Biochemical characterization of a Naegleria TET-like oxygenase and its application in single molecule sequencing of 5-methylcytosine

June E. Pais, New England Biolabs Inc
Nan Dai, New England Biolabs Inc
Esta Tamanaha, New England Biolabs Inc
Romualdas Vaisvila, New England Biolabs Inc
Alexey I. Fomenkov, New England Biolabs Inc
Jurate Bitinaite, New England Biolabs Inc
Zhiyi Sun, New England Biolabs Inc
Shengxi Guan, New England Biolabs Inc
Ivan Correa, Emory University
Christopher J. Noren, New England Biolabs Inc

Only first 10 authors above; see publication for full author list.

Journal Title: PNAS
Volume: Volume 112, Number 14
Publisher: United States National Academy of Sciences | 2015-04-07, Pages 4316-4321
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.1073/pnas.1417939112
Permanent URL: https://pid.emory.edu/ark:/25593/vm9vg

Final published version: http://dx.doi.org/10.1073/pnas.1417939112

Copyright information:
© 2015 National Academy of Sciences.

Accessed August 30, 2023 5:15 PM EDT
Biochemical characterization of a Naegleria TET-like oxygenase and its application in single molecule sequencing of 5-methylcytosine

June E. Pais1, Nan Dai2, Esta Tamanaha3, Romualdas Vaisvila3, Alexey I. Fomenkov3, Jurate Bitinaite3, Zhiyi Sun3, Shengxi Guan4, Ivan R. Corrêa Jr.4, Christopher J. Noren5, Xiaodong Cheng6, Richard J. Roberts7, Yu Zheng8,9, and Lana Saleh1,1

*Research Department, New England Biolabs, Ipswich, MA 01938; and 2Department of Biochemistry, Emory University School of Medicine, Atlanta, GA 30322

Edited by Anjana Rao, Sanford Consortium for Regenerative Medicine and La Jolla Institute for Allergy and Immunology, La Jolla, CA, and approved March 2, 2015 (received for review September 17, 2014)

Modified DNA bases in mammalian genomes, such as 5-methylcytosine (5mC) and its oxidized forms, are implicated in important epigenetic regulation processes. In human or mouse, successive enzymatic conversion of 5mC to its oxidized forms is carried out by the ten-eleven translocation (TET) proteins. Previously we reported the structure of a TET-like 5mC oxygenase (NgTET1) from Naegleria gruberi, a single-celled protist evolutionarily distant from vertebrates. Here we show that NgTET1 is a 5-methylpyrimidine oxygenase, with activity on both 5mC (major activity) and thymidine (T) (minor activity) in all DNA forms tested, and provide unprecedented evidence for the formation of 5-formyluridine (5fU) and 5-carboxyuridine (5caU) in vitro. Mutagenesis studies reveal a delicate balance between choice of 5mC or T as the preferred substrate. Furthermore, our results suggest substrate preference by NgTET1 to 5mCpG and TpG dinucleotide sites in DNA. Intriguingly, NgTET1 displays higher T-oxidation activity in vitro than mammalian TET1, supporting a closer evolutionary relationship between NgTET1 and the base J-binding proteins from trypanosomes. Finally, we demonstrate that NgTET1 can be readily used as a tool in 5mC sequencing technologies such as single molecule, real-time sequencing to map 5mC in bacterial genomes at base resolution.

TET proteins | NgTET1 | 5-methylcytosine | SMRT sequencing | bacterial methylethone

M odified DNA bases exist in all forms of life, from viruses to mammals with many different biological roles. Accordingly, diverse mechanisms have evolved to “write,” “read,” and “erase” these modifications. In mammals, 5-methylcytosine (5mC) is the major form of DNA modification and is implicated in many crucial developmental processes. In human and mouse, 5mC can be successively oxidized into 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxycytosine (5caC) by the ten-eleven translocation (TET) family of oxygenases (1–4). The bases of 5fC and 5caC can be excised by thymine DNA glycosylase (4). The 5mC-oxidation–coupled base-excision repair pathway provides a plausible route for active demethylation in mammalian cells. Many other species, from simple to complex, maintain DNA methylation machinery throughout their life cycle that may contribute to epigenetic regulation. Therefore, an interesting perspective is to examine shared and distinct features of TET oxygenases in diverse eukaryotes (5, 6).

The human and mouse genomes encode three paralogous TET proteins, TET1, TET2, and TET3, which presumably carry out both redundant and distinct functions (7, 8). TET proteins belong to the diverse group of α-ketoglutarate (αKG) and Fe(II)-dependent oxygenases (5). Subgroup classification based on sequence similarity links the TET proteins to base J-binding proteins (JBP1 and JBP2), which are primarily present in trypanosomes and possess thymidine (T)-hydroxylation activity (1). Further bioinformatic analysis revealed eight paralogous TET/JBP-like genes in the genome of Naegleria gruberi, a single-celled amoebobflagellate protist that is a distant cousin of the parasitic trypanosomes, evolutionarily far removed from vertebrates (5, 9). Interestingly, genetic components for “writing” 5mC (i.e., homologs of mammalian DNA methyltransferases) are also present in the genome (9). These components may parallel the methylation/oxidation processes in mammalian cells.

We show here that new insights into the TET family of enzymes can be obtained by studying a representative from a protist that may have shared an ancestral TET enzyme with mammals, but then evolved separately from vertebrates for much of eukaryotic evolution. We have previously reported the in vitro biochemical activity and structure of an active N. gruberi TET/JBP-like protein, termed NgTET1 (10). We showed that like mammalian TETs, NgTET1 is capable of catalyzing the oxidation of 5mC to 5hmC, 5fC, and 5caC in vitro. The crystal structure of NgTET1 in complex with a symmetrically methylated oligonucleotide (oligo) reveals a base-flipping mechanism in which the DNA is bent and the flipped 5mC is positioned in the catalytic binding pocket (10). The hydrogen-bond networks between NgTET1 and substrate are specific to the flipped 5mCpG dinucleotide, and we reported a substrate preference for 5mCpG-containing oligo DNA (10). Here we extend this observation by reporting the activity of NgTET1 at direct methylome profiling using single-molecule, real-time sequencing.

Significance

The discovery that 5-methylcytosine (5mC) can be iteratively oxidized by mammalian ten-eleven translocation (TET) proteins marks a breakthrough in the field of epigenetics. To better understand the evolutionary and functional linkage of TET family members, we characterized NgTET1 from the protist Naegleria gruberi, which bears homology to both TET and base J-binding protein, a thymidine hydroxylase in trypanosomes. We show that NgTET1 performs iterative oxidation of both 5mC and thymidine (T) (minor activity) on various DNA forms, and that these activities can be modulated by mutagenesis. We also present evidence for the effect of sequence context on both 5mC- and T-oxidase activities. Finally, we show the utility of NgTET1 at direct methylome profiling using single-molecule, real-time sequencing.


The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

1To whom correspondence may be addressed. Email: saleh@neb.com or yu.zhengyu@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1417939112/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1417939112

4316-4321 | PNAS | April 7, 2015 | vol. 112 | no. 14
NgTET1 on various types of DNA containing different methylation motifs and in different conformations. Importantly, we show that it exhibits T-oxygenase activity, similar to JBP1 and JBP2, but can catalyze the formation of further oxidized T species, 5-formyluridine (5FUr) and 5-carboxyuridine (5caU), in addition to 5-hydroxymethyluridine (5hmU). We compare the in vitro activities of NgTET1 and the catalytic domain of mouse TET1 (mTET1CD) on various substrates and show that the two enzymes exhibit similar 5mC-oxygenase activities but vary in the extent of their T-oxygenase activities, with NgTET1 displaying notably higher T-oxygenase activity than mTET1CD. Finally, we demonstrate the utility of NgTET1 in methylome sequencing applications, such as single molecule, real-time (SMRT) sequencing.

**Results**

**NgTET1 Is an Fe(II)/α-KG-Dependent 5-Methylpyrimidine Oxygenase.** Full-length NgTET1 was expressed and purified to homogeneity and tested for activity on DNA containing 5mC. First, a restriction enzyme (RE)-based assay was used to test protection of prRS(M.HpaII), a linear plasmid in which all internal Cs in a CCGG recognition site are methylated by the endogenously expressed M.HpaII methyltransferase, upon treatment with NgTET1 (Fig. 1A). The plasmid is cleaved at the same recognition sequence (C5mC-CCGG) by MspI for both 5mC and 5hmC sites. When oligo substrates contain 5mC or symmetrical 5C, MspI cleavage does not occur (Fig. S1 and Table S1). Fig. 1A illustrates that the observed MspI protection is dependent on the concentration of NgTET1 used in a 30-min reaction at 34 °C, the optimal temperature for the NgTET1 reaction (Fig. S2). Full protection from MspI digestion is achieved at 0.01 μM plasmid DNA (equivalent to 0.3-μM 5mC sites) and an NgTET1 concentration of 1 μM and higher (Fig. 1A).

We also used a liquid chromatography-mass spectrometry (LC-MS)-based assay as a more sensitive method to detect and quantify each species in the oxidation reaction, as previously described (10). Fig. 1B shows a representative chromatogram from an LC-MS–based activity assay in the absence or presence of NgTET1 and genomic DNA (gDNA) from human cells (IMR90) as a substrate. 5mC of IMR90 is completely converted to 5hmC (major product) with small amounts of 5FUr and 5caC remaining (Fig. 1B). The plasmid is cleaved at the same recognition sequence (C5mC-CCGG) by MspI for both 5mC and 5hmC sites. When oligo substrates contain 5mC or symmetrical 5C, MspI cleavage does not occur (Fig. S1 and Table S1). Fig. 1B illustrates that the observed MspI protection is dependent on the concentration of NgTET1 used in a 30-min reaction at 34 °C, the optimal temperature for the NgTET1 reaction (Fig. S2). Full protection from MspI digestion is achieved at 0.01 μM plasmid DNA (equivalent to 0.3-μM 5mC sites) and an NgTET1 concentration of 1 μM and higher (Fig. 1B).

In addition to its oxygenase activity on dsDNA, NgTET1 oxidizes 5mC on hemimethylated (hemiDNA) and single-strand DNA (ssDNA) (Fig. 1D). After a 1-h incubation, the amounts of 5mC, 5hmC, 5fC, and 5caC as quantified by the LC-MS assay are comparable for all three types of substrates (Fig. 1D and Table S1). The ability of NgTET1 to catalyze oxidation of hemiDNA or ssDNA is consistent with the observation that NgTET1 forms hydrogen-bond contacts with 5mC on only one strand in the crystal structure of the enzyme, in complex with a symmetrically methylated dsDNA oligo substrate (10).

The relatively permissive substrate specificity of NgTET1, as indicated by its activity on these various substrates, raised the question of whether similar promiscuity is observed with the mammalian TET proteins. The activity of the C-terminal catalytic domain of mTET1CD on hemiDNA and ssDNA has been reported previously (11). Here we compare the activity of mTET1CD on ssDNA, dsDNA, and gDNA, using an LC-MS–based activity assay to measure the amount of 5mC, 5hmC, 5fC, and 5caC after a 1-h incubation. Indeed, we found that mTET1CD can convert 5mC to 5caC in all substrates tested, with similar efficiency as NgTET1 (Fig. 1E).

**T-oxygenase activity.** Bioinformatic analysis suggests an evolutionary linkage between the TET proteins and JBP, which catalyze the hydroxylation of the methyl group in T to form 5hmC, 5fC, and 5caC. We confirmed this hypothesis by designing a competition assay using oligos with the same sequence bearing either a single T or 5mC site have been unsuccessful because of...
The T-oxygenase activity of NgTET1 is dependent on nucleotide-sequence context. Dis-

lack of detection of T oxidation. Nonetheless, we conclude that the T-oxygenase activity of NgTET1 is minor compared with its $^{5mC}$-oxygenase activity.

Although T oxidation appears to be minor, this activity may have some physiological relevance as $^{5mC}$ formation through T oxidation was recently reported for the mammalian TET proteins in mouse embryonic stem cells (13). We compared the in vitro T-oxygenation activity of NgTET1 and mTET1CD on various gDNA and oligo substrates (Figs. 2B and 3D). Intriguingly, significantly higher levels of $^{5mU}$ and $^{5U}$ formed in the reaction of NgTET1 compared with mTET1CD, and $^{5mC}$ was detected only in the NgTET1 reaction. To further characterize this activity, we first tested whether T oxidation is dependent on $^{5mC}$ methylation of the substrate DNA (i.e., if cytosine methylation is required for binding or recruitment of the DNA to the active site of NgTET1 before T oxidation). We compared T-oxygenation activity on pUC19 produced in a DNA cytosine-C5-methyltransferase ($dcm$) Escherichia coli strain (methylated at CCGG sites) compared with that from a $dcm^+$ strain (without cytosine methylation), isolated under identical conditions. As expected, LC-MS analysis of the reaction products shows peaks corresponding to $^{5mC}$, $^{5C}$, and $^{5aC}$ for pUC19 ($dcm^+$) but not pUC19 ($dcm^-$) in the presence of NgTET1 (Fig. 2C). On the other hand, both substrates form comparable amounts of oxidized T products ($^{5mU}$ and $^{5U}$) in the presence of NgTET1 (Fig. 2C), suggesting that this activity is not dependent on the presence of $^{5mC}$ in the DNA substrate.

We next probed the role of specific amino acid residues in the recognition of $^{5mC}$ and T in the active site of NgTET1. In the crystal structure, the flipped $^{5mC}$ is situated in the active-site pocket and stabilized by hydrogen-bonding interactions with the side chains of three key residues: aspartic acid 234 (D234), histidine 297 (H297), and asparagine 147 (N147) (Fig. S3) (10). Here we describe both $^{5mC}$- and T-oxidation activities of several site-directed variants of these residues with a 56-bp dsDNA substrate $^{5mC}$CGG (Table S1), plotting the ratio of the amount of $^{5mC}$ and T remaining after a 1-h reaction compared with that of the WT NgTET1 reaction (Fig. 2D). We also performed a kinetic time-course analysis for selected variants to monitor the decay of both $^{5mC}$ and T over time (Fig. S4). Alteration of any of these residues resulted in a decrease (relative to WT) in the total amount of $^{5mC}$ converted after 1 h, ranging from 17-fold for D234A to 1.5-fold for H297Q (Fig. 2D). The observed effects of these alterations on NgTET1 activity are in agreement with our

![Fig. 2.](image-url) T-oxygenase activity of NgTET1. (A) Kinetic time course depicting the decay of $^{5mC}$ or T for a reaction with 4 μM NgTET1 and 2 μM oligo $^{5mC}$CGG. Reaction species were detected and quantified by LC-MS (Agilent 1200). The data are fit to a single exponential and the observed rate constants with SEM are provided. (B) Quantification of oxidized T reaction species using 6.7 μM mTET1CD or NgTET1 (in Mops buffer pH 6.9) as measured by LC-MS (6490 Triple Quad LC-MS) for 0.2 μg sheared (1.5-kb) gDNA substrates. The error bars (in black) represent the SE (SEM) ($n \geq 3$). (C) LC-MS (Agilent 1200) kinetic traces, with species fraction determined by LC-MS (Agilent 1200), using excess enzyme (20 μg) and unmethylated pUC19 plasmid DNA (2.5 μg). (D) LC-MS (Agilent 1200) quantification of $^{5mC}$ or T after a 1-h reaction of 4 μM NgTET1 WT or variant proteins with 2 μM oligo $^{5mC}$CGG. Error bars (in black) represent the SEM ($n \geq 3$).

![Fig. 3.](image-url) NgTET1 activity is dependent on nucleotide-sequence context. Distribution of NgTET1 reaction species: (A and B) As quantified by LC-MS (Agilent 1200), using excess enzyme (20 μM) with (A) genomic (2.5 μg, sheared to 1.5-kb) or (B) plasmid (2.5 μg) DNA containing different methylation sequences as indicated in red (Table S2); (C and D) As quantified by 6490 Triple Quad LC-MS for (C) 30-min reaction of 8 μM NgTET1 with 4 μM oligo DNA or (D) 6.7 μM NgTET1 (in Mops buffer pH 6.9) or mTET1CD with 1.6 μM oligo DNA. For A–D, error bars (in black) represent the SEM ($n \geq 3$). (E) Kinetic traces, with species fraction determined by LC-MS (Agilent 1200), of NgTET1 with $^{5mC}$, $^{5aC}$, or $^{5C}$-containing oligos hemimethylated at a single CpX site (Table S1). Reactions were done in Mops buffer (pH 6.75) using 8 μM NgTET1 and 4 μM DNA. The data are fit to a single exponential and the observed rate constants with SEM are provided.
previously published results (10). Interestingly, T oxidation was decreased for some, but not all, of the variants. Alteration of D234 to A, most strikingly, greatly diminishes 5mC activity, but actually increases T activity by approximately threefold compared with WT. This pattern is also observed for D234N and H297Q, although to a lesser extent (Fig. 2D). The same trends are observed in the overall kinetic time courses of the NgTET1 variants (Fig. S4). D234 is proposed to make interactions specific for a C by interacting with the exocyclic amino group N4 of 5mC (Fig. S3), rather than T, which carries a carbonyl oxygen at the corresponding position. By disrupting this interaction and substituting D with the much smaller A, or to a lesser extent N, the active site pocket may more easily accommodate a T. However, the activity on T is still low (total conversion ∼7%) relative to the total amount of Ts present in the DNA molecule. It is yet unclear how the H297 to Q substitution may increase T oxidation, in the absence of any structural information for this variant.

The Extent of 5mC Oxidation Is Dependent on Sequence Context. We reported previously that NgTET1 has a strong preference for 5mCpG sites in oligo substrates (10). Here we test the activity of NgTET1 on other types of DNA, both genomic and plasmid, and confirm that NgTET1 activity is dependent on the methylation sequence context (Fig. 3A and B and Table S2). The amount of residual 5mC after a 1-h reaction varies greatly, with the most complete conversion of 5mC observed for substrates containing 5mCpG methylation (e.g., mammalian gDNA) and the least conversion for substrates with non-5mCpG methylation (e.g., MG1655 (Fig. 3A and B and Table S2). A similar context preference is observed for both NgTET1 and mTET1CD on M.Fnu4HI gDNA, which is methylated at 5GpCNNC sites (Fig. 1E).

To examine the effect of the methylation sequence context on the activity of NgTET1 more closely, we designed oligo substrates bearing an N5mCpCNpGG methylation motif, where N5 and N7 are maintained at C when N5 is A, T, C, or G, and N7 is maintained at G when N7 is A, T, C, or G (Table S1). Our results reflect that the nucleotide 2′ upstream of 5mC plays a minor role in the extent of 5mC oxidation by NgTET1, whereas the nucleotide 3′ downstream exhibits a much more pronounced effect on this activity with a substantial increase in the intermediate species (5hmC and 5FmC) when N5 is not a G (Fig. 3C). The relative amount of 5FmC formed after a 30-min reaction for non-5mCpG-containing substrates is only 6–20%, compared with 72% for the 5mCpG-containing substrate (Fig. 3C). These results are consistent with those of plasmid and genomic substrates (Fig. 3B and C), as well as previous observation (10). We next tested substrate preference for NgTET1 T-oxidase activity. A set of ssDNA oligo substrates containing nine Ts, each followed by either a G, A, or C was compared (Fig. 3D and Table S1), and the results reflect a TpG sequence preference.

Interestingly, each oxidative step of the reaction seems to be differentially sensitive to sequence context. For example, much lower amounts of 5CaC are formed on G5mCpG substrates (M.AluI and M.HacIII) than G5mCpCpG substrates (MG1655 and pUC19), despite approximately the same amount of oxidized 5mC. These differences prompted us to further investigate the observed sequence context preference when the starting substrate contains 5hmC or 5FmC, rather than 5mC. Fig. 3E shows the decay of 5mC, 5hmC, or 5FmC over time for hemiDNA substrates containing a single modification. The data demonstrate the faster kinetics of the first step of the overall NgTET1 reaction (i.e., 5mC to 5hmC) compared with the two subsequent steps (Fig. 3E). In addition, the decay of 5mC proceeds to completion with nearly all of it gone after a 1-h reaction, whereas a considerable amount of starting material remains for the 5hmC- and 5FmC-containing substrates (~20% and ~40%, respectively). Despite these differences in kinetics between each step of the reaction, the substrate preference for CpG sites is observed for the 5hmC- and 5FmC-containing oligos as well, and CpG-modified oligos are clearly the poorest substrates (Fig. 3E). Overall, these results suggest that the reaction kinetics of each oxidative step by NgTET1 not only depends on the type of cytosine modification but also on the sequence context flanking the modified cytosine.

Mapping 5mC Using NgTET1 in SMRT Sequencing. SMRT sequencing has been shown to readily detect most modifications in a DNA template, such as N6-methyl-adenine and N4-methyl-cytosine, but is limited in its detection of 5mC (14–16). In SMRT sequencing, the DNA polymerase kinetics are monitored by measuring the interpulse duration (IPD), the length of time between two successive nucleotide incorporation events (14, 15). The effect on the IPD ratio, a measurement of the IPD compared with an unmodified control template, is indicative of the modification of a specific base. By using NgTET1 to convert 5mC to 5FmC, the newly modified base gives a stronger signal enabling better detection of 5mC (17).

We first used NgTET1 to treat pRS(M.HpaII), containing 15 C5mCGG recognition sites on each strand. Under our reaction conditions, all of the 5mC is reacted and 87% 5mC is generated (Fig. 1C). Fig. S5 shows the plasmid-wide view of IPD ratio data for pRS(M.HpaII) treated with NgTET1 and a representative IPD ratio profile in a 50-bp window. An increase in IPD ratio is detected at each predicted 5mC site, and the observed primary IPD ratio peak at the +2 position and a less prominent peak at the +6 position relative to the modification are consistent with the kinetic signature observed for a 5FmC modification (17). As a result, all 30 5mCpG sites in pRS(M.HpaII) were readily detected (Fig. S5).

We then used NgTET1 to map the methylome of Helicobacter pylori strain 26695 gDNA, in which there are three known active cytosine-5-methylases with the specificities: G5mCCG, 5mCCTC, and C5mCTTC (18, 19). The reaction of H. pylori gDNA with NgTET1 results in an overall 5mC conversion of ~95% but with only 52% 5CaC formed (Fig. 3A). The relatively low conversion efficiency to 5FmC may be because of the fact that two of the three known 5mC methylation motifs in H. pylori are in a non-CpG context. Nonetheless, all three methylated motifs were detected and the sequence contexts around the methylated sites show that there is no significant bias, suggesting that NgTET1 did not preferentially convert a subset of these sites (Fig. 4A).

Because of the high levels of 5hmC and 5FmC formed in the reaction with H. pylori genome gDNA (Fig. 3A), we wanted to improve the detection of 5mC by using sodium borohydrate (NaBH4) to reduce 5FmC to 5hmC, and T4-β-glucosyltransferase (T4-JGT) to convert all 5hmC to β-glucosyl-oxy-5-methylcytosine (5gmc), a modification that would readily be detected by SMRT sequencing (20, 21). This protocol results in a product mixture of 5gmc and 5gmc with negligible amounts of 5hmC, 5FmC, or 5FmC (Fig. S6). Using this approach followed by SMRT sequencing, we observed an improved signal in the genome-wide average IPD ratio profile, leading to a higher percentage detection of all three methylated motifs (Fig. 4B). Note that in addition to the IPD ratio increase in the +2 position, 5hmC increases the IPD ratio at the modified cytosine position.

Using this method, 89%, 98%, and 57% of the sites genome-wide are reported as methylated for the 5mCCTC, 5gmcCTTC, and G5mCGC motifs, respectively (Table S4). These detection percentages are slightly improved from earlier results for H. pylori 26695 using mTET1CD (80%, 92%, and 50%) (19). However, the low detection for the G5mCGC motif is puzzling, especially because NgTET1 should not exhibit bias on this site (Fig. 4A). We therefore compared the IPD profiles between the GCGC sites detected as methylated and the sites detected as unmethylated, as reported by the Pacific Biosciences analysis pipeline, and noticed that the algorithm appears to ignore those
NgTET1 can be applied to SMRT sequencing to assist in mapping bacterial methylomes, and that further optimization in the modification detection algorithm can be made to increase performance.

**Discussion**

Members of the TET/JBP family are distributed over a wide phylogenetic distance and often show patterns of lineage-specific expansion. Among them, JBP1 and JBP2 have been shown to possess T-hydroxylation activity, whereas multiple paralogous genes in mammals, mushroom (Coprinus cinereus), and honey bee (Apis mellifera) have been shown to possess 5mC-oxygenase activity (1, 6, 22, 23). Our results demonstrate that NgTET1 and the evolutionarily distant mTET1CD possess both 5mC- and T-oxygenase activities, which may initially have been shared by a single ancestral enzyme. Although 5mC-oxygenase activity in both enzymes is almost identical (Fig. 1E), mTET1CD T-oxygenase activity is significantly lower than that of NgTET1 (Figs. 2B and 3D). Indeed, the highest-level oxidation product, 5caU, is obtained only in the reaction of NgTET1. This is, to the best of our knowledge, the first in vitro evidence for oxidation of T on DNA to 5caU by any DNA modifying enzyme. Based on these observations, we hypothesize that unlike the heterolobosean NgTET1, and the evolutionarily close kinetoplastid JBP1, the mammalian TETs may have gradually lost most of their T-oxygenase activity, possibly to accommodate the emergence of multiple 5hmU glycosylases in the mammalian genome. More work is required to elucidate whether the oxidized Ts are epigenetically relevant or merely a promiscuous activity of these enzymes.

Interestingly, the mutant D234A (and H297Q, to a lesser extent) of NgTET1 appears to have reversed its substrate preference with almost no 5mC-oxygenase activity and increased T-oxygenase activity (Fig. 2D). Structure-based sequence alignments reveal that the residues D234 and H297 are conserved among the methylated GCG motifs detected by SMRT sequencing of gDNA using NgTET1. (A) Sequence logos for 5mCCTC, 5mCCTC and GmCGC motifs detected by SMRT sequencing of NgTET1-treated gDNA. (B) IPD ratio plots corresponding to the 5mCCTC motif in gDNA treated in the absence of NgTET1 (Top), with NgTET1 (Middle), or with NgTET1/NaBH4/T4-4GT (Bottom). (C) IPD ratio plots for the sequences detected (Upper) versus undetected (Lower) as belonging to the 5mCGC motif for gDNA treated with NgTET1/NaBH4/T4-4GT. For B and C, the error bars (in black) represent the SEM. (D) Scatter plot of IPD ratio values at the methylated and +2 positions for 5mCGC and CCGG sequences for gDNA treated with NgTET1/NaBH4/T4-4GT. (E) Scatter plot at undetected positions only. All observations were made for the 5mCCTC motif (Fig. S7). Here we explored a simple hard decision boundary for both methylated and +2 positions, and plotted the corresponding detection sensitivity, as well as specificity in Fig. 4E. Note that we used the genome-wide CCGG sites, which are known to be unmethylated in the genome, as the true negative set to calculate the detection specificity. It can be seen that by imposing the same IPD ratio cut-off (e.g., 1.75) at both modified and +2 positions, it is possible to dramatically increase the detection sensitivity to over 90% while maintaining high specificity (false-discovery rate < 5%) for all motifs (Fig. 4E). Overall, the above results demonstrate that
Tet-like dioxygenase in complex methylomes

We thank Derrick Xu and Meg Mabuchi for initial efforts in the development of oxBS-seq (26), in the fact that it does not include the relatively harsh bisulfite treatment that could result in DNA degradation (27–29). Although this method does not currently distinguish between 5mC and 5hmC (both TAB-seq and oxBS-seq distinguish between these two modifications), NgTET1 could be potentially used in a manner analogous to that of mammalian TET1 in TAB-seq to map 5hmC.

Using NgTET1 coupled with SMRT sequencing, we show comprehensive characterization of 5mC motifs in both plasmid and gDNA. As previously reported for SMRT sequencing of cytosine modifications (17), a prominent increase in signal at the +2 position is observed along with an increase at the modification position (0 position). Our analysis indicates that improvement in sensitivity can be made by incorporating both the methylated and +2 position into the 5mC calling algorithm (90% detection compared with 57% for the G^3CCG motif), while ensuring high specificity (Fig. 4E). It is important to note that NgTET1 efficiency, in its oxidation of 5mC on various substrates as well as its effectiveness as a tool for comprehensive genome-wide mapping of 5mC modification using SMRT sequencing, matches that of mTET1C (19). The smaller size of NgTET1 renders it an ideal enzyme for production, sequencing applications, and future engineering efforts for the purposes of relaxing substrate specificity and enhancing 5mC conversion in all types of gDNA. We envision the use of NgTET1 in a variety of methylome sequencing applications for the critical understanding of the epigenetic function of 5mC and its oxidized forms.

Materials and Methods

Protein purification and preparation of DNA substrates are described in detail in SI Materials and Methods. Detailed descriptions of DNA substrates used are provided in Tables S1 and S2. See Table S5 for a list of ds-oligo substrates containing the MspI site, used in Fig. S1.

The NgTET1 reaction conditions and LC–MS–based activity assay were performed as described previously (10) and are described in detail in SI Materials and Methods. For the RE-based NgTET1 activity assay, purified DNA (300 ng) from each NgTET1 reaction was digested with 20 units (U) of BamHI (New England Biolabs) to linearize the plasmid and 50 U of MspI (New England Biolabs) in New England Biolabs CutSmart buffer (pH 7.9) for 1 h at 37°C in 20-μl total volume. The reaction products were resolved on a 1.8% agarose gel.

Preparation of NgTET1 and NgTET1NaBH4/T4-jGT reaction samples for SMRT sequencing and preparation of SMRTbell template libraries, sequencing, and analysis are detailed in SI Materials and Methods.

ACKNOWLEDGMENTS. We thank Derrick Xu and Meg Mabuchi for initial work on the project; Rick Morgan and Yvette Luyten for providing the pPR5 (M.Hpall) plasmid and M.Fnu4HI gDNA; Chandler Fulton and Elaine Lai for helpful discussions; and Bill Jack for critical review of the manuscript. This work is supported by New England Biolabs, and by National Institutes of Health Grants GM105132 (to Y.Z.) and GM049245-21 (to X.C.). Funding for the open access charge is from New England Biolabs.

References