HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2015 March 18.

Published in final edited form as:

Nature. 2014 February 20; 506(7488): 391–395. doi:10.1038/nature12905.

Structure of a *Naegleria* Tet-like dioxygenase in complex with 5-methylcytosine DNA

Hideharu Hashimoto[1], **June E. Pais**[2], **Xing Zhang**[1], **Lana Saleh**[2], **Zheng-Qing Fu**[3], **Nan Dai**[2], **Ivan R. Corrêa Jr.**[2], **Yu Zheng**[2,*], and **Xiaodong Cheng**[1,*]

[1]Departments of Biochemistry, Emory University School of Medicine, 1510 Clifton Road, Atlanta, GA 30322, USA

[2]New England Biolabs, 240 County Road, Ipswich, MA 01938, USA

[3]Department of Biochemistry & Molecular Biology, University of Georgia, Athens, GA 30602 USA, and Sector 22, Advanced Photon Source, Argonne National Laboratory, Argonne, IL 60439, USA

Abstract

Cytosine residues in mammalian DNA occur in five forms, cytosine (C), 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). The ten-eleven translocation (Tet) dioxygenases convert 5mC to 5hmC, 5fC and 5caC in three consecutive, Fe(II)- and α-ketoglutarate-dependent oxidation reactions[1–4]. The Tet family of dioxygenases is widely distributed across the tree of life[5], including the heterolobosean amoeboflagellate *Naegleria gruberi*. The genome of *Naegleria*[6] encodes homologs of mammalian DNA methyltransferase and Tet proteins[7]. Here we study biochemically and structurally one of the *Naegleria* Tet-like proteins (NgTet1), which shares significant sequence conservation (approximately 14% identity or 39% similarity) with mammalian Tet1. Like mammalian Tet proteins, NgTet1 acts on 5mC and generates 5hmC, 5fC and 5caC. The crystal structure of NgTet1 complexed with DNA containing a 5mCpG site revealed that NgTet1 uses a base-flipping mechanism to access 5mC. The DNA is contacted from the minor groove and bent towards the major groove. The flipped 5mC is positioned in the active site pocket with planar stacking contacts, Watson–Crick polar hydrogen bonds and van der Waals interactions specific for 5mC. The sequence conservation between NgTet1 and mammalian Tet1, including residues involved in structural integrity and functional significance, suggests structural conservation across phyla.

Correspondence should be addressed to: Xiaodong Cheng (xcheng@emory.edu; Tel: +1 404 727 8491; Fax: +1 404 727 3746), Yu Zheng (zhengy@neb.com; Tel: +1 978 380 7441; Fax: +1 978 921 1350).

Author Contributions HH performed antibody-based and TDG-based activity assays, crystallographic experiments and expression of mouse Tet1 in HEK293T cells. XZ made the overexpression construct in E. coli, developed (together with HH) assay conditions and performed NgTet1-8 sequence analysis. JP and LS performed kinetic assays using LC-MS method and JP characterized the mutants. Z-Q.F performed crystallographic phasing calculations and generated an initial poly-alanine model. ND and IC developed the LC-MS method for detection of modified cytosine residues. XZ, YZ, and XC organized and designed the scope of the study, and all were involved in analyzing data and preparing the manuscript.

Accession Number: The X-ray structures (coordinates and structure factor files) of NgTet1 with bound DNA have been submitted to PDB under accession number 4LT5.

The free-living amoeboflagellate *Naegleria gruberi* has eight Tet/JBP-like dioxygenases (NgTet1-8; Extended Data Fig. 1). The NgTet proteins vary in length, but all contain a conserved core region of ~210 residues including the invariant Fe(II)-binding histidines and aspartate (the HxD…H motif). We measured NgTet1 activity using various double-stranded DNA as substrates, each containing a single modified base X within a G:X pair in a CpG sequence. We used antibodies specific for 5hmC, 5fC and 5caC (Extended Data Fig. 2a–c). Using 5mC-containing DNA as substrate, 5hmC (the first reaction product) and 5caC (the last reaction product) are detected in the presence of α-ketoglutarate (αKG), but not with *N*-oxalylglycine (NOG) (Fig. 1a). NgTet1 initially produces 5hmC at 5min, 5fC between 5 to 10min and finally 5caC at 15min under the assay conditions (Fig. 1b). NgTet1 is active on all three DNA substrates containing 5mC, 5hmC or 5fC, generating 5caC (Fig. 1c). We applied quantitative mass spectrometry to monitor the kinetics of product formation (Fig. 1d and Extended Data Fig. 2d). When the amount of 5mC rapidly disappears (2–5min), a peak of 5hmC forms transiently before being converted to 5fC and 5caC products (Fig. 1d). The first conversion from 5mC to 5hmC is faster (k_{obs}=21h^{-1}) than the second conversion from 5hmC (k_{obs}≈3h^{-1}). In addition, we used human thymine DNA glycosylase to probe the products generated by NgTet1 (Extended Data Fig. 2e).

We determined the crystal structure of NgTet1 with a 14-base-pair (bp) oligonucleotide containing a single methylated CpG site in the presence of Mn^{2+} and NOG to form a catalytically inert complex, at 2.9Å resolution (Extended Data Table 1). Like other structurally characterized αKG-dependent dioxygenases[8], NgTet1 has a core double-stranded β-helix fold that binds Fe(II) and αKG (Fig. 2a). Two twisted β-sheets (a four-stranded minor sheet and an eight-stranded major sheet) pack together with five helices on the outer surface of the major sheet to form a three-layered structure (Fig. 2a–b). The unequal number of strands of the two sheets creates an active site located asymmetrically on the side of the molecule where the extra strands of the major sheet are located. A 3_{10}-helix (h3 or h7) marks the end of each sheet and sits at the entrance to the active site. Two long loops associated with the 3_{10}-helices provide most of the functionally important residues. The hairpin loop (L1) between β5 and h3 of the major sheet recognizes the intrahelical guanine opposite to the target 5mC via Ser148, and the extended loop (L2) connecting h7 from the minor sheet to the β7 of the major sheet (Fig. 2b) is responsible for binding of the metal ion (His229 and Asp231) and the flipped-out 5mC (Asp234).

The DNA is bound to the basic surface of the protein with substantial protein-induced distortions from B-form DNA (Fig. 2c and Extended Data Fig. 3). The phosphate backbone flanking the CpG site is kinked ~65° and concurrently, one of the 5mC nucleotides flips out. Phosphate-protein contacts are concentrated on the four phosphates surrounding the flipped 5mC (Extended Data Fig. 3a–b), involving residues of the 3_{10}-helices h3 (Ala156) and h7 (Arg224) (Fig. 2d–f).

The enzyme approaches DNA from the minor groove, which is markedly widened near the flipped 5mC to ~10Å in groove width due to severe bending of the DNA. The tip of the hairpin loop, Ser148, hydrogen bonds with the intrahelical orphaned guanine (Fig. 2e), while the side chain of Gln310 of the C-terminal helix α10 makes bifurcated hydrogen bonds with the 3'-guanine of the flipped 5mC (Fig. 2d). Such base specific interactions would account

for the preference of NgTet1 for 5mCpG as substrate. Replacing the 3'-guanine with adenine, thymine or cytosine resulted in reduction of the rate of 5mC conversion by a factor of ~1.75, 3.8 and 5.8, respectively (Fig. 2g). Similarly, mutating Gln310 to alanine (Q310A) resulted in ~60% reduction of 5mC conversion (Fig. 1e). No direct interaction was observed for the 5mC in the opposite strand (Fig. 2d), consistent with NgTet1 being active on both fully and hemi-methylated CpG sites (Extended Data Fig. 2f).

The extrahelical 5mC is bound in a cage-like active site via stacking of the flipped base in between Phe295 and the guanidino group of Arg224 (Fig. 2e). Superimposition of a normal intrahelical 5mC onto the flipped 5mC suggests a very small rotation around the glycosidic bond (Extended Data Fig. 3d). The polar groups of the 5mC ring that normally form the Watson–Crick pairings with guanine now form hydrogen bonds with the side-chain amide group of Asn147 (interacting with the O2 oxygen), the side-chain imidazole ring of His297 (interacting with the N3 nitrogen), and the side chain carboxylate oxygen atoms of Asp234 (interacting with the N4 nitrogen) (Fig. 2h). Interactions with the exocyclic amino group N4 (NH_2) define the binding pocket specificity for a cytosine rather than thymine. Mutations of Asn147, His297 or Asp234 resulted in much reduced (N147D, H297Q, H297N, D234N) or nearly abolished activity (D234A) on 5mCpG (Fig. 1e). The target methyl group – wedged between the hydrophobic side chains of Ala212 and Val293 (Fig. 2i) - is ~5.2Å from the metal ion, which is similar to the observed distance (~4.5Å) between the substrate atom to be oxidized and the iron in most structurally characterized αKG-oxygenases[8]. An additional hydroxyl or formyl or carboxylate group attached to the C5 methyl could fit into the space, consistent with 5hmC or 5fC or 5caC being a substrate/product of NgTet1 (Extended Data Fig. 4a–e).

The metal ion Mn^{2+} has six ligands in an octahedral coordination (Fig. 2j). The NOG molecule is involved in extensive polar and hydrophobic interactions with the protein (Fig. 2k–l). The importance of these interactions is underscored by the fact that NOG-interacting residues are invariant or highly conserved among the eight NgTet-like homologs examined (Extended Data Fig. 1b). The NOG carboxylate group at the C5 position projects towards the interior hydrophobic core sandwiched between the two β-sheets (Fig. 2k), while the negatively charged carboxylate is balanced by the interaction with the invariant Arg289 (Fig. 2l). The deep binding pocket of NOG (which is concealed in the NgTet1-DNA complex) suggests that the cofactor αKG binding precedes that of the DNA substrate (Extended Data Fig. 5), and stabilizes the NgTet1 structure by interacting with Arg289 buried in the hydrophobic core.

The αKG dioxygenase family[8,9] includes members of the AlkB-like DNA/RNA repair enzymes[10]. We compared the complex structure of NgTet1-DNA-NOG-Mn^{2+} to that of *Escherichia coli* AlkB-DNA-αKG-Mn^{2+} (Fig. 3) and its human homolog ABH2 (Extended Data Fig. 6)[11,12] (the only other dioxygenases acting on nucleic acids structurally characterized in complex with DNA). The structures of NgTet1 and AlkB can be superimposed via the core elements of the jelly-roll fold (colored in Fig. 3a–b). Both enzymes contain the hairpin loop (L1) after strand β5 and the active-site loop (L2) prior to strand β7. Besides the N-terminal and C-terminal additions (Extended Data Fig. 6a), NgTet1 has, within the core region, extra helices α5 and α6, immediately after the kinked helix α4

(owing to Pro72 located in the middle of the helix). In the places of h3 and h7, two 3_{10}-helices unique to NgTet1 (Fig. 3a), AlkB has two additional β-strands, adjacent to β5 of the major sheet and β11 of the minor sheet, respectively (Fig. 3b). Unique to AlkB is an additional 12-residue-long loop (L3) prior to strand β5 making DNA backbone contacts, whereas the corresponding loop L3 in NgTet1 is a 4-residue short loop containing an invariant Lys137 among the eight NgTet proteins (Extended Data Fig. 1c).

The most striking difference between NgTet1 and AlkB is that the bound DNA molecules lie nearly perpendicular to each other relative to the proteins (Fig. 3c–d). Both DNA molecules are bound against the basic surface of the protein (Fig. 3c–d), composed partly from the positively charged residues of the minor sheet unique to AlkB or the C-terminal helix α10 unique to NgTet1. We note that the C-terminal additions of all NgTet proteins (Extended Data Fig. 1b) and mammalian Tet enzymes are heavily enriched with basic residues that could also potentially interact with DNA. The vastly different protein-DNA interactions may reflect the fact that AlkB recognizes a damaged base pair whereas NgTet1 recognizes a normal Watson–Crick base pair during the initial protein-DNA encounter. Like DNA methyltransferases[13] and DNA base excision repair enzymes[14], NgTet1 and AlkB (and ABH2) use a base flipping mechanism to access the DNA bases where modification or repair occurs[15].

The perpendicular DNA binding orientation also dictates how the flipped target base binds in the active site. The target nucleotide is simply rotated along the phosphodiester backbone (Extended Data Fig. 3d)[16], probably due to extensive protein–phosphate pinches[17] surrounding the flipped nucleotide. Thus, the flipped target bases, 5mC in NgTet1 and 3mC in AlkB, are also nearly perpendicularly positioned in their respective active sites (Fig. 3e). Yet, the distance between the target methyl group and the metal ion remains the same (~5Å), consistent with a conserved chemical reaction. Also conserved is the ion-pair interaction of an active site arginine with the C1 carboxylate group of NOG of NgTet1 or αKG of AlkB - which is nearly superimposable (Extended Data Fig. 6c). However, the position of this arginine is different in the two enzymes in accordance with the perpendicular orientation of the target bases (Fig. 3f–g). Therefore, the two enzymes approach the DNA substrates differently resulting in distinct conformations of flipped target bases yet maintaining the ion-pair interaction with NOG/αKG.

Here we described the first structure of a Tet-like dioxygenase, NgTet1, which is capable of converting 5mC to 5hmC, 5fC and 5caC. In mammalian genomes, the products of Tet enzymes include 5hmC in both CpG and non-CpG sequence context[18–20]. Likewise, NgTet1 is active on 5mCpG and 5mCpA (in a reduced rate). Structurally, NgTet1 represents the core structure of the catalytic domain of the mammalian Tet enzymes. The mammalian Tet proteins have their catalytic domains located in the C-terminal part of the proteins[1] with an atypical insertion of ~300 residues not found in other αKG-dioxygenases (Fig. 4a). The insertion separates the two halves of the ferrous binding motif, HxD…H. In addition, a stretch of ~50 residues containing a unique symmetrically spaced four cysteine residues CX_7CXCX_7C is located in the N-terminal portion of the catalytic domain. Removing these two insertions shows that NgTet1 and mammalian Tet1 share ~14% identity or ~39% similarity (Fig. 4b), the highest conservation among the pairwise comparisons of NgTet1

and other αKG-oxygenases examined (Extended Data Table 2 and Extended Data Fig. 6–7). The sequence conservation is scattered throughout the entire region, including the residues involved in structural integrity and those with functional significance (DNA binding, base specific interactions, metal ion and αKG bindings) (Extended Data Table 3). The conservation extends beyond the core of the jelly-roll fold shared by NgTet1 and AlkB/ABH2 (for example, Lys86-Glu108 ion pair in Fig. 4c), indicating NgTet1 and mammalian Tet1 share an overall higher degree of structural conservation owing to their common substrate and enzymatic properties. Another structural conservation between NgTet1 and mammalian Tet1 involves an invariant proline located in the middle of helix α4, causing a kink (Fig. 4d) – a unique feature which might be conserved among the Tet/JBP family as the kinked helix α4, together with the following helices α5 and α6, is composed of a stretch of residues predicted to be Tet/JBP specific[1]. No corresponding helices α5 and α6 are present in other structurally characterized αKG-oxygenases examined (Extended Data Fig. 6–7).

The two large insertions of mammalian Tet1 lie in the loop (L3) between helix α2 and strand β5 (Cys-rich region) and the loop between strands β8 and β9 (Fig. 4c). The Cys-rich insertion is in the DNA binding interface and thus might play roles in DNA binding. The large 300-residue insertion, which shares significant sequence similarity to the C-terminal domain (CTD) of RNA polymerase II[21], points away from the catalytic core and has a potential regulatory function. Like mammalian Tet proteins, histone lysine-specific demethylase LSD1 has an atypical insertion of the Tower domain into the catalytic amine oxidase domain, whereas the closely related LSD2 is devoid of the insertion and is active[22]. Similarly, deletions of the CTD-insertion in mouse Tet1 catalytic domain retain activity when expressed in HEK293T cells (Fig. 4e), providing further support of the evolutionary conservation between NgTet1 and mammalian Tet proteins.

Full Methods

We designed synthetic NgTet1 by optimizing codon set for *Escherichia coli* and assembling the gene by overlapping oligonucleotides. We generated a hexahistidine–SUMO (small ubiquitin-like modifier)-tagged construct containing full length NgTet1 (pXC1010). The tag was cleaved and the NgTet1 was crystallized with fully methylated 14-bp DNA in the presence of $MnCl_2$, and *N*-oxalylglycine (NOG). The structure was determined by single anomalous diffraction using bromine-labeled (5-BrdU) DNA. The 5mC dioxygenase activity of NgTet1 was assayed by three methods including specific antibodies, base excision by TDG and liquid chromatography–mass spectrometry.

Protein expression and purification

We generated a hexahistidine–SUMO-tagged construct (pXC1010) of full length NgTet1 (321 amino acids; XP_002667965.1). The protein was expressed in *E. coli* BL21 (DE3)-Gold cells with the RIL-Codon plus plasmid (Stratagene). Cultures were grown at 37°C until the OD_{600} reached 0.5; the temperature was then shifted to 16°C, and isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to 0.4 mM to induce expression. Cells were re-suspended with 4 volumes of 500 mM NaCl, 20 mM sodium phosphate, pH 7.4, 20 mM imidazole, 1 mM dithiothreitol (DTT) and 0.25 mM phenylmethyl-sulphonyl fluoride and

sonicated for 5 min (1s on and 2s off). The lysate was clarified by centrifugation at 38,000 *g* for 60 min. Hexahistidine fusion protein was isolated on a nickel-charged chelating column (GE Healthcare). The His-SUMO tag was removed by incubating with Ulp1 for 16 h at 4°C. The cleaved protein was further purified by a tandem HiTrap Q and SP column (GE-Healthcare) and eluted from SP column and concentrated. The protein was then loaded onto a Superdex 75 (16/60) column (equilibrated with 150 mM NaCl, 20 mM HEPES, pH 8.0, 1 mM DTT) where it eluted as a single peak corresponding to a monomeric protein.

Mutagenesis of NgTet1 was performed with the Q5 Site-Directed Mutagenesis Kit (NEB) using custom oligonucleotide primers that incorporated the desired amino acid changes. Wild-type and variant proteins were expressed as N-terminal 6XHis-tagged constructs (pTXB1) from *E. coli* T7 Express competent cells (NEB #C2566), as described above except induced with 50 μM IPTG. Cells were suspended in 20 mM Tris-HCl, pH 7.5, and 20 mM NaCl and lysed by sonication. The clarified lysates were loaded onto a HiTrap Heparin HP column (GE-Healthcare) and eluted with a 0.02–1.0 M NaCl gradient over 30 column volumes. Fractions were pooled and diluted 2-fold in 20 mM Tris, pH 7.5, 400 mM NaCl, and 10 mM imidazole, and loaded directly onto a HisTrap HP column (GE-Healthcare) and eluted with a 10–500 mM imidazole gradient. Fractions were pooled and dialyzed against 20 mM Tris, pH 7.5, and 300 mM NaCl, concentrated and stored in 50% glycerol at −20°C.

NgTet1 activity assay

Various FAM-labeled 32-bp DNA molecules containing a hemi-modified CpG dinucleotide were used as substrates (Fig. 1a–c and Extended Data Fig. 2a–c, e): FAM-5'-TCG GAT GTT GTG GGTCAG **XG**C ATG ATA GTG TA -3' and 3'–AGC CTA CAA CAC CCA GTC **GC**G TAC TAT CAC AT-5', where X=5mC, 5hmC or 5fC (synthesized by the New England Biolabs, Inc.). Reactions were carried out at 34°C for 1h with 2 μM of NgTet1 and 1 μM of DNA in 50 mM BisTris-HCl, pH 6.0, 75 μM $(NH_4)_2FeSO_4$, 1 mM α-ketoglutarate (αKG) or *N*-oxalylglycine (NOG), 2 mM ascorbic acid, and 100 mM NaCl. After the reaction, 1 μg of Proteinase K per 20 μL was added into the reaction mixture and incubated at 23°C for 30 min. DNA was precipitated by ethanol and dissolved in 2 mM Tris-HCl, pH 7.0 for analysis by antibody and TDG glycosylase.

5hmC, 5fC and 5caC detection by antibodies

DNA samples were mixed with 2× loading buffer (98% formamide, 1 mM EDTA and 1 mg ml^{-1} of bromophenol-blue and xylene cyanol) and loaded onto 15% denaturing gel (15% acrylamide, 7 M urea and 24% formamide in 1× TBE buffer). 10 pmol of 5hmC, 5fC or 5caC control oligos and their 2 fold serial dilutions were used as standard (Fig. 1a–c and Extended Data Fig. 2a–c, e) along with DNA reacted with NgTet1 in the presence of NOG or αKG. The gels were run at 200 V for 75 min. FAM-labeled single stranded DNA was visualized by UV exposure to estimate the amounts of sample loading. DNA was transferred to Zeta-Probe® GT Blotting Membrane (BIO-RAD) by Trans-Blot® SD Semi-Dry Electrophoretic Transfer Cell (BIO-RAD) at 2 mA cm^{-2} for 1h, then cross-linked to the membrane by UV for 10 min. The membrane was blocked with 5% skim milk for 30 min at 4°C, and incubated with anti-5hmC polyclonal antibody (Active Motif catalog no. 39792;

1:5000 dilution), anti-5fC polyclonal antibody (Active Motif catalog no. 61227; 1:4000 dilution) or anti-5caC polyclonal antibody (Active Motif catalog no. 61225; 1:2000 dilution) in 5% skim milk containing TBS-T at 4°C for 16h. The membranes were washed twice with TBS-T for 5 min, then incubated with HRP conjugated anti-rabbit antibody (Southern Biotech, 1:2000) in TBS-T for 45 min at 23°C. After two TBS-T washes twice for 5 min, the membranes were treated with western lightning plus ECL (Perkin Elmer) and exposed for 4s or 15s.

The Active Motif polyclonal antibodies were used by others for dot blot[2,23–25], immunohistochemistry[23,26,27], immunofluorescence[2,24], MeDIP[25] and MeDIP-Seq[28].

5fC and 5caC detection by TDG

A 20 μL of reaction mixture containing 0.25 μM of 32-bp FAM labeled DNA and 25 μM of TDG catalytic domain[29] was incubated at 30°C for 15 min in 10 mM BisTris-HCl, pH 6.0, 100 mM NaCl, 1 mM EDTA and 0.1% BSA. After the reaction, 0.1 N of NaOH were added and the samples boiled for 5 min at 95°. The samples were mixed with an equal volume of 2× Loading Dye (98% formamide, 1 mM EDTA and 1 mg $ml^{-1}$of bromophenol-blue/xylene cyanol), and loaded onto a 15% denaturing gel. FAM fluorescence was visualized by Typhoon Trio (GE Healthcare).

Substrate specificity analysis using liquid chromatography–mass spectrometry (LC-MS)

To prepare the samples for LC-MS measurements, each reaction mixture (as described below) was incubated at 34°C for the specified time, and subsequently quenched by heating at 95 °C for 3 min. The samples were then placed on ice for 3 min followed by digestion using proteinase K (NEB) at a final concentration of 1 μg/μL for 1h at 50°C. The DNA was recovered by using QIAquick® Nucleotide Removal Kit (QIAGEN, Valencia, CA). A mixture of nuclease P1 (Sigma-Aldrich, St. Louis, MO), Antarctic phosphatase (NEB) and DNAse I (NEB) was used to digest the recovered DNA. LC-MS was performed on an Agilent 1200 series (G1316A UV Detector, 6120 Mass Detector, Agilent, Santa Clara, CA) with Waters Atlantis T3 (4.6×150 mm, 3 μm, Waters, Milford, MA) column with in-line filter and guard. The data points were best fitted by a single exponential equation to follow the disappearance of 5mC (GraphPad Prism software) (Fig. 1d and 2g). The k_{obs} value for 5hmC disappearance is estimated from the 10 min time point (when nearly all 5mC to 5hmC conversion has been completed) and beyond (Fig. 1d).

For quantitative analyses of various 5mC oxidative species, either the 56-bp hemi-methylated dsDNA-1 (Fig. 1d) or genomic DNA (gDNA) of Hela cells (NEB #N4006S) (Extended Data Fig. 2d) were used as substrates. The reaction mix, incubated for 1h at 34°C, contained 20 μl of 4 μM NgTet1, 2 μM dsDNA, 50 mM Bis-Tris pH 6, 50 mM NaCl, 1 mM DTT, 2 mM ascorbic acid, 1 mM αKG, 100 μM $FeSO_4$. For reaction with gDNA, 20 μM NgTet1 and 2.5 μg gDNA were used and the experiment repeated three times. For substrate specificity analyses (Fig. 2g), the 56-bp dsDNA-1 containing a hemi-methylated single 5mCpN (N=G, A, T, C) was used. For activity analyses of mutant proteins (Fig. 1e), the

fully-methylated 56-bp dsDNA-2 was used. The reaction conditions are the same as the above except the reaction time was 10 min and the experiment repeated three times.

Hemi-methylated dsDNA-1 (56-bp, M=5mC, X:Y=G:C, A:T, T:A or C:G):

5’-
CGGCGTTTCCGGGTTCCATAGGCTCCGCCC**M**__**X**__GGCTCTGATGACCAGGGCA
TCACA-3’

3’-
GCCGCAAAGGCCCAAGGTATCCGAGGCGGGG**Y**CCGAGACTACTGGTCCCGT
AGTGT-5’

Fully-methylated dsDNA-2 (56-bp, M=5mC):

5’-
CGGCGTTTCCGGGTTCCATAGGCTCCGCCC**M**__GGACTCTGATGACCAGGGCA
TCACA-3’

3’-
GCCGCAAAGGCCCAAGGTATCCGAGGCGGGG**M**__CTGAGACTACTGGTCCCG
TAGTGT-5’

Crystallography

We used various lengths of DNA for co-crystallization. Starting from a 13-bp DNA with five and six bps on either side of a CpG dinucleotide, we fixed the 5-bps side and lengthened at the 6-bps side one bp at a time to obtain 14-bp, 15-bp and 16-bp oligonucleotides. Crystals complexed with either the 14-bp or the 16-bp DNA were obtained, and the crystals with the 14-bp DNA (5’-TGG AA(5mC) GCA ATT CT-3’ and 5’-AGA ATT G(5mC)G TTC CA-3’) diffracted X-rays to higher resolution. An equimolar mixture of protein and DNA (0.5 mM) were incubated in 2 mM NOG, 2 mM $MnCl_2$, 100 mM NaCl, and 20 mM HEPES-NaOH, pH 8.0, for 16h at 4°C. Crystallization was carried out in a 2 μl sitting drop with equal volume of the complex solution and well solution. Crystals appeared within 2–7 days at 16°C under the conditions of 2 M ammonium sulfate, 0.1 M BisTris-HCl, pH 5.4–5.6. The complex with brominated DNA, 5’-XGG AA(5mC) GGA AXT CX-3’ and 5’-AGA ATT C(5mC)G TTC CA-3’ (X=5-BrdU), was crystallized under the condition of 20% (w/v) polyethylene glycol 3350, and 0.2 M sodium citrate, pH 8.0. Crystals were cryoprotected by soaking in mother liquor supplemented with 20% (v/v) glycerol or ethylene glycol and by plunging into liquid nitrogen.

X-ray diffraction datasets were collected at the SER-CAT beamline at the Advanced Photon Source, Argonne National Laboratory and processed using HKL2000[30]. We used single anomalous diffraction (SAD) to obtain crystallographic phases using 5-BrdU containing crystals. Two date sets from two crystals were collected at 100K, each containing 360 frames of 2-degree oscillation at a wavelength of 0.91931Å, at a slightly higher energy (100 eV) than the Bromine (Br) absorption edge. Each individual data set showed only an anomalous signal to ~6Å, but merging the two data sets during data reduction resulted in a high redundancy (~55) and an anomalous signal to 4.5Å. Three Br sites were subsequently

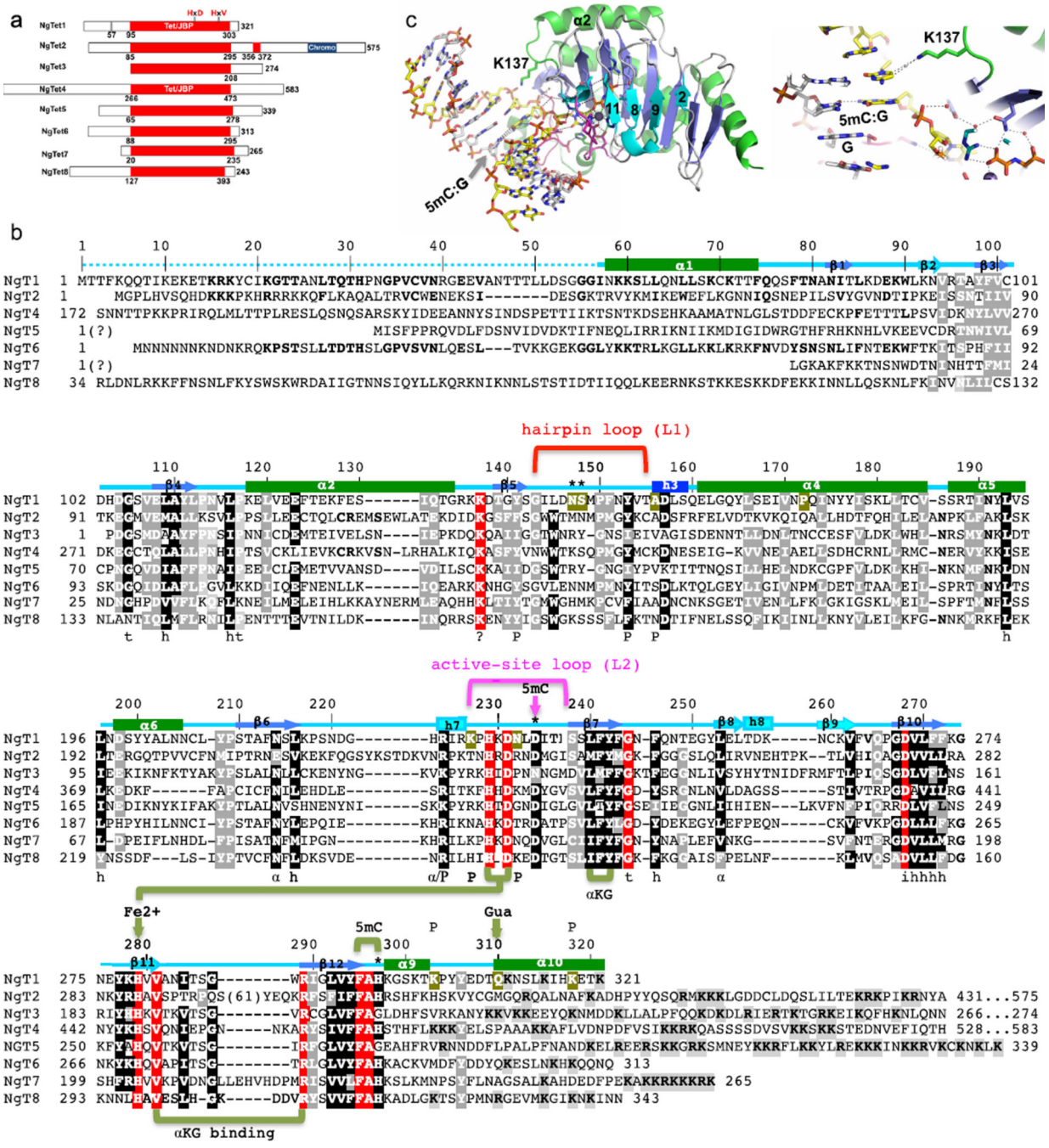
found by Parallel-SHEXLD[31]. The SGXPro program[32] was used to calculate experimental phases to 4.5Å and a poly-alanine model were built. The PHASER module of the PHENIX software suite[33] was used to do phase extension with partial structure refinement and model building, which generated a more traceable map to 3.5Å and an improved the poly-alanine model of 264 residues. The resulting electron density map for DNA was easily visible and the model was built using program COOT[34]. Finally, PHENIX scripts were used for model refinement against the native data set to 2.9Å resolution (Extended Data Table 1), with an optimized weight for the X-ray target and the stereochemistry or the Atomic Displacement Parameters during the last refinement cycles. The first 56 residues were not modeled in the final structure due to lack of continuous electron density. In addition, the side chains of Ile57 and Lys321 (the last residue) were not modeled. The crystal contains one protein-DNA complex per asymmetric unit. The crystallographic thermal B-factors (~50Å^2) for the central DNA base pairs including 5mCpG are comparable to that of the protein. The outer bases, without any protein contacts, have higher thermal B-factors (~90Å^2 and ~67Å^2, respectively, for both ends), resulting in an averaged B-factor for DNA 1.7 fold higher than that of the protein (Extended Data Table 1 and Extended Data Fig. 3e). The MolProbity statistics[35] for the final structure include 97% favored, 3% allowed and 0% outliers in the Ramachandran plot, 0% of the rotamer outliers, and 4.9 of all-atom clashscore.

The Secondary Structure Matching (SSM) script in COOT[34] generated initial pairwise alignments, followed by visual inspections, between structures of NgTet1 (PDB 4LT5) and *E. coli* AlkB (PDB 3O1M), hABH2 (PDB 3BUC), hABH3 (PDB 2IUW), FTO (PDB 3LFM), P4H (PDB 2JIJ) or TYW5 (PDB 3AL6) (Extended Data Fig. 6–7).

Analysis of the genomic 5hmC in HEK293T cells overexpressing the mouse Tet1 catalytic domain

Flag tagged mouse Tet1 catalytic domain (residues 1367-2039; NP_001240786.1) and its internal deletion constructs were sub-cloned into pcDNA3.1 and transformed into HEK293T cells. Plasmid DNA (2.7 μg) and 4.1 μg of 25 kDa, branched polyethyleneimine (Sigma catalog No. 408727) were incubated in 300 μL of no serum medium for 10 min and added into 80–90% confluent cells in 1.3 mL D-MEM media containing 2% fetal calf serum in 6-well plates. After 48h of transfection, cells were harvested and the genomic DNA was purified with GeneEluteTM mammalian genomic DNA miniprep kit (Sigma catalog No. G1N70). Purified genomic DNA was blotted to a Zeta-Probe® GT Blotting Membrane (BIO-RAD) and then cross-linked to the membrane by UV for 10 min. The membrane was subjected to Western analysis using anti-5hmC antibody (Active Motif catalog no. 39792; 1:10000 dilution) and HRP conjugated anti-rabbit antibody (Southern Biotech, 1:2000). Whole cell lysate (~5×10^5 cells) were loaded onto 10% SDS-PAGE gel, and blotted onto PVDF membrane (BIO-RAD). Flag tagged proteins were detected by anti-Flag M2 antibody (Sigma cat. No. F1804; 1:2000), HRP-conjugated anti-mouse IgG antibody (ROCKLAND cat. No. 610-1319-0500; 1 μg/mL) and Plus-ECL (PerkinElmer cat. No. NEL103001EA).

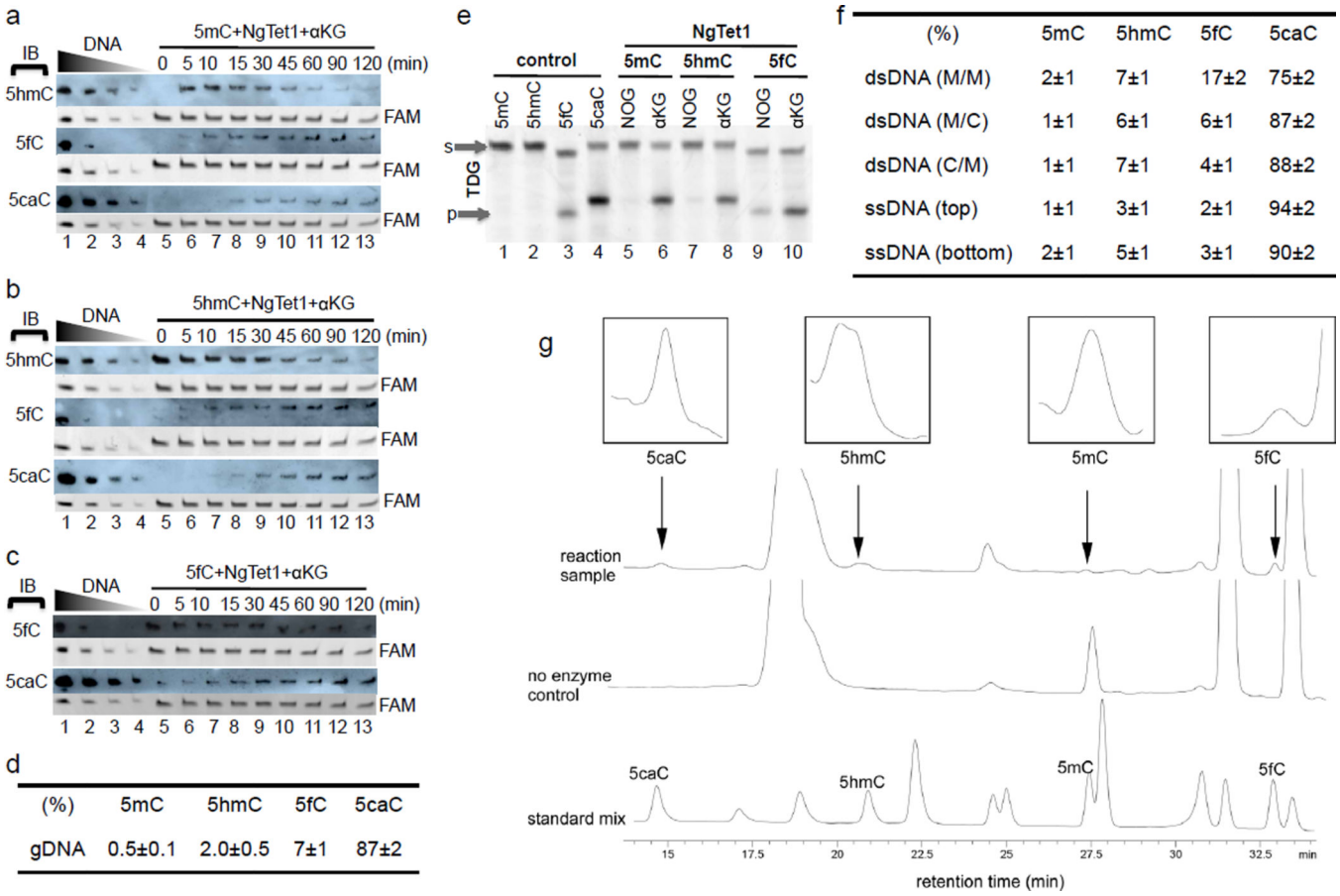
Extended Data



Extended Data Figure 1. Sequence alignment of *Naegleria* Tet-like dioxygenases (1 to 8)

a, Schematic representation of NgTet1-8. **b**, Secondary structural elements are indicated in green (helices), blue (the major sheet) and cyan (the minor sheet). Numbering above the sequences corresponds to NgTet1. White-on-red residues are invariant among the eight sequences examined, while black or gray-highlighted positions are conserved substitutions). Positions highlighted are responsible for various functions as indicated: t for structural turn,

h for hydrophobic core, 5mC for 5mC binding, α or αKG for binding of αKG or NOG, P for DNA phosphate interaction, i for intra-molecular polar interaction (D268) and ? for side chain of K137 pointing to the DNA major groove (panel c). Sequences included are NgTet1 (XP_002667965.1), NgTet2 (XP_002682154.1), NgTet3 (XP_002668005.1), NgTet4 (XP_002676528.1), NgTet5 (XP_002668409.1), NgTet6 (XP_002674105.1), NgTet7 (XP_002668594.1), and NgTet8 (XP_002676954.1). However, the N-terminal sequences for NgTet3, 5, 7, 8 are extended in frame to include more conserved sequence elements until either a putative initiation methionine or the end of a sequence contig (i.e., until a sequencing gap is encountered), therefore the exact N-terminus is unknown. For NgTet6, the sequence of XP_002674105 (177 residues) is likely incomplete at the N-terminus. Extending scaffold 42_31984-32508 to 32980 and allowing for a proper splicing junction results in a protein of 313 residues that shares 51% identity with NgTet1 across the whole protein except for the first 20 residues. Of the five NgTet proteins tested (NgTet1-5), two of them (NgTet1 and NgTet4) have 5mC dioxygenase activities. **c**, An invariant Lys137 among the eight NgTet dioxygenases, located in the loop between helix α2 and strand β5, points to the major groove of DNA with the terminal ε-amino group approximately 4.3 Å away from the base 3’ to the target 5mCpG site. An exchange of a C:G to G:C pair at this position does not affect crystallization. The corresponding loop in AlkB is the long loop L3 (see Fig. 3b) that makes DNA backbone contacts. In mammalian Tet1, the Cys-rich region is predicted to insert within the corresponding loop L3 (see Fig. 4c).



Extended Data Figure 2. Activity of NgTet1 on various DNA substrates

a–c, The time courses (lanes 5–13) of the reactions using 32-bp DNA substrates containing 5mC (panel **a**), 5hmC (panel **b**) or 5caC (panel **c**). Lanes 1–4: antibody sensitivity against 10 pmol of control oligonucleotides and 2 fold serial dilutions. Lanes 5–13: the rate of conversion appears to be the fastest for the reaction of 5mC to 5hmC, and decreases with each subsequent reaction: 5mC to 5hmC > 5hmC to 5fC > 5fC to 5caC. **d**, Activities of NgTet1 (20 μM) on genomic DNA (gDNA) of Hela cells (2.5 μg). After 1 h reaction, 87% of the products are 5caC in gDNA with the remaining being 5fC and 5hmC. The percentages were estimated from integration of the peaks in LC-MS traces. The mean and standard deviation (±s.e.m.) were estimated from three repeated experiments. **e**, Human thymine DNA glycosylase (TDG) excises 5fC and 5caC (but not 5mC and 5hmC) when paired with a guanine in a CpG sequence (lanes 1–4) (He et al., 2011; Maiti and Drohat, 2011; Hashimoto et al., 2012). After NgTet1 reactions with DNA substrates containing 5mC, 5hmC or 5fC, in the presence of αKG, the product DNA containing 5fC and 5caC becomes a substrate for TDG (lanes 6, 8 and 10), but not with NOG (lanes 5 and 7), again demonstrating the production of 5fC and 5caC by NgTet1. **f**, Activities of NgTet1 on 56-bp double-stranded (ds) DNA-2 with methylation on both strands (M/M) or single strand (hemi-methylated either on top M/C or bottom C/M strand) or single-stranded (ss) DNA (Reaction time 1 h and ±s.e.m. estimated from three repeats). We note that an *in vitro* activity of the mouse Tet1 catalytic domain on single-stranded DNA has also been observed (Zhang et al., 2012).

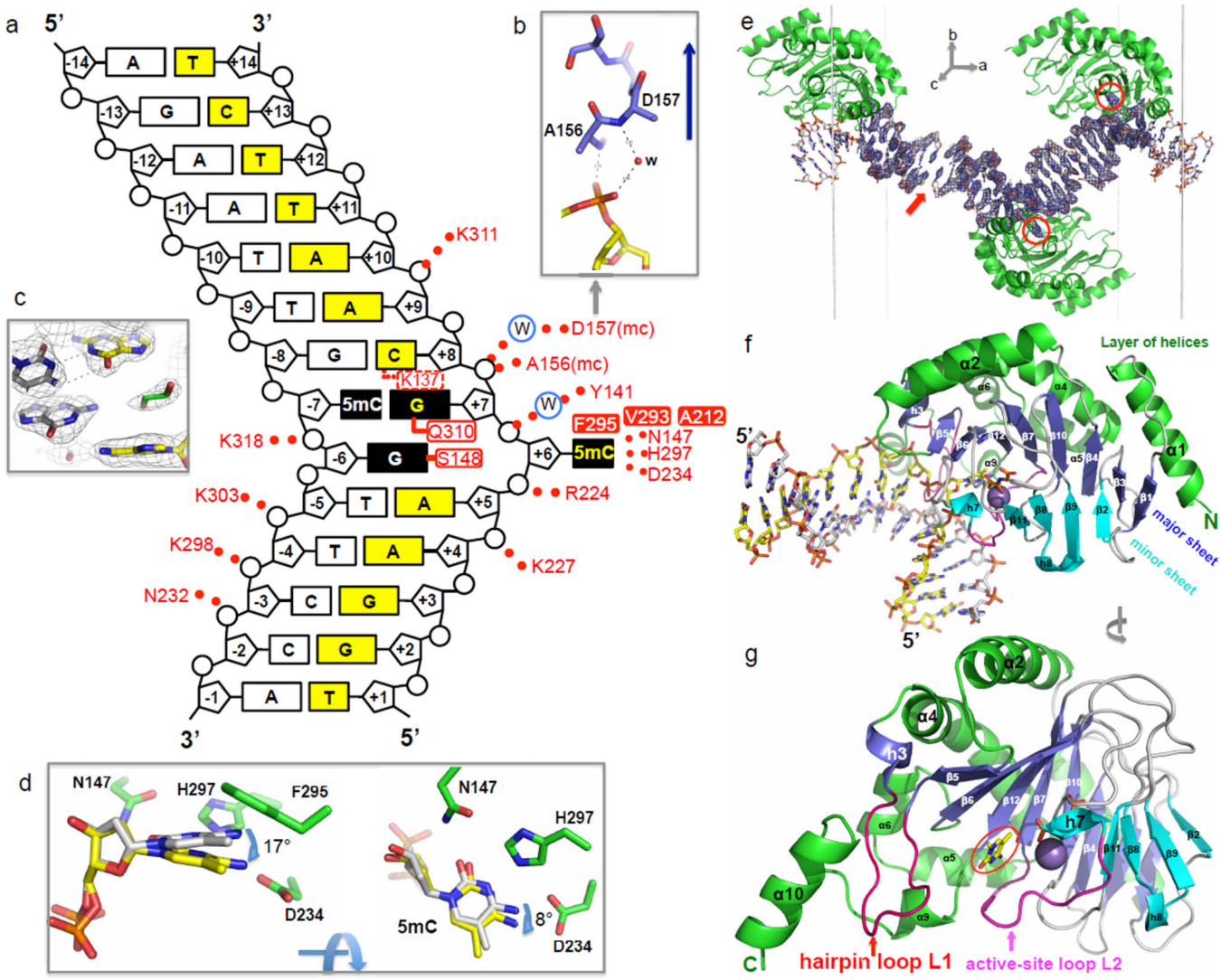
g, LC-MS traces of a sample reaction mix on the hemi-methylated 5mCpG dsDNA-1 (top panel), reaction control with no enzyme (middle panel), and the standard deoxyribonucleoside mix (bottom panel). Arrows indicate peaks of 5mC, 5hmC, 5fC and 5caC. Identities of the peaks are confirmed by comparing the retention time with the standard as well as by mass spectrometry.

Hashimoto, H., Hong, S., Bhagwat, A. S., Zhang, X. & Cheng, X. Excision of 5-hydroxymethyluracil and 5-carboxylcytosine by the thymine DNA glycosylase domain: its structural basis and implications for active DNA demethylation. *Nucleic Acids* Res **40**, 10203–10214 (2012).

He, Y. F. *et al*. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 (2011).

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

Zhang, L., Yu, M. & He, C. Mouse Tet1 protein can oxidize 5mC to 5hmC and 5caC on single-stranded DNA. *Acta Chimica Sinica* **70**, 2123–2126 (2012).



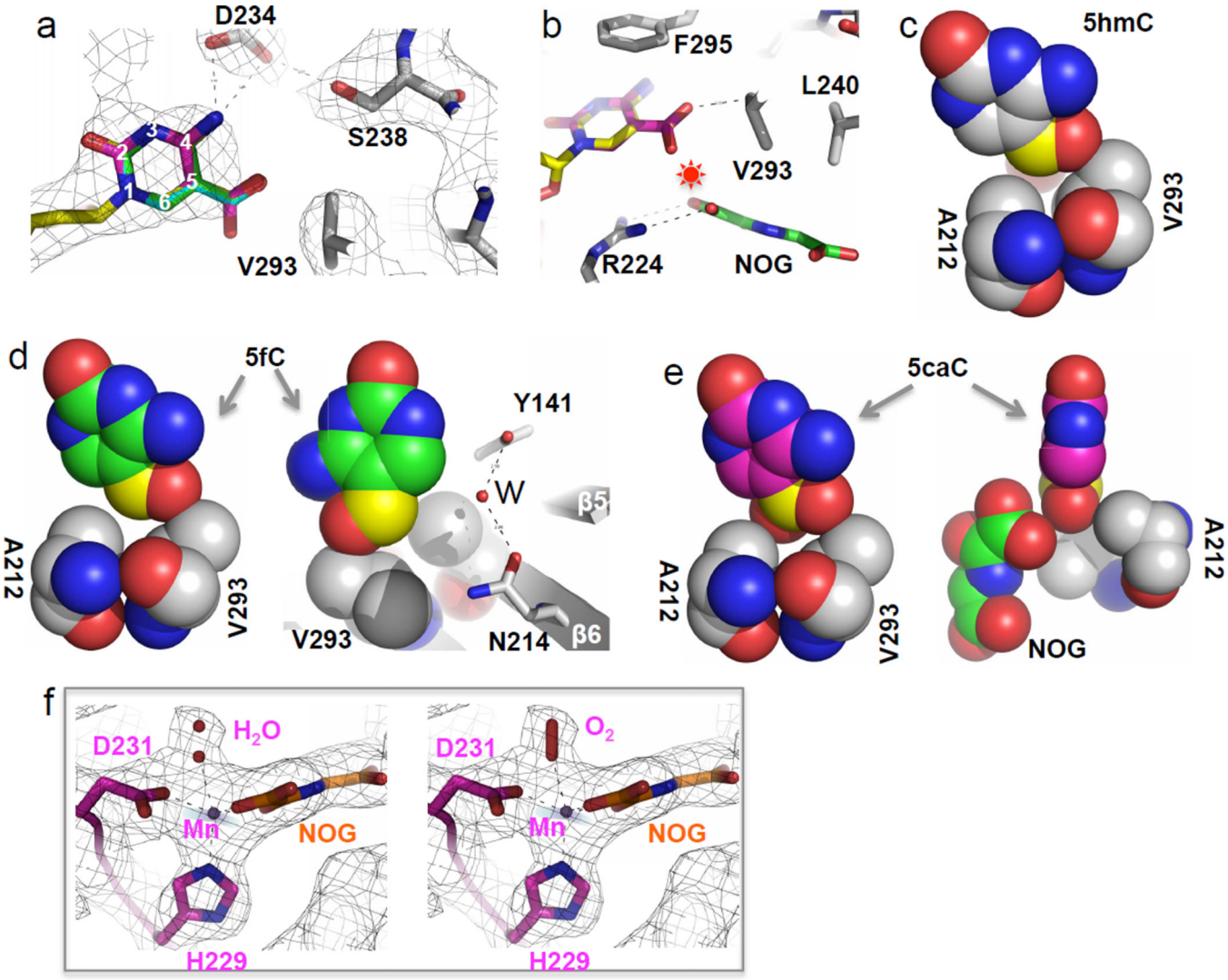
Extended Data Figure 3. Structure of NgTet1-DNA complex

a, Schematic NgTet1-DNA interactions. **b**, The amino end of the 3_{10} helix h3 interacts with the DNA backbone phosphate 3' to the 5mCpG site. An arrow indicates the helical dipole. **c**, Unlike other DNA base flipping enzymes such as DNA methyltransferases (Klimasauskas et al., 1994) and DNA repair glycosylases (Slupphaug et al., 1996), NgTet1 lacks a finger residue to occupy the space left by the everted 5mC. Instead, solvent molecules maintain the base stacking surrounding the flipped nucleotide. An ethylene glycol and a water molecule (behind ethylene glycol) occupy the space left by the everted 5mC. **d**, Superimposition of a normal intrahelical 5mC (colored in grey) onto the flipped 5mC suggests a small rotation around the glycosidic bond. **e**, The simulated annealing omit electron density, contoured at 2.5σ above the mean, by omitting entire 14-bp DNA (approximately 21% of total content in the crystal). The density is shown for the length of the unit cell along the **a** axis (indicated by vertical grey lines). The bent DNA molecules mediate crystal packing contacts along the **a** axis by 2-fold symmetry. The flipped 5mC is clearly visible in the active site (indicated by red circles). The broken density for the outer DNA bases in one end (as indicated by an

arrow), in the absence of any protein contacts, correlates with higher crystallographic thermal B-factors (~90 Å^2) than that for the central DNA base pairs including 5mCpG (~50 Å^2) or those of the other end (~67 Å^2). **f–g**, Enlarged panels showing NgTet1 structure in two views.

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

Slupphaug, G. *et al*. A nucleotide-flipping mechanism from the structure of human uracil-DNA glycosylase bound to DNA. *Nature* **384**, 87–92 (1996).



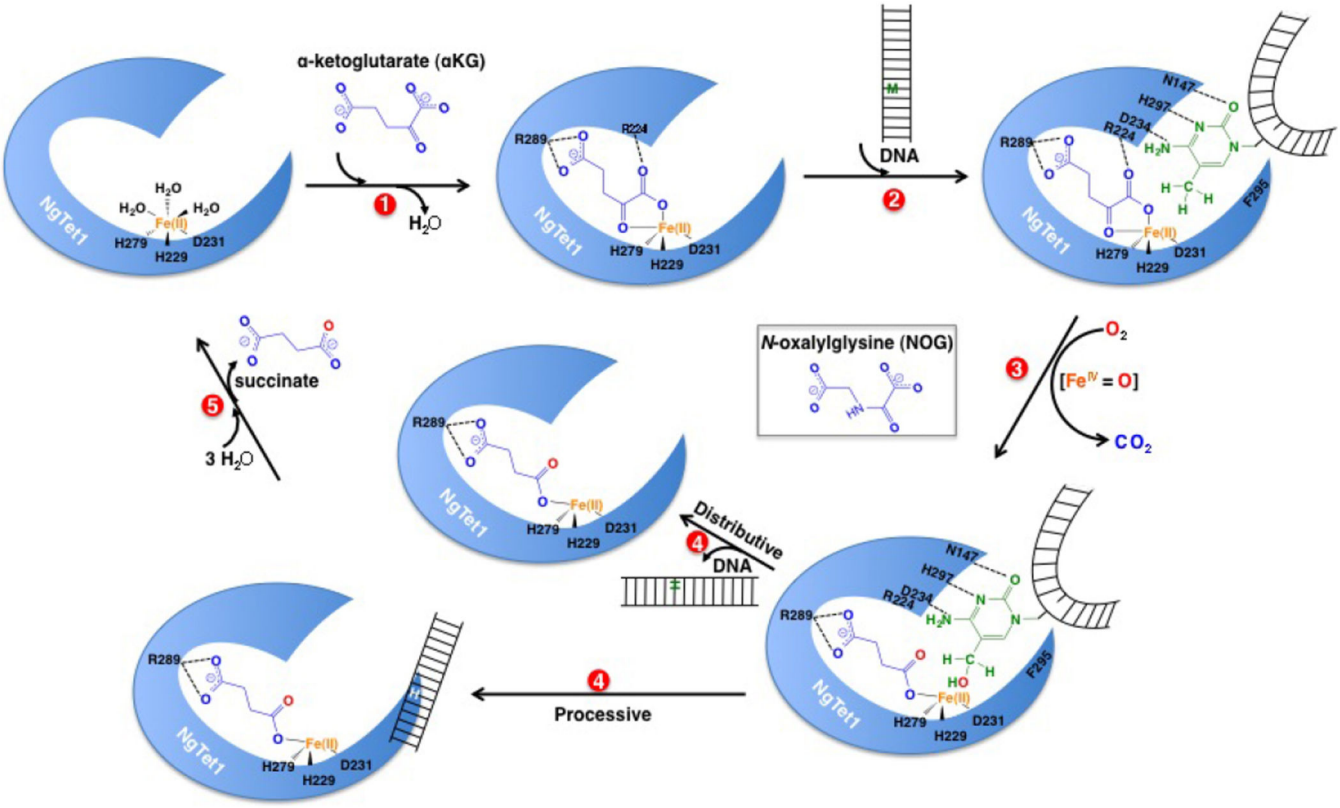
Extended Data Figure 4. Modeling of 5hmC, 5fC and 5caC in the active site of NgTet1

a, Superimposition of bases of 5hmC (cyan), 5fC (green), and 5caC (magenta) onto the flipped 5mC (yellow) in the NgTet1 active site. The closest residue to the 5-position modifications is Val293 of strand β12 and Ala212 of strand β6 (see panels **c–e**). The interaction involving hydrophobic Ala212 and Val293 might be the reason that the enzyme prefers 5mC (carrying a hydrophobic methyl group) over 5hmC (carrying a hydroxyl oxygen) or 5fC (carrying a carbonyl oxygen atom). **b**, The carboxylate group of 5caC (the

final product of oxidation reaction by NgTet1) would be in the vicinity of the C1 carboxylate group of NOG (it would be succinate during the reaction cycle – see Extended Data Fig. 5), resulting in repulsion. **c**, Space filling model of 5hmC. The atoms are colored with blue for nitrogen, red for oxygen, grey for carbon. The hydroxymethyl moiety of 5hmC is colored in yellow (CH_2) and its hydroxyl oxygen atom is in close contact with either the side chain of Val293 (as shown) or Ala212 (not shown). **d**, Space filling model of 5fC. The carbon atoms of 5fC are colored either as green (ring carbon) or yellow (the formyl carbon). A study suggested the existence of a hydrated form of 5fC in DNA containing synthetic 5fC at a level of about 0.5% (Pfaffeneder, *et al*. 2011). Because further oxidation of 5fC to 5caC would require the addition of water to the formyl group, the hydrated form of 5fC might be the real substrate during the oxidation of 5fC to 5caC (Yu and He, 2012). Our structure may provide evidence in support of this hypothesis. A water molecule, held in place by Asn214 and Tyr141, might provide the water molecule needed for the formation of 5fC hydrate. **e**, Space filling model of 5caC. The carbon atoms of 5caC are colored either as magenta (ring carbon) or yellow (the carboxylate carbon). The negatively charged carboxylate groups of 5caC and the carboxylate group of NOG would result in repulsion (right panel). **f**, We could model a water molecule with two alternative positions (left panel) or a di-oxygen O_2 molecule (right panel) as the sixth metal ligand as observed in the electron density 2Fo-Fc, contoured at 1 σ above the mean. Previously, we studied a Jumonji PHF2-metal interaction (PDB 3PU8), where a water molecule was modeled as the sixth ligand. Comparing the two structures, we concluded that the density observed in NgTet1 active site is more than a water molecule and the density was best fit with either a water molecule with dual positions or an O_2 molecule or a mixture of both. However, we do note that the observation of a dioxygen molecule needs to be confirmed independently by other methods.

Pfaffeneder, T. *et al*. The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angewandte Chemie* **50**, 7008–12 (2011).

Fu, Y. & He, C. Nucleic acid modifications with epigenetic significance. *Curr Opin Chem Biol* **16**, 516–24 (2012).



Extended Data Figure 5. Proposed mechanism of 5mC oxidation by Fe(II)- and αKG-dependent NgTet1

We suggest an ordered binding of αKG (step 1) followed by DNA (step 2), DNA bending and base flipping by NgTet1. Like many base-flipping enzymes, NgTet1 might use a multi-step flipping pathway to distinguish substrate (5mC, 5hmC and 5fC) from non-substrate (unmodified C). The discrimination step could occur either before flipping when the C:G pair is intrahelical, during the flipping or after flipping where the nucleotide becomes extrahelical. The hydroxylation reaction involves an peroxide intermediate that also covalently activates αKG and a reactive Fe(IV) intermediate (step 3) (Aik et al., 2012; Fu and He, 2012). The hydroxylated DNA is subsequently released (step 4), followed by exchange of succinate with αKG (step 5 and step 1) for next round of reaction. We do not know whether NgTet1 acts on DNA substrates distributively or processively (step 4) for the three consecutive, oxidation reactions that convert 5mC to 5caC. Metal ions Zn(II), Mn(II) or Co(II) have been used to replace Fe(II) in the studies of other dioxygenases, for example FIH (Elkins et al., 2003) and AlkB (Yu et al., 2006; Yang et al., 2008); they occupy Fe(II)-binding site but does not support catalysis. Like αKG, NOG (shown in the middle), initially used as an inhibitor in the study of FIH (Elkins et al., 2003), is ligated to Fe(II) or Mn(II) in a didentate manner but does not support catalysis due to decreased susceptibility to attack by an iron bound peroxide. We used the combination of Mn(II) and NOG in a very similar fashion as Zn(II) and NOG used in the FIH study.

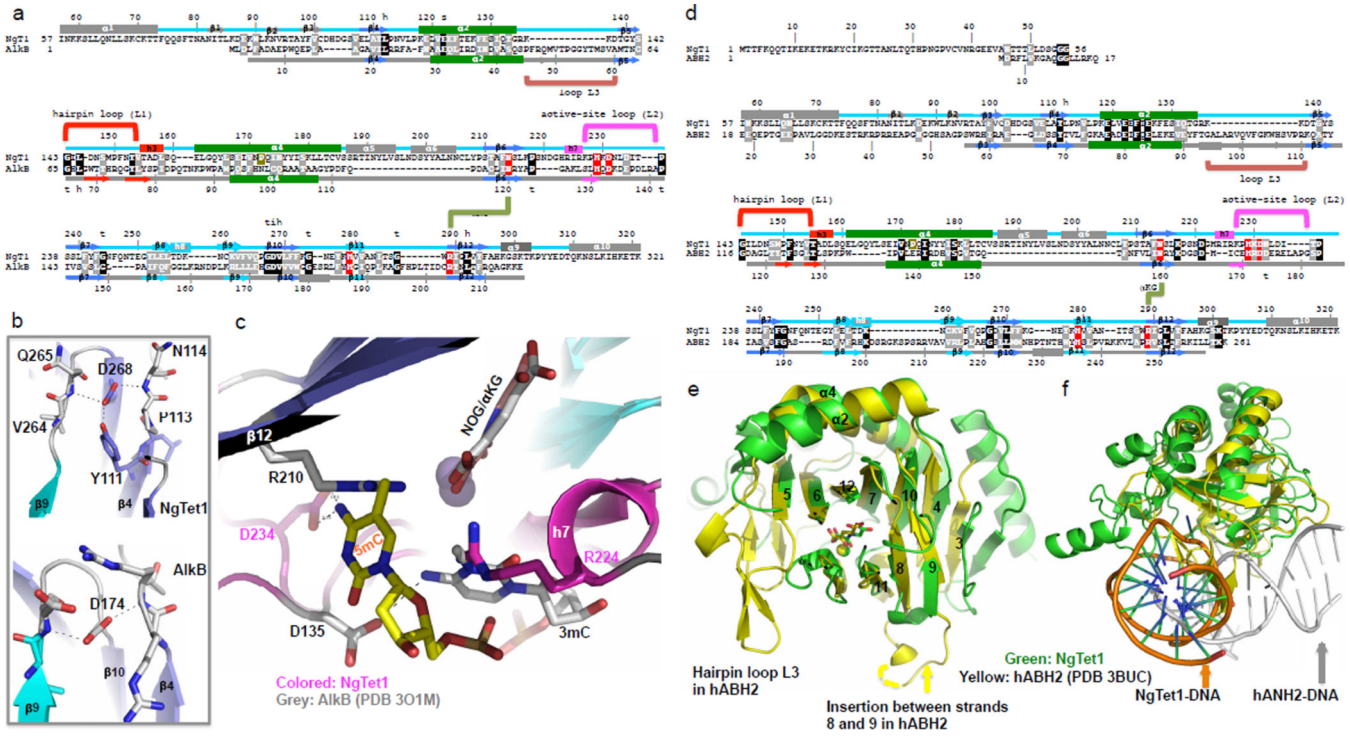
Aik, W., McDonough, M. A., Thalhammer, A., Chowdhury, R., Schofield, C. J. Role of the jelly-roll fold in substrate binding by 2-oxoglutarate oxygenases. *Curr Opin Struct Biol*. **22**, 690–700 (2012).

Elkins, J.M., Hewitson, K. S., McNeill, L. A., Seibel, J. F., Schlemminger, I., Pugh, C. W., Ratcliffe, P. J., Schofield, C. J. Structure of factor-inhibiting hypoxia-inducible factor (HIF) reveals mechanism of oxidative modification of HIF-1 alpha. *J. Biol. Chem*. **278**, 1802–1806 (2003).

Fu, Y. & He, C. Nucleic acid modifications with epigenetic significance. *Curr. Opin. Chem. Biol*. **16**, 516–524 (2012).

Yang, C. G., Yi, C., Duguid, E. M., Sullivan, C. T., Jian, X., Rice, P. A., He, C. Crystal structures of DNA/RNA repair enzymes AlkB and ABH2 bound to dsDNA. *Nature*, **452**, 961–965 (2008).

Yu, B., Edstrom, W. C., Benach, J., Hamuro, Y., Weber, P. C., Gibney, B. R., Hunt, J. F. Crystal structures of catalytic complexes of the oxidative DNA/RNA repair enzyme AlkB. *Nature* **439**, 879–884 (2006).



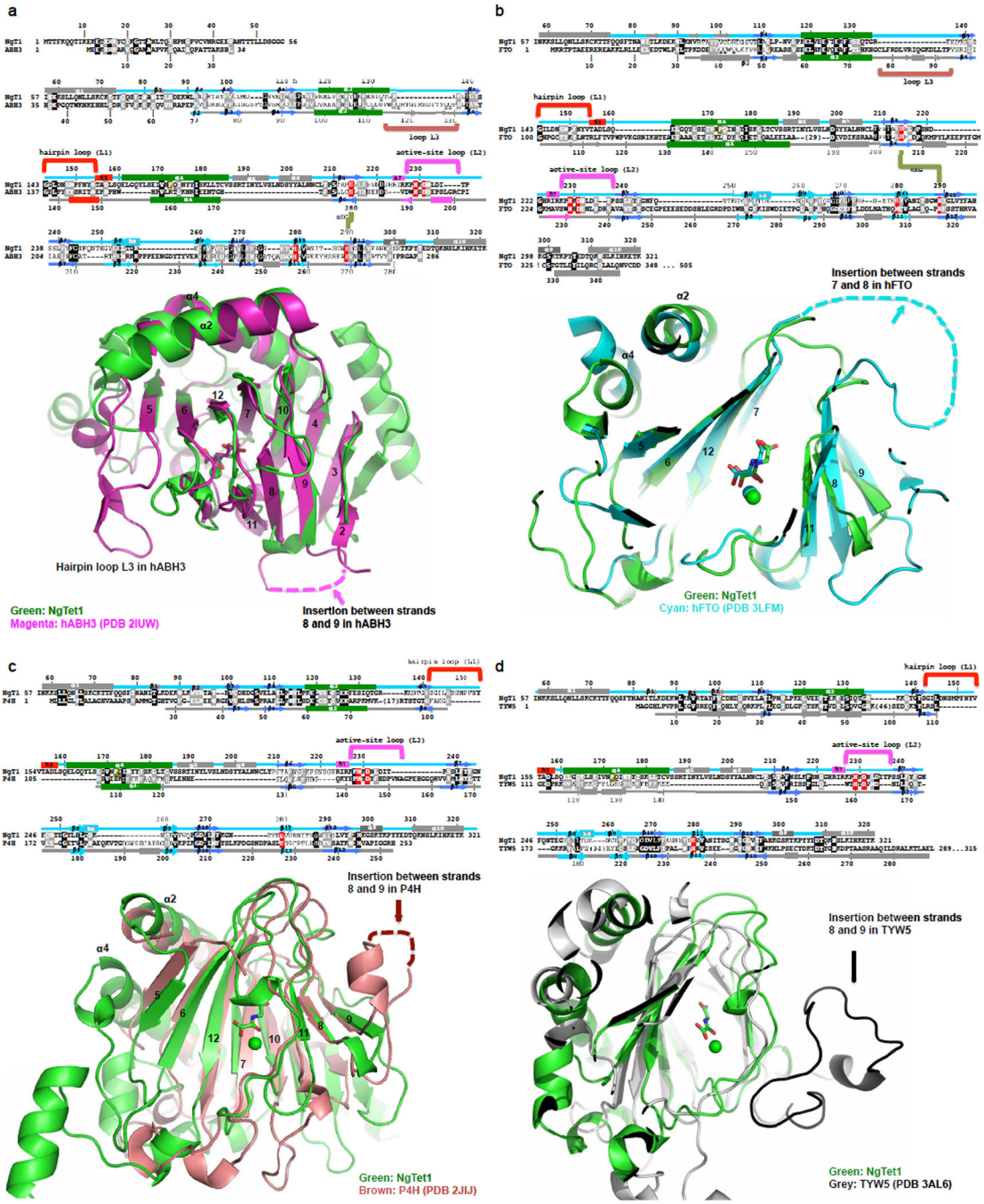
Extended Data Figure 6. Pairwise comparison of *Naegleria* NgTet1 and *E. coli* AlkB (panels a–c) or human ABH2 (panels d–f)

a, Structure-based sequence alignment of NgTet1 (PDB 4LT5) and AlkB (PDB 3O1M). NgTet1 has N-terminal as well as C-terminal additions. Secondary structural elements and residue numbering are indicated above (NgTet1) or below (AlkB) the sequences. Shared structural elements are colored in green (helices), blue (the major sheet) and cyan (the minor sheet). White-on-red residues are invariant residues between the two, important for binding of metal ion and αKG, white-on-black are invariant for the hydrophobic core (h), structural turns (t) before or after β strands (glycine and proline residues), and intra-molecule interaction (see panel **b**). Gray-highlighted positions are conserved substitutions. The two proteins share 19 invariant residues that are important for metal ion coordination, αKG binding, hydrophobic packing and intramolecular interactions, as well as glycine and proline

residues essential for structural turns before or after β strands. **b**, An invariant aspartate (Asp268 in NgTet1 and Asp174 in AlkB), located in strand β10, performs a network of stabilizing polar interactions with the main-chain amide nitrogen atoms immediately after strand β4 and β9. **c**, Superimposition of active sites of NgTet1 and AlkB indicate a co-variation of the binding site of the target base (5mC or 3mC) and the location of an arginine (Arg224 of NgTet1 and Arg210 of AlkB) that suggest conserved reaction chemistry and a conserved ion-pair interaction with the C1 carboxylate group of NOG of NgTet1 or αKG of AlkB. Arg224 of NgTet1, located in the 3_{10}-helix h7, interacts with the C1 carboxylate group of NOG in a bidentate manner. Superimposition of NgTet1 and AlkB indicated that the flipped 3mC occupies the space of Arg224 of NgTet1. Instead, AlkB uses Arg210 of strand β12 of the major sheet, from the opposite direction of Arg224 of NgTet1, to interact with the C1 carboxylate group or αKG. In NgTet1, the NOG molecule is involved in extensive interactions with the protein, including the carboxylate groups at C1 and C5 positions interacting with two arginine residues (Arg224 of h7 and Arg289 of β12), respectively, hydrophobic interactions with the side chains of Ile225 of h7, Leu240 of β7 and Leu253 of β8 and polar interactions with the side chains of Asn214 of β6 and Tyr242 of β7 (see Fig. 2k–l). **d**, Structure-based sequence alignment of NgTet1 and hABH2. **e**, ABH2 (colored in yellow; PDB 3BUC) has a hairpin loop insertion L3 between helix α2 and strand β5 and a 12-residue insertion between strands β8 and β9. NgTet1 (colored in green) has the TET/JBP1-specific structural element (helices α5 and α6) and a C-terminal addition (helices α9 and α10). The two proteins share 28 invariant residues. **f**, The DNA molecules, bound with NgTet1 or hABH2, lie nearly perpendicular to each other relative to the proteins. We also note that the AlkB-DNA and ABH2-DNA complexes were captured by chemical cross-linking between an engineered mutant S129C, located in the AlkB-specific strand (colored magenta in Fig. 3b) next to β11 as part of the minor sheet, and a disulphide-modified cytosine two nucleotides 3’ to the target base (Yang et al., 2008; Yi et al., 2010).

Yang, C. G. *et al*. Crystal structures of DNA/RNA repair enzymes AlkB and ABH2 bound to dsDNA. *Nature* **452**, 961–965 (2008).

Yi, C. *et al*. Iron-catalysed oxidation intermediates captured in a DNA repair dioxygenase. *Nature* **468**, 330–333 (2010).



Extended Data Figure 7. Pairwise comparison of *Naegleria* NgTet1 (colored in green) and (a) human ABH3 (magenta), (b) human FTO (cyan), (c) *Chlamydomonas* P4H (brown) or (d) Human TYW5 (grey)

Among them, NgTet1 has the TET/JBP1-specific structural element (helices α5 and α6). **a**, Like ABH2, ABH3 (PDB 2IUW) has a hairpin loop L3 insertion between helix α2 and strand β5 and a 13-residue insertion between strands β8 and β9. The two proteins share 30 invariant residues. **b**, FTO (PDB 3LFM) has a hairpin loop insertion between helix α2 and strand β5, a ~30-residue insertion in the location corresponding to the helices α5 and α6 of

NgTet1 and a 15-residue insertion between strands β7 and β8. The two proteins share 26 invariant residues. **c**, *Chlamydomonas reinhardtii* prolyl-4 hydroxylase type I (P4H, PDB 2JIJ) has insertions prior to strand β5, between strands β6 and β7, strands β8 and β9 and strands β10 and β11. The two proteins share 29 invariant residues. **d**. The tRNA Wybutosine (yW)-synthesizing enzyme 5 (TYW5, PDB 3AL6) has large insertions between helix α2 and strand β5 and between strands β8 and β9, in addition to a C-terminal domain. The two proteins share 30 invariant residues.

Extended Data Table 1

X-ray data collection and refinement statistics (values in parentheses are for the highest resolution shell)

Data collection	**Native**	**Merged from 2 crystals**
DNA	14-bp DNA	5-BrdU (3 sites)
Space group	$I2_12_12_1$	
Cell	α = β = γ = 90°	
a (Å)	84.0	83.2
b (Å)	108.6	107.3
c (Å)	166.4	166.5
Beamline	APS 22-BM	APS 22-ID
Wavelength (Å)	1.00000	0.91931
Resolution *	30.00 - 2.89	100-3.50
	(2.99-2.89)	(3.56-3.50)
R_{merge} *	0.077 (0.783)	0.124 (0.557)
<I/sigma I> *	22.8 (2.1)	61.3 (12.1)
Completeness (%) *	99.9 (99.5)	99.9 (100.0)
Redundancy *	6.4 (6.1)	56.2 (54.3)
Observed reflections	110,392	535,602
Unique reflections *	17,328 (1699)	9527 (493)
		(8469 have both I+ and I−)
		0.73 (FOM)
Refinement		
Resolution	29.5-2.89	
	(2.99-2.89)	
No. reflections	17,321	
R_{work}/R_{free}	0.193/0.215	
No. of atoms		
protein	2146	
DNA	570	
Mn^{2+}	1	
NOG	10	
Others	58	
water	99	

Data collection	Native	Merged from 2 crystals
B-factors (Å^2)		
Wilson B	68.0	
protein	44.5	
DNA	72.1	
Mn^{2+}	31.9	
NOG	39.8	
Others	64.4	
water	32.8	
r.m.s deviations		
Bond length (Å)	0.006	
Bond angles (°)	0.7	

Extended Data Table 2

Summary of pairwise comparisons of NgTet1 with other αKG-dependent dioxygenases (Extended Data Figs 5 and 6)

NgTet1 (PDB 4LT5)	Identity Number of residues (%)	Similarity Number of residues (%)
hTet1	43 (13.4%)	125 (38.9%)
mTet1	46 (14.3%)	128 (39.9%)
AlkB (PDB 3BI3)	19 (5.9%)	65 (20%)
ABH2 (PDB 3BUC)	28 (8.7%)	68 (21.2%)
ABH3 (PDB 2IUW)	30 (9.3%)	85 (26.5%)
FTO (PDB 3LFM)	26 (8%)	66 (20.6%)
P4H (PDB 2JIJ)	29 (9%)	63 (19.6%)
TYW5 (PDB 3AL6)	30 (9.3%)	65 (20%)

The percentage is calculated as 100 ×(number of residues/321), where 321 is the length of NgTet1.

Extended Data Table 3

Conserved residues with functional significance

Function	NgTet1	hTet1
Metal Fe (II)	H229, D231, H279	H1672, D1674, H2028
αKG/NOG	R289, L253	R2043, L1705
5mC	H297, D234, A212, V293, F295	H2051, N1677, A1645, V2047, Y2049
3’ Gua to 5mC	Q310	N2064
Orphaned Gua	S148	S1582
DNA Phosphates	K298, K311	K2052, K2065

Acknowledgements

Dr. Richard J. Roberts was originally an author of this manuscript, however, as a staunch supporter of the open access movement, he will not author a paper that is not open access. We thank Dr. John R. Horton for critical comments and Brenda Baker at the organic synthesis unit of New England Biolabs for synthesizing the oligonucleotides. YZ thanks Dr. Chandler Fulton at Brandeis University for the teachings of *N. gruberi* biology.

The Department of Biochemistry of Emory University School of Medicine supported the use of SER-CAT beamlines. This work was supported by grants from the National Institutes of Health GM049245 to X.C (who is a Georgia Research Alliance Eminent Scholar) and GM095209 and GM105132 to Y.Z.

References

1. Tahiliani M, et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science. 2009; 324:930–935. [PubMed: 19372391]
2. Ito S, et al. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature. 2010; 466:1129–1133. [PubMed: 20639862]
3. Ito S, et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science. 2011; 333:1300–1303. [PubMed: 21778364]
4. He YF, et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science. 2011; 333:1303–1307. [PubMed: 21817016]
5. Iyer LM, Zhang D, Maxwell Burroughs A, Aravind L. Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA. Nucleic Acids Res. 2013; 41:7635–7655. [PubMed: 23814188]
6. Fritz-Laylin LK, et al. The genome of Naegleria gruberi illuminates early eukaryotic versatility. Cell. 2010; 140:631–642. [PubMed: 20211133]
7. Iyer LM, Abhiman S, Aravind L. Natural history of eukaryotic DNA methylation systems. Progress in molecular biology and translational science. 2011; 101:25–104. [PubMed: 21507349]
8. Aik W, McDonough MA, Thalhammer A, Chowdhury R, Schofield CJ. Role of the jelly-roll fold in substrate binding by 2-oxoglutarate oxygenases. Curr Opin Struct Biol. 2012; 22:691–700. [PubMed: 23142576]
9. McDonough MA, Loenarz C, Chowdhury R, Clifton IJ, Schofield CJ. Structural studies on human 2-oxoglutarate dependent oxygenases. Curr Opin Struct Biol. 2010; 20:659–672. [PubMed: 20888218]
10. Trewick SC, Henshaw TF, Hausinger RP, Lindahl T, Sedgwick B. Oxidative demethylation by Escherichia coli AlkB directly reverts DNA base damage. Nature. 2002; 419:174–178. [PubMed: 12226667]
11. Yang CG, et al. Crystal structures of DNA/RNA repair enzymes AlkB and ABH2 bound to dsDNA. Nature. 2008; 452:961–965. [PubMed: 18432238]
12. Yi C, et al. Iron-catalysed oxidation intermediates captured in a DNA repair dioxygenase. Nature. 2010; 468:330–333. [PubMed: 21068844]
13. Klimasauskas S, Kumar S, Roberts RJ, Cheng X. HhaI methyltransferase flips its target base out of the DNA helix. Cell. 1994; 76:357–369. [PubMed: 8293469]
14. Slupphaug G, et al. A nucleotide-flipping mechanism from the structure of human uracil-DNA glycosylase bound to DNA. Nature. 1996; 384:87–92. [PubMed: 8900285]
15. Roberts RJ, Cheng X. Base flipping. Annu Rev Biochem. 1998; 67:181–198. [PubMed: 9759487]
16. Horton JR, et al. Caught in the act: visualization of an intermediate in the DNA base-flipping pathway induced by HhaI methyltransferase. Nucleic Acids Res. 2004; 32:3877–3886. [PubMed: 15273274]
17. Werner RM, et al. Stressing-out DNA? The contribution of serine-phosphodiester interactions in catalysis by uracil DNA glycosylase. Biochemistry. 2000; 39:12585–12594. [PubMed: 11027138]
18. Sun Z, et al. High-resolution enzymatic mapping of genomic 5-hydroxymethylcytosine in mouse embryonic stem cells. Cell reports. 2013; 3:567–576. [PubMed: 23352666]
19. Yu M, et al. Base-resolution analysis of 5-hydroxymethylcytosine in the Mammalian genome. Cell. 2012; 149:1368–1380. [PubMed: 22608086]
20. Ficz G, et al. Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature. 2011; 473:398–402. [PubMed: 21460836]
21. Upadhyay AK, Horton JR, Zhang X, Cheng X. Coordinated methyl-lysine erasure: structural and functional linkage of a Jumonji demethylase domain and a reader domain. Curr Opin Struct Biol. 2011; 21:750–760. [PubMed: 21872465]

22. Fang R, et al. LSD2/KDM1B and its cofactor NPAC/GLYR1 endow a structural and molecular model for regulation of H3K4 demethylation. Mol Cell. 2013; 49:558–570. [PubMed: 23260659]

References

23. Szulwach KE, et al. 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. Nat Neurosci. 2011; 14:1607–1616. [PubMed: 22037496]
24. Inoue A, Shen L, Dai Q, He C, Zhang Y. Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. Cell Res. 2011; 21:1670–1676. [PubMed: 22124233]
25. Nestor CE, et al. Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. Genome Res. 2012; 22:467–477. [PubMed: 22106369]
26. Munzel M, et al. Quantification of the sixth DNA base hydroxymethylcytosine in the brain. Angewandte Chemie. 2010; 49:5375–5377. [PubMed: 20583021]
27. Haffner MC, et al. Global 5-hydroxymethylcytosine content is significantly reduced in tissue stem/progenitor cell compartments and in human cancers. Oncotarget. 2011; 2:627–637. [PubMed: 21896958]
28. Stroud H, Feng S, Morey Kinney S, Pradhan S, Jacobsen SE. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. Genome Biol. 2011; 12:R54. [PubMed: 21689397]
29. Hashimoto H, Hong S, Bhagwat AS, Zhang X, Cheng X. Excision of 5-hydroxymethyluracil and 5-carboxylcytosine by the thymine DNA glycosylase domain: its structural basis and implications for active DNA demethylation. Nucleic Acids Res. 2012; 40:10203–10214. [PubMed: 22962365]
30. Otwinowski Z, Borek D, Majewski W, Minor W. Multiparametric scaling of diffraction intensities. Acta Crystallogr A. 2003; 59:228–234. [PubMed: 12714773]
31. Fu Z-Q, Chrzes J, Sheldrick GM, Rose J, Wang B-C. A paralle program using SHELXD for quick heavy-atom partial structureal solution on high-performance computers. Journal of Applied Crystallography. 2007; 40:387–390.
32. Fu ZQ, Rose J, Wang BC. SGXPro: a parallel workflow engine enabling optimization of program performance and automation of structure determination. Acta Crystallogr D Biol Crystallogr. 2005; 61:951–959. [PubMed: 15983418]
33. Adams PD, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr. 2010; 66:213–221. [PubMed: 20124702]
34. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr. 2010; 66:486–501. [PubMed: 20383002]
35. Davis IW, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. 2007; 35:W375–W383. [PubMed: 17452350]

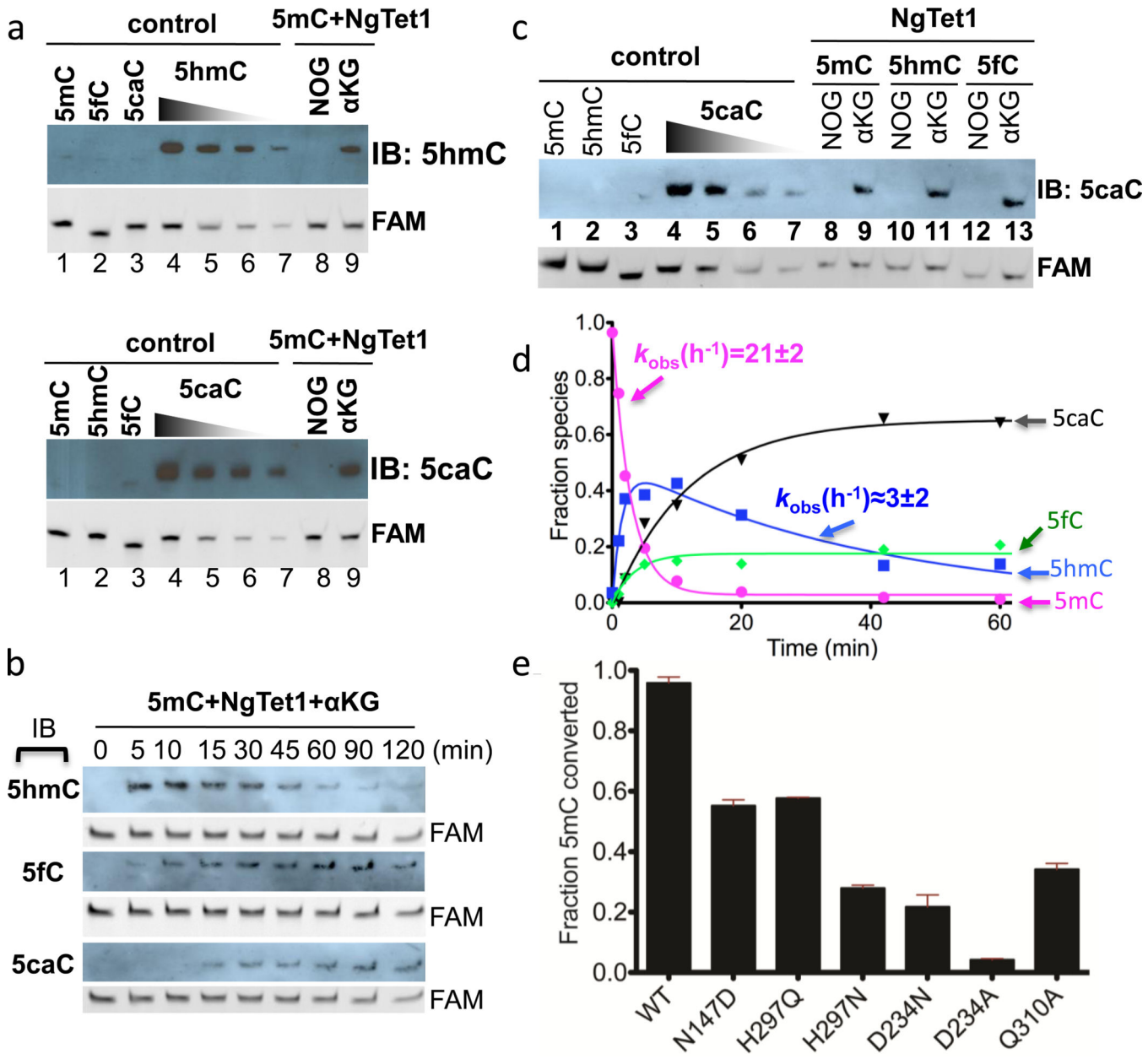


Figure 1. Activity of NgTet1

a, Detection of 5hmC (top) and 5caC (bottom) by antibodies. **b**, The relative amount of each reaction product was sequentially observed over the full time course of the reaction. **c**, NgTet1 is active on all three DNA substrates, producing 5caC. **d**, Quantitative LC-MS measurement of 5mC disappearance and formation of 5hmC, 5fC and 5caC. **e**, The effects of mutations on the conversion of 5mC. Error bars indicate s.d. of the mean value from three independent experiments.

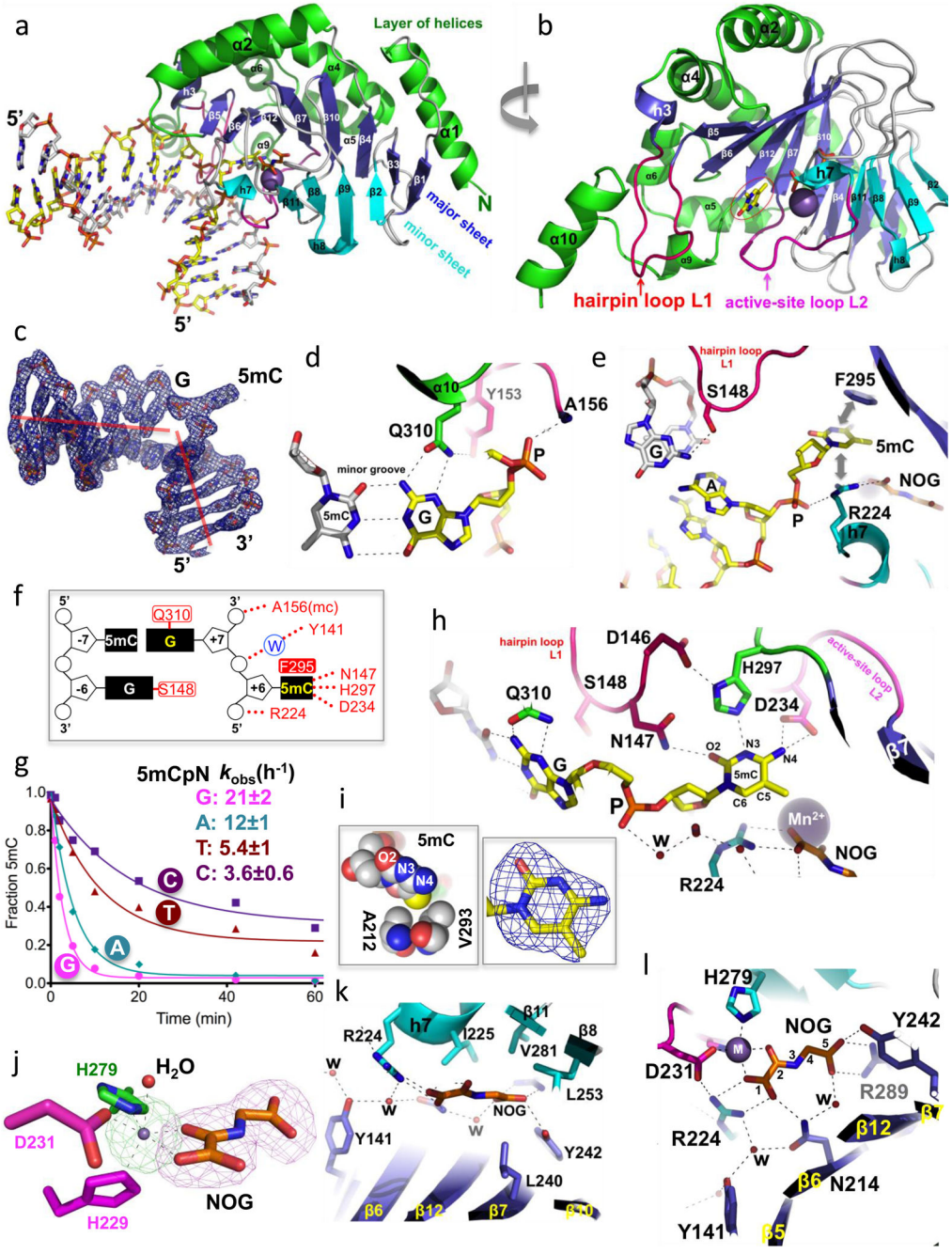


Figure 2. Structure of NgTet1-DNA complex

a, The NgTet1 protein folds in a three-layered jelly-roll structure. **b**, Rotated ~90° from the view of panel **a**. **c**, Electron density 2Fo–Fc, contoured at 1σ above the mean, is shown for the entire 14-bp DNA with a flipped out 5mC. **d**, Q310 interacts with 3'-Gua in the minor groove. **e**, S148 interacts with the intrahelical orphaned guanine. F295 and R224 form planar π stacking contacts with the extrahelical 5mC. **f**, Summary of the NgTet1–DNA interactions focusing on 5mCpG dinucleotide: mc, main-chain-atom-mediated contacts; w, water-mediated contacts. **g**, Substrate preference of 5mCpN (N=G, A, T or C) of NgTet1. **h**, The

hydrogen bond interactions with the polar atoms of 5mC. Inserted is the simulated annealing omit electron density, contoured at 4.5σ above the mean, for omitting 5mC. **i**, The hydrophobic side chains of A212 and V293 border the methyl group (in yellow) of 5mC. Other atoms are colored as blue for nitrogen, red for oxygen and grey for carbon. **j**, The octahedral coordination of Mn^{2+} observed in the NgTet1–NOG–metal interactions (Extended Data Fig. 4f). Simulated annealing omit electron densities, contoured at 10σ and 5σ above the mean, are shown for the Mn^{2+} (green mesh) and NOG (magenta mesh), respectively. **k–l**, Two views of NOG-NgTet1 interactions.

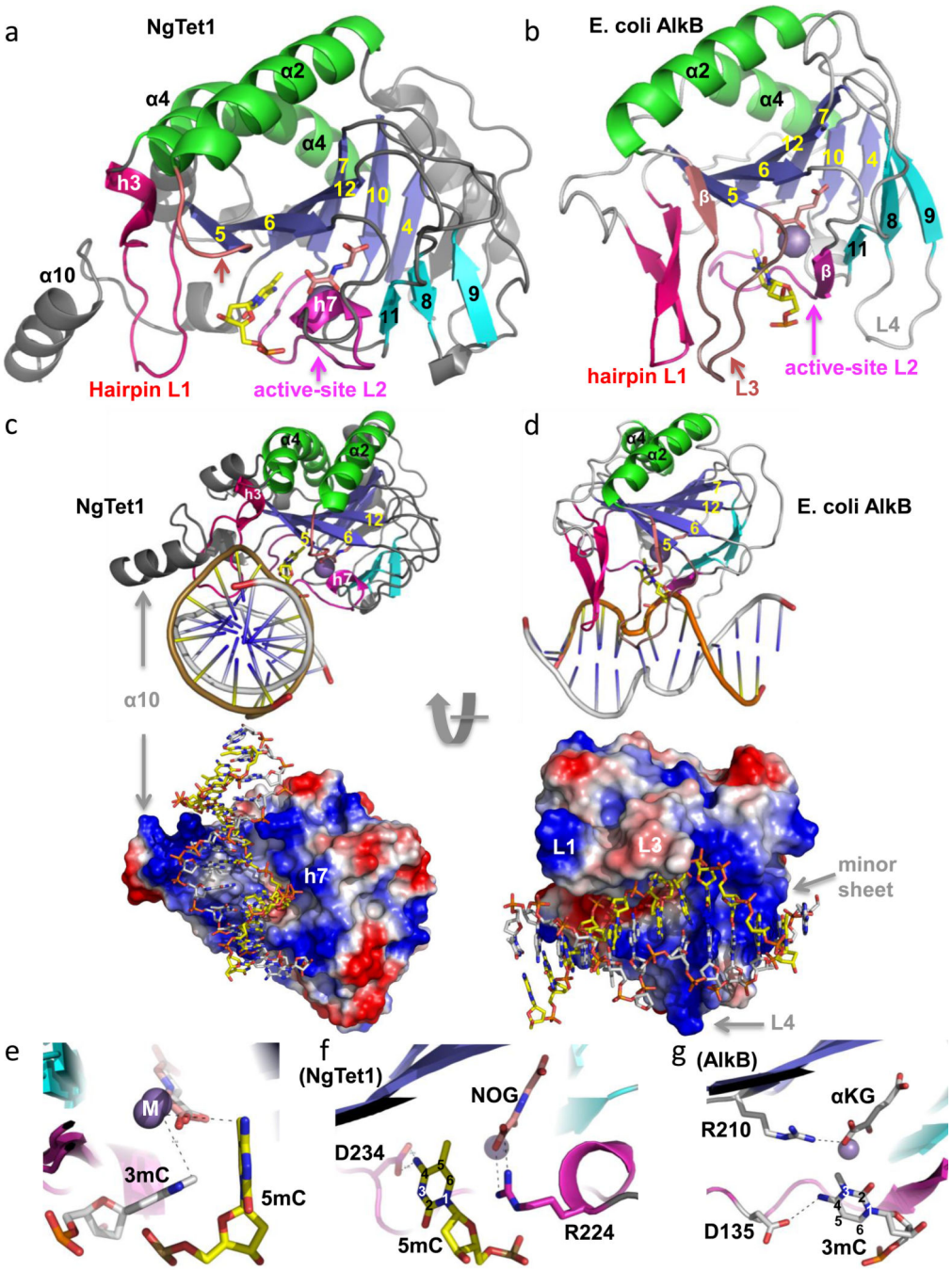


Figure 3. Comparison of NgTet1 and AlkB

a–b, Structures of NgTet1 and AlkB aligned in a similar orientation. **c–d**, NgTet1 (**c**) and AlkB (**d**) are shown in relatively similar orientations. The surface charge at neutral pH is displayed as blue for positive, red for negative, and white for neutral. **e**, Superimposition of NgTet1 (5mC) and AlkB (3mC) in the active sites. The metal ions (M) are shown as balls and NOG or αKG (in the back) as sticks. **f–g**, Co-variation between the location of the target base (5mC in NgTet1 and 3mC in AlkB) and the NOG/αKG-interacting arginine (R224 of NgTet1 and R210 of AlkB).

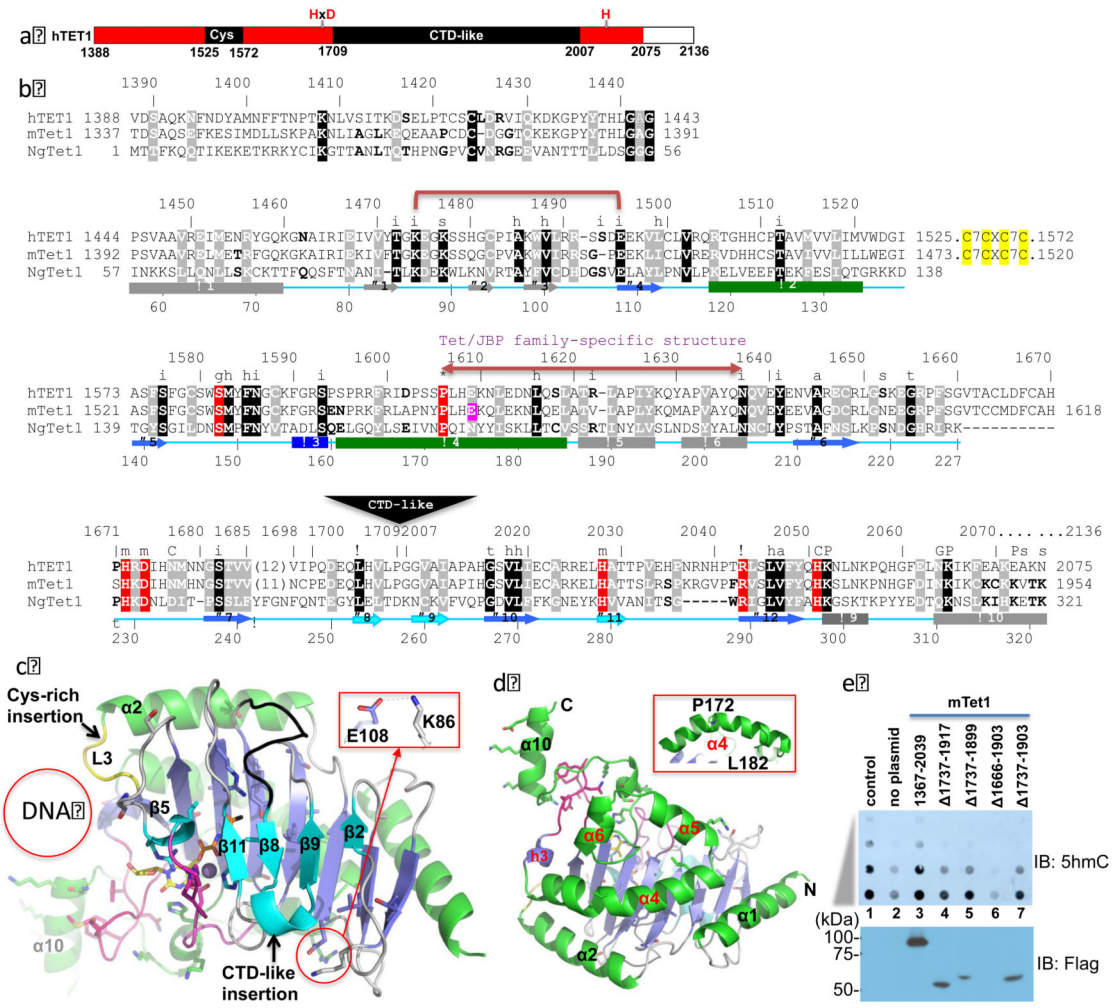


Figure 4. Pairwise comparison of NgTet1 and mammalian Tet1

a, Schematic representation of hTet1 C-terminal catalytic domain. **b**, Sequence alignment of NgTet1, hTet1 and mTet1. Labels above the sequences indicate that i for intra-molecular polar interaction; s for exposed surface residue; h for hydrophobic core; t for structural turn; α for αKG binding; m for metal ion coordination; P for DNA phosphate interaction; g for DNA base interaction with the orphaned guanine; G for DNA base interaction with the 3' guanine to 5mC; C for 5mC interaction; a for active site residues (A212 and V293) near the methyl group of 5mC. **c**, Structure of NgTet1 with arrows indicating the two large insertions of mammalian Tet1. Highlighted is the charge-charge interaction between invariant K86 and E108. **d**, A kinked helix α4, owing to P172 (conserved among NgTet1, human and mouse Tet1, Tet2 and Tet3) located in the middle. **f**, Antibody detection of 5hmC in genomic DNA of HEK293T cells (top panel) expressing Flag tagged mouse Tet1 catalytic domain or its internal deletions (bottom panel). Top panel: Lane 1 is the 32-bp oligonucleotide containing a single 5hmC (20 pmol and 2 fold serial dilutions) and lanes 2–7 are the genomic DNA (500 ng and 2 fold serial dilutions). Bottom panel: Lane 1 is the molecular weight marker and Lanes 2 and 7 are the whole cell lysates with approximately equal amount of protein.