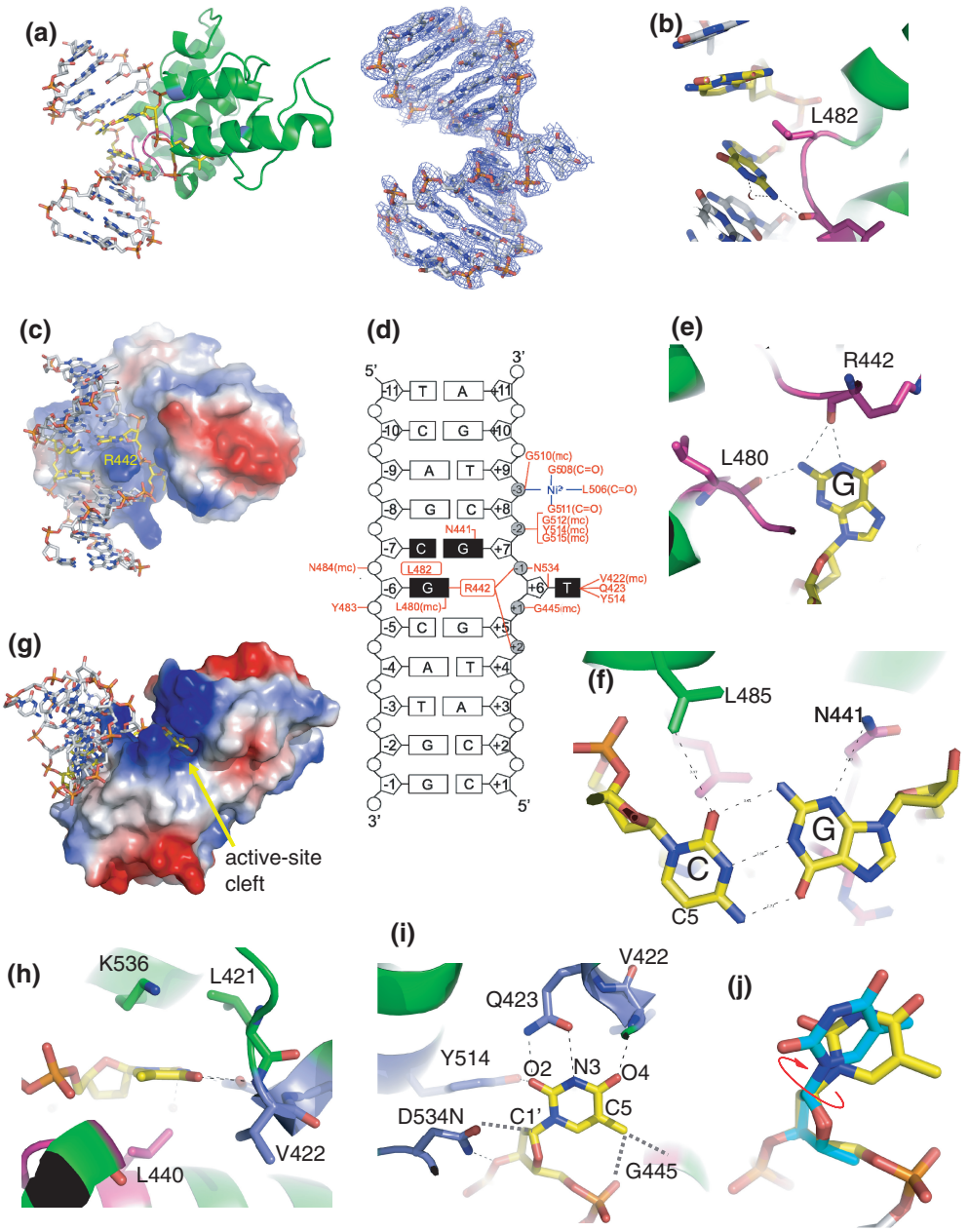


**Figure 3.** Structures of MBD4 D534N in complex with G:T mismatch. (**a**) MBD4 (colored in green) approaches the DNA from the minor groove side and bends the DNA at the central G:T mismatch and C:G pair. Right panel shows 2Fo–Fc electron density, contoured at 1σ above the mean, for the entire 11-bp DNA. (**b**) Leu482 intercalates between the central Cyt and Gua of the unmodified strand. (**c**) Arg442 penetrates into the DNA helix from the minor groove. (**d**) Summary of the MBD4–DNA interactions; black boxes represent the CpG recognition sequence and the extrahelical thymine; mc, main-chain-atom-mediated contacts; C = O, main-chain carbonyl oxygen mediated metal interactions. The metal mediated interaction with the 3′ phosphate at the −3 position from the flipped nucleotide is also a conserved feature in AlkA (21). (**e**) The three hydrogen bonds formed with the intrahelical orphaned guanine. (**f**) The neighboring C:G base pair in the CpG context interacts with MBD4 in the minor groove. No interaction was observed in the major groove, where the $C_5$ atom of Cyt is located. (**g**) The flipped thymine is bound in the open active-site cleft. (**h**) The flipped thymine is stacked between Lys536 and Leu440. The terminal amino group of Lys536 is close to the main chain carbonyl oxygen atom of F419 (~3.6 Å). (**i**) Thymine-specific interactions in MBD4. (**j**) Superimposition of a normal intrahelical thymine (colored in cyan) onto the flipped thymine suggests a rotation around the glycosidic bond.

The $C_5$ methyl group of the flipped thymine is 3.6 Å away from one of the 5′ phosphate oxygen atoms, and 4.0 Å away from the Cα atom of Gly445 (Figure 3i), suggesting that an additional hydroxyl group attached to the $C_5$ methyl could fit into the space, consistent with 5hmU being a substrate of MBD4. Indeed, 5hmU flips completely out of the DNA helix by MBD4 and can be well overlaid with thymine in the same binding pocket (Figure 4a). However, the hydroxyl group points away from Gly445 and forms an inter-molecular polar interaction with the terminal ε-amino group of Lys536 (Figure 4b), whose aliphatic portion of side chain stacked with the ring of the extrahelical nucleotide.

### The structure of a stalled intermediate formed by WT MBD4 glycosylase domain

To obtain a product complex structure, WT MBD4 protein were mixed with substrate DNA containing a G:T mismatch and co-concentrated (~3 h at 4°C) before setting up crystallization trials. As expected, the thymine base was already cleaved and possibly released via the open-active-site cleft connected to the solvent (Figure 3g and Supplementary Figure S1d). Superimposition of the protein components of the D534N mutant and WT structures (with rms deviation of ~0.5 Å when compared 144 pairs of Cα atoms) reveals a conformational change of side chain of Lys536, moving toward the base and possibly pushing out the cleaved base (Figure 4c). Mutating Lys536 to alanine (K536A) reduces the activity by a factor of ~5–10 (Figure 4d, left panel), comparing to that of WT enzyme (Figure 2b). While K536A mutation is expected to affect product release, it also reduces the reaction rate probably due to the involvement of Lys536 side chain in stabilizing the flipped base in a twisted state (Figure 4b).

Unexpectedly, the electron density clearly indicates that the sugar ring of the cleaved nucleotide does not yet contain a hydroxyl oxygen atom at the $C_1'$ carbon atom (Figure 5a). Rather, the ribose ring $C_1'$ carbon is in close contact with one of the carboxylate oxygen atoms of the catalytic residue Asp534 (refined to ~2.6 Å) and the hydroxyl oxygen atom of Tyr514 (~3.1 Å) (Figure 5b), leaving no room for a water nucleophile between the aspartic acid and the substrate nucleotide. However, a water molecule is observed, coordinated by the second carboxylate oxygen atom of Asp534 and the hydroxyl oxygen atom of Tyr514 and is ~3.1 Å away from $C_1'$ (Figure 5b), a position approximately corresponding to the $O_2$ position of the uncleaved thymine. Perhaps this water molecule will be able to serve as the nucleophile to attack the $C_1'$ and eventually complete the reaction (Supplementary Figure S3).

MBD4 is known for exhibiting single-turnover, pre-steady state kinetics (27). It appears that the crystal structure of the MBD4 product complex represents a stalled reaction intermediate after the cleavage of the base but before the final nucleophilic attack of the $C_1'$ atom (Supplementary Figure S3). To investigate whether this is due to the crystallization condition being unfavorable for MBD4 reaction, we incubated the protein–DNA complex (in 20 mM HEPES pH 7.0 and 100 mM NaCl) overnight at room temperature, to assure that the first round of base excision reaction was completed (under 1 h, Supplementary Figure S1e), before setting up crystallization. We then repeated the crystallographic study and the resulting structure is essentially identical to the structure without the extended pre-incubation, showing that the $C_1'$ lacks the hydroxyl group while the base has already been cleaved (data not shown).

The time course of an MBD4 glycosylase reaction with a 4:1 DNA to protein molar ratio, under the same condition as pre-crystallization incubation (100 mM NaCl), shows that minimal enzyme turnover occurred during the 20-h incubation period, while the first round of base excision was completed before 1 h, suggesting that indeed the reaction was stalled for a significant amount of time after base excision (Supplementary Figure S1e). When no salt was added to reaction, MBD4 has a turnover rate of ~4 h (Supplementary Figure S1f).

## DISCUSSION

### Initiation of base excision by MBD4

MBD4 does not have the equivalent of Lys249 of hOGG1, which is proposed to attack the deoxyribose $C_1'$ atom (22). Instead, MBD4 has Tyr514 in the corresponding position, whose hydroxyl group makes a hydrogen bond to the $O_2$ of thymine or 5hmU (Figures 3i and 4a). Together with the $O_2$-interacting amino group of Gln423, Tyr514 could serve to activate the leaving group by providing a hydrogen bond to the developing negative charge on a pyrimidine $O_2$ in the transition state (Figure 4a), as suggested for the interaction between uracil DNA glycosylase His157 and uracil $O_2$ (28). Indeed, mutating Try514 to phenylalanine (Y514F) nearly abolishes MBD4 activity (Figure 4d, right panel).

### Potential reaction mechanism(s)

The stalled complex shows that MBD reaction is a stepwise $S_N1$-type cleavage reaction where the base

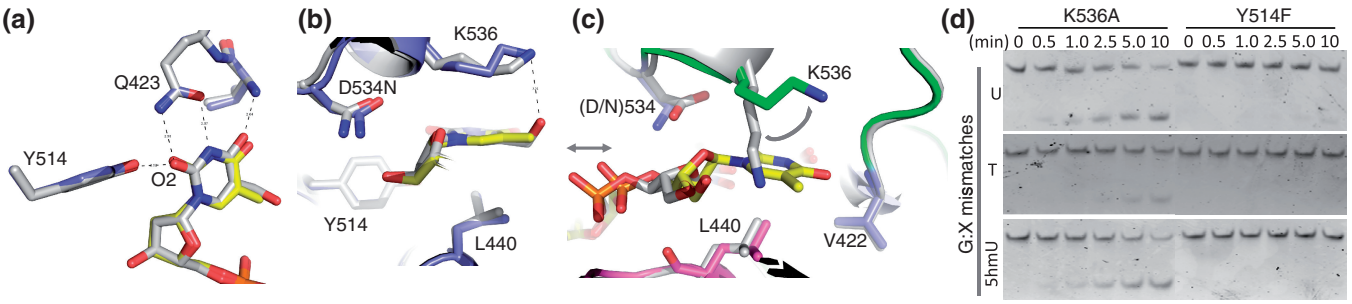


**Figure 4.** Comparison of the 5hmU conformation with that of thymine in MBD4. (**a**) Superimposition of 5hmU-bound (in gray) and thymine-bound (in color) structures. (**b**) The hydroxyl group of 5hmU interacts with Lys536. We note that the interaction does not form an ideal hydrogen bond. Rotation of the C–C bond between the ring $C_5$ and the methyl hydroxyl ($CH_2$–OH) could position the hydroxyl group in several alternative conformations (Supplementary Figure S2b). (**c**) Superimposition of the WT (in gray) and mutant MBD4 (in color) structures indicates a conformational change of Lys536 in conjunction with the release of the cleaved base. (**d**) The activity of MBD4 mutants, K536A (left) and Y514F (right), on G:U (top panel), G:T (middle panel) and G:5hmU (bottom panel) substrates under the same single turnover conditions as that of the wild-type enzyme (see Figure 2b).

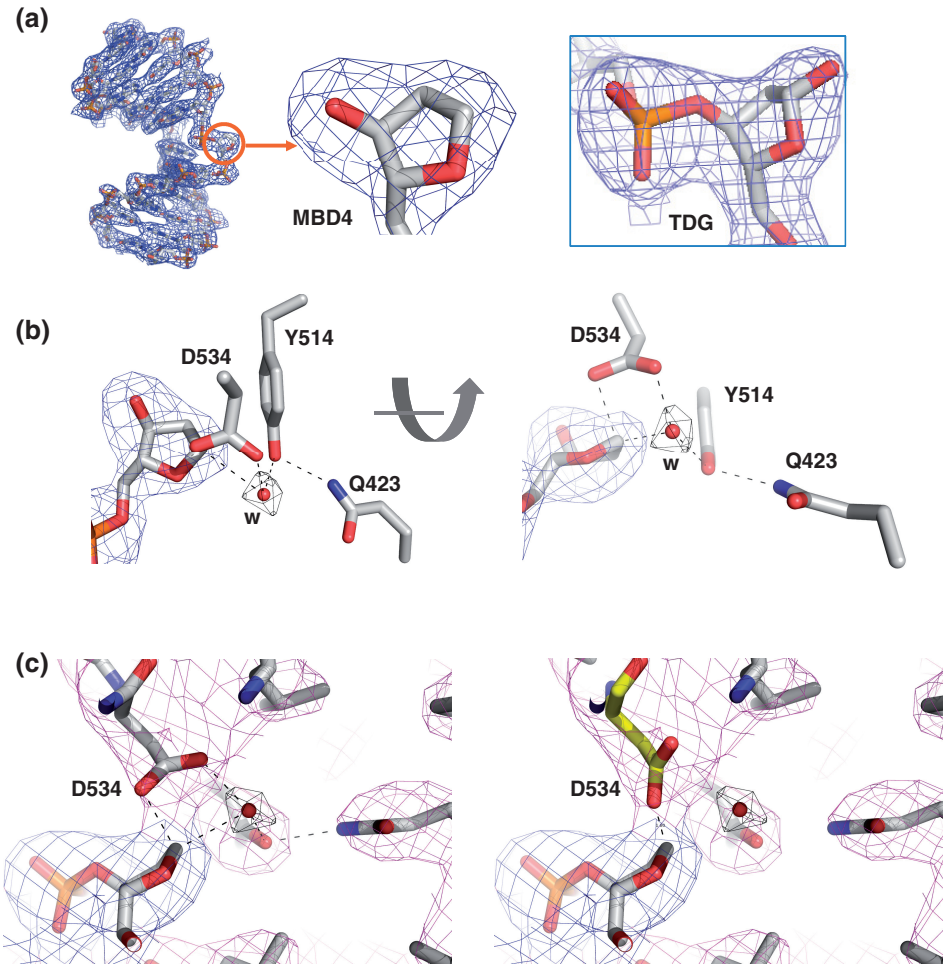


**Figure 5.** A stalled intermediate. (**a**) Electron densities (2Fo–Fc), contoured at 1σ above the mean, are shown for the 11-bp DNA containing the ribose ring in the MBD4 WT structure (left panel). An enlarged view of the sugar is provided (middle panel). Included for comparison is an abasic sugar with a hydroxyl group attached to $C_1'$, generated by thymine DNA glycosylase (H. Hashimoto, X. Zhang and X. Cheng, our unpublished data) (right panel). (**b**) In MBD4, the ribose ring $C_1'$ is in direct contact to the carboxylate of the catalytic residue Asp534. A water molecule, coordinated by the side chains of Asp534, Tyr514 and Gln423, is in a position that approximately corresponds to the $O_2$ position of the uncleaved thymine. (**c**) The electron densities (2Fo–Fc), contoured at 1σ above the mean, are connected between the ribose ring and the side chain of Asp534. Structural refinement positioned the two carboxylate oxygen atoms of Asp534 within tight hydrogen bonding distances from the $C_1'$ of ribose, the main chain amide nitrogen atom of Leu537 and the water molecule (left panel). A simple rotation around the χ2 torsion angle could move one of the carboxylate oxygen atoms of Asp534 as close as 1.8 Å to the $C_1'$ of ribose (right panel).

excision and nucleophilic attack of $C_1'$ are clearly separated (29). In this mechanism, the $C_1'$ should exist as a positively charged carbocation or an oxacarbenium ion intermediate following base removal (29). The half-life of a highly reactive oxacarbenium ion is estimated to be only $\sim 10^{-10}$–$10^{-12}$ s (30,31), too short-lived to be observed in the time scale of crystallization. However, it is potentially possible that the oxacarbenium ion is stabilized by negatively charged Asp534, forming a long-lived ion-pair intermediate at the active site of MBD4. A similar mechanism has been proposed for *Bacillus stearothermophilus* MutY, where Asp144 of MutY was in the vicinity of $C_1'$ of a non-cleavable substrate analogue and proposed to stabilize positive charge accumulation on an oxacarbenium ion as well as coordinate a water nucleophile (32,33) (Supplementary Figure S4a). In addition to Asp144, a second acidic residue Glu43 of MutY interacts with the substrate adenine-$N_7$ that was thought to facilitate glycosidic bond scission (32). The equivalent functional group in MBD4 might be the $O_2$-interacting Tyr514.

An alternative mechanism would be a covalent glycosyl-enzyme intermediate, like those used by retaining O-glycosylases such as hen egg-white lysozyme (HEWL) (34). HEWL uses two carboxylate residues (Glu35 and Asp52) during the catalysis. When the catalytic acid/base Glu35 is mutated to the corresponding amide (E35Q), the covalent glycosyl-Asp52 intermediate is observed because hydrolysis of this intermediate is slowed enormously as the attack of water no longer benefits from the base catalysis ordinarily provided by Glu35. MBD4 only has the Asp52 equivalent carboxylate (Asp534 in MBD4) and lacks the Glu35 equivalent residue in the active site, which might have allowed us to capture the covalent intermediate between Asp534 and the $C_1'$ atom in the absence of a base activated water

nucleophile. A similar interaction has been observed in the structure of *E. coli* AlkA complexed with DNA containing 1-azaribose abasic inhibitor, where Asp238 directly interacts with the positively charged $N_1'$ (equivalent to $C_1'$) of the 1-azaribose ring (21) (Supplementary Figure S4b). It was proposed that Asp238 of AlkA directly assists the removal of positively charged bases, natural substrates for AlkA. However, MBD4 only removes uncharged bases and it remains to be determined whether Asp534 carboxylate initiates the attack on $C_1'$ that breaks the *N*-glycosidic linkage.

Unfortunately, our current structure at the resolution of 2.8 Å is insufficient to determine whether a covalent bond (~1.45 Å) formed between the $C_1'$ and the carboxylate oxygen of Asp534 or to determine the details of the sugar conformation. The distance between the two interacting atoms could be significantly reduced by a simple rotation around the side chain $\chi 2$ torsion angle of Asp534 (Figure 5c) and/or by relaxation of the sugar conformation. In the 1.64 Å-resolution structure of HEWL E53Q mutant, such movement caused the $C_1$ of the six-membered pyranose ring to be 1.75 Å closer to the carboxylate oxygen of Asp52, resulting in a covalent bond (34). For MBD4, an atomic resolution structure will be required to settle whether the covalent bond forms.

The loss of activity by mutating Asp534 to the corresponding amide (D534N) (Figure 2a) suggests that the negative charge on Asp534 is critical for catalysis. In addition, the pH profile of MBD4 activity is also consistent with the requirement of a deprotonated carboxylate for catalysis: the activity is similar in the pH range of 6.5–9.4, but drops precipitously below that until total loss of activity at pH 4.0 (Supplementary Figure S1g).

In summary, we show that the mouse MBD4 glycosylase domain binds to G:X mismatched substrate DNA and flips out the target nucleotide into the active-site pocket. Many structural features are conserved among the HhH DNA glycosylases [including the human MBD4 glycosylase domain (35), Supplementary Figure S5]: the phosphodiester backbone pinching (36) caused by extensive protein–phosphate contacts surrounding the flipped nucleotide, the significant DNA bending due to hydrophobic intercalation of protein residues between the Cyt and Gua of the unmodified strand, and the use of an arginine finger to penetrate DNA from the minor groove and fill the space left by the flipped nucleotide. However, the recognition of the flipped nucleotide and catalytic mechanism in the active site could be very different among the HhH enzymes. MBD4 specifically recognizes the Watson–Crick polar edge of thymine and 5hmU (Figures 3i and 4a), thus restricting its activity to thymine/uracil-based modifications. The observation of a long-lived intermediate is unique for DNA *N*-glycosylases. If the involvement of a covalent bond between enzyme and substrate can be confirmed, it would offer significant new insight for the catalytic mechanism of DNA *N*-glycosylases. The tight binding of the glycosylase may have an important biological function for protection of the cleaved site and prevents its non-specific processing until subsequent repair activities are recruited to the site.

## ACCESSION NUMBERS

Protein Data Bank: The coordinates and structure factors of the mouse MBD4 catalytic domain—DNA complexes have been deposited with accession numbers 4EVV (D534N and G:T mismatch), 4EW0 (D534N and G:5hmU mismatch) and 4EW4 (WT and G:T resulting in a stalled intermediate).

## SUPPLEMENTARY DATA

Supplementary Data are available NAR Online: Supplementary Table 1 and Supplementary Figures 1–5.

## ACKNOWLEDGEMENTS

We thank sincerely Brenda Baker of New England Biolabs for DNA oligo synthesis, Drs John R. Horton and John Chrzas for help with X-ray data collection, Dr Robert M. Blumenthal of the University of Toledo College of Medicine for critical comments and editing the manuscript and Dr Paul J. Berti of McMaster University for discussion. H.H. performed all experiments. X.Z. and X.C. organized and designed the scope of the study, and all were involved in analyzing data and writing and revising the manuscript.

## FUNDING

U.S. National Institutes of Health [GM049245-18]; Georgia Research Alliance Eminent Scholar (to X.C.) and The Department of Biochemistry at the Emory University School of Medicine (use of the Southeast Regional Collaborative Access Team synchrotron beamlines at the Advanced Photon Source of Argonne National Laboratory). Funding for open access charge: National Institutes of Health.

*Conflict of interest statement*. None declared.

## REFERENCES

1. 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.
2. Hajkova,P., Jeffries,S.J., Lee,C., Miller,N., Jackson,S.P. and Surani,M.A. (2010) Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. *Science*, **329**, 78–82.
3. Guo,J.U., Su,Y., Zhong,C., Ming,G.L. and Song,H. (2011) Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell*, **145**, 423–434.
4. Cortellino,S., Xu,J., Sannai,M., Moore,R., Caretti,E., Cigliano,A., Le Coz,M., Devarajan,K., Wessels,A., Soprano,D. *et al.* (2011) Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell*, **146**, 67–79.
5. 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.
6. Bhutani,N., Brady,J.J., Damian,M., Sacco,A., Corbel,S.Y. and Blau,H.M. (2010) Reprogramming towards pluripotency

requires AID-dependent DNA demethylation. *Nature*, **463**, 1042–1047.
7. Popp,C., Dean,W., Feng,S., Cokus,S.J., Andrews,S., Pellegrini,M., Jacobsen,S.E. and Reik,W. (2010) Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature*, **463**, 1101–1105.
8. Morgan,H.D., Dean,W., Coker,H.A., Reik,W. and Petersen-Mahrt,S.K. (2004) Activation-induced cytidine deaminase deaminates 5-methylcytosine in DNA and is expressed in pluripotent tissues: implications for epigenetic reprogramming. *J. Biol. Chem.*, **279**, 52353–52360.
9. Tahiliani,M., Koh,K.P., Shen,Y., Pastor,W.A., Bandukwala,H., Brudno,Y., Agarwal,S., Iyer,L.M., Liu,D.R., Aravind,L. *et al.* (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, **324**, 930–935.
10. 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.
11. Kriaucionis,S. and Heintz,N. (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, **324**, 929–930.
12. Bhutani,N., Burns,D.M. and Blau,H.M. (2011) DNA demethylation dynamics. *Cell*, **146**, 866–872.
13. Millar,C.B., Guy,J., Sansom,O.J., Selfridge,J., MacDougall,E., Hendrich,B., Keightley,P.D., Bishop,S.M., Clarke,A.R. and Bird,A. (2002) Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. *Science*, **297**, 403–405.
14. Riccio,A., Aaltonen,L.A., Godwin,A.K., Loukola,A., Percesepe,A., Salovaara,R., Masciullo,V., Genuardi,M., Paravatou-Petsotas,M., Bassi,D.E. *et al.* (1999) The DNA repair gene MBD4 (MED1) is mutated in human carcinomas with microsatellite instability. *Nat. Genet.*, **23**, 266–268.
15. Otwinowski,Z., Borek,D., Majewski,W. and Minor,W. (2003) Multiparametric scaling of diffraction intensities. *Acta Crystallogr. A*, **59**, 228–234.
16. 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. *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 213–221.
17. 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. *J. Biol. Chem.*, **278**, 5285–5291.
18. Emsley,P. and Cowtan,K. (2004) Coot: model-building tools for molecular graphics. *Acta. Crystallogr. D. Biol. Crystallogr.*, **60**, 2126–2132.
19. Petronzelli,F., Riccio,A., Markham,G.D., Seeholzer,S.H., Genuardi,M., Karbowski,M., Yeung,A.T., Matsumoto,Y. and Bellacosa,A. (2000) Investigation of the substrate spectrum of the human mismatch-specific DNA N-glycosylase MED1 (MBD4): fundamental role of the catalytic domain. *J. Cell. Physiol.*, **185**, 473–480.
20. 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 Res.*, **40**, 4841–4849.
21. 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. *EMBO J.*, **19**, 758–766.
22. 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.
23. 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. *Nat. Struct. Biol.*, **5**, 1058–1064.
24. He,Y.F., Li,B.Z., Li,Z., Liu,P., Wang,Y., Tang,Q., Ding,J., Jia,Y., Chen,Z., Li,L. *et al.* (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science*, **333**, 1303–1307.
25. Parikh,S.S., Mol,C.D., Slupphaug,G., Bharati,S., Krokan,H.E. and Tainer,J.A. (1998) Base excision repair initiation revealed by crystal structures and binding kinetics of human uracil-DNA glycosylase with DNA. *EMBO J.*, **17**, 5214–5226.
26. Mol,C.D., Arvai,A.S., Begley,T.J., Cunningham,R.P. and Tainer,J.A. (2002) Structure and activity of a thermostable thymine-DNA glycosylase: evidence for base twisting to remove mismatched normal DNA bases. *J. Mol. Biol.*, **315**, 373–384.
27. Petronzelli,F., Riccio,A., Markham,G.D., Seeholzer,S.H., Stoerker,J., Genuardi,M., Yeung,A.T., Matsumoto,Y. and Bellacosa,A. (2000) Biphasic kinetics of the human DNA repair protein MED1 (MBD4), a mismatch-specific DNA N-glycosylase. *J. Biol. Chem.*, **275**, 32422–32429.
28. Drohat,A.C. and Stivers,J.T. (2000) Escherichia coli uracil DNA glycosylase: NMR characterization of the short hydrogen bond from His187 to uracil O2. *Biochemistry*, **39**, 11865–11875.
29. Berti,P.J. and McCann,J.A. (2006) Toward a detailed understanding of base excision repair enzymes: transition state and mechanistic analyses of N-glycoside hydrolysis and N-glycoside transfer. *Chem. Rev.*, **106**, 506–555.
30. McCann,J.A. and Berti,P.J. (2008) Transition-state analysis of the DNA repair enzyme MutY. *J. Am. Chem. Soc.*, **130**, 5789–5797.
31. McCann,J.A. and Berti,P.J. (2007) Transition state analysis of acid-catalyzed dAMP hydrolysis. *J. Am. Chem. Soc.*, **129**, 7055–7064.
32. Lee,S. and Verdine,G.L. (2009) Atomic substitution reveals the structural basis for substrate adenine recognition and removal by adenine DNA glycosylase. *Proc. Natl Acad. Sci. USA*, **106**, 18497–18502.
33. Brinkmeyer,M.K., Pope,M.A. and David,S.S. (2012) Catalytic contributions of key residues in the adenine glycosylase MutY revealed by pH-dependent kinetics and cellular repair assays. *Chem. Biol.*, **19**, 276–286.
34. Vocadlo,D.J., Davies,G.J., Laine,R. and Withers,S.G. (2001) Catalysis by hen egg-white lysozyme proceeds via a covalent intermediate. *Nature*, **412**, 835–838.
35. Manvilla,B.A., Maiti,A., Begley,M.C., Toth,E.A. and Drohat,A.C. (2012) Crystal Structure of Human Methyl-Binding Domain IV Glycosylase Bound to Abasic DNA. *J. Mol. Biol*, **420**, 164–175.
36. Werner,R.M., Jiang,Y.L., Gordley,R.G., Jagadeesh,G.J., Ladner,J.E., Xiao,G., Tordova,M., Gilliland,G.L. and Stivers,J.T. (2000) Stressing-out DNA? The contribution of serine-phosphodiester interactions in catalysis by uracil DNA glycosylase. *Biochemistry*, **39**, 12585–12594.