NUCLEOBASE, NUCLEOSIDE, AND NEIGHBORING NUCLEOTIDES: INTRINSIC PREFERENCES FOR TET ENZYME-MEDIATED OXIDATION OF 5-METHYLCYTOSINE

Jamie E. DeNizio

A DISSERTATION

in

Biochemistry and Molecular Biophysics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

_____

Rahul M. Kohli, M.D., Ph.D., Assistant Professor of Medicine

Graduate Group Chairperson

_____

Kim A. Sharp, Ph.D., Associate Professor of Biochemistry and Biophysics

Dissertation Committee:

Kristen W. Lynch, Ph.D., Professor of Biochemistry and Biophysics

Marisa S. Bartolomei, Ph.D., Perelman Professor of Cell and Developmental Biology

Ben E. Black, Ph.D., Professor of Biochemistry and Biophysics

Zhaolan (Joe) Zhou, Ph.D., Associate Professor of Genetics

To my husband, Michael. You'll never walk alone.

ABSTRACT

NUCLEOBASE, NUCLEOSIDE, AND NEIGHBORING NUCLEOTIDES: INTRINSIC PREFERENCES FOR TET ENZYME-MEDIATED OXIDATION OF 5-METHYLCYTOSINE

Jamie E. DeNizio

Rahul M. Kohli

The Ten-eleven-translocation (TET) family of enzymes can oxidize the fifth base of DNA, 5-methylcytosine (mC) sequentially, to 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), and 5-carboxycytosine (caC). The biochemical preference of TET enzymes for these substrates, in the canonical cytosine guanine dinucleotides (CpG), mimics the order in which they are generated and is reflected in levels of these oxidized modifications (oxmCs) detected in various genomes. Other than this exception, there is conflicting or limited data concerning intrinsic substrate preferences of TET, particularly with regards to different nucleic acid structures, sequence contexts, and extent to which TET mediates oxmCs in clustered proximity to one another. Thus, in this thesis, I present our efforts to determine intrinsic substrate preferences of TET enzymes, and in doing so expand upon our understanding of mechanisms driving these relative activities and the functional significance of observed levels of oxmCs *in vivo*. After a review of the field, in Chapter 2, I present our work comparing TET activity on different DNA and RNA structures *in vitro*. We found that TET is relatively promiscuous on a variety of DNA/RNA structures but prefers DNA, a specificity that is dictated by nucleic acid identity of the target base, as well helical conformation of the substrate. In Chapter 3, I newly expose the relative tolerance of TET activity on hmC with a non-G at the +1 base, although mCpG is still largely preferred. This tolerance for hmC oxidation by TET and fC and caC excision by TDG, regardless of the +1 base, supports a model explaining hmCpH depletion relative to mCpH and hmCpG in some genomes. In Chapter 4, I narrate our efforts to quantify clusters of oxmCs using modification-specific sequencing methods and observe that TET is intrinsically capable of clusters of at least fC and

caC. Also, we explore the possibility that these clusters are mediated by either strand processivity of TET or underlying sequence context preferences. Finally, I propose two kinetics-based experiments to test our hypotheses regarding mechanisms driving these substrate preferences, along with ways to exploit our knowledge of TET enzymes for creation of more efficient and specific epigenetic editing tools.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1: Introduction

Portions of this chapter have been adapted from:

Liu, M.Y., DeNizio, J.E., Schutsky, E.K., Kohli, R.M. The expanding scope and impact of epigenetic cytosine modifications. *Curr. Opin. Chem. Biol.* 2016, 33, 67-73.

1.1: Epigenetic cytosine modifications

1.1.1: 5-methylcytosine, the fifth base of DNA

In a single cell, there are more than 6 billion deoxyribonucleotides. Simultaneously, there can be as much as 8X more RNA material. In this congested milieu of information-carrying molecules, epigenetic modifications provide an instruction manual to ensure the reliable interpretation of the genetic code. In eukaryotes, covalent modifications of DNA are a critical component of this regime. The prototypical modification, 5-methylcytosine (mC), is broadly conserved in species ranging from vertebrates to fungi and protists and is considered the fifth base of DNA (Smith and Meissner, 2013). It is considered an epigenetic mark both because it does not alter the Watson-Crick base-pairing of cytosine and because it can be heritable across cell generations.

DNA has been shown to be methylated at the 5-position of cytosine in both cytosine-guanine dinucleotides (CpGs) and CpH (H=A,G,T) sequence contexts. Each CpH dinucleotide makes up roughly 4% of the genome, as expected by probability. Although CpH methylation (mCpH) has been detected at as much as 2-6% of CpHs in embryonic stem cells (ESCs) and neurons (Jang et al., 2017), it is sparse in differentiated cells. In contrast, CpGs are relatively depleted in vertebrate genomes, accounting for less than 1% of dinucleotides, but 60-80% of these CpGs are methylated. The reduction of CpGs is due to the high rate of spontaneous deamination of mC, which results in a G•T mismatch (Bird, A. P., 1980; Cooper and Youssoufian, 1988; Shen, J. C. et al., 1994). Unlike spontaneous C deamination, which results in a uracil that is

quickly repaired via base excision repair (BER), G•T mismatches are often erroneously resolved to A•T (Bellacosa and Drohat, 2015). Interestingly, CpGs tend to occur in clusters, primarily in regions termed CpG Islands (CGIs), which are CpG-dense regions typically devoid of mC that occur in or near 70% of promoters in vertebrates (Deaton and Bird, 2011).

The presence of cytosine methylation is primarily associated with transcriptional repression. In particular, methylation of CGIs near promoters often results in gene silencing due to impaired transcription factor binding or the recruitment of other proteins such as histone modifiers that are thought to promote a closed, repressive chromatin state (Deaton and Bird, 2011). There have been a couple instances in which a single mCpG site has been implicated in affecting transcription, both in an imprinting (Choi et al., 2018) and cell-type specific manner (Furst et al., 2012). However, in general, impactful changes in methylation occur at multiple CpGs. When these regions exist with different methylation statues among multiple samples, such as in different states of development, various cell types, or diseased versus healthy tissues, these regions are formally considered Differentially Methylated Regions (DMRs).

Although most commonly considered a repressive mark, the downstream effect of methylation can vary depending on the biological context. For example, there is loci-specific variation of whether the methylation state acts as the catalyst for or a component of silencing (Schubeler, 2015). Meanwhile, non-methylated CGIs have also been associated with inactive genes, and mCpGs have also been shown to exist in activating transcription regulatory regions. CpG methylation has also been identified in intergenic DNA, exons of genes, and satellite DNA, but the role of methylation in these regions is still somewhat undefined. The function of CpH methylation has been shown to be similarly variable. mCpHs can cause transcriptional repression *in vitro*, as well as in mouse neurons (Guo et al., 2014). In contrast, actively transcribed genes in neurons tend to lack mCpH (Lister et al., 2013), while those actively transcribed in ESCs tend to be hyper-methylated at CpHs (Lister et al., 2009).

Regardless of the downstream effects on transcription, the symmetrical nature of CpG dinucleotides allows for the stable inheritance of the relative methylation statuses. While methylation is primarily established during embryonic development by DNA methyltransferase (DNMT) 3a/3b, irrespective of opposite strand methylation, it is maintained after cell division by DNMT1, which specifically copies mC, converting hemi-methylated DNA after replication to symmetrically methylated DNA (Jurkowska et al., 2011). This mechanism of maintenance, allowing for epigenetic "memory," results in the long-term stability of methylation patterns, which are important for maintaining cell fates, tumor suppression, silencing retrotransposons, X-chromosome inactivation, and establishing genomic imprinting (Bird, A., 2011; Dor and Cedar, 2018). While both DNMT1 and DNMT3a/3b exhibit greater activity for CpG sites, DNMT3a/b are considered to be solely responsible for both the establishment and maintenance of mCpH *in vivo*, as DNMT1-specific methylation has been shown to localize only to CpGs (Guo et al., 2014).

In addition to DNA, cytosine methylation has also been detected in different nucleic acid structures. 5-methylcytosine in RNA is much less prevalent, accounting for only at most 1% of cytosines in RNA, but has been identified in several kinds of RNA, including mRNA, tRNA, rfRNA, and ncRNA (Huang, W. et al., 2016; Motorin et al., 2010; Squires et al., 2012). Also, unique RNA methyltransferases, NSun2 and DNMT2, have been identified and characterized (Bohnsack et al., 2019). Although the functional role of mC in RNA remains largely unclear, there is evidence that it plays a role in nuclear transport and enhancing mRNA translation. Further, deletion of the RNA methyltransferase NSun2 has been shown to impair cellular differentiation pathways and is associated with neurological deficiencies in mammals (Sanchez-Vasquez et al., 2018).

1.1.2: Pathways of cytosine demethylation

Just as there is a biological need for methylation, there is a requirement for the need for the removal of DNA methylation, or DNA demethylation, as well. There are two methods by which DNA demethylation, *i.e.* the ultimate removal of the 5-methyl group, can occur. First, DNA

replication can lead to a passive reduction in methylation when the maintenance activity of DNMT1 is diminished or inhibited; the mC signal is diluted over time upon sequential replications. This can occur for mCpGs, but as mentioned above, it likely occurs on a more significant scale for mCpHs due to DNMT1 discrimination. In addition to this passive, temporally-dependent mechanism, evidence in germ cells and early embryos dictate that an active mechanism must also exist for rapid, regulated demethylation (Kohli and Zhang, 2013). In plants, there are a family of DNA glycosylases, Repressor of Silencing 1 (ROS1)/DEMETER (DME), that can excise mC directly (Gong et al. 2002; Jang et al. 2014; Hong et al. 2014). In the absence of a similar enzyme in mammals, several indirect avenues for demethylation have been hypothesized. At one point, there was strong interest in the possibility that demethylation could occur after repair of deamination of mC. It was proposed that the AID/APOBEC family of enzymes, known for deaminating cytosine, could deaminate mC to thymine for excision by thymine DNA glycosylase (TDG) or methyl-binding domain protein 4 (MBD4), followed by base excision repair (BER) to reintroduce unmodified cytosine (Rai et al., 2008; Bhutani et al., Nature 2010; Cortellino et al., 2011). Although our lab has recently shown that at least one family member, APOBEC3A, is capable of efficiently deaminating mC (Schutsky et al., 2017), other family members have poor activity on mC (Nabel et al., 2012). Further, the AID/APOBEC family, in general, have strong sequence preferences in single-stranded DNA (Ito, F. et al., 2017), making this pathway an unlikely candidate for the full extent of demethylation observed.

Ten-eleven translocation (TET) family of enzymes, which are Fe(II)/α-ketoglutarate (α-KG)-dependent dioxygenases, and are now considered one of the major players of active demethylation. The three mammalian TET homologs (TET1, 2, 3) were initially found to oxidize mC to 5-hydroxymethylcytosine (hmC) (Ito, S. et al., 2010; Tahiliani et al., 2009). Later discoveries revealed TET enzymes could sequentially oxidize hmC to 5-formylcytosine (fC) and 5-carboxylcytosine (caC) (He, Y. F. et al., 2011; Ito, S. et al., 2011; Pfaffeneder et al., 2011)

(Figure 1-1). The conversion of mC to the oxidized bases (collectively referred to as oxmCs), fC and caC in particular, is thought to invoke passive demethylation upon replication as DNMT1 is largely inhibited from methylating opposite oxmCs (Hashimoto et al., 2012; Ji et al., 2014; Seiler et al., 2018). However, proposing a role for passive demethylation in any step of the pathway would not account for the rate at which demethylation occurs in germ cells and embryonic cells. Currently, the most compelling model for active demethylation involves the oxidation of mC to fC or caC by TET, base excision of fC and caC by TDG, followed by BER to regenerate unmodified cytosine (Chen, H. et al., 2014; He, Y. F. et al., 2011) (Figure 1-1). In support of this model, TDG knockout mice exhibit higher levels of 5mC (Cortazar et al., 2011; Cortellino et al., 2011), and the overexpression of TDG in HEK293 cells causes lower levels of fC (Nabel et al., 2012) and caC (He, Y. F. et al., 2011). In addition, this pathway has been reconstituted *in vitro* (Weber et al., 2016). Notably, it has not been explored whether active demethylation occurs at CpHs or in other likely TET substrates, such as RNA.

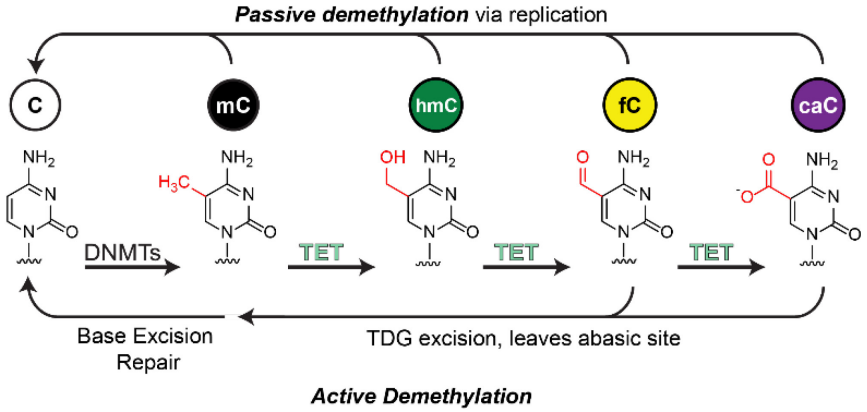


Figure 1-1. TET oxidation is involved in two possible pathways for cytosine demethylation.

DNMTs transfer a methyl group to the 5-position of cytosine. TET enzymes are capable of sequentially oxidizing mC to either hmC, fC, or caC. All four modified cytosine bases can be passively demethylated after replication due to maintenance failure or inhibition. Only fC and caC are possible intermediates for active demethylation, as they are excised by TDG, then ultimately replaced with unmodified cytosine via base excision repair.

1.1.3: Stability of oxmCs in epigenome

The bases resulting from TET oxidation are not merely transient intermediates in demethylation but have been stably mapped in many cell lines (Booth et al., 2015; Wu, H. and Zhang, 2015). In some cell types, 5hmC can account for 1-10% of 5mC, depending on the sequencing methodology, with levels of ~10% reported in ESCs and as high as 40% in Purkinje neurons (Globisch et al., 2010). The higher order modifications, fC and caC, are also stably detected; however, they are found at levels at least 10-fold lower than hmC— approximately 1 in 10^5–10^6 nucleotides. In general, oxmCs are generally considered activating marks, as they are enriched in active enhancers and other regulatory regions.

This observed stability detected by sequencing localization begs the question whether these modifications are maintained after replication, like mC, or independently re-established via an alternate mechanism. Isotopic labeling experiments support potentially long-lived modifications (Bachman et al., 2014; Bachman et al., 2015). Despite the necessity for mC to act as a precursor, it seems unlikely that the maintenance of oxmCs occurs through the same mechanisms as mC. First, it would require targeted methylation across from an oxidized base, which is unlikely as DNMT1 has been shown to be very inefficient at methylating opposite an oxmC (Hashimoto et al., 2012; Ji et al., 2014; Seiler et al., 2018). Second, a TET enzyme would be required to both recognize the existing oxmC and the mC in the CpG•CpG pairs to generate the same base across from it. In addition, single-molecule fluorescent labeling of both mC and hmC indicated that hmC frequently occurs opposite mC in CpG nucleotide pairs and that dually hydroxymethylated CpG•CpG base pairs are rare (Song, C. X. et al., 2016). In agreement with this observation, *in vitro* work within our lab showed that mouse TET2 has *de novo* activity

(Crawford et al., 2016), oxidizing independently of the identity of the opposite strand CpG. However, it remains unclear if other TET homologs have maintenance activity.

1.2: Functional epigenetic roles of oxmCs

1.2.1: Chemical effects of oxmCs

There are still many important questions remaining concerning how oxmCs are generated. That said, it is helpful to understand the functional consequences of these modifications on gene expression when considering where and when TET may be regulated to mediate these marks. For example, there is some evidence that oxmCs themselves can directly impact transcription. Ox-mCs form high-fidelity, Watson-Crick base pairs with guanine and are generally not prone to spontaneous deamination and oxidation events (Renciuk et al., 2013; Schiesser et al., 2013). The 5-modified group occupies the major groove of B-form DNA and appears to have a subtle but potentially significant impact on helical thermodynamics and stability (Renciuk et al., 2013; Szulik et al., 2015). For example, the electron-withdrawing character of 5-formyl and 5-carboxyl groups weakens the N-glycosidic bond and decreases the pK_a of the base, resulting in less stable base pairing and possibly promoting base excision by TDG (Dai et al., 2016; Maiti et al., 2013). Also, DNA containing as few as one fC or two hmC displayed greater flexibility in a single-molecule cyclization assay, and this flexibility correlated with enhanced nucleosome stability (Ngo et al., 2016). However, hmC-containing DNA has also been shown to have increased binding affinity to the histone core but ultimately results in looser chromatin packing due to the weakened interaction between hmC and the H2A-H2B dimer (Mendonca et al., 2014). In support of this potential effect of oxmCs, looser chromatin packing has been shown to correlate with high levels of hmC (Mahe et al., 2017).

1.2.2: Protein readers of oxmCs

Overall, however, the evidence suggests that 5-modification of cytosines largely maintains the structural and sequence integrity of DNA while providing an accessible handle for epigenetic readouts. Through proteomic analyses, several oxmC-specific reader proteins have been identified (Chen, C. C. et al., 2012; Liutkeviciute et al., 2009; Spruijt et al., 2013). Transcription factors have been one area of focus; for example, methyl-CpG-binding protein 2 (MeCP2) was identified as a major hmC-binding protein in the brain, with similar affinity for mC and hmC (Iurlaro et al., 2013), while Wilms tumor protein 1 (WT1) can recognize caC as well as C and mC (Hashimoto, Olanrewaju et al., 2014; Spruijt et al., 2013). The yeast RNA polymerase II elongation complex was reported to form hydrogen bonds specifically with caC, resulting in transient transcriptional pausing (Kellinger et al., 2012; Wang, L. et al., 2015). Also, through altered CTCF binding, the identity of the modified cytosine has been shown to dictate splicing by affecting the rate of RNA polymerization: The presence of mC at a CTCF site resulted in exclusion of an exon while an hmC or caC caused exon inclusion (Shukla et al., 2011).

Thus, there is evidence to suggest both that oxmCs themselves can play a role in gene regulation and that they can be read by proteins that affect transcription downstream. The mechanisms discussed in this section are likely more significant for stable oxmCs, but even the transient presence of these modifications may elicit similar effects temporarily. Regardless, it appears that the signature of oxmCs is very context-specific. Understanding the scope of TET reactivity, and further, how TET is intrinsically tuned to regulate this activity, can potentially help us link context and phenotype to function.

1.3: Scope of TET reactivity

There are several layers to the regulation of the catalytic activity of TET enzymes (Figure 1-2). While they are all undoubtedly interconnected, it is informative to discuss them as discreet reactivities. From a wide lens in, they are as follows: type of nucleic acid (DNA versus RNA, Chapter 2), sequence context (CpG versus CpH, Chapter 3), C versus T, and modified 5-position preference. In the following section, I will discuss what we know about these levels of reactivity from the biological contexts in which TET homologs and oxmCs occur and previous structural and biochemical data.

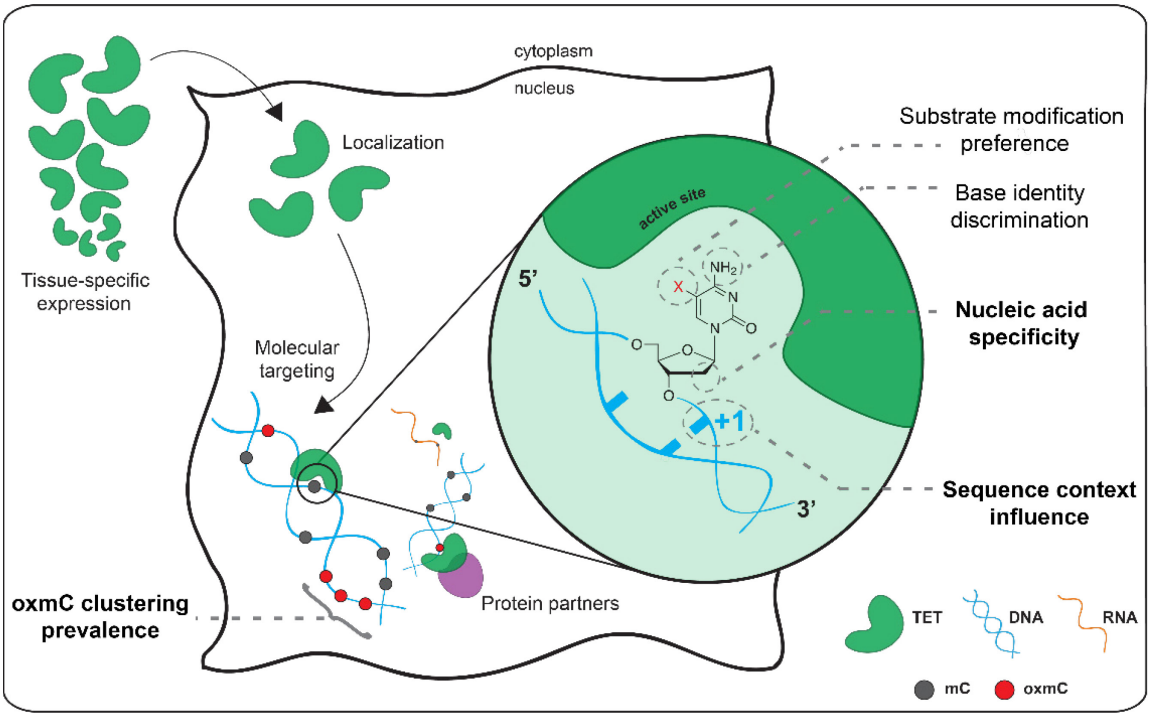


Figure 1-2. Summary cartoon of factors influencing TET reactivity.

1.3.1: External factors influencing TET reactivity

Prior to discussing the intrinsic factors that influence TET oxidation, it is important to highlight the external factors that can dictate TET reactivity, such as post-translational modifications (PTMs), protein recruitment, and chromatin accessibility. PTMs have been shown to

affect localization and chromatin binding (Wu, X. and Zhang, 2017). Several PTMS have also been demonstrated to alter the enzymatic activity of TET enzymes. For example, acetylation of two N-terminal lysine residues of TET2 enhances activity and stabilizes the protein (Zhang, Y. et al., 2017). O-GlcNAcylation stimulates the activity of TET1 to a greater extent than the OGT-TET interaction alone (Hrit et al., 2017). TET is also the target of both covalent and noncovalent PARylation; while the former is stimulating, the latter decreases the catalytic activity of TET (Ciccarone et al., 2015). There has also been some evidence that metabolic pathways can influence TET activity via modification of the TET protein: In starvation conditions, it has been shown that AMP-activated kinase phosphorylates TET2 residue S99. Along with the finding that TET2 in cells cultured in high-glucose media exhibit a reduced half-life, it is proposed that S99 phosphorylation protects TET2 from degradation (Wu, D. et al., 2018).

Although a classic chaperone has not been identified, the TET enzymes do interact with a variety of proteins (Spruijt et al., 2013), a number of which have been shown to aid in recruiting TET to specific loci. For example, several transcription factors, including NANOG, PU.1, and WT1, have been shown to recruit TET to their target genes to enhance the efficiency of reprogramming, ensure monocyte-to-osteoclast differentiation, and suppress acute myeloid leukemia (AML), respectively (Costa *et al.*, Nature 2013; de la Rica *et al.*, Genome Biol 2013; Wang *et al.*, Mol Cell 2015). It has also been proposed that Lin 28, which binds both ssDNA and RNA, binds active transcriptional bubbles, and recruits TET1 to certain loci to generate hmC (Zeng et al., 2016). However, no direct protein interaction was shown. TET2 has also been shown to be recruited to specific promoters by co-activator SMAD nuclear interacting protein 1 (SNIP1),

which also facilitates the interaction of TET2 with many different sequence-specific DNA-binding factors (Chen, L. et al., 2018).

The level of chromatin compaction is also proposed to affect both TET activity and the capacity for active demethylation. In *in vitro* assays using reconstituted nucleosomes, TET1 prefers the linker region rather than the histone core (Kizaki, Seiichiro, Zou et al., 2016). With regards to the capacity for demethylation, TDG activity has also been shown to be inhibited to some extent by the nucleosome histone core; while both U and T can be excised from a nucleosome core particle at the wobble base pair in dyad region, there is still some proportion that is inaccessible to cleavage, either due to inhibition of the base extrusion or TDG intercalation, or inhibition of initial TDG binding (Tarantino et al., 2018). However, fC- and caC- containing nucleosomes have not been tested.

1.3.2: Biological contexts of TET enzymes

The biological contexts where TET homologues and oxmCs are found can be informative of both the substrate and sequence determinants of TET enzymes (Figure 1-2). Computational studies have mapped a large family of TET-related proteins spanning the evolutionary tree (Iyer et al., 2013). Some species harbor a bewildering number of TET homologues—47 in the fungus *C. cinerea*, many catalytically active and associated with transposons (Iyer et al., 2014; Zhang, L. et al., 2014). There is also a TET homologue in *D. melanogaster* (DmTET), but interestingly, mC is absent from genomic DNA but present in RNA, along with oxmCs (Delatte et al., 2016; Fu et al., 2014). It is untested whether DmTET lacks activity on deoxyribomethylcytosine or if the absence of this substrate in the fly genome obscures this possible activity. Further, it is unclear what the comparative activity levels of DmTET versus mammalian TET homologs are on RNA.

On the other hand, oxidized RNA bases are detected in Tet-null embryonic stem cells (Fu et al., 2014), as well as in organisms that lack TET homologues (Huber et al., 2015), indicating that TET-independent mechanisms could contribute.

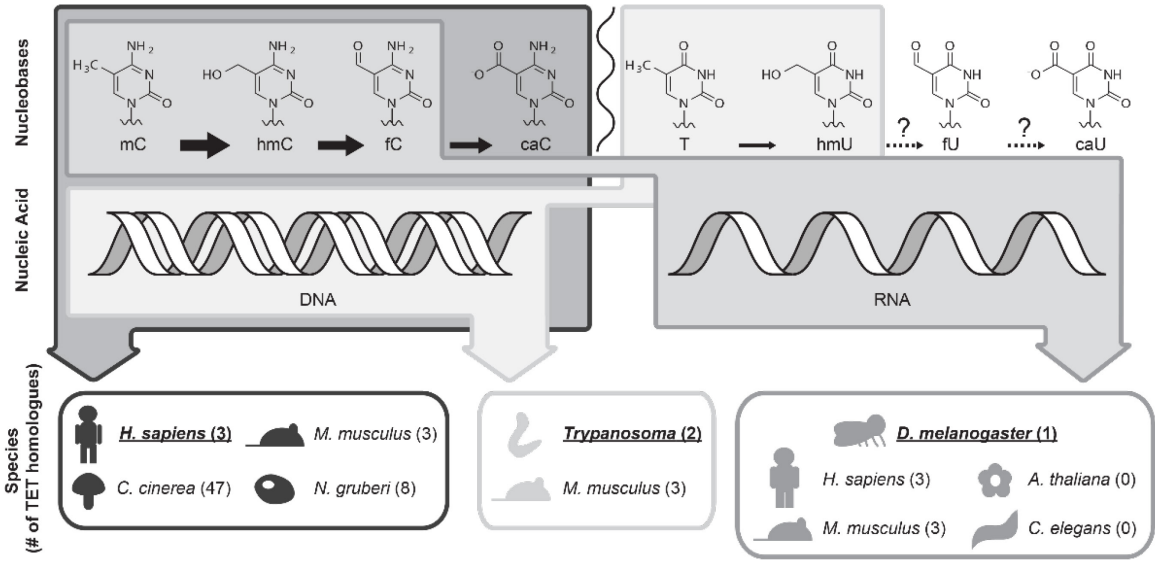


Figure 1-3. The broad scope of TET reactivity.

Summary of the generation of various nucleobases *(top row)* in both DNA and RNA *(middle)* of diverse species *(bottom)*, with emphasis on the strongest exemplar *(underlined)*. Canonically, TET enzymes iteratively oxidize mC in dsDNA to hmC, fC, and caC, with decreasing reactivity on more highly oxidized substrates. In addition, mC, hmC, and fC have been detected in RNA, with evidence for TET-dependent oxidation, as well as potential alternative mechanisms. Finally, TET enzymes can, to a limited extent, convert T to hmU in DNA, similar to the well-known JBP1/2 enzymes in Trypanosoma.

Analyzing different TET homologs can also highlight whether substrate specificities are varied or conserved. For example, the preference for CpGs in DNA is conserved between mammalian TET homologs and *Naegleria gruberi* TET1 (NgTET1) (Pais et al., 2015). In contrast, human TET2 has been previously shown to have minimal activity on mCpHs (Hu, L. et al., 2013), while NgTET1 has robust activity on certain modified CpHs *in vitro* (Pais et al., 2015). Several of

the TET homologs also display differential activity on cytosine versus thymine. Several dioxygenases in the Fe(II)/αKG-dependent family, such as the trypanosomal J-binding proteins (JBP1 and JBP2), naturally hydroxylate thymine, resulting in 5-hydroxymethyluracil (hmU), rather than cytosine (Cliffe et al., 2009). NgTET1 can also oxidize thymine *in vitro* but has at about 30-fold less activity on T compared to mC (Pais et al., 2015). Meanwhile, mammalian TET enzymes have very limited activity on thymine both *in vivo*, via the detection of hmU in cells (Pfaffeneder et al., 2014), and *in vitro* (Pais et al., 2015). Shared between all species is the preference for oxidizing mC over the other two substrates, hmC and fC. Although it will be discussed further in subsection 1.3.4, this shared preference likely occurs because the key TET residues responsible are largely conserved across the phylogenetic tree.

The human TET homologs are arguably the best-characterized among the enzyme family and comparing the different targeting domains in both the full-length and truncated isoforms can illuminate TET's intrinsic localization mechanisms. First, it should be noted that the human TET homologs (TET1-3) are expressed differentially across development, as well as in different tissues. For example, TET1 and TET2 are expressed in embryonic stem (ES) cells, while TET2 and TET3 are found in most differentiated cells. Sequencing studies of TET knockdown tissue or cell lines have shown that the different TET enzymes have both overlapping and distinct targets in the genome. Further, both TET1 and TET3 also have truncated isoforms with unique expression profiles and potentially unique biological functions (Melamed et al., 2018).

Some of the different biological functions of the different mammalian homologs and isoforms could be due to the presence or absence of a Cys-X-X-Cys (CXXC) domain, historically known for binding unmethylated CpGs, particularly in CGIs. It is thought that this domain targets

TETs to CGIs in order to assist in the maintenance of the unmethylated state, by oxidizing, excising, and repairing any mC that may erroneously arise; however, there is no direct evidence to support this. The full-length isoform of TET2 lacks a CXXC domain while TET1 and TET3 both have CXXC domains in their non-catalytic, N-terminal domains. Interestingly, while the CXXC domain of TET3 binds to unmethylated CpGs, it binds most strongly to caCpGs (Jin et al., 2016). TET2, which has been identified only as a single isoform and does not have a CXXC domain, has been shown to interact with the CXXC-containing protein IDAX (Ko et al., 2013). Similarly, the truncated isoforms of TET1 and TET3 all lack CXXC domains and have been shown to utilize other protein interactions for altered DNA targeting (Melamed et al., 2018). Although TET may have an intrinsic targeting mechanism via the CXXC domain, we know that mC oxidation occurs in regions other than CGIs, so there are likely many other factors contributing to its localization.

1.3.4: Structural and biochemical evidence of specificities regulating TET activity

TET enzymes, Fe(II)/α-ketoglutarate (α-KG)-dependent dioxygenases, are generally accepted to act via generation of a reactive Fe(IV)-oxo intermediate, driving proton abstraction and subsequent rearrangements that result in oxidation of the 5-position of cytosine and regeneration of Fe(II) (Figure 1-3) (Lu, Zhao & He, Chem Rev 2015). The original crystal structure was solved using a variant of the catalytic domain of TET2 in complex with mC-containing duplex DNA (Hu, L. et al., 2013). At the same time, a similar structure of NgTET1 was also solved (Hashimoto, Pais et al., 2014). In both structures, TET forms a compact globular structure in which a double-stranded beta helix (DSBH) core is surrounded by a Cys-rich domain on both sides. The DNA sits above the DSBH core, packing into a shallow groove formed by two loops from a Cys-subdomain, one of which forms a hydrophobic ridge and inserts into the minor groove. The target mC is flipped out of the DNA helix into a cavity of TET, in proximity to the

necessary cofactors. In agreement with these structures, it has been shown that only the C-terminal catalytic domain of of the enzyme is necessary for oxidation both *in vitro* and *in vivo* (Hu, L. et al., 2013).

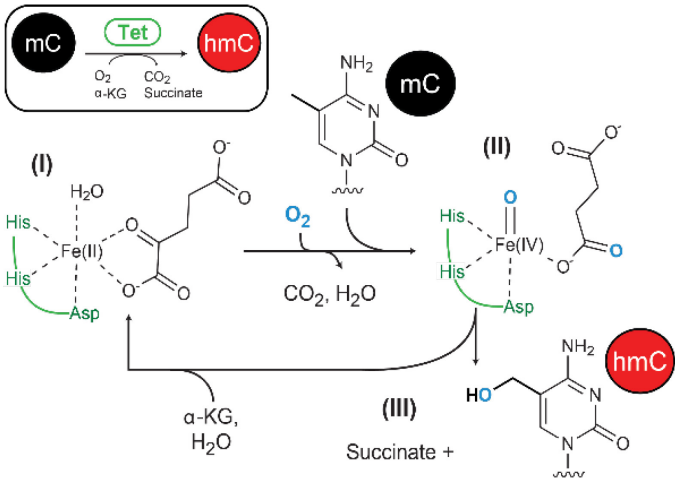


Figure 1-4. Scheme of proposed catalytic reaction steps of TET enzymes.

(I) TET harnesses electrons from Fe(II) and coordinated α-ketoglutarate (αKG) to split molecular oxygen (O2), while αKG undergoes oxidative decarboxylation resulting in succinate and a Fe(IV)-oxo intermediate (II) This intermediate activates the target C-H bond of mC for oxidation, resulting in hmC and regenerated Fe(II) (III) (Lu et al., 2015).

In the active site of the structure with human TET2, the 5-methyl group, which has no direct contacts with TET, is positioned towards N-oxalylglycine (NOG, an inactive analog of αKG) and Fe(II), primed for catalysis (Hu, L. et al., 2013). Importantly, there is steric availability surrounding the 5-methyl group to allow for larger chemical groups[1]. Likely responsible for both the specific recognition of cytosine and positioning the base for catalysis, the endocyclic nitrogen atoms N3 and N4 of the mC base form two hydrogen bonds with residues H1904 and N1387,

[1] Our lab exploited the observed space surrounding the methyl group to develop activity-based probes in Ghanty *et al.*, JACS 2018.

respectively. Also, TET residue Y1902 forms a base-stacking interaction with mC. Previous work from our lab identified that Y1902 along with T1372 comprise an active site scaffold that is essential for catalysis (Liu, M. Y. et al., 2017). We also found that mutating T1372 to E/Q/N/D/V effectively halted oxidation at hmC, while having no effect DNA-binding affinity.

Two years after the initial crystal structures, the same groups solved the structures of either TET2-CS or NgTET1 in complex with hmC- and fC-containing DNA (Hashimoto et al., 2015; Hu, L. et al., 2015). In general, the structures were very similar to the one containing mC and did not provide an immediate explanation for the observed difference in TET activity on mC compared to hmC. As a result of applying molecular dynamic simulations to the three structures containing each of the oxmCs, He *et al.* concluded that the rate-limiting hydrogen abstraction step in the TET2-containing complex results in slower decay of the ferryl-oxo intermediate with hmC and fC compared to mC (Hu, L. et al., 2015). In line with this observation, all TET enzymes to date exhibit a catalytic preference for mC over hmC and hmC over fC (Hashimoto et al., 2015; Ito, S. et al., 2011; Pfaffeneder et al., 2011). Despite having similar binding affinities, hTET2 displayed a ~2-5-fold difference in k_{cat} and K_m for mC over the other substrate bases (Hu, L. et al., 2015).

In addition to the relative preference for the three substrate bases, an important initial question in the field was whether the TET enzymes are capable of iterative oxidation, or catalytic processivity—that is, whether TET is capable of generating fC and caC from a mC substrate without fully releasing the hmC intermediate. The crystal structure of TET2 indicates that some conformational change must occur after catalysis to allow for exchange of fresh cofactors, but that does not exclude the possibility that TET remains either specifically or non-specifically bound

to DNA (Crawford et al., 2016). To test whether TET can act iteratively, our lab performed highly sensitive isotope-based experiments and found that mouse TET2 can generate fC and caC from a single encounter with mC-containing DNA, without releasing hmC prior to doing so (Crawford et al., 2016). In contrast, another group has proposed that TET is distributive, meaning it randomly catalyzes individual turnover events. Further, they also called our experimental approach into question (Tamanaha et al., 2016), but we would also argue that their conditions favored distributive behavior. Nevertheless, iterative oxidation provides a mechanism for the efficient production of highly oxidized fC and caC bases, which could encode distinct epigenetic functions. However, it remains possible that not at all fC and caC are generated in this manner.

Several other possible determining factors of TET's substrate preferences, such as nucleic acid identity, substrate secondary structure, and sequence context, are not readily clear from the crystal structure data alone. First, the majority of TET contacts with the target strand other than those with the cytosine base are with the phosphodiester backbone and there appears to be sufficient room surrounding the 2'-position of the sugar to accommodate a hydroxyl group (Hu, L. et al., 2013). Thus, despite a relatively greater abundance of oxmCs in DNA compared to RNA, the structural data supports relative promiscuity for different nucleic acids.

Next, if we examine the contacts that TET makes with the opposite strand of DNA in the crystal structure, we observe that TET2 residue Y1294 is inserted into where the guanine of the target mC•G base pair would occur in a conventional B-form helix (Hu, L. et al., 2013). The authors of the study proposed that the Y1294 and M1293 push out this guanine, breaking its base-stacking interactions, allowing mC to break its Watson-Crick base pair and flip out of the DNA helix and into the active site of TET. However, in Chapters 2-4, we provide evidence that

TET is biochemically active on single-stranded DNA, suggesting that this displacement of guanine is not essential for the engagement of the target base, although it is possible that it provides a stabilizing effect. Nevertheless, a pathological mutation of one of the residues that forms this hydrophobic ridge, W1291R, has been shown to result in a loss of activity *in vitro*, reaffirming the importance of the positioning of this loop.

Finally, Y1294 has also been implicated as critical for CpG recognition, as it forms a base-stacking interaction with the opposite strand mC, which base pairs with the +1 guanine (Hu, L. et al., 2013). To test this, the authors assessed the activity of a double-mutant TET2 variant with Y1294A and M1293A and observed significantly diminished catalytic activity *in vitro*. Notably, however, the authors do not perform biochemical analysis on CpG or any of the CpH contexts with this TET2 variant or that with the single Y1294A mutation.

1.3.5 Intrinsic ability of TET to catalyze clusters of oxmCs

In addition to the efforts to catalog and expose the substrate preferences intrinsic to TET and the underlying mechanisms regulating this specificity, there is a need within the field to understand TET activity in the presence of multiple target sites on a single DNA molecule. As mentioned previously, most transcriptionally activating events of TET activity occur from the oxidation (or full demethylation pathway) of multiple CpGs. Thus, in Chapter 4, we investigate whether TET has the propensity to catalyze multiple reactions in clusters. One way in which this could occur is if TET were intrinsically strand processive; that is, if TET is capable of catalyzing sequential reactions on a single strand of DNA without releasing the substrate. In the crystal structure of hTET2, there are several sequence-non-specific contacts between positively charged and polar residues of TET and the phosphodiester backbone of the target strand of DNA that

support the notion that TET may be capable of loosely maintaining these interactions and sliding along DNA (Hu, L. et al., 2013), exhibiting strand processivity between potential target sites. Likely due to the challenging nature of addressing this question, only one group has asked whether TET enzymes can act processively. Although they proposed that several TET homologs do not exhibit strand processivity, we do not believe their experimental design was set up to adequately to test this (Tamanaha et al., 2016). Thus, in Chapter 4, we also set out to tackle the question of whether TET can act in a processive manner on a strand of DNA.

1.4: Thesis Objectives

The reductionist power of biochemistry allows us to remove the multitude of factors involved in the full pathways of epigenetic regulation. This dissertation aims to exploit a pared-down environment to ask solely about the influence of the intrinsic properties of TET on its observed activities. In Chapter 2, we combine biochemistry with molecular modeling to define the substrate preferences of human TET enzymes for DNA, RNA, and DNA:RNA hybrids. In addition, we identify the strongest determinant for these relative preferences and provide a possible structural and thermodynamic mechanistic explanation. In Chapter 3, we performed rigorous biochemistry with both TET and TDG on substrates containing each of the three possible substrate bases (mC, hmC, fC) in either a CpG or CpH sequence context. Our findings allow us to propose a model for the biological observation of why there is not a proportionate amount of hmCpH in cells with higher levels of mCpH. In Chapter 4, we use BS-seq to localize TET-generated higher order oxmCs (hoxmCs, fC and caC) on DNA that initially had all mCpGs to quantify TET enzymes' ability to modify DNA in clusters. Further, we test whether intrinsic processivity of the enzyme is a possible mechanism for this activity. Overall, this dissertation

contributes to our understanding of the biochemical basis for the patterns of oxmCs found in the genome.

CHAPTER 2: Selectivity and Promiscuity in TET-mediated oxidation of 5-methylcytosine in DNA and RNA

This chapter has been adapted from the following manuscript:

DeNizio, J.E.[2], Liu, M.Y.[2], Leddin, E.M., Cisneros, G.A., Kohli, R.M. Selectivity and Promiscuity in TET-Mediated Oxidation of 5-Methylcytosine. Biochemistry 2019, 58, 411-421.

2.1: Abstract

Enzymes of the ten-eleven translocation (TET) family add diversity to the repertoire of nucleobase modifications by catalyzing the oxidation of 5-methylcytosine (5mC). TET enzymes were initially found to oxidize 5-methyl-2'-deoxycytidine in genomic DNA, yielding products that contribute to epigenetic regulation in mammalian cells, but have since been found to also oxidize 5-methylcytidine in RNA. Considering the different configurations of single- and double-stranded DNA and RNA that co-exist in a cell, defining the scope of TET's preferred activity and the mechanisms of substrate selectivity is critical to better understand the enzymes' biological functions. To this end, we have systematically examined the activity of human TET2 on DNA, RNA, and hybrid substrates *in vitro*. We found that, while ssDNA and ssRNA are well tolerated, TET2 is most proficient at dsDNA oxidation and discriminates strongly against dsRNA. Chimeric and hybrid substrates containing mixed DNA and RNA character helped reveal two main features by which the enzyme discriminates between substrates. First, the identity of the target nucleotide alone is the strongest reactivity determinant, with a preference for 5-methyldeoxycytidine, while both DNA or RNA are relatively tolerated on the rest of the target strand. Second, while a complementary strand is not required for activity, DNA is the preferred partner, and complementary RNA diminishes reactivity. Our biochemical analysis, complemented by molecular dynamics simulations, provides support for an active site optimally configured for dsDNA reactivity but permissive for various nucleic acid configurations, suggesting a broad range of plausible roles for TET-mediated 5mC oxidation in cells.

[2] These authors contributed equally to this work.

2.2: Introduction

Long thought to be the only significant DNA modification in mammalian genomes, 5-methyl-2'-deoxycytidine (5mdC) is now known to be a substrate for further oxidation. Oxidative modifications to 5mdC, largely in the context of cytosine guanine dinucleotides (CpGs), are catalyzed by enzymes of the ten-eleven translocation (TET) family, which are Fe(II) and α-ketoglutarate (α-KG) dependent dioxygenases. Step-wise oxidation of 5mdC by TET enzymes can yield 5-hydroxymethyl-2'-deoxycytidine (5hmdC), 5-formyl-2'-deoxycytidine (5fdC), and 5-carboxyl-2'-deoxycytidine (5cadC) (He, Y. F. et al., 2011; Ito, S. et al., 2011; Tahiliani et al., 2009). These DNA modifications can function as potential intermediates in the long-sought pathway for erasure of 5mdC, referred to as demethylation, and can also serve independent functions regulating gene expression (Liu, M. Y., DeNizio, Schutsky et al., 2016). For example, 5hmdC is thought to activate loci silenced by methylation (Mellen et al., 2017), and intragenic 5hmdC in DNA may also impact RNA splicing (Marina et al., 2016). The distinctive localization patterns of 5hmdC, 5fdC and 5cadC in sequencing studies further fuels speculation about additional independent roles in epigenetic regulation (Wu, H. and Zhang, 2015).

Just as 2'-deoxycytidine (dC) modifications are widely prevalent in DNA, modifications to the 5-position in cytidine (rC) in RNA are also found across all domains of life (Motorin et al., 2010; Squires et al., 2012). 5-methylcytidine (5mrC) and its oxidized analogs (5hmrC, 5frC, and 5carC) are detectable in many types of RNA, including mRNA, tRNA, rRNA, and ncRNA (Huang, W. et al., 2016; Motorin et al., 2010; Squires et al., 2012). Levels of 5mrC in total RNA vary across tissues and organisms but can approach levels of 5mdC in DNA; for example, in mouse brain, 5mrC comprises approximately 1% of total cytosines in RNA (Fu et al., 2014; Huber et al., 2015), while 5mdC is 2-5% of cytosines in DNA (Globisch et al., 2010; Wagner et al., 2015). By contrast, hydroxymethylation is approximately 2-3 orders of magnitude lower in RNA than in DNA (Globisch et al., 2010; Wagner et al., 2015). Unlike for 5mdC in DNA, the function of 5mrC in RNA is not well established, although potential roles in nuclear export of mRNA have been demonstrated (Yang et al., 2017). In Drosophila, 5hmrC was found to be enriched in the coding regions and may favor translation of mRNA transcripts, as more ribosomes were bound to mRNA

containing 5hmrC (Delatte et al., 2016). With regards to demethylation, pathways for 5mrC "removal" have not been found; however, the loss of 5hmrC has been shown to increase 5mrC levels, suggesting the possibility of functional parallels (Shen, Q. et al., 2018).

Although DNA and RNA have unique methyltransferases involved in methylation (Schapira, 2016), TET enzymes appear to be involved in oxidation of both DNA and RNA. In mammalian systems, all three TET enzymes (TET1, UniProtKB Q8NFU7; TET2, UniProtKB Q6N021; TET3, UniProtKB O43151) exhibited activity on 5mrC *in vitro* and in transfected cells (Fu et al., 2014). TET1 and TET2 were also detected in an unbiased screen for RNA-binding nuclear proteins in mouse embryonic stem cells (He, C. et al., 2016). Further support for TET oxidation in both DNA and RNA comes from studies demonstrating that activity on both substrates can be inhibited by the same means: Mutations in isocitrate dehydrogenase (IDH) which result in accumulation of the competitive inhibitor 2-hydroxyglutarate, previously known to impede DNA oxidation, have also been shown to reduce RNA oxidation (Xu, Q. et al., 2016). Phylogenetic analysis also highlights a broader role for TET enzymes in acting on different nucleic acids (Iyer et al., 2009; Iyer et al., 2013). Drosophila again provide a compelling example in this regard; they lack the canonical 5mdC substrate in their DNA, but have a TET enzyme homologue that appears to act on RNA as a bona fide substrate (Delatte et al., 2016; Raddatz et al., 2013).

The observation of both DNA and RNA oxidation, as well as the diversity of potential functions of 5mdC and 5mrC oxidation products, makes it a priority to understand how the nature of the nucleic acid impacts TET activity. No systematic comparison of TET activity on various configurations of DNA versus RNA has yet been performed, but previous biochemical experiments and crystal structures provide a strong starting point for speculating about the mechanistic features involved in DNA versus RNA oxidation (Hu, L. et al., 2013; Hu, L. et al., 2015). In the structure of a truncated active human TET2 variant (TET2-CS) bound to a 12-mer double-stranded (ds) DNA substrate, one of the most striking features is that 5mdC on the target strand is flipped out of the duplex and into the active site of TET, raising questions about the conformational flexibility required for dC- versus rC-based substrates. Outside of the target

nucleotide, key contacts on the target DNA strand are largely confined to the phosphate backbone, suggesting the possibility of promiscuity towards nucleic acid identity. Regarding the impact and necessity of a complement strand, the structure and biochemistry present conflicting data. Mutations of TET2-CS that could disrupt contacts with the complementary strand decrease activity on dsDNA, but the effect on single-stranded (ss) DNA has not been tested in parallel(Hu, L. et al., 2013). One study showed that dsDNA is more reactive than ssDNA (Fu et al., 2014), while another suggested that ssDNA is more reactive (Kizaki, S. and Sugiyama, 2014). The relative activities of TET enzymes on various DNA and RNA configurations are even less well-established. Early evidence suggested that the catalytic domain of mouse TET1 greatly prefers dsDNA to a sequence-matched ssRNA substrate. TET-mediated oxidation to 5frC and 5carC has also been demonstrated and could potentially occur in dsRNA, though oxidation to these forms appeared to occur at very low efficiency (Basanta-Sanchez et al., 2017; Fu et al., 2014).

To better understand the breadth of TET enzyme capabilities, here we have performed a comprehensive examination of the impact of nucleic acid identity on TET activity, by systematic variation of the target base, target strand, and complementary strand of the substrates. Our results show a dominant role for the target nucleotide in dictating the efficiency of oxidation. Outside of this preference, TET2 has a surprising tolerance for most substrate configurations, except for dsRNA, arguing that TET enzyme promiscuity makes these enzymes suited to diverse biological roles. By placing our biochemical analysis in the context of targeted molecular dynamics simulations, we further identify structural features that support this promiscuity as well as potential mechanisms that can explain the observed spectrum of activity on DNA versus RNA substrates.

2.3: Results

2.3.1: Design of Substrate Series

To probe TET's capacity for reacting with various configurations of deoxyribo- versus ribonucleic acids, we designed oligonucleotides that permitted systematic analysis of substrate requirements for TET activity. Two different substrate series were generated for these

experiments, each containing a 16-bp all-DNA oligonucleotide, as well as a matched all-RNA oligonucleotide, which contain uridine in place of thymidine. These substrates contain a single 5-methylcytosine (5mC) in a CpG sequence context: 5mdC in the all-DNA substrate (designated as D(M)) or 5mrC in the all-RNA substrate (designated as r(m)) (Figure 2-1A). Given the potential relevance of the sugar identity of the target base, which is extruded into the active site during oxidation, we also designed chimeric substrates that could distinguish between the impact of the identity of the target 5mC and that of other nucleotides on the target strand. In the DNA substrate, 5mdC was replaced with 5mrC (yielding D(m)), and vice versa for the RNA substrate (yielding r(M)). Thus, Series 1 and 2 each contain four total oligonucleotides: D(M), r(m), D(m), and r(M).

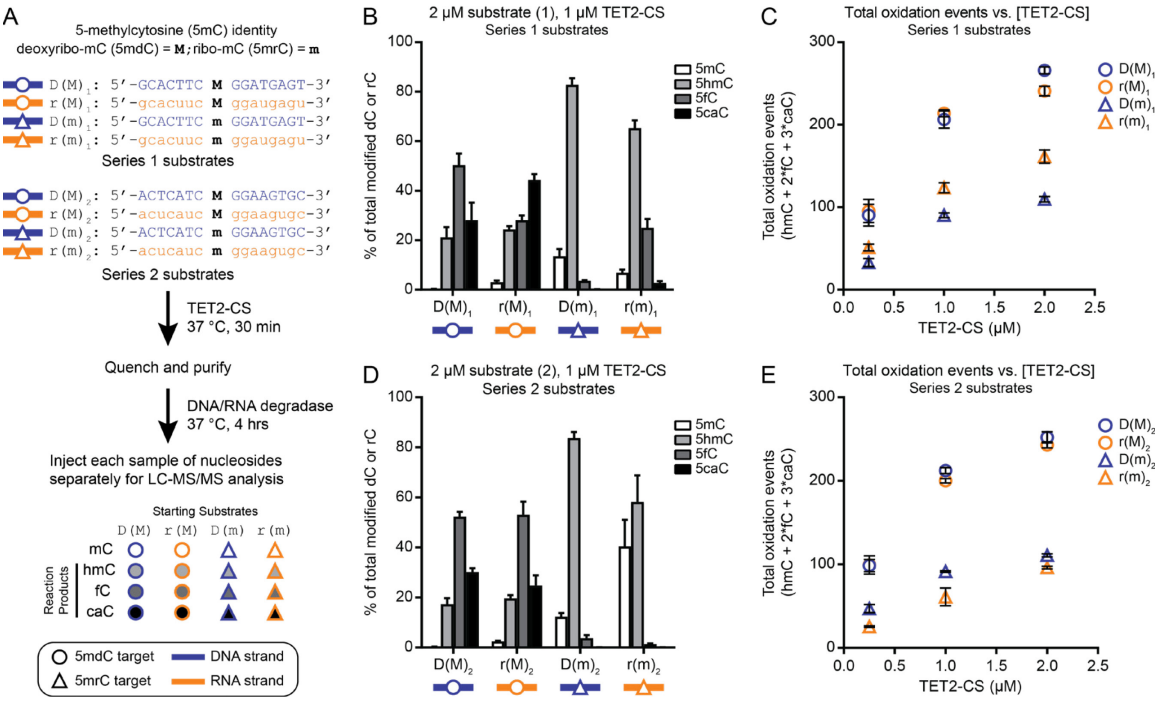


Figure 2-1. TET2 discriminates mostly based on the sugar identity of the target nucleotide rather than that of the flanking target strand.

A) Workflow showing the substrates and procedures used in the experiments. **B)** Activity of TET2-CS with the Series 1 substrates. Reaction products were quantified by LC-MS/MS and expressed as the percentage of each base relative to the amount of total modified cytosine (DNA or RNA). **C)** The reaction in (B) was repeated at three enzyme concentrations, and the reaction products were expressed as total oxidation events: the generation of 5hmC counts once, 5fC counts twice, and 5caC counts three times. **D-E)** The same experiments using Series 2 substrates. Data for individual oxidation products used to generate (C and E) can be found in Figure S2-2A,B. All error bars represent standard deviation for 3-5 independent replicates.

Notably, we designed Series 1 and 2 to be complementary to one another. For Series 2, we generated two additional oligonucleotides that contained unmodified cytosine, rather than 5mC, in the all-DNA or all-RNA context (D(C) or r(c)). Comparing Series 1 substrates in the absence or presence of either D(C) or r(C) complementary strands could therefore permit us to examine single-stranded versus double-stranded nucleic acids. Furthermore, complexing Series 1 and 2 substrates that both contain reactive 5mC sites could permit an analysis of activity on different strands in a duplex.

2.3.2: TET activity on 5mdC versus 5mrC

TET's activity was first analyzed on the simplest iteration of the series, the all single-stranded substrates. *In vitro* reactions were performed using recombinant TET2-CS (Hu, L. et al., 2013), and the reacted substrates were purified, digested to nucleosides and analyzed by LC-MS/MS (Figure 2-1A, Figure S2-1). We first examined the product distribution at a 2:1 ratio of substrate to enzyme. By displaying the data as the relative percentages of each modified cytosine, the amount of remaining 5mC reflects the substrate consumption, while the flux through the oxidative pathway is given by the relative distribution of the three forms of TET-oxidized products. Under these conditions with Series 1, the DNA substrate ($D(M)_1$) is entirely reacted, while nearly all the RNA substrate ($r(m)_1$) is converted (7% 5mrC remaining), indicating that the RNA is a relatively proficient substrate. The DNA and RNA substrate reactions do differ, however, in the *extent* to which they are oxidized: Roughly 80% of the DNA substrate is converted to the higher oxidized modifications (5fdC and 5cadC), while the RNA substrate is primarily converted to the first oxidative product, 5hmrC (65%) (Figure 2-1B). With the Series 2 substrates, the general trends all hold, with the exception that TET less proficiently oxidizes $r(m)_2$, with approximately 40% of the 5mrC substrate unreacted, relative to $D(M)_2$ where 5mdC is completely consumed (Figure 2-1D).

Given that driving conditions could obscure the comparison between relative reactivities of the all-DNA and all-RNA substrates, we also examined reactivity at an 8:1 ratio of substrate:enzyme. Under these conditions, there remains 2.6 and 5.2-fold as much RNA as DNA

substrate for Series 1 and Series 2, respectively (Figure S2-2A,B). To see how far we could push the oxidation of RNA, we also analyzed the reactivity at a 1:1 ratio of substrate to enzyme. For both Series 1 and 2, while roughly 60% of 5mdC is oxidized to 5cadC, there is very little 5carC generated from 5mrC (12% for Series 1 and none detectable for Series 2) despite near complete consumption of 5mrC (Figure S2-2A,B). Interestingly, while only 12% of 5mrC remains unreacted for $r(m)_2$, it is almost exclusively converted to 5hmrC (81%), suggesting a greater barrier to oxidation from 5hmC to 5fC for RNA than for DNA.

To facilitate a more rigorous comparison of the overall reactivity of substrates, we also analyzed the data for total oxidation events for each substrate. Considering that 5fC and 5caC both represent multiple catalytic events, total oxidation can be defined as the sum of the turnover events, namely: (fraction of 5hmC) + 2×(fraction of 5fC) + 3×(fraction of 5caC) (Crawford et al., 2016). Analyzing the data this way, the preference for ssDNA over ssRNA is evident and maintained across different conditions. Specifically, we observed an average of 1.7 and 3.3-fold more turnovers, for Series 1 and Series 2, respectively, with D(M) substrates relative to r(m) substrates across the different concentration ratios (Figure 2-1C,E).

We focused our initial analysis on TET2-CS given that the variant is well-expressed and more active than full catalytic domain variants of TET2 (TET2-CD). To test the generality of the preference for ssDNA over ssRNA across the mammalian TET enzymes, we also examined the reactivity of $D(M)_1$ and $r(m)_1$ with TET1-CD, TET2-CD and TET3-CD. Consistent with prior studies, all three TET enzymes are capable of oxidizing both the ssDNA and ssRNA substrates, and a consistent preference for oxidation of 5mdC in $D(M)_1$ over 5mrC in $r(m)_1$ was observed. The proportion of reacted 5mdC, *i.e.* sum of the percentages of 5hmC, 5fC, and 5caC, in $D(M)_1$ is 6.6-fold that of 5mrC in $r(m)_1$ for TET1-CD, 1.5-fold for TET2-CD, and 5-fold for TET3-CD (Figure S2-2C). Thus, activity can reliably be detected with both DNA and RNA, across two substrate series, a range of substrate:enzyme ratios, and different TET variants, and ssDNA appears to be consistently preferred over ssRNA for human TET enzymes.

To further explore the mechanistic basis for discrimination between ssDNA and ssRNA, we next focused on the nature of the target nucleotide relative to the flanking residues of the

substrates. After reacting TET2-CS with $D(m)_1$, which differs from $D(M)_1$ only by the addition of a single 2'-hydroxyl at the target base, we quantified 5mrC and the 5hmrC, 5frC, and 5carC oxidation products. With this substrate, the reactivity decreases significantly, with a turnover that averaged 2.4-fold less across the substrate:enzyme ratios examined relative to $D(M)_1$, and is more comparable to the all-RNA substrate, $r(m)_1$ (Figure 2-1C). The reciprocal change in the $r(M)_1$ substrate, which differs from $r(m)_1$ only by the removal of a single 2'-hydroxyl at the target base, shows a rescue of activity, with $r(M)_1$ showing largely comparable activity to the all-DNA substrate, $D(M)_1$. Like the Series 1 substrates, the Series 2 substrates display a similar pattern, where the distribution of oxidized products from $r(M)_2$ more closely resembles that of $D(M)_2$, while that of $D(m)_2$ is more similar to $r(m)_2$ (Figure 2-1D). Thus, analysis of these chimeric substrate oligonucleotides indicates that the difference in activity between ssDNA and ssRNA is largely attributable to the target base identity, rather than the DNA or RNA character of the surrounding bases.

2.3.3: Impact of complement strand on reactivity

We next sought to evaluate the impact of a complement strand and its nucleic acid identity on TET activity. The Series 1 substrates, including the all-DNA strand, all-RNA strand, and the two chimeric oligonucleotides, were each annealed to an unmethylated complement, either the all-DNA strand, $D(C)_2$, or the all-RNA strand, $r(c)_2$. Given the wide range of substrate reactivity observed across these permutations, the substrates were compared at a 1:1 substrate:enzyme ratio using TET2-CS. Under these conditions, the dsDNA ($D(M)_1$:$D(C)_2$) substrate reacts to near completion, with almost complete conversion of 5mdC to 5cadC (Figure S2-3A). Similarly, single-stranded $D(M)_1$ shows complete consumption of 5mdC, and total oxidation events are 92% of those observed with dsDNA (Figure 2-2A), as oxidation to 5cadC is incomplete (Figure S2-3A). This small decrease in activity on ssDNA compared to dsDNA is also be observed at a lower substrate:enzyme ratio where the total oxidation events of ssDNA are also 92% of dsDNA (Figure S2-3B), even though the oxidized products of both the ss- and dsDNA substrates are predominantly 5hmdC and 5fdC (Figure S2-3B). Notably, with the full catalytic

domains, TET2-CD shows a similar small reactivity preference for dsDNA over ssDNA, while TET1-CD and TET3-CD are essentially agnostic to the presence or absence of the complementary strand (Figure S2-4). As prior studies have hinted at a length dependence to substrate reactivity (Hu, L. et al., 2015; Kizaki, S. and Sugiyama, 2014), we generated two additional 5mdC-containing substrates containing the embedded $D(M)_1$ 16-mer sequence in a 27-mer or a 60-mer oligonucleotide and compared ssDNA versus dsDNA reactivity. Although reactivity of these substrates does decrease with length, the observation that dsDNA is more reactive than ssDNA is maintained across all three substrates in the series (Figure 2-2B, Figure S2-5). The observation that ssDNA reactivity is slightly enhanced by a DNA complement also carries forward to the chimeric substrate $D(m)_1$, though notably the target RNA base in $D(m)_1$ has the greater effect of lowering overall activity on both single and double stranded substrates (Figure 2-2A).

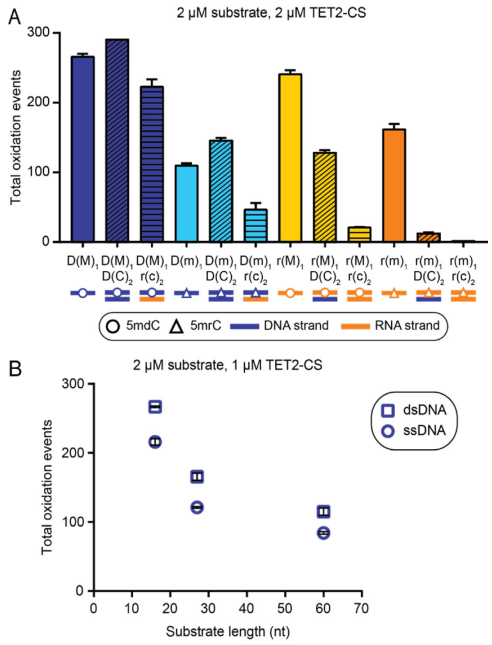


Figure 2-2. Double-stranded DNA is the preferred TET substrate, while dsRNA is strongly disfavored.

A) Activity of TET2-CS on Series 1 substrates, either single-stranded or duplexed to a non-reactive DNA or RNA complementary strand. **B)** Comparison of TET2-CS activity on sequence-matched ssDNA versus dsDNA substrates of different lengths. Error bars represent standard deviation for 2-3 independent replicates.

In contrast to the improved reactivity when duplexing DNA with DNA, duplexing any target strand with RNA dampens reactivity. TET2-CS showed reactivity on the 5mdC of the DNA:RNA hybrid duplex ($D(M)_1$:$r(c)_2$), but total oxidation events are decreased relative to the dsDNA substrate $D(M)_1$:$D(C)_1$ (23% reduction); in comparison, $D(m)_1$ experiences an even greater reduction in reactivity (68%) when complexed to RNA (Figure 2-2A). Interestingly, for the chimeric $r(M)_1$ substrate, whose reactivity rivaled the all-DNA $D(M)_1$ substrate when single-stranded, reactivity decreases when duplexing this substrate to the DNA complement and decreases further when duplexing to the RNA complement. The disfavored nature of an RNA complementary strand is most evident with the all-RNA substrate, $r(m)_1$, where reactivity is nearly eliminated on the dsRNA substrate, contrary to prior suggestions that dsRNA may be a substrate (Basanta-Sanchez et al., 2017). Thus, although the target base identity is a critical determinant of reactivity with single-stranded substrates, these preferences can be overridden by disfavored duplex structures in the rest of the substrate.

2.3.4: Activity on symmetrically methylated duplexes

Our analysis of duplex substrates initially focused on a single reactive 5mC site in the target strand, but we recognized that these substrates are well suited to examining a more complex question. Considering that most CG:CG dinucleotide pairs in genomic DNA are symmetrically methylated (Schubeler, 2015), TET enzymes likely encounter duplex substrates where both strands contain reactive 5mC nucleobases. The DNA:RNA hybrids, $D(M)_1$:$r(m)_2$ and $D(M)_2$:$r(m)_1$, as well as the chimera hybrids, $D(m)_1$:$r(M)_2$ and $D(m)_2$:$r(M)_1$, are instructive experimental substrates, as in each case, activity on each strand can be assessed independently, given that one strand contains 5mdC, while the other contains 5mrC. With $D(M)_1$:$r(m)_2$ and $r(m)_1$:$D(M)_2$, we observed nearly symmetrically-reciprocal results, with high activity on the all-DNA strand and less activity on the all-RNA strand in both cases (Figure 2-3A). Notably, comparing these two DNA:RNA hybrids with the dsDNA duplex (5mC on both strands) shows a greater proportion of the 5mdC converted to 5cadC for the all-DNA substrate (Figure 2-3B). This result suggests that having an RNA complement strand, whether it is reactive or non-reactive as above,

decreases activity on the DNA strand. With the chimeric substrates, $D(m)_1$:$R(M)_2$ and $r(M)_1$:$D(m)_2$, the RNA strand containing 5mdC is more reactive than the DNA strand containing 5mrC (Figure 2-3). This observation is consistent with a dominant influence for the identity of the target nucleotide over that of the flanking nucleic acids in the target strand.

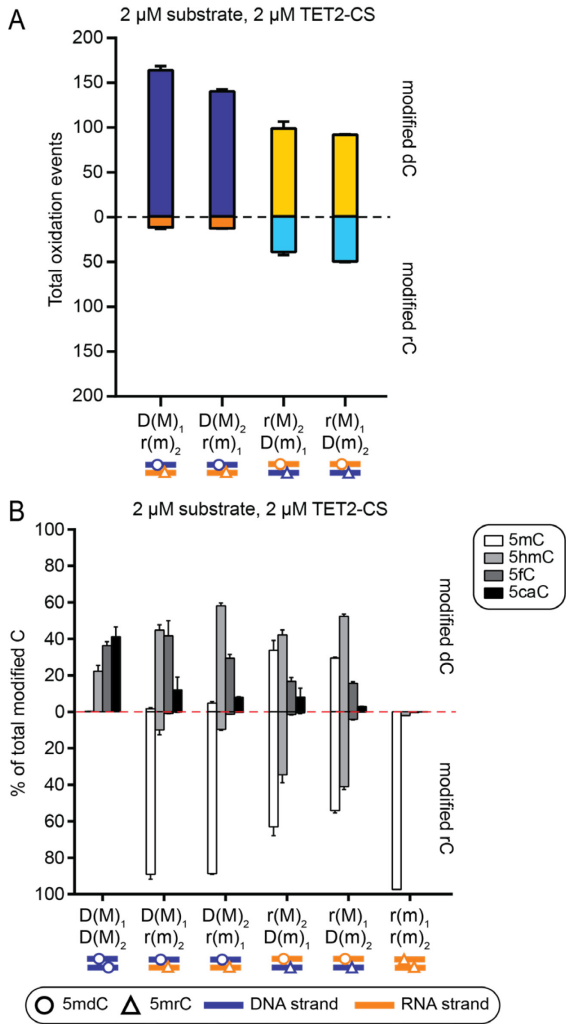


Figure 2-3. For duplexes with two reactive strands, TET2 activity on each strand is largely independent of the other, complementary strand.

A) Series 1 and 2 substrates, each containing 5mC, were duplexed to one another. Reaction products are shown as the total oxidation events on each strand. **B)** Reaction products from the same reactions are also shown as percentages of each modification for DNA (top half of graph) or RNA (bottom half of graph) targets. Also included for comparison are results for symmetrically methylated dsDNA and dsRNA, showing the reaction products for both strands combined. Error bars represent the standard deviation for 3 independent replicates.

2.3.5: Modeling of 5mrC in TET-dsDNA structure.

We have used classical molecular dynamics (MD) simulations to examine the TET2-CS structure complexed to dsDNA (Liu, M. Y. et al., 2017) in order to understand the mechanisms that could contribute to nucleic acid selectivity. Given the importance of the target site identity revealed by our biochemical analysis, we converted 5mdC to 5mrC by inserting a single 2'-hydroxyl at the target 5mdC base and then examined its impact on non-bonded interactions and H-bond dynamics. The MD results provide a qualitative assessment that help explain the tolerance of TET2 toward 5mrC at the target site, as well as the preference for oxidation of 5mdC over 5mrC.

The mutation of 5mdC to 5mrC results in both global and local changes in the structure and dynamics of the system that help explain the decreased reactivity of 5mrC (Figure 2-4, S2-6-8). At a global level, the inclusion of the 2'-OH results in an overall destabilizing interaction of the DNA with the protein ($\Delta E_{5mdC\ system\ -\ 5mrC\ system}$=-47.4 kcal/mol) (Figure 2-4A). Additionally, the normal mode analysis of both systems suggests that the dynamics are also affected by the inclusion of the 2'-OH (Figure S2-6A,B). Interestingly, the substitution of 5mdC with 5mrC in the target strand creates a partial shift from B-form DNA, most distinctly shown by the change in shift and slide toward A-form character in two base pairs upstream (5') of the target base (Figure S2-7), as well as altered intermolecular interactions spanning the R1260-R1269 region of the enzyme (Figure S2-8). The A-form is common in RNA and DNA-RNA duplexes and in protein-DNA complexes for polymerases and endonucleases (Kulkarni and Mukherjee, 2017). These alterations could partially explain the decreased reactivity of the 5mrC chimeric substrates as well as the more profound defect in oxidation of dsRNA that was observed.

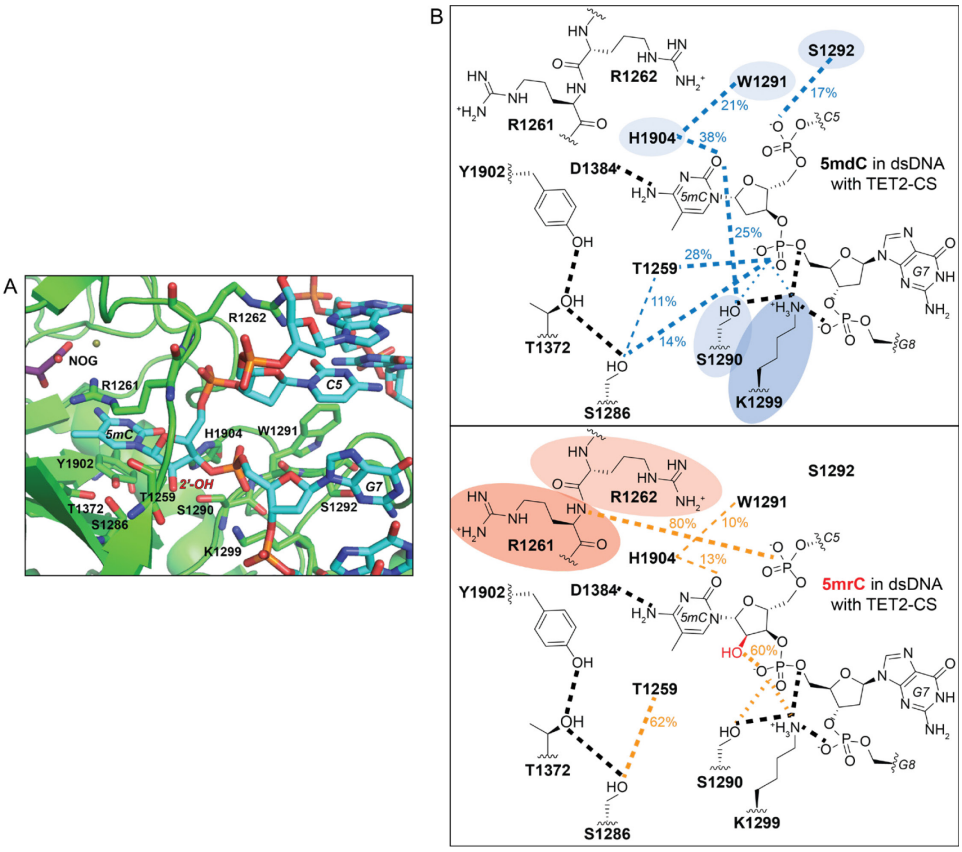


Figure 2-4. Modeling 5mrC as the target base in dsDNA bound to hTET2 results in dynamic and structural changes.

A) The active site of TET2-CS (green) bound to dsDNA (teal) containing 5mrC. **B)** H-bond network and energetic differences in models of the active site with either 5mdC (top) or 5mrC (bottom). H-bonds (present >10% simulation time in at least one system) are designated by dashed lines, including those that occur in both models (<10% difference, black). When there is more than one bond that contributes to an interaction between two molecules, it is shown as a dotted line and the simulation time present is not shown (see Figure S2-8 for all contributing data). For H-bonds that occur in both systems but at different levels ($\geq$10% difference), the bond line is thinner in the system in which it is less prevalent. Significantly different non-bonded interactions (sum of Coulombic and van der Waals energies $\geq$|1| kcal/mol) between a TET residue and the target 5mC in the two systems are highlighted according to the relative magnitude of the difference (see Table S2-1 for values).

A more local view of the active site sheds light on differences that can explain the decreased reactivity of 5mrC, but also the preservation of core elements that permit 5mrC to be oxidized. The 2'-OH of 5mrC does not introduce any evident steric conflicts but results in the formation of an H-bond with the phosphate backbone of G7, the downstream base in the CpG dyad (Figure 2-4, Figure S2-8, Table S2-1). This DNA structural change alters the interactions of

the substrate in the active site, including an increase in the predominance of the H-bonds between G7 and S1290 and K1299 with the simultaneous loss of H-bonds from T1259 and S1286 to G7. Residues R1261 and R1262, which form key interactions with the target base backbone (Figure S2-8) and with the αKG analog, respectively, are stabilized via additional Coulomb and van der Waals interactions in the 5mrC system. Nonetheless, despite these interactions with the backbone of the target strand, there are changes suggesting decreased engagement of the nucleobase to the enzyme: In addition to loss of an H-bond between S1290 and the nucleobase, the H-bond interactions on the Watson-Crick face of the nucleobase from H1904 are reduced by a third of the overall simulation time in the 5mrC-containing structure compared with the 5mdC system. Notably, other core components of the active site are largely intact. In particular, we have previously demonstrated the importance of an active site scaffold in modeling and mutagenesis experiments, whereby the active site T1372 hydrogen bonds with Y1902, which stacks with the target cytosine nucleobase to position it optimally for oxidation (Liu, M. Y. et al., 2017). The introduction of 5mrC does not have a noticeable impact on the intermolecular interactions of the scaffold.

2.4: Discussion

Given the expanded recognition of the role of oxidized 5mC bases in DNA and RNA biology, here we employed a systematic approach to explore the nucleic acid determinants of TET reactivity. Our results offer a view of the selectivity and promiscuity that characterizes TET-mediated oxidation and allow for substantiation of a model (Figure 2-5) that links reactivity to the nature of the target nucleotide, the flanking nucleotides on the target strand, and the complementary strand.

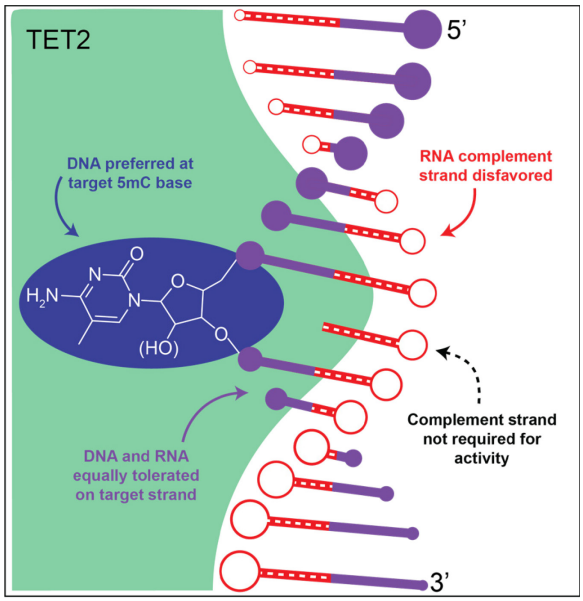


Figure 2-5. Summary of substrate determinants of TET reactivity from biochemical findings.

The 2'-OH present in 5mrC, but absent in 5mdC, is shown in parentheses. The complement strand is shown as a dashed line to indicate that it is not required for TET activity.

Our results uncover a strong association between the nucleic acid identity of the target nucleotide and TET reactivity. We find that TET enzymes are tolerant to both ssDNA and ssRNA substrates, but ssDNA substrates are preferred, consistent with past studies (Fu et al., 2014). Strikingly, the diminished reactivity of the ssRNA can be rescued by removing the single 2'-OH from the target nucleotide. Correspondingly, the addition of a single 2'-OH at the target nucleotide decreases the reactivity of ssDNA to the level observed with the ssRNA substrate. These results align with findings from an *in vitro* study showing that TET can oxidize substrates containing 2'-(R)-fluorinated modified cytosines, albeit to a lesser extent than 5mdC-containing DNA (Schroder et al., 2016). In support of our biochemical findings, molecular dynamic simulations suggest that the addition of a single 5mrC in the active site of a TET2-dsDNA structure can be accommodated, allowing for key nucleobase-orienting interactions to be maintained, but results in significant energetic and conformational changes surrounding the target base relative to 5mdC which could explain its altered reactivity. These findings are interesting to consider in the context of other DNA/RNA modifying enzymes. Some enzymes, such as the base excision repair enzyme uracil

DNA glycosylase, potently discriminate against a ribonucleotide target using steric discrimination against the 2'-OH (Pearl, 2000). For other enzymes, such as AID/APOBEC family DNA deaminase enzymes, a preference for a deoxynucleotide over ribonucleotide target appears linked to the preferred sugar puckers accessible to the DNA substrates (Nabel et al., 2013). The modeling with TET2-CS suggests an alternative mechanism that can explain tolerance for 5mrC, but a preference for 5mdC. The 5mrC substrate can be accommodated from a steric and conformational perspective, but the interactions of the reactive base appear disfavored relative to 5mdC.

Our biochemical data also inform our understanding of the role of the flanking and complementary strand nucleotides on activity. dsDNA is the preferred substrate, but DNA:RNA hybrids are also tolerated when the DNA is the substrate (Figure 2-5). The only conformation that appears to be strongly discriminated against is dsRNA. While a single 5mdC can alleviate some of the negative effects of the RNA duplex, it is still strongly disfavored, possibly due to the discrimination against A-form conformations. The proficiency for oxidation of ssRNA suggests that the conformational flexibility available to single-stranded substrates, but not dsRNA, facilitates promiscuous oxidation of 5mC. In considering the impact of the nucleotides outside of the target site, TET2 offers an interesting comparison to another nucleotide-flipping enzyme, the RNA methyltransferase, DNMT2. When methylation activity was examined on tRNA substrates, a chimera that contains a dC in place of rC in the loop region of the tRNA substrate was more, rather than less, efficiently methylated than the all-RNA natural analog (Kaiser et al., 2017). Therefore, rather than having selectivity primarily dictated by the target nucleotide as we observe for TET2, the flanking nucleotides and their secondary structure determines reactivity for DNMT2.

The evident promiscuity in TET family enzymes raises important points with regards to their physiological function. dsDNA, ssDNA, ssRNA, and DNA:RNA hybrids co-exist in mammalian cells and represent four of the five most preferred substrates across our tested series, which lends biochemical credence to the possibility that TET oxidation products could be generated in any of these settings. At the same time, the observed promiscuity *in vitro* raises questions about how TET-mediated oxidation could be targeted *in vivo*. For functions when

activity on DNA predominates, spatial regulation is likely to play an important role. TET1 and TET2 have been shown to primarily localize to the nucleus in both overexpression studies and in cells, while TET3 localizes to both the cytoplasm and the nucleus (Arioka et al., 2012; Di Stefano et al., 2014; Huang, Y. et al., 2016; Muller et al., 2012). Regulation could also involve the non-catalytic domains and/or partner proteins. Our studies focused on the crystal structure version of TET2 and the catalytic domains of TET1-3, but full-length TET1 and TET3 have CXXC-domains that can mediate targeting to non-methylated CpG islands in DNA, while TET2 is thought to utilize its ancestral gene neighbor, IDAX, for this purpose (Ko et al., 2013). For functions where activity on RNA is thought to predominate, major questions remain open regarding the location and timing of cytosine methylation in RNA and how selectivity for TET-mediated oxidation could be achieved.

More broadly, beyond the mammalian TET1-3 enzymes, it will be interesting to see if promiscuity towards DNA and ssRNA is a shared feature of the TET enzymes found across the phylogenetic tree, especially in organisms that offer a restricted or unusual selection of nucleic acid substrates (Iyer et al., 2009; Iyer et al., 2013). TET homologs in some species have been associated with transposons and retroelements, suggesting the possibility that their activity could be associated with less conventional nucleic acid intermediates associated with mobile genetic elements (Iyer et al., 2014). Among the species with distinct TET homologs, Drosophila stand out in particular since they lack methylated DNA yet have retained a TET homolog that appears to act on RNA. Whether the proposed impact on protein translation applies to other species, or whether this represents a unique branch point in the evolution of TET activity, remains to be explored. Our findings are also interesting to consider in the context of the broader nucleic acid dioxygenase family, which also includes AlkB family enzymes. The promiscuity of AlkB enzymes towards DNA and RNA has been thought broaden their function in nucleic acid repair (Falnes et al., 2004) and may have been exploited in the evolution of diverse physiological functions potentially associated with TET enzymes.

The observed promiscuity of TET activity adds to the challenges of studying DNA and RNA cytosine modifications. In knockouts, phenotypes attributed to activity on 5mdC in DNA, may

instead be related to altered oxidation of 5mrC in RNA or vice versa. Engineering of TET2-CS has been able to yield variants with altered selectivity (Liu, M. Y. et al., 2017; Sudhamalla et al., 2018). Our MD simulations suggest that it may similarly be possible to find mutants with altered selectivity by manipulating the interactions that differentiate 5mdC from 5mrC engagement in the active site. For example, mutations of key residues such as S1290 or K1299 in TET2 could potentially enhance or diminish the activity on DNA versus RNA. Indeed, beyond improving our understanding of the molecular determinants of enzyme selectivity, building enhanced specificity into TET enzymes could offer tools to better explore their biological roles as well as expand the power of CRISPR/dCas9-based epigenetic editing biotechnology (DeNizio et al., 2018; Komor et al., 2017).

2.5: Methods

2.5.1: Oligonucleotide preparation

DNA oligonucleotides were purchased from Integrated DNA Technologies (IDT). RNA oligonucleotides and DNA/RNA chimeras were purchased from Trilink. All oligonucleotides were HPLC purified and the masses confirmed. The oligonucleotides were resuspended in 10 mM Tris-Cl pH 7.5 and quantified by UV absorbance spectroscopy, using the extinction coefficient at 260 nm for each oligonucleotide. All substrates were diluted to 10 μM (of reactive, methylated substrate, whether single- or double-stranded) in 10 mM Tris-HCl pH 7.5 and 50 mM NaCl. Duplexed oligos were annealed by heating to 95 °C for 30 s and cooling at 1 °C increments, for 15 s per step, to 4 °C.

2.5.2: Purification of TET enzymes from Sf9 insect cells

The crystalized human TET2 variant (TET2-CS) includes residues 1129–1936 Δ1481–1843 (Hu, L. et al., 2013). The full catalytic domains of TET2 (TET2-CD, residues 1129–2002), TET1 (TET1-CD, 1418–2136), and TET3 (TET3-CD, 689–1660) were also purified. These constructs, with an N-terminal FLAG tag, were subcloned into a pFastBac1 vector for Sf9 insect cell expression, as described previously (Liu, M. Y., DeNizio, and Kohli, 2016). Expression was

carried out for 24 hr. Cells from a 1 L culture were collected and resuspended in lysis buffer (50 mM HEPES, pH 7.5, 300 mM NaCl, 0.2% (v/v) NP-40) containing cOmplete, EDTA-free Protease Inhibitor Cocktail (Roche, 1 tablet/10 mL). Cells were lysed by three passes through a microfluidizer at 15,000 psi. The lysate was cleared by centrifugation at 20,000*g* for 30 min. The supernatant was then passed two to three times over a 500 µL or 1 mL packed column of anti-FLAG M2 affinity resin (Sigma), prepared according to the manufacturer's instructions. The column was washed three times with 10 mL of wash buffer (50 mM HEPES, pH 7.5, 150 mM NaCl, 15% (v/v) glycerol). To elute the bound protein, one column volume of elution buffer (wash buffer containing 100 µg/mL of 3× FLAG peptide (Sigma)) was incubated on the column for 10 min, and serial elutions were collected until no more protein was detected by Bio-Rad Protein Assay and SDS-PAGE. The three most concentrated fractions were pooled, aliquoted, and stored at −80 °C. Protein concentrations were measured using a Qubit 4 Fluorometer (Invitrogen) prior to use in *in vitro* activity assays.

2.5.3: TET reactions on DNA and RNA substrates

Substrates were diluted to the designated concentrations in reaction buffer (50 mM HEPES, pH 6.5, 100 mM NaCl, 1 mM α-KG, 1 mM DTT, and 2 mM sodium ascorbate). Fresh ammonium iron(II) sulfate (Sigma) was added to 75 µM prior to initiation, and purified enzyme was added last to start the reaction (t = 0), bringing the total volume to 25 µL. This mixture was incubated at 37 °C for 30 min. Reactions were quenched by addition of pre-mixed quenching solution (25 µL H_2O, 100 µL Oligo Binding Buffer (Zymo), and 400 µL ethanol). Oligonucleotide products were purified using the Zymo Oligo Clean & Concentrator kit, eluted in 10 µL of Millipore water, and analyzed by LC-MS/MS.

2.5.4: Quantification of reaction products by LC-MS/MS

The purified TET reaction products were degraded to component DNA and/or RNA nucleosides using 1 µL of Nucleoside Digestion Mix (New England Biolabs) in 10 µL total volume for 4 hours at 37 °C. The mixture was diluted 10-fold into 0.1% formic acid, and approximately 2

pmol was loaded onto an Agilent 1200 Series HPLC equipped with a 5 μm, 2.1 × 250 mm Supelcosil LC-18-S analytical column (Sigma) equilibrated to 45 °C in Buffer A (0.1% formic acid). The nucleosides were separated in a gradient of 0–10% Buffer B (0.1% formic acid, 30% (v/v) acetonitrile) over 8 min at a flow rate of 0.5 mL/min. Tandem MS/MS was performed by positive ion mode ESI on an Agilent 6460 triple-quadrupole mass spectrometer, with gas temperature of 225 °C, gas flow of 12 L/min, nebulizer at 35 psi, sheath gas temperature of 300 °C, sheath gas flow of 11 L/min, capillary voltage of 3,500 V, fragmentor voltage of 70 V, and delta EMV of +1,000 V. Collision energies were 10 V for all bases, aside from C and 5hmC for which the collision energies were 25V. MRM mass transitions were {dC 228.1→112.1 m/z; dT 243.1→127.1; 5mdC 242.1→126.1; 5hmdC 258.1→124.1; 5fdC 256.1→140.0; 5cadC 272.1→156.0; rC 244.1→112.1; 5mrC 258.1→126.1; 5hmrC 274.1→124.1; 5frC 272.1→140.0; 5carC 288.1→156.0}. Standard curves were generated from standard deoxyribonucleosides (Berry & Associates) and ribonucleotide triphosphates (Trilink), first digested to nucleosides using the Nucleoside Digestion mix under the same conditions as above prior to serial dilution, from 1,250 fmol to 0.614 fmol (Figure S2-1); sample peak areas were fit to the standard curve to determine amounts of each modified cytosine in the sample. Results are expressed as the percentage of each modified cytosine out of all modified DNA or RNA bases detected. Alternatively, to quantify total oxidation events, we considered that TET enzymes catalyze one oxidation reaction to generate 5hmC, two reactions to generate 5fC, and three reactions to generate 5caC. Therefore, we expressed total oxidation events as 1×(% 5hmC) + 2×(% 5fC) + 3×(% 5caC) (Crawford et al., 2016).

2.5.5: Molecular dynamics simulations

Molecular dynamics (MD) simulations were performed using the ff99SB force field with the pmemd.cuda program from the AMBER16 software suite(Case et al., 2017; Salomon-Ferrer et al., 2013). The 5mrC system was created from a crystal structure of the human TET2-DNA complex (PDBID: 4NM6) (Hu, L. et al., 2013). Similar to our previous simulations (Liu, M. Y. et al., 2017), the current system includes magnesium as a surrogate for iron in the active site,

coordinated His residues protonated on ND1, and a linker inserted using Modeller (Fiser et al., 2000; Sali and Blundell, 1993).

The previously published deoxyribonucleotide parameters for cytosine derivatives, α-ketoglutarate, Mg(II), and Zn were used (Liu, M. Y. et al., 2017), and ribonucleotide parameters for 5mrC were generated using PyRED (Bayly et al., 1993; Dupradeau et al., 2010; Vanquelef et al., 2011; Wang, F. et al., 2013). The system with 5mrC was generated using the LEaP module (Frisch et al., 2016) and neutralized to a net charge of zero with potassium ions and then solvated in a truncated octahedron of TIP3P water extending at least 12 Å from the complex surface (Jorgensen et al., 1983). SHAKE was used for all bonds involving hydrogen, and long-range Coulomb interactions were treated using the smooth particle mesh Ewald method with a 9 Å cutoff (Essmann et al., 1995). Simulations were carried out with the Berendsen thermostat in the NVT ensemble with a 2 fs time-step for 100 ns (Berendsen et al., 1984).

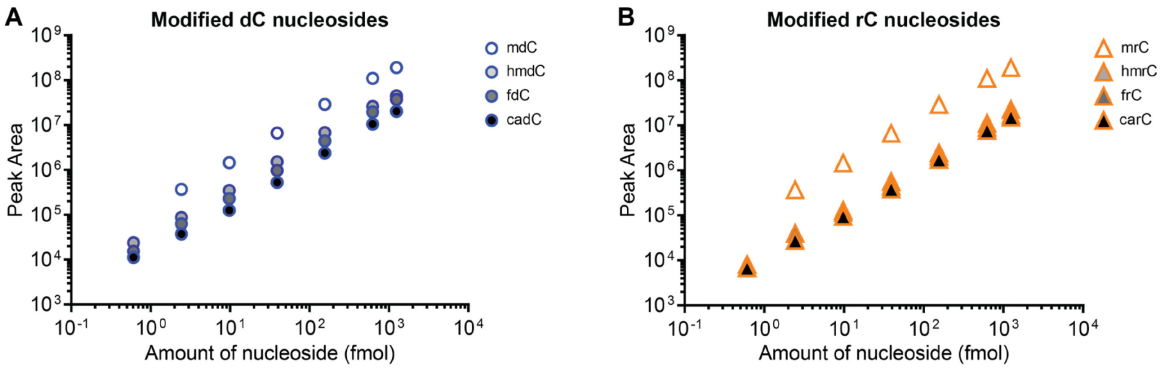
All dsDNA 5mdC reference calculations were performed with Langevin dynamics. Two tests were performed in order to validate that meaningful comparisons could be made between the 5mrC system and the previous work with the 5mdC system. These tests included a single 100 ns run of the 5mrC system with Langevin dynamics (minimization in the NVT ensemble and production in the NPT ensemble with the Berendsen barostat; collisional frequency of 1.0 ps^{-1}) and a 1 fs time-step, and a single 100 ns run of the 5mdC system with the Berendsen thermostat in the NVT ensemble and a 2 fs time-step used for the dynamics reported here.

Upon confirming compatibility, each simulation was carried out three times and data were averaged across replicates. The cpptraj program in AMBER was used to analyze trajectories for correlations, hydrogen bonding, atomic fluctuations, root mean square deviations, residue distances, and nucleotide angular parameters (Babcock et al., 1994; Olson et al., 2001; Roe and Cheatham, 2013). Energy decomposition analysis (EDA) was performed using an in-house FORTRAN90 program to investigate the Coulomb and van der Waals interactions between individual residues throughout the simulation (Dewage and Cisneros, 2015; Elias and Cisneros, 2014; Graham et al., 2012). VMD and the ProDy interface were used in principal component (PCA) and normal mode (NMA) analyses (Bakan et al., 2011; Humphrey et al., 1996). The first

100 PCA modes were calculated from each trajectory using 5,000 snapshots. UCSF Chimera was used to generate images (Pettersen et al., 2004).

2.6: Supplementary Information

2.6.1: Supplementary Figures



C

	0.6 fmol	2.4 fmol	9.8 fmol	39.1 fmol	156.3 fmol	625.0 fmol	1250.0 fmol	*Slopes*
mdC	-	3.69E+05	1.45E+06	6.65E+06	2.91E+07	1.11E+08	1.92E+08	1.59E+05
hmdC	2.38E+04	8.76E+04	3.50E+05	1.52E+06	6.78E+06	2.61E+07	4.49E+07	3.72E+04
fdC	1.55E+04	6.32E+04	2.30E+05	9.74E+05	4.44E+06	1.95E+07	3.75E+07	3.03E+04
cadC	1.13E+04	3.75E+04	1.27E+05	5.25E+05	2.39E+06	1.05E+07	2.03E+07	1.63E+04
mrC	-	3.69E+05	1.45E+06	6.65E+06	2.91E+07	1.11E+08	1.92E+08	1.59E+05
hmrC	-	2.76E+04	1.15E+05	4.51E+05	2.06E+06	8.95E+06	1.71E+07	1.38E+04
frC	8.06E+03	3.95E+04	1.27E+05	5.57E+05	2.44E+06	1.10E+07	2.24E+07	1.78E+04
carC	6.78E+03	2.78E+04	9.55E+04	3.86E+05	1.75E+06	7.75E+06	1.51E+07	1.22E+04

Figure S2-1. Representative nucleoside standard curves for LC-MS/MS analysis.

There is little to no discrepancy between the detection levels of dC **(A)** and rC **(B)** nucleosides. **(C)** Raw data used to generate the standard curves and the resulting slope of the line, which is used to normalize the raw sample data. Standards were collected on each day of MS collection and used to normalize the temporally appropriate data.

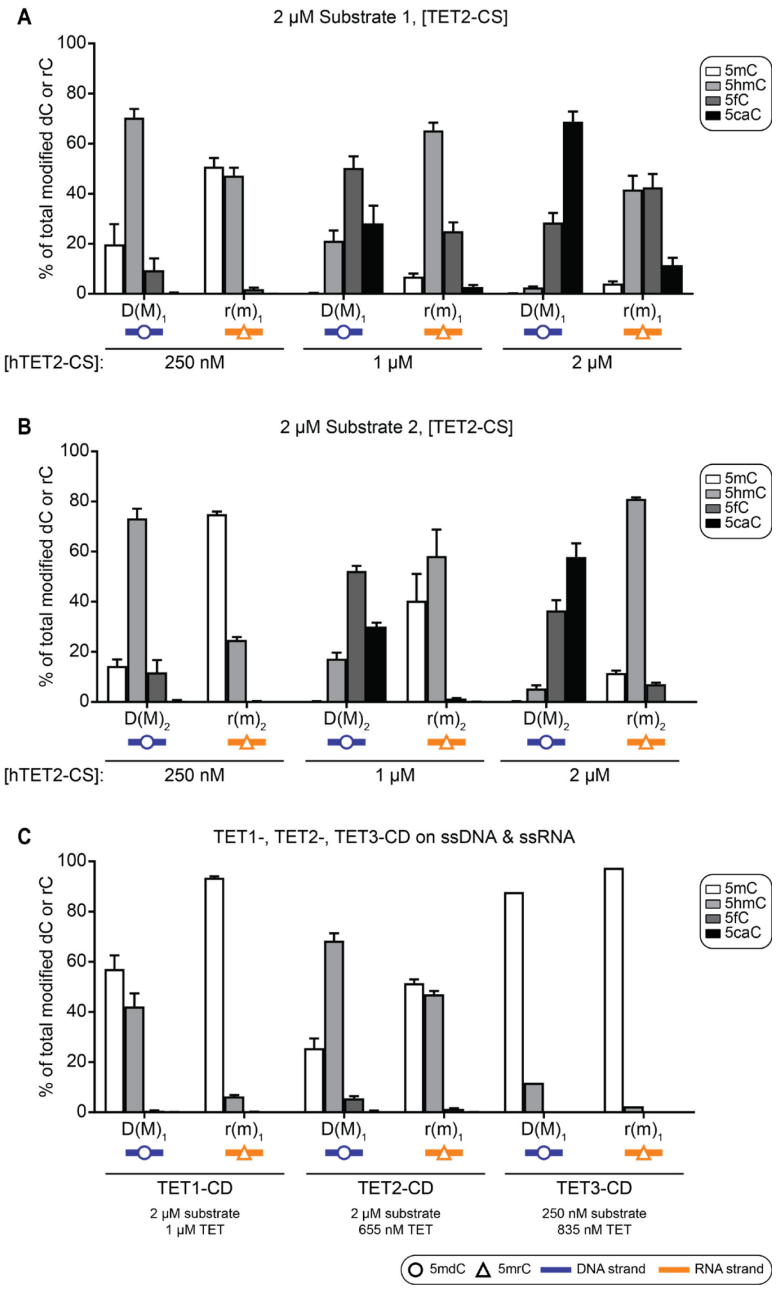


Figure S2-2. All human TET isozymes can oxidize ssDNA and ssRNA.

(A-B) Activity of TET2-CS on ssDNA and ssRNA substrates from Series 1 and 2. The data shown here for 1 μM TET2-CS is the same as in Figure 2-1B,D, and the data are alternatively expressed as total oxidation events in Figure 2-1C,E. **(C)** Activity of the full catalytic domains of TET1-3 on ssDNA and ssRNA substrates from Series 1. Error bars represent standard deviation for 3 replicates.

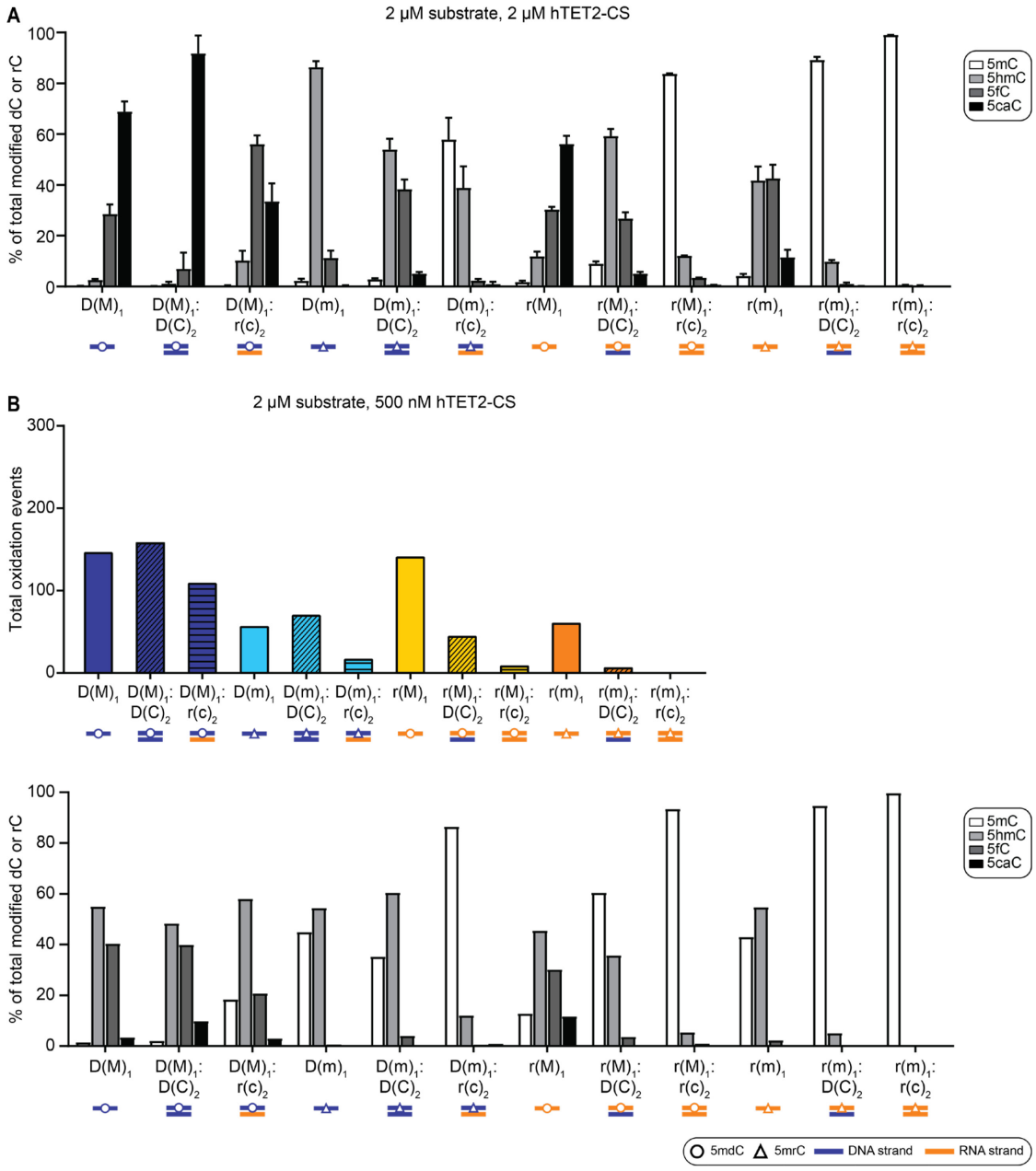


Figure S2-3. Double-stranded DNA is preferred, while dsRNA is strongly disfavored for two different reaction conditions.

(A) Activity of 2 μM TET2-CS on ss- and ds-substrates. These reaction products are alternatively expressed as total oxidation events in Figure 2-2A. Error bars represent standard deviation for 2-3 replicates. **(B)** Total oxidation events (top) and relative reaction products (bottom) of reactions with 500 nM TET2-CS and the same series of ss- and ds-substrates.

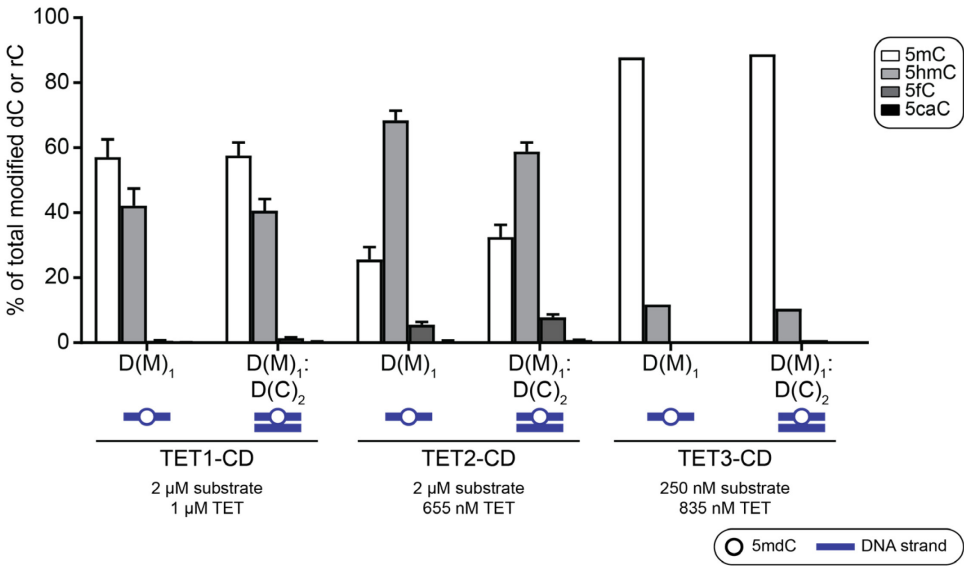


Figure S2-4. TET2-CD shows slight preference for ds- over ssDNA, while TET1- and TET3-CD are equally reactive on these substrates.

Error bars represent standard deviation for 3 replicates. The ssDNA data are replicated from Figure S2C for clarity.

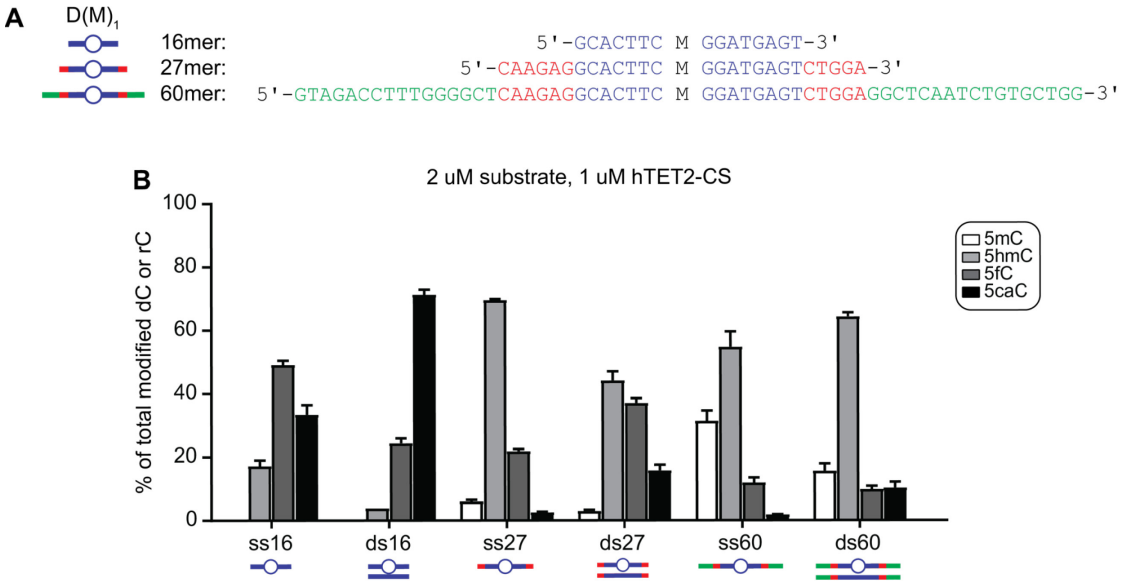


Figure S2-5. Preference for ds- over ss-DNA is consistent at various DNA lengths.

(A) Design of sequence-matched DNA substrate length series. **(B)** Activity of TET2-CS on the length series (total oxidation events shown in Figure 2-2B). Error bars represent standard deviation for 2-3 replicates.

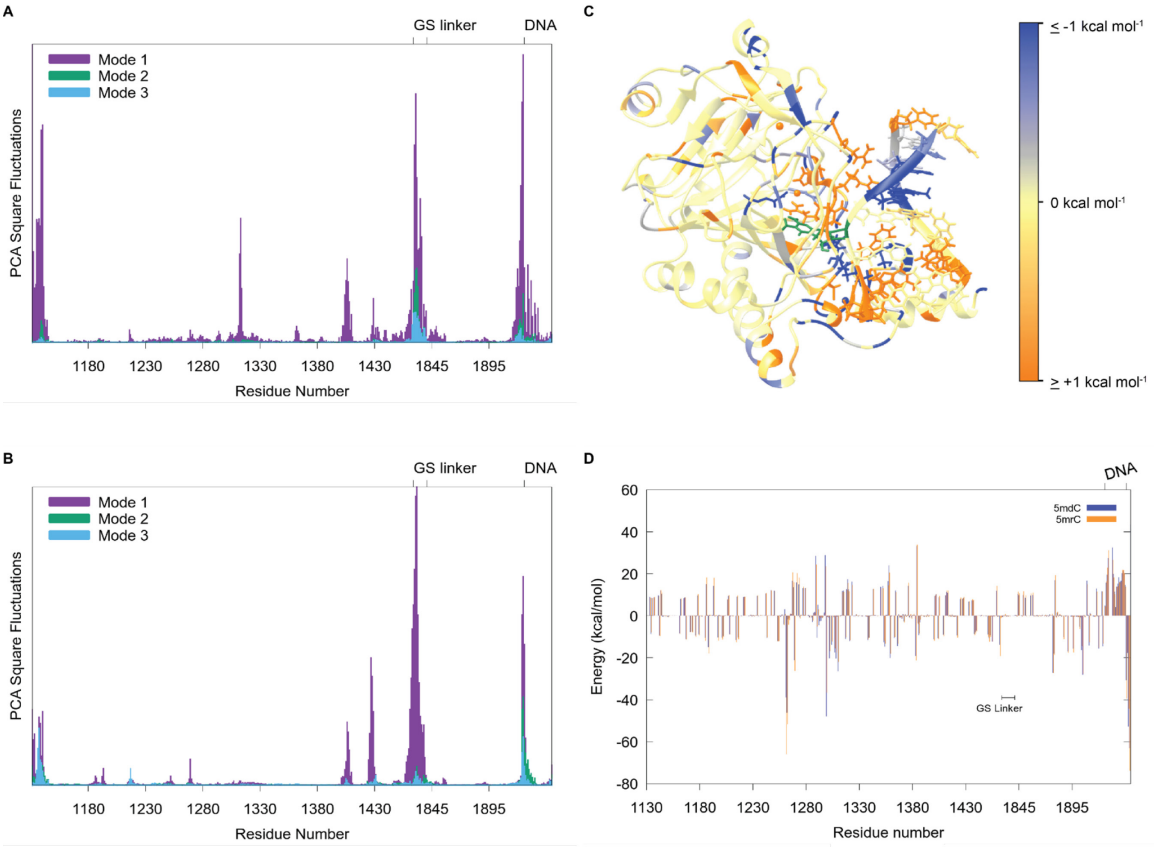


Figure S2-6. Global changes in dynamics and energetics of TET2 bound to dsDNA with 5mrC.

(A-B) Normal mode analyses, showing mode 1 (purple), mode 2 (green), mode 3 (blue) of 5mdC- **(A)** and 5mrC-containing systems **(B)**. **(C)** Structure of TET2 colored by changes in total interaction energy (sum of Coulomb and van der Waals energies) of each residue or base relative to the target 5mC between the models containing 5mdC and that containing 5mrC in dsDNA. TET residues showing significant changes in H-bonding and non-bonded interactions, as well as surrounding DNA bases, are shown as sticks. **(D)** Average total interaction energies >|1 kcal mol^{-1}|, used to calculate the differences displayed in Panel C, shown side-by-side here for each residue.

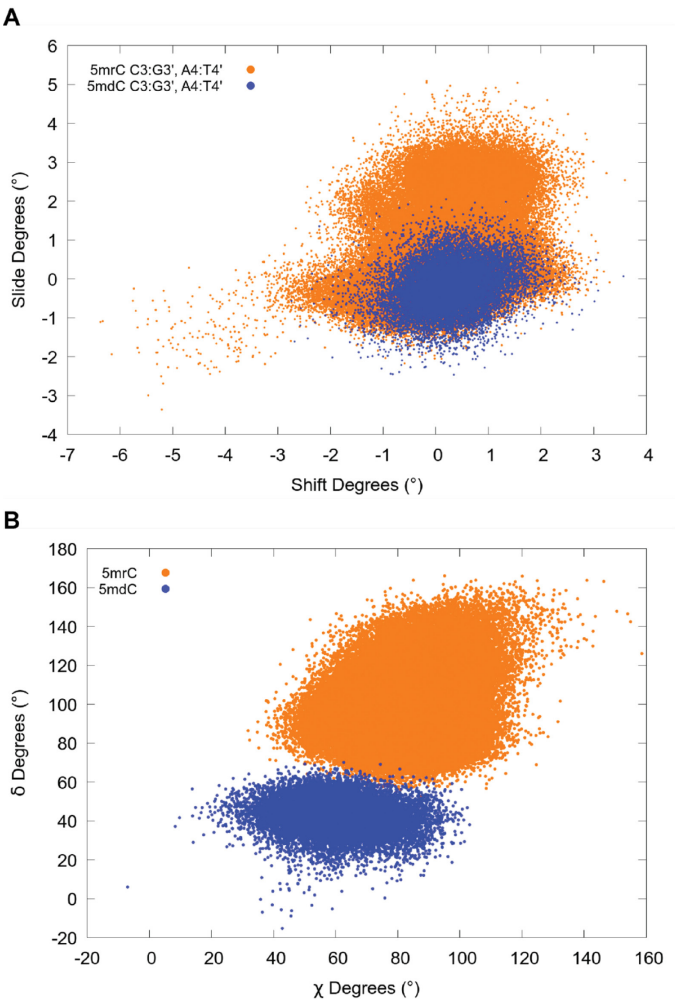


Figure S2-7. 5mrC-containing system displays shift from B-form to A-form DNA.

(A) Shift and slide for C3•G3' and A4•T4' in dsDNA with either 5mdC or 5mrC. Parameters from all replicates are shown. **(B)** Delta and chi angles for the target 5mC in dsDNA. Angles from all replicates are shown.

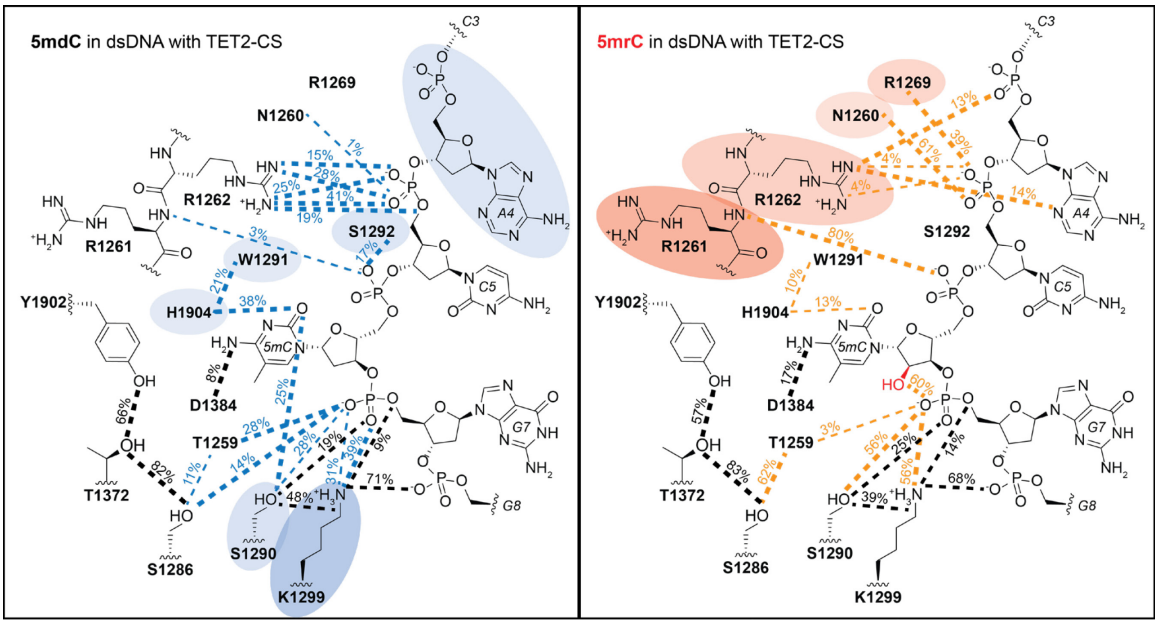


Figure S2-8. Expanded view of H-bond network and energetic differences in the active site of TET2 with dsDNA containing either 5mdC or 5mrC.

Expanded data from Figure 2-4 of TET2 in complex with dsDNA containing either 5mdC (left) or 5mrC (right). H-bonds (>10% simulation time in at least one system) are designated by dashed lines, including those that occur in both systems (<10% difference, black). When a H-bond occurs unequally (>10% difference) in both systems, it is shown as a thinner dashed line in the system in which it is less prevalent. Significantly different non-bonded interactions (sum of Coulomb and van der Waals energies >|1 kcal mol^{-1}|) between a TET residue and the target 5mC in the two systems are highlighted according to the relative magnitude of the difference (see Table S2-1 for values).

2.6.2: Supplementary Tables

$\Delta E = E_{5mdC} - E_{5mrC}$	
A4	-3.73 ± 0.01
N1260	2.43 ± 0.08
R1261	27.1 ± 1.7
R1262	5.48 ± 1.67
R1269	5.11 ± 0.83
C1289	4.16 ± 0.34
S1290	-5.32 ± 2.13
W1291	1.97 ± 0.72
S1292	-3.39 ± 2.20
K1299	-11.2 ± 0.4
H1382	1.54 ± 0.39
D1384	-0.64 ± 3.97
H1904	-1.92 ± 0.23

Table S2-1. Differences in the total non-bonded interactions between the 5mC-containing nucleotide and the specified base or residue (in kcal mol-1; ΔE ± avg. st. dev.).

CHAPTER 3: Biochemical basis for depletion of CpH hydroxymethylation in genomes rich in CpH methylation

3.1: Introduction

Both the hereditary propagation and dynamic nature of cytosine modifications are critical for transcriptional regulation in mammals, influencing cell development and maintenance. The most common epigenetic mark, 5-methylcytosine (mC), primarily occurs at cytosine-guanine dinucleotides (CpGs), with 60-80% of CpGs methylated in most cell types. In differentiated cells, CpHs (H=A,C,T) are rarely methylated, but in embryonic stem cells (ESCs) and brain cells, as much as 2-6% of CpHs can be methylated (Jang et al., 2017). In these cell types, due to the relative scarcity of CpG dinucleotides across the genome as a whole, mCpHs can account for up to half of all methylated cytosines in the genome despite the lower fraction of modification sites (Kinde et al., 2015).

Importantly, mC is not the only relevant mark that occurs on cytosine bases in mammalian genomes. In all cell types, 5-hydroxymethylcytosine (hmC) has also been identified as a stable modification; it is prevalent in ESCs and can account for up to 1% of all nucleotides in neurons. The Ten-eleven translocation (TET) family of enzymes, which are Fe(II)/α-ketoglutarate(αKG)-dependent dioxygenases, are responsible for introducing this mark by oxidizing mC; the hmC product can then be further oxidized to yield 5-formylcytosine (fC), or yet again to 5-carboxycytosine (caC). While independent roles for each of the oxidized mC bases (oxmCs) remain subject to ongoing studies, oxmCs have been implicated in pathways for DNA demethylation. That is, the presence of an oxmC at one cytosine in CpG•CpG base pairs can prohibit the maintenance methylation of the other cytosine, promoting passive demethylation. The higher oxidation states, fC and caC, can also be actively excised by the base excision repair enzyme, thymine DNA glycosylase (TDG); repair of the excised base leads to the installation of unmodified cytosine in place of the modified base, yielding a complete biochemical pathway for active demethylation of mC.

Although mCpHs are robustly detected in ES and neuronal cells, the presence of oxmCpHs is more controversial. Different sequencing methodologies have offered differing results regarding the presence of hmCpHs *in vivo*. With Pvu-Seal-seq, which utilizes an hmC-specific restriction enzyme to localize hmC at single-base resolution, 24% of hmCs detected were in CpH contexts in ESCs (Sun et al., 2015). In contrast, TET-assisted bisulfite-sequencing (TAB-seq) (Yu et al., 2012) and APOBEC-coupled epigenetic sequencing (ACE-seq) (Schutsky et al., 2018), both of which can specifically localize hmC at base resolution, detected very few hmCpHs in both ESCs and neurons.

Regardless of these discrepant results, in each instance there is not a proportional amount of hmCpH relative to mCpH, compared to the ratio of hmCpG to mCpG. There are two, non-mutually exclusive explanations to account for these observations: The first, that TET enzymes may be resistant to oxidation of mCpH (Masser et al., 2018). The second, that hmCpH may be particularly susceptible to active demethylation via the TET/TDG pathway, restoring unmodified CpH. In this chapter, I will examine the intrinsic substrate preferences of both TET and TDG to test these hypotheses.

Structural and biochemical studies of TET enzymes also provide conflicting evidence with regards to the tolerance for oxidation at sites with non-G bases 3' downstream of the target site, or at the +1 position. Using the human TET2 variant that allowed for crystallization with duplex DNA (TET2-CS), Hu *et al.* compared TET oxidation on mCpG versus mCpA and mCpC (Hu, L. et al., 2013). While the mCpG is almost fully oxidized, very little of both the mCpA and mCpC are converted. The associated crystal structure with mCpG-containing DNA led the authors to propose that a non-G base at the +1 position would impede the base-stacking interaction between it and the TET active site residue Y1294, resulting in diminished activity.

Despite this claim, there is *in vitro* evidence that a TET homolog can efficiently oxidize a target modified base when in a CpH sequence context. Specifically, although a TET homolog from *N. gruberi*, NgTET1, generally prefers CpG for all three substrates (mC/hmC/fC), it can oxidize DNA containing mCpG and mCpA with comparable efficiency (Pais et al., 2015). Further, there is only about a 3-fold decrease in activity on DNA containing mCpT and mCpC. With hmC-

containing DNA, there is only about a 2-fold decrease in TET reactivity compared to hmCpG for all three CpH sequence contexts. Thus, it will be informative to observe the activity of human TET2 on the full range of possible modified cytosine substrates in each of the four dinucleotide sequence contexts.

TDG has also been shown to be active in a variety of non-CpG contexts, including on bases other than a mismatched T. Although TDG activity on a G•T mismatch is the most impacted by the identity of the +1 base, all of the other substrates tested also show some level of discrimination (Morgan et al., 2007). The crystal structure of a truncated form of TDG (residues 82-308) in complex with double-stranded DNA containing a G•U mismatch reveals several TDG interactions with the 3'-G base (Coey et al., 2016). It was proposed that these interactions likely stabilize the TDG insertion loop within the DNA helix, aiding nucleotide flipping into the active site. To support this theory, a recent study showed that the nucleotide flipping of T in a G•T mismatch is sequence-dependent, particularly with regards to the +1 base; that is, the flipped state occurs less frequently for T in CpH contexts, likely to prevent abberant removal of T from normal A•T pairs (Dow et al., 2019). Although nucleotide flipping of fC and caC have not been measured, the TDG interactions with the 3'-G are maintained in a structure with fCpG (Pidugu et al., 2016), as well as in one with caCpG (Zhang, L. et al., 2012), suggesting that TDG may exhibit +1 base discrimination as well.

Here, we test whether the intrinsic substrate preferences of TET and TDG enzymes can possibly explain the depletion of hmCpH observed in relevant genomes. Steady-state measurements of TET oxidation products revealed that TET2 is impacted by the identity of the +1 base when oxidizing mC but, surprisingly, not hmC. Also, we observed that TDG can excise fC and caC efficiently regardless of sequence context. Thus, our evidence supports a biochemical pathway that favors the oxidation and eventual excision of hmCpH, despite the relative preference for mCpGs.

3.2: Results

3.2.1: Impact of +1 base on TET oxidation

To investigate the impact of the +1 base on TET reactivity, we chose to use the crystal structure variant of human TET2 (TET2-CS). TET2-CS is a more stable construct, allowing for profiling of a larger range of activity. Importantly, substrate preferences with this construct have been previously shown to be consistent with that of the TET2 construct with a full catalytic domain (DeNizio et al., 2019; Liu, M. Y. et al., 2017). To assess the relative substrate preferences of TET2 here, we devised a matched substrate series with twelve different oligomers, each having one of the three substrate, modified cytosine bases (mC, hmC, and fC) in each of the four possible sequence contexts (CpG, CpC, CpA, and CpT) (Figure 3-1A). Reactions were performed in conditions that have been previously optimized for overall activity (Crawford et al., 2016; Liu, M. Y. et al., 2017).

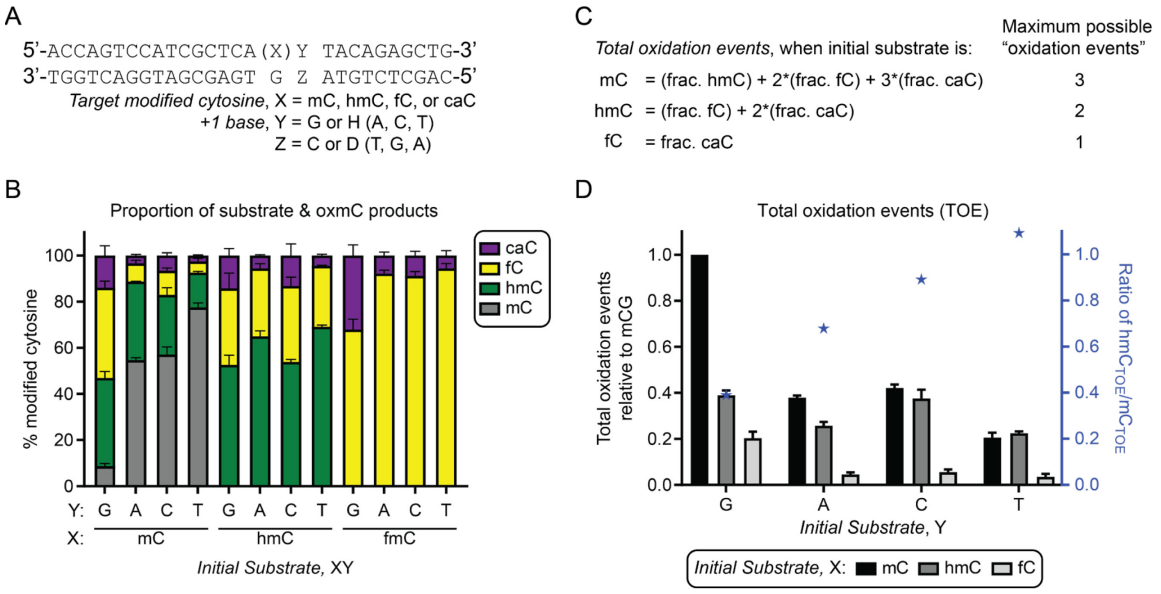


Figure 3-1. TET oxidation of XpY dinucleotides.

A) Substrates were designed to contain a single target modified cytosine (X) in a dinucleotide pair with one of the four bases (Y), base paired to its unmodified, respective partner (Z). **B)** Activity of TET2 with each of the twelve substrates. Reaction products were quantified by LC-MS/MS and expressed as the percentage of each base relative to the amount of total modified cytosine. The reaction products of each sample in B) are expressed as total oxidation events (TOE) using the respective equations in **C)**. **D)** TOE are normalized to the respective maxima of oxidation events and then presented relative to the mCG

(left y-axis). The ratios of these values for hmC/mC are also shown as blue stars for each respective sequence context *(right y-axis)*.

TET2 activity on the mC-containing substrates, under conditions with both excess and limiting enzyme, showed that mCpG is the most reactive substrate, as predicted. The mCpH substrates are less oxidized (Figure 3-1, 3-2); only 9% of mCpG remains unreacted, while the most converted of the three mCpHs, mCpA, has 55% substrate remaining (Figure 3-1B). Notably, however, all three mCpH substrates generate caC, with the mCpH substrates oxidizing to as much as half the proportion of caC that mCpG does, despite there being about a 6-fold difference in substrate consumed. Given the challenges posed by assessing the linked products, we also examined product formation by considering the total oxidation events (TOEs), whereby: for an mC substrate, hmC represents one turnover, fC represents two turnovers, and caC represents three turnovers (Figure 3-1C). Following that same logic, the TOE can be calculated for hmC and fC substrates as well. By normalizing the data to that of mCpG, we see that mCpA only experiences 37% of the oxidation events that mCpG does, with mCpT having only 21% (Figure 3-1D). Interestingly, when we compare the amount of mCpG oxidation to that at a mCpT on the same single-strand of DNA, the preference for mCpG persists for the full catalytic domain of TET2, as well as TET1 (Figure 3-3).

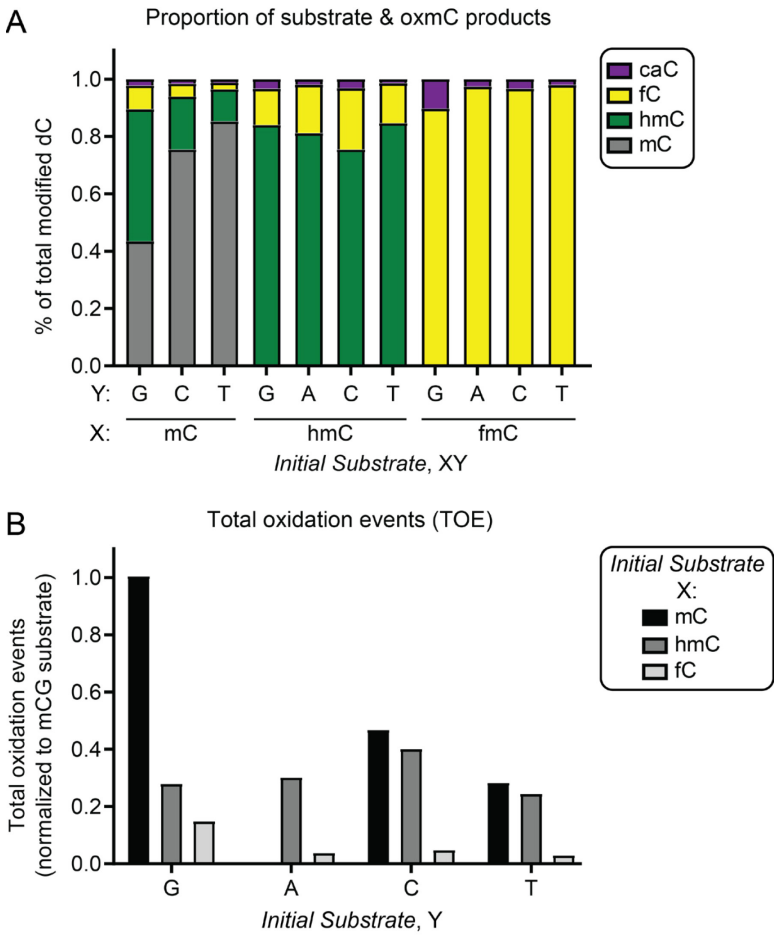


Figure 3-2. TET2 activity on XpY-containing substrates in limiting enzyme conditions.

Activity of 0.75 uM TET2-CS incubated with 1 uM XpY-containing DNA, with the exception that data for mCpA is not included. **A)** Reaction products were quantified by LC-MS/MS and expressed as the percentage of each base relative to the amount of total modified cytosine. The reaction products of each sample in A) are expressed as total oxidation events (TOE) using the respective equations in Figure 3-1C. **B)** TOE are normalized to the respective maxima of oxidation events and then presented relative to mCG

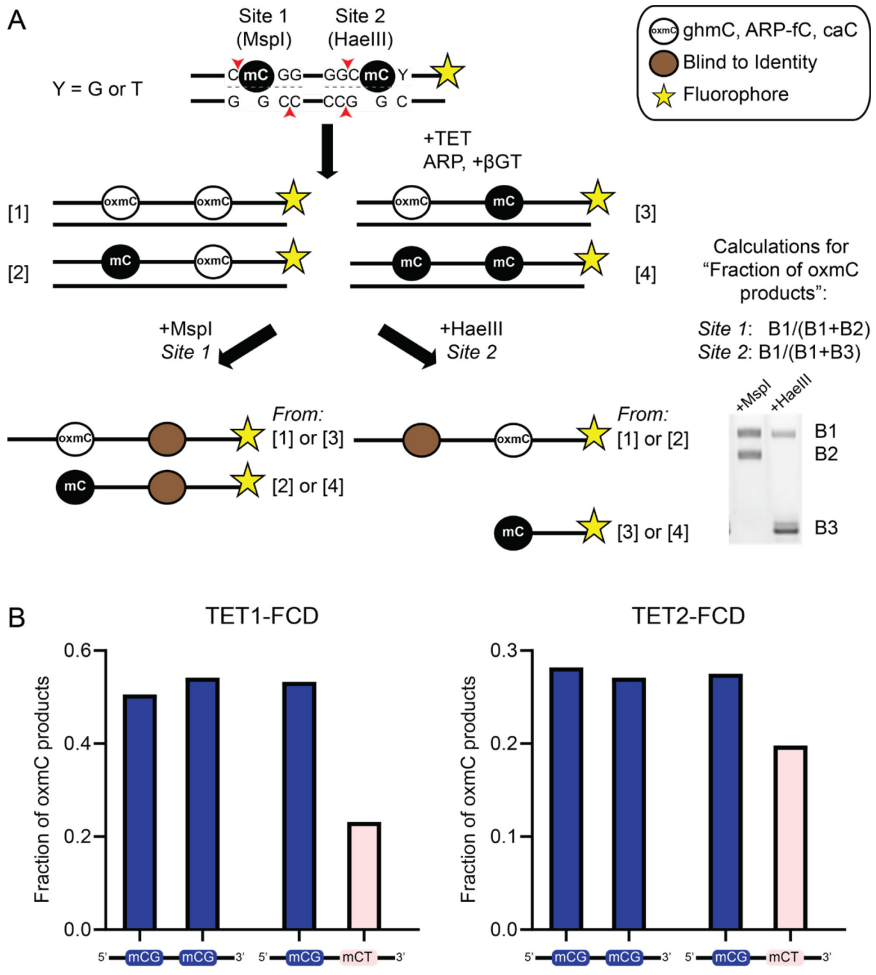


Figure 3-3. TET1 and TET2 have greater activity on mCpG than mCpT in the same single-stranded DNA substrate.

The TET reactions in these experiments were carried out as detailed in Chapter 2 (See section *2.5.3*) but with 250 nM of either the mCpG/mCpG or mCpG/mCpT oligomer substrate (CTAGCACmCGGTAGAATTCAGGCmCYTACGCCAATCT, Y=G or T) and 250 nM of TET1-CD or TET2-CD. **A)** Restriction enzyme-, gel-based assay for detection of oxidation at two mC sites on the same DNA strand (See Chapter 4, Figure 4-6 and Methods subsection *4.4.5* for more details). **B)** Relative levels of oxidation by TET1 and TET2 on a single-stranded DNA with either two mCpG sites or a mCpG at Site 1 and a mCpT at Site 2. Importantly the "-TET" cleavage controls samples were both fully cleaved by HaeIII at Site 2 (not shown).

In agreement with previous studies, the double-stranded mCpG-containing substrate is also more reacted by TET2-CS than the hmCpG-containing substrate. Surprisingly, however, hmCpG is of similar reactivity to the hmCpH substrates, with the proportion of remaining hmC substrate being within 16% for all sequence contexts (Figure 3-1B). Further, the hmCpHs all

convert nearly as well as the mCpH substrates, evidenced by both their similar substrate consumption (Figure 3-1B) and TOE relative to mCpG (Figure 3-1D). Importantly, the ratios of relative TOE of hmCpH substrates to their mCpH counterparts exceed 0.68, nearly doubling that of the ratio of hmCpG to mCpG, which is 0.39 (Figure 3-1D). Finally, fCpG is not an efficient substrate, but converts about twice as efficiently as each of the fCpH substrates, which are minimally reacted (Figure 3-1C). These trends are consistent even when a more limiting amount of enzyme is used, and 43% of the mCpG substrate remains unreacted (Figure 3-2).

3.2.1: Impact of +1 base on TDG excision activity

To assess the likelihood of active demethylation occurring from mCpHs, we also wanted to investigate the impact of the +1 base on rates of TDG excision of fC and caC. Using the same substrate series as the TET experiments above, single turnover kinetics experiments were performed with TDG on G•T, G•fC and G•caC-containing substrates under saturating enzyme conditions. Importantly, this allows us to operate under the assumption that the rate constants are not impacted by substrate engagement, product release, or product inhibition; thus, the observed rates of excision reflect the maximal rate of product formation. As previously published, TDG excision of T is greatest with G at the +1 base (G•TG), exhibiting a 75-fold difference in activity over the least active substrate, G•TT (Dow et al., 2019) (Figure 3-2, Table 3-1). Also, fC and caC are both efficient substrates for excision by TDG, with caCpG excised at a similar rate to T, while the rate of fCpG excision is 2-fold greater. Surprisingly, TDG does not exhibit the same level of discrimination against fC or caC when in a CpH context as it does against T: Compared to fCpG, the k_{max} is only reduced by 1.4-, 1.8-, and 4.7-fold for fCpA, fCpC, and fCpT, respectively. Compared to caCpG, the k_{max} is only reduced by 1.3-, 1.7-, and 3.4-fold for caCpA, caCpC, and caCpT, respectively (Figure 3-2, 3-5, Table 3-1). Therefore, TDG appears to be relatively unaffected by the identity of the +1 base.

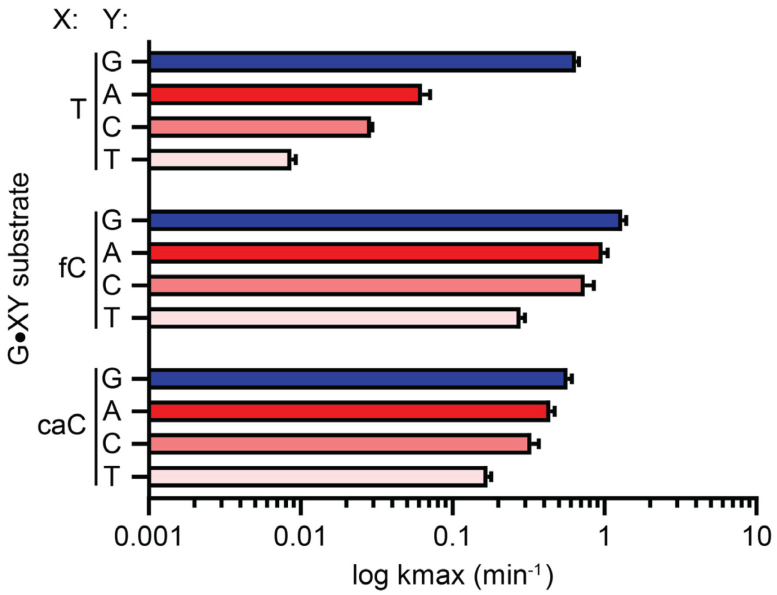


Figure 3-4. Excision activity of TDG on mispaired T, fC, or caC in XpY dinucleotides.

Excision rates (k_{max}) were determined from fitting curves of product accumulation over time (Figure 3-5). Values are reported in Table 3-1.

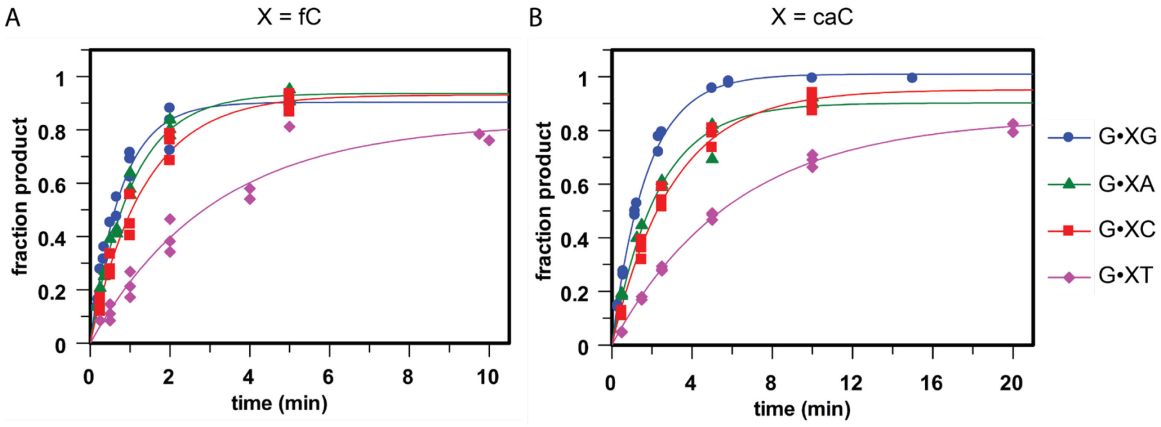


Figure 3-5. Excision activity of TDG on fC, or caC in XpY sequence context.

Single turnover kinetics experiments were performed at 37 °C using a saturating concentration of TDG. Data were fitted to a single exponential equation and the resulting rate constants (k_{max}) are shown in Table 3-1. The corresponding curves for excision of G•TY have been previously published in (Dow et al., 2019).

Substrate	k_{max} (min^{-1})	k_{max} relative to X**G**
T**G**	0.65 ± 0.03	-
T**A**	0.063 ± 0.008	10
T**C**	0.029 ± 0.0007	22
T**T**	0.0087 ± 0.0006	75
fC**G**	1.31 ± 0.08	-
fC**A**	0.97 ± 0.08	1.4
fC**C**	0.74 ± 0.11	1.8
fC**T**	0.28 ± 0.02	4.7
caC**G**	0.57 ± 0.04	-
caC**A**	0.44 ± 0.03	1.3
caC**C**	0.33 ± 0.04	1.7
caC**T**	0.17 ± 0.01	3.4

Table 3-1. Relative rates of excision by TDG.

The k_{max} (k_{obs}) values are the mean and standard deviation for at least three independent single turnover kinetics experiments, at 37 °C. These values are graphed in Figure 3-2 on a log scale.

3.3: Discussion

The biochemical experiments presented here expose the relative tolerances of both TET and TDG for CpH contexts. In the conditions tested, TET2 exhibits a preference for mCpG, but displays much greater activity on the mCpH contexts than previously observed with TET2-CS (Hu, L. et al., 2013). Surprisingly, TET2 does not similarly discriminate against hmCpH substrates compared to hmCpG. In general, hmCpH and hmCpG, as well as mCpH, are oxidized to a similar extent. Meanwhile, TDG excision of fC and caC proceeds with relatively little dependence on the identity of the +1 base, suggesting that TDG can process fC and caC generated for DNA demethylation regardless of context.

3.3.1: Possible mechanisms of TET regulating observed substrate preferences

By analyzing our gel-based data and the available crystal structures, we can speculate about the underlying mechanisms regulating the sequence preferences of TET observed here. First, in addition to the LC-MS/MS data, a TET oxidation experiment was performed with two

substrates each containing two mCs on the top strand; one substrate has two mCpGs while the other has one CpG and one CpT (Figure 3-3). In the tested conditions, TET2 oxidized each mCpG site to a greater extent than the mCpT, even when on the same strand of DNA. This result illuminates possible mechanisms driving the relative substrate preferences of TET2 for mCpG over mCpH. First, the finding that substrate preference is maintained even when there is not a complement strand present seems to contradict the notion proposed from the crystal structure that TET contacts opposite base paired with the +1 base, driving CpH discrimination. Second, considering that base-stacking interactions are maintained even in single-stranded DNA, one can imagine that the CpG base-stacking interaction, which is known to be the strongest of any dinucleotide pair, aids in mC substrate engagement in some way that results in enhanced activity.

Meanwhile, we can also gain insight into the relative tolerance of hmCpH compared to hmCpG by analyzing the crystal structures of TET2-CS in complex with either mC- or hmC-containing DNA. In the hmC-containing structure, TET2 makes additional contacts with the phosphodiester backbone surrounding the flipped-out base compared to the similar structure with mC. If we consider our hypothesis from above that the strength of the base-stacking interaction is important, one can envision a scenario in which the balance of contributing forces results in relatively similar hmCH versus hmCG oxidation. There is precedent to support this general hypothesis. Using molecular dynamic simulations, we have previously shown that changing the target deoxyribo-mC to ribo-mC in a model of the crystal structure results in an altered H-bond network between TET2 and the DNA backbone surrounding the target base. Interestingly, this simulated change in interactions correlated with decreased activity on ribo-mC (DeNizio et al., 2019).

For TDG, there is some kinetic data to suggest a mechanism for the observed fC and caC sequence tolerance. Previously, using ^{19}F NMR method, it was shown that regulation of thymine excision by the +1 base largely manifests by restricting flipping of thymine into TDG's active site (Dow et al., 2019). Uracil, meanwhile, appeared to stably flip into the active site of TDG for all UpX contexts. Current ^{19}F NMR methods can only detect the non-flipped state if it is ~5% of the total signal. Although it is possible that small differences in flipping equilibria contribute to the

differences in k_{max}, it is more likely that the small differences in k_{max} are due to post-flipping mechanisms, such as formation of a productive enzyme-substrate complex and/or the chemical steps of the reaction. Unfortunately, nucleotide-flipping equilibria for fC and caC are currently unknown. Also, the environment of fC and caC in TDG in their respective crystal structures (Pidugu et al., 2016; Zhang, L. et al., 2012) is not likely to be very informative because it is only a snapshot of the most stable conformation. Thus, NMR or molecular modeling experiments in the future will be useful in determining if TDG activity on fC and caC is unaffected by the +1 base due to enhanced nucleotide-flipping.

3.3.2: Biological implications

Considering that TET activity on mCpH is relatively robust, even when activity is limited on mCpG (Figure 3-2), the notion that TET is intrinsically resistant to oxidizing mCpH does not seem likely. Further, our biochemical data shows that for every mCpH reacted to hmCpH, a greater proportion of those hmCpHs are converted to fC/caC than the proportion of hmCpG generated from mCpG. The fact that TDG is largely agnostic to the identity of the +1 base supports the idea that the fCpH/capCH generated in this scenario can be ultimately replaced by unmodified cytosine. Thus, from our data it is seems likely that the lack of hmCpH observed in the genome is due, in part, to greater relative turnover by of hmCpHs by TET (Figure 3-6).

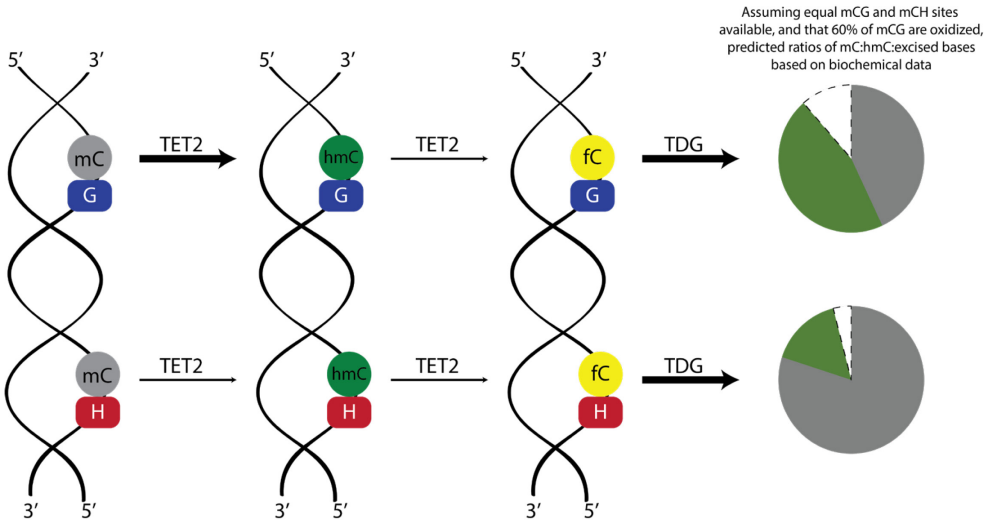


Figure 3-6. Model of modified cytosine conversion, based on intrinsic enzyme preferences, to explain depletion of hmCpH.

The biochemical data presented here supports a model in which hmCpH and hmCpG are equally oxidized by TET and subsequently excised by TDG, despite a strong preference for mCpG oxidation over that of mCpH. This would result in relatively greater proportions of hmCpGs and mCpHs.

It is still possible, however, that disfavored mCpH oxidation is partly responsible for the lack of hmCpH observed *in vivo*. It seems unlikely that differential localization would be responsible because that the majority of mCpH occur near mCpGs (Lee, J. H. et al., 2017). However, the lack of *in vivo* hmCpH may be due to either mCpH-specific proteins that block oxidation, similar to methyl-CpG binding proteins, or a lack of hmCpH-specific proteins, considering that hmCpG-specific proteins do exist (Spruijt et al., 2013). Also, it could be possible that other domains of TET located in the N-terminal domain regulate a preference for hmCpG that we do not observe in these experiments using only the catalytic domain; one could imagine a mechanism similar to how the CXXC and BAH1 domains within DNMT1 inhibit methylation of unmethylated CpGs (Song, J. et al., 2011).

In the future, we plan on confirming the biochemical similarity of TET on hmCpG and hmCpHs with a second, non-MS based assay. In addition, although a TET-TDG-BER pathway has been reconstituted from mCG *in vitro*, we would like to show that this is also biochemically feasible at CpHs, as well.

3.4: Methods

The oligonucleotides and TET proteins used in this Chapter, as well as the LC-MS/MS methods, were prepared as detailed in Chapter 2 (See sections *2.5.1, 2.5.2, and 2.5.4*, respectively).

3.4.1: TET reactions on XpY-containing substrates

The reactions performed prior to downstream analysis by LC-MS/MS were carried out as detailed in Chapter 2 (See section *2.5.3*), but in this chapter, specifically, 1 μM substrate was incubated with either 1.5 μM or 0.75 μM of purified hTET2-CS (Figure 3-1 and 3-2, respectively).

3.4.2: TDG protein expression and purification

TDG was expressed and purified by the laboratory of Alexander Drohat at the University of Maryland, as previously described (Coey et al., 2016; Maiti et al., 2009). The enzyme concentration was determined measuring absorbance at 280 nm, using an extinction coefficient of $\varepsilon 280 = 17.4\ mM^{-1}cm^{-1}$.

3.4.3: TDG excision assay

The excision activity of TDG on fCpY- and caCpY-containing substrates was monitored in a manner similar to previous studies (Dow et al., 2019). Briefly, the experiments were performed under saturating enzyme conditions ([E] > [S], [E] >> K_D), so that the rates observed (k_{obs}) are reflective of the maximal rate of product formation (k_{max}). Reactions were initiated by adding TDG to 0.5 μM of DNA substrate in HEN.1 buffer (0.02 M HEPES pH 7.5, 0.1 M NaCl, 0.2 mM EDTA) at 37 °C. At each timepoint, aliquots were removed and immediately quenched by adding 50% (v:v) quench solution (0.3 M NaOH, 0.03 M EDTA). The samples were then heated for 5 min at 85 °C to cleave the DNA backbone at abasic sites quantitatively. The resulting DNA fragments allowed for quantification of substrate and product by HPLC. Rate constants were determined from fitting progress curves (Figure 3-5) to the equation, $A(1–\exp(–k_{obs}t))$ using non-linear regression (A is the amplitude, k_{obs} is the rate constant, and t is the reaction time).

CHAPTER 4: Examining TET-oxidized clusters of oxmCs *in vitro* and the possible contributing mechanisms intrinsic to TET

4.1: Introduction

CpGs represent less than 1% of the dinucleotides in the human genome but are often highly clustered. Cytosine methylation in these clustered CpG motifs has been studied extensively, with an emphasis on the transcriptional regulation of regions that are differentially methylated across development and in disease. These CpGs are often in localized, clustered regions (which can include CpG Islands or CGIs) and are referred to as differentially methylated regions (DMRs) (Deaton and Bird, 2011; Wang, L. et al., 2014). DMRs represent hotspots of methylation dynamics, which are critical for genomic imprinting and regulation of gene expression during development), as well as in many diseases (Bergman and Cedar, 2013; Smith and Meissner, 2013). A major effort in the field has been to understand how and why cytosine modifications are introduced or maintained, particularly within these dynamic regions.

As is the case with differential methylation at DMRs, TET-oxidized bases also cluster at specific loci on both a population (Booth et al., 2012; Sun et al., 2015; Wen et al., 2014; Yu et al., 2012) and single-cell level *in vivo*. For TET enzymes, oxmC clustering could be limited by either the underlying CpG signature or the available mC bases at those CpGs. Although there is some overlap with methylated regions, the distribution of hmC is unique such that it cannot solely be explained by mC availability (Pastor et al., 2011). Notably, when TDG is knocked down in ES cells, the higher order TET modifications fC and caC are also enriched in localized clusters, including at unmethylated CGI promoters (Neri et al., 2015; Wu, H. et al., 2014).

In order to eventually contextualize the impact of external clustering factors on the observed signature of oxmCs *in vivo,* we wanted to first assess the intrinsic capacity for TET to generate clusters of oxidized modifications in a pared-down reaction environment, beginning with substrates homogenously CpG methylated. To do so, we needed a way to localize multiple oxmCs on large stretches of DNA, so we decided to utilize a single-base resolution sequencing technique.

There are several sequencing methods that are now regularly used to assess the oxmC content of DNA. While there are many variations depending on the specific modified base of interest (Wu, H. and Zhang, 2015), those that could prove most useful to us in this chapter are reviewed below (Figure 4-1). In general, each of these methods exploits the differential chemical or enzymatic reactivity of cytosine bases. The resulting bases are identified on individual sequencing reads and aligned to a reference genome. The localization of a modified cytosine is "called" when the frequency of that oxmC at a particular base exceeds a pre-designated threshold, typically 20%. Thus, the reported signature of a particular oxmC is only representative of the population average.

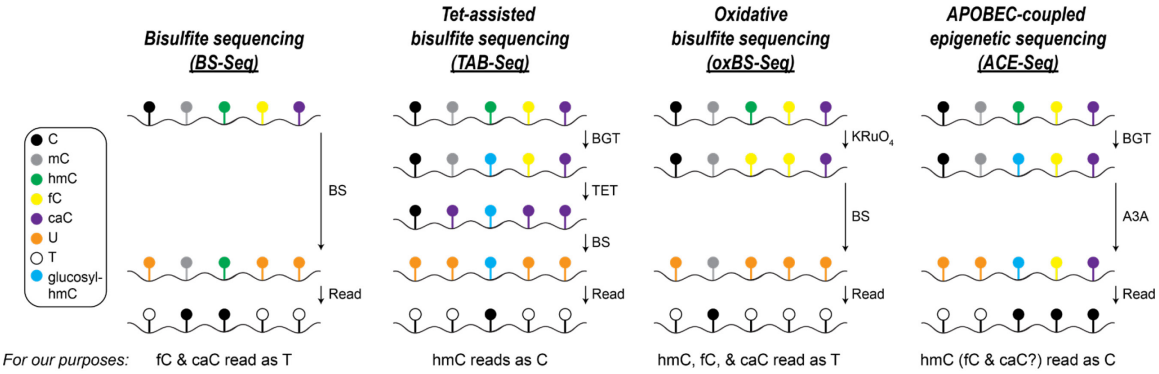


Figure 4-1. Chemical and Enzymatic Deamination-based sequencing methods for localizing specific oxmCs.

Bisulfite (BS) sequencing is considered the gold standard for detecting mC; bisulfite can chemically deaminate C, but not mC, to U, which is read as T by current sequencing platforms. Upon the discovery of hmC, it was shown that although bisulfite creates a sulfonated product with hmC (referred to as CMS), it is also read as C during sequencing (Huang, Y. et al., 2010). In contrast, fC and caC are deaminated by bisulfite and converted to U (Wu, H. and Zhang, 2015). Thus, both mC and hmC are read as "non-converted" C, while unmodified C, fC, and C are converted and read as T in sequencing.

To identify hmC specifically, separate from mC, TET-assisted bisulfite sequencing (TAB-seq) was developed (Yu et al., 2012). In this method, hmC is first glucosylated (to ghmC) via T4

β-glucosyltransferase (βGT). Then, TET is used to oxidize mC to either fC or caC, which, like C, can be converted by bisulfite. Meanwhile, ghmC is protected from TET oxidation and remains unconverted by bisulfite and is thus the only base read as C upon sequencing. Another bisulfite-based method for hmC-specific detection is oxidative bisulfite sequencing (oxBS-seq) (Booth et al., 2012). In this method, an oxidant, typically potassium perruthenate ($KRuO_4$), is used to chemically oxidize hmC to fC, to then be converted by bisulfite. A BS-only experiment is also performed on the sample in tandem to identify the true C signal; this is then subtracted from the oxBS-seq signal so that true hmCs may be called. Our lab has developed an enzymatic-based method for specifically localizing hmC (Schutsky et al., 2018). In APOBEC-coupled epigenetic sequencing (ACE-seq), hmC is also protected by glucosylation via βGT, then single-stranded DNA is generated via snap-cooling so that C and mC can be deaminated by the APOBEC3A enzyme. Notably, each of these methods was developed for analysis of genomic DNA, where fC and caC typically occur at very low levels (Wu, H. et al., 2016), so they are not thought to obscure the interpretation of hmC levels. However, for our purposes, we anticipated *in vitro* generation of fC and caC at relatively high levels, so oxBS-seq was identified as the best method for identifying clusters of oxmCs, as it yields deamination as a single readout for all three oxmCs on single sequencing reads.

In addition to assessing TET's ability to elicit clusters of oxmCs, in this chapter, we also wanted to explore the mechanism by which this behavior could be mediated, specifically testing if TET is intrinsically capable of acting processively on a strand of DNA. The processivity of DNA-modifying enzymes is defined as the capacity to repeatedly catalyze sequential reactions without releasing their substrates (Van Dongen Stijn F M et al., 2014). This can refer to sequential activity at multiple sites (strand processivity), or, unique to TET enzymes, the capability of iterative oxidation, *i.e.* repetitive activity at a single site, which is also often referred to as catalytic processivity. Our lab showed that mouse TET2 can act iteratively on a substrate with a single mC, generating fC and caC in a single encounter (Crawford et al., 2016). However, another group proposed that human TET1 and TET2 and *N. gruberi* TET1 do not exhibit strand nor catalytic processivity but act distributively (Tamanaha et al., 2016). However, the experimental design and

activity levels of their recombinant enzymes were not sufficient to discern between strand processive or distributive mechanisms. Thus, the processive capacity of TET enzymes remains unresolved, as does the interplay between catalytic and strand processivity. In this chapter, I utilized measurements that allowed me to negate the possibility of catalytic processivity for the time being because I assess all three oxmCs as “product.”

Classic studies of DNA-modifying enzymes, such as DNMTs (Holz-Schietinger and Reich, 2010; Vilkaitis et al., 2005), restriction endonucleases (Jeltsch et al., 1996), and DNA repair enzymes (Stivers and Jiang, 2003), provide precedent for enzymes with processive mechanisms that underlie the generation of clustered products. Functionally, strand processivity encompasses facilitated diffusion mechanisms of both 1D diffusion (sliding) and short hops/jumps (Halford and Marko, 2004; Stanford et al., 2000). In contrast, enzymes that act distributively randomly catalyze reactions following the association/conversion/dissociation pattern mediated by 3D diffusion. While distributive enzymes often accumulate products slowly over time or as the result of localized cellular density, processive enzymes can mediate concerted product accumulation by catalyzing reactions over a few or up to thousands of bases with enhanced speed and efficiency (Breyer and Matthews, 2001).

Features revealed in the crystal structure of human TET2 bound to 5mCpG-containing DNA (Hu, L. et al., 2013) suggest that TET belongs to the family of processive enzymes. As discussed in previous chapters, TET engages its catalytic substrate via a base-flipping mechanism, with specific residues contacting the cytosine base. Strikingly, all other interactions with DNA appear non-specific: a hydrophobic ridge inserts into the minor groove and a number of residues form a network of hydrogen bonds and electrostatic interactions with the phosphodiester DNA backbone. This structure is reminiscent of other processive DNA modifying enzymes, which are thought to maintain electrostatic interactions with DNA during sliding; upon target site recognition, these enzymes rearrange to form stronger interactions to initiate catalysis. Although the mechanisms of substrate recognition and engagement are still unresolved, the structure does support the possibility that TET can act processively.

In this chapter, I therefore aimed to address two questions regarding the intrinsic capabilities of TET. First, what is the extent of clustering elicited by TET *in vitro?* Second, can TET act processively? Addressing both questions presented the similar challenge of how to link the information between two or multiple mCpGs. This objective required the development of both unique metrics for quantifying oxmC clustering and a novel, gel-based assay for assessing TET processivity. Although each of these questions remain somewhat unresolved, both experimental approaches yielded other insights into possible mechanisms of TET substrate regulations.

4.2: Results

4.2.1: Capacity for TET-generated oxmC clustering in vitro

To first assess the extent of clustering intrinsic to TET, we performed *in vitro* reactions with the full catalytic domain of either TET1 or TET2 on a 2.7 kb enzymatically-CpG-methylated double-stranded pUC19 substrate. An advantage of this substrate is that there are no unmethylated CpGs to complicate the interpretation of the sequencing data. The original plan was to use oxBS-seq to analyze TET activity because this method would allow us to localize unreacted mCpGs and reacted oxmCs separately. Although we did achieve eventual success in optimizing the assay (Figure 4-2), most of the data collected was analyzed using a BS-seq protocol. Although we are unable to separately detect mC and hmC, we can analyze the clustering of the higher oxidized modifications, fC and caC (hoxmCs), together. We performed numerous experiments in which we varied both the TET concentration and incubation time, and in each instance saw varying degrees of clustering (Figure 4-3).

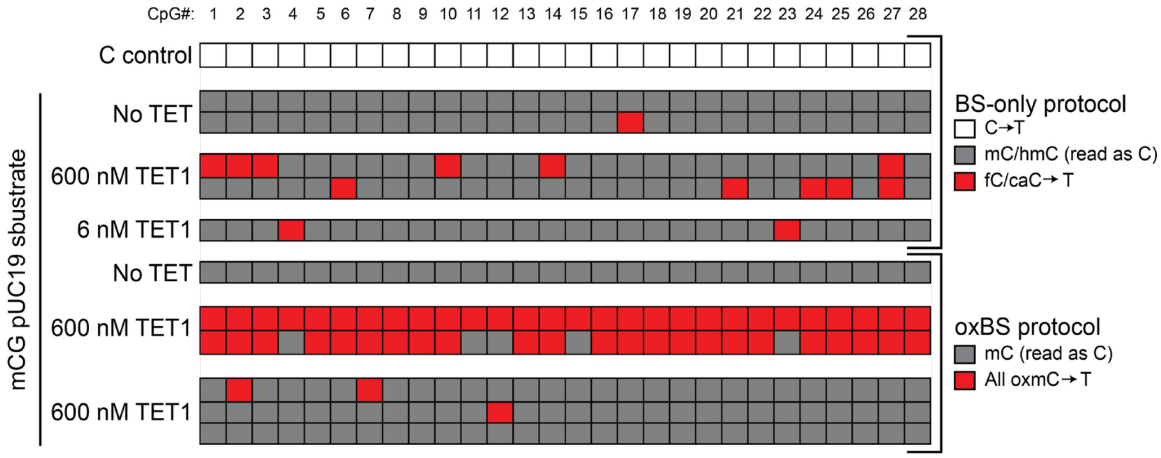


Figure 4-2. Preliminary experiment using oxBS-seq to localize TET1-generated hmC, fC, and caC together, separately from mC.

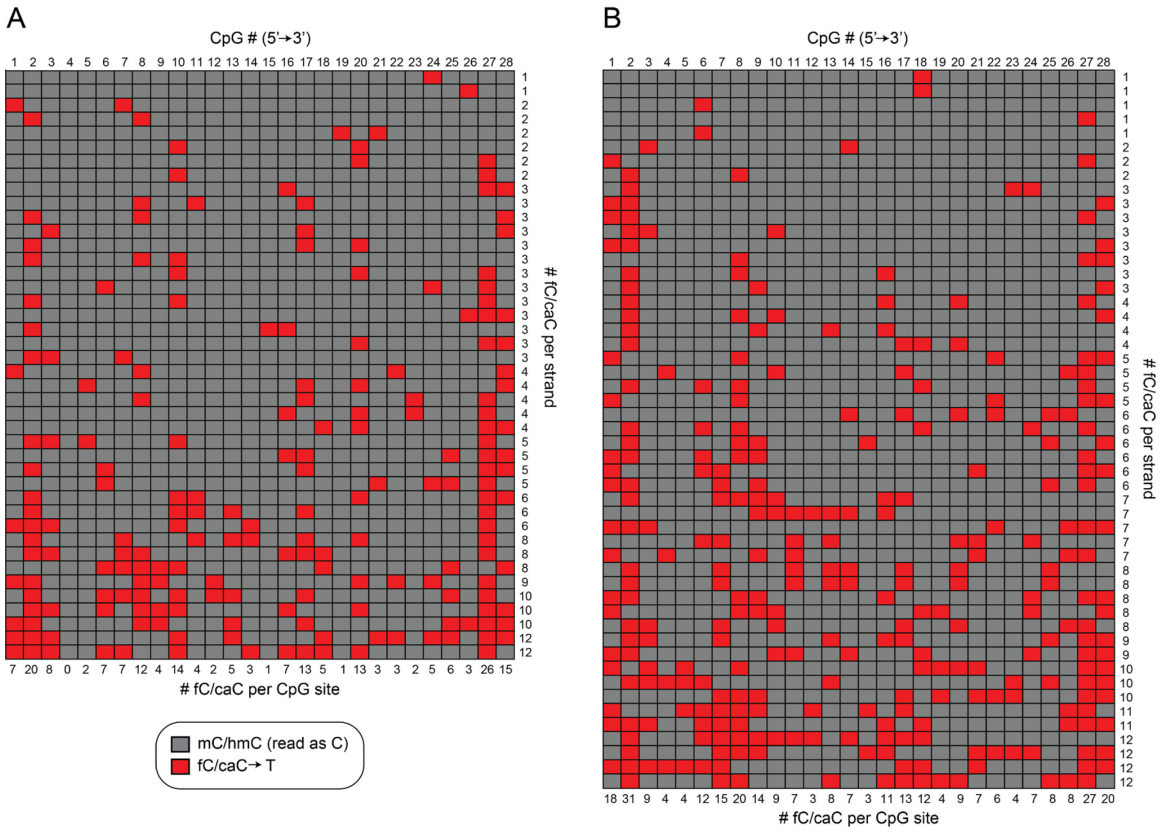


Figure 4-3. TET1 and TET2 mediate clusters of hoxmCs from homogenously CpG-methylated single strands of DNA

An exemplary set of TET2-reacted, BS-converted data is shown in Figure 4-4. In the grid in Panel A, each row represents the data from a single Sanger sequencing read of a single clone from a particular DNA molecule. The boxes in each column represent the 28 CpGs in the 362-base-pair PCR-amplified region that was sequenced. The DNA experienced varied levels of TET oxidation and exhibited varying levels of clustering (Figure 4-4A). Qualitatively, for the clones with nine hoxmCs or less, there are several instances in which two oxmCs occur at neighboring CpGs. Meanwhile, for clones with eleven or twelve oxidation events, one can observe instances in which 4-7 hoxmCs occur in a row on a particular DNA strand (Figure 4-4A).

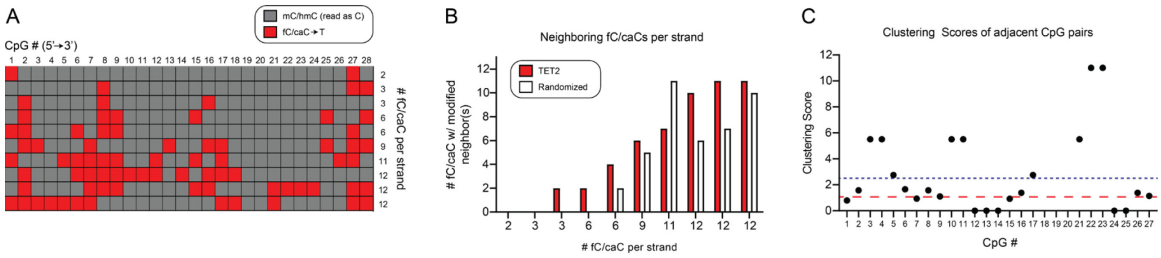


Figure 4-4. TET2 generates clusters of hoxmCs on a fully CpG-methylated substrate.

A) Localization of hoxmCs (fC/caC) on DNA from a single TET reaction. Each row represents the data from a single Sanger sequencing read, representing a single molecule of DNA, and each box represents a single CpG site. There are 28 CpGs in the 362 bp region analyzed. In the included data, all CpH were successfully converted by bisulfite, but that is not represented here.
B) Comparison of the number of hoxmCs with at least one neighboring hoxmCs in either direction between the dataset in A) and a randomized data set, for each individual DNA strand. **C)** Clustering scores for each relative pair of CpGs, with the average score for the 27 CpG pairs indicated by a blue dashed line. The red dashed line at the clustering score of 1 delineates the threshold for which clustering is considered to occur at a non-random frequency.

In addition to this qualitative assessment, we have also developed metrics by which clustering can be measured semi-quantitatively (Figure 4-4B) and quantitatively (Figure 4-4C). Although the larger stretches of clustered hoxmCs are inarguably significant, we have determined that it is both easier but still effective to quantify the non-random nature of the clustering on a two-CpG-site scale. First, for different levels of overall activity, we can assess the number of hoxmCs that occur with an hoxmC at either neighboring CpG (Figure 4-4B). Further, by generating a

horizontally randomized data set, we can determine whether the frequency of neighboring hoxmCs occurs to a greater extent in the actual data set than by chance. Although there is an outlier for the sample that has eleven hoxmCs, in general there are less hoxmC sites with oxidized neighbors in the randomized data set than the actual TET2 dataset. We have also developed a metric to assign each CpG a clustering score. This score compares the observed frequency at which two hoxmCpGs occur next to each other to the frequency at which the event is likely to occur due to the underlying, supposedly-unlinked reactivities at each CpG site individually. If the first CpG is designated CpG_1 and the second CpG_2, on a sliding scale from 5' to 3', then the *clustering score* can be written as:

$$\frac{\text{(Freq. of hoxmCs at both } CpG_1 \text{ and } CpG_2)}{\text{(Freq. of hoxm}CpG_1\text{) x (Freq. of hoxm}CpG_2\text{)}} \qquad \text{(Eq. 1)}$$

When the clustering score is greater than 1, it indicates that the observed clustering of two hoxmCpGs occurs more frequently than probability would suggest based on their underlying individual reactivities alone. In the dataset in Figure 2-1A, there is an average clustering score of 2.5. It should be noted that the rarity of oxmCs at each site has an impact due to the nature of the metric. That is, if two neighboring CpGs are very frequently oxidized in most of the analyzed DNA molecules, there is more likely to be a lower clustering score as there is a relatively high probability that both CpGs will be oxidized. For example, even though CpG numbers 7 and 8 are heavily modified in this particular data set (45% and 64%, respectively) and may appear to be clustered by eye, CpG number 7 only has a clustering score of 0.94 because the frequency of dual oxidation does not exceed the already high likelihood that both would be oxidized. In contrast, CpG pairs that are not frequently oxidized but do occur in the same DNA molecule when they are, such as CpG numbers 22 and 23, will have a relatively high clustering score. One advantage of this score is that it implicitly accounts for any underlying preferences that TET may have and provides an assessment of clustering independent from that.

We would have ultimately preferred to analyze a large dataset using oxBS-seq and next-generation sequencing, but there is utility in analyzing all of the data we have collected from separate BS-seq experiments (Figure 4-3), so that we may benefit from improved depth when

assessing the clustering capacity of TET1 and TET2. The criteria for the reads included were that they 1) had hoxmCs at less than 50% of the CpG sites and 2) were from experiments in which the unreacted mCpG control samples showed no detectable BS-conversion (data not shown). Although there are a different number of sequencing reads included in the two data sets, both are similarly reacted, with 17% and 21% hoxmC, respectively, of the possible sites reacted for the TET1- and TET2-treated samples (Figure 4-3). When examining the total oxidation at each CpG site, we see that TET1 and TET2 exhibit somewhat similar reactivity profiles (Figure 4-5A,B). Despite this, TET2 has generated a greater proportion of clusters of three or more hoxmCs than TET1 (Figure 4-3). Further, while there is only one cluster of four and one of five hoxmCs in the TET1 data, there are several clusters of four to seven hoxmCs in the TET2 dataset. Also, in a TET2-treated sample, when there are seven hoxmCs observed, six of those occur sequentially. In addition to these qualitative observations, the clustering scores for these combined data sets showed that TET2 has a slightly greater propensity to elicit clusters of hoxmCs than TET1 (Figure 4-5C,D); there are both fewer data points with scores less than 1 for TET2 compared to TET1, and the average clustering score for TET2 is 2.1 compared to 1.4 for TET1.

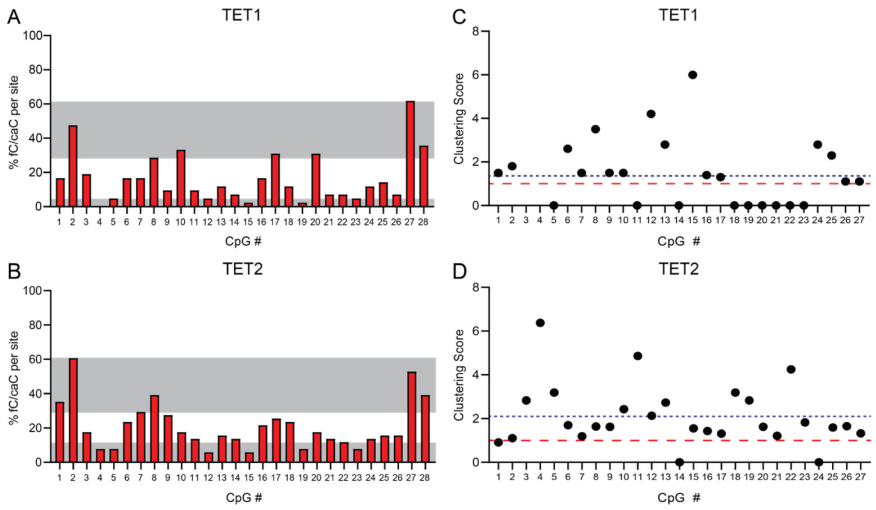


Figure 4-5. Overall frequency of hoxmCs and clustering scores at each CpG in TET1 and TET2 treated BS-seq samples.

A-B) Total hoxmCs observed at each CpG in the amalgamated BS-seq datasets (Figure 4-3) of TET1- (A) and TET2- (B) treated samples. The grey boxes indicate the upper and lower quartiles of the most and least reacted CpG sites,

respectively. **C-D)** Clustering scores for each relative pair of CpGs in the TET1 (C) and TET2 (D) datasets, with the average score for the 27 CpG pairs indicated by a blue dashed line. The red dashed line at the clustering score of 1 delineates the threshold for which clustering is considered to occur at a non-random frequency.

4.2.2: Measuring TET processivity

As mentioned previously, processivity is a common mechanistic feature of DNA-modifying enzymes that can act on sequential target sites. As utilized in the denominator of the clustering score above, if an enzyme acts in a distributive manner, then the probability of two oxidation events occurring is the same as the product of their independent probabilities (Figure 4-7A). To determine if TET can act processively, I designed a novel, enzyme-coupled assay that allows me to quantify any/all oxidation at two neighboring sites on a DNA oligomer (Figure 4-6A). Critically, the two mC sites are embedded in MspI and HaeIII modification-sensitive restriction sites. MspI and HaeIII specifically cleave sequences containing mC, hmC, and fC but not caC. I leveraged the lack of cleavage at caC, together with selective chemical protection of hmC and fC, to interrogate the identity of the modifications at each site. Specifically, I incubate the substrate with TET under conditions of low product formation to operate under the assumption that only one TET enzyme interacts with one DNA molecule. I attach a bulky aldehyde reactive probe (ARP) to fC (Koh et al., 2011) and then glucosylate hmC using T4 β-glucosyltransferase (βGT) (Terragni et al., 2012) so that all TET reaction products (hmC, fC, and caC) are fully protected from restriction cleavage, while unreacted mC is cleaved to give fragments of distinct lengths. By treating each sample with MspI, HaeIII, or both restriction enzymes together and quantifying cleavage products on a denaturing PAGE gel, I can compare the relative amounts of uncleaved DNA (*i.e.* oxidized reaction products). Treating with one enzyme informs us of the frequency of oxidation at a single site and using both enzymes tells us the frequency that oxidation occurs at two sites on the same strand.

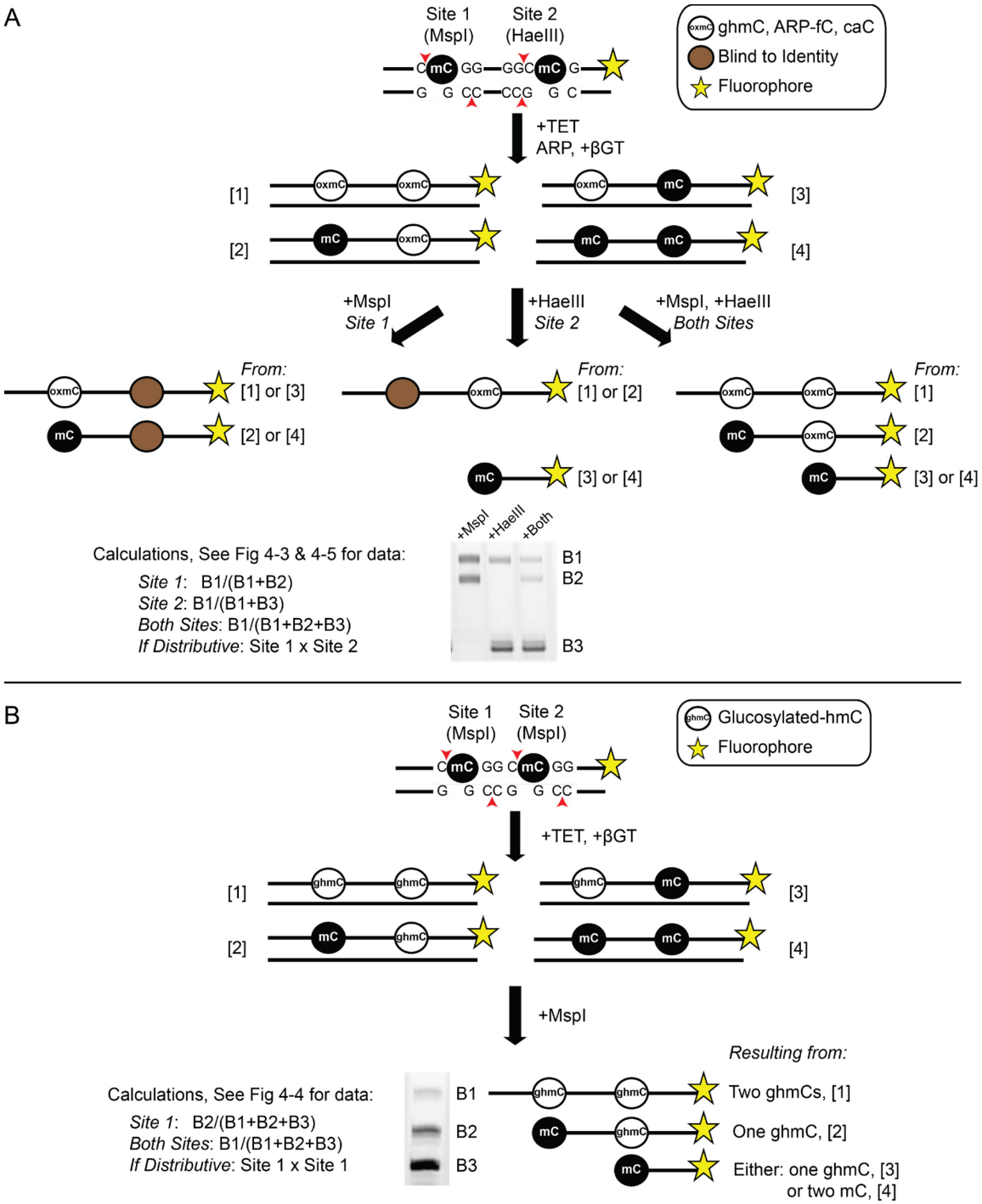


Figure 4-6. R.E.-, gel-based assays for measuring oxidation at two mCpG sites with the goal of assessing strand processivity

Using this assay, we analyzed reactions of substrate (1) with TET1 and TET2 (Figure 4-7B). At 1:1 DNA:enzyme, both TET1 and TET2 oxidize about half of the substrate molecules.

Interestingly, the two mCpG sites are not equally reactive for either TET1 or TET2, varying by 2- to 3-fold. Also surprisingly, the two TET enzymes exhibit different preferences for site 1 and site 2, with TET1 oxidizing more mC at site 2 and TET2 oxidizing more at site 1. For both TET1 and TET2, the fraction of DNA with oxidation at both sites is proportionally similar to the fraction that has oxidation at the less reactive site. This suggests that that oxidation at the two sites is often linked, as there is not a large fraction of DNA with only the less reactive site oxidized. Nevertheless, using the frequencies of oxidation at site 1 and site 2 separately, we can calculate the likelihood that both sites would be oxidized if the enzyme were distributive. For both TET1 and TET2, the amount of DNA with both sites oxidized is greater than that of the theoretical amount that the enzyme would "If Distributive," suggesting that the enzymes are acting processively in these assays.

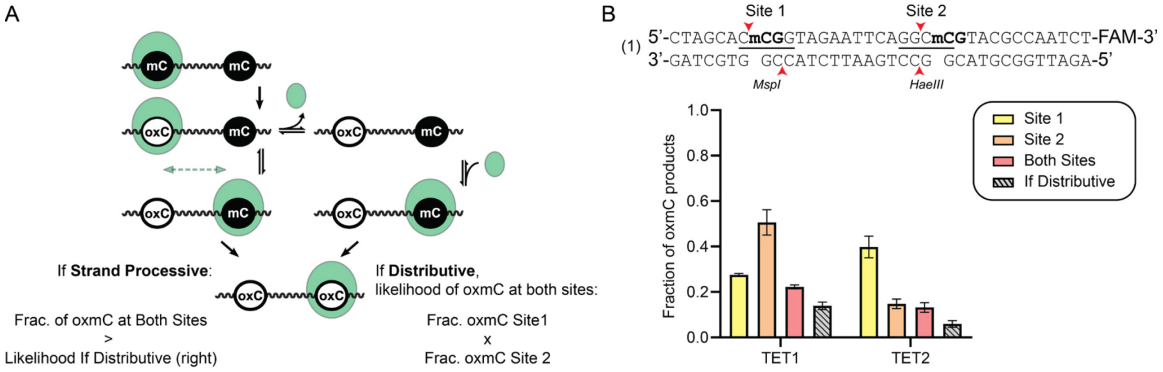


Figure 4-7. TET1 and TET2 exhibit strand processivity when comparing oxidation levels calculated using an indirect, R.E.-based assay

A) Cartoon scheme of simplified mechanisms for strand processive *(left)* and distributive *(right)* oxidation, with corresponding mathematical probabilities. **B)** Relative levels of oxmC products on substrate (1), measured at each site separately and both sites simultaneously via a restriction enzyme-, gel-based assay (see Figure 4-6A for details). The fraction of DNA with oxmC products at each site separately are multiplied to generate the theoretical value for the fraction of DNA that would have two oxmCs "If Distributive."

In order to test for processivity, it is essential that conditions are such that allow for the assumption that only one TET molecule interacts with one DNA molecule. Although the 1:1 DNA:enzyme conditions used are not ideal for making this assumption, we were limited by the

sensitivity of the assay. Nevertheless, despite being confident our analysis of the data in Figure 4-7 is mathematically sound, we wanted to develop another assay to robustly prove that the TET enzymes were acting processively.

Therefore, we performed a modified (but similar) gel-based assay (Figure 4-6) in tandem with whole oligo ESI-MS analysis of the same TET-treated sample (Figure 4-8). In the modified, gel-based assay, the two mCpG sites are each embedded within the target sequence of the MspI restriction enzyme (Figure 4-8A, 4-6B). For ease of analysis, we only protected hmC from cleavage prior to MspI digestion or MS analysis. Due to the design of the oligomer substrate, we can only assess the fraction of DNA with hmC(s) at site 1 and both sites, but not that at site 2 alone (Figure 4-8A,C). By assuming an identical proportion of DNA contained hmC at site 2 as at site 1, we were able to generate the theoretical fraction of DNA with hmCs at both sites if TET were acting distributively (Figure 4-8B). Despite performing the reaction with a more limiting amount of TET1 (2:1 DNA:enzyme), we observed a similar result to the previous assay: The amount of DNA with oxidation at both sites was greater than what we would expect to see if TET were acting distributively. Meanwhile, the whole oligo ESI-MS assay allowed us to detect masses corresponding to the 1) unreacted, dually-methylated substrate strand, 2) a singly glucosyl-hydroxymethylated strand, and 3) a dually-glucosylhydroxymethylated strand (Figure 4-8D). Assuming that the peak corresponding to a single glucosyl-hydroxymethylated strand represents the proportion of sample with a ghmC at either site 1 or site 2, we can calculate the fraction of DNA with a single hmC by dividing half of that signal by the sum of the signals from the substrate peak (with two mC) and two product peaks (one ghmC and two ghmCs). Similar to the gel-based assay, if we assume there are equal proportions of DNA with a single hmC at either site, we can calculate the theoretical fraction of DNA that would have hmCs at both sites if TET were acting distributively (Figure 4-8B). Unlike the gel-based method, however, in this assay the observed and theoretical fractions of DNA with hmCs at both sites are relatively equal, suggesting that TET is acting distributively. These conflicting results gave us pause and we decided that more rigorous work needed to be done in order to confidently report whether the TET enzymes can act processively.

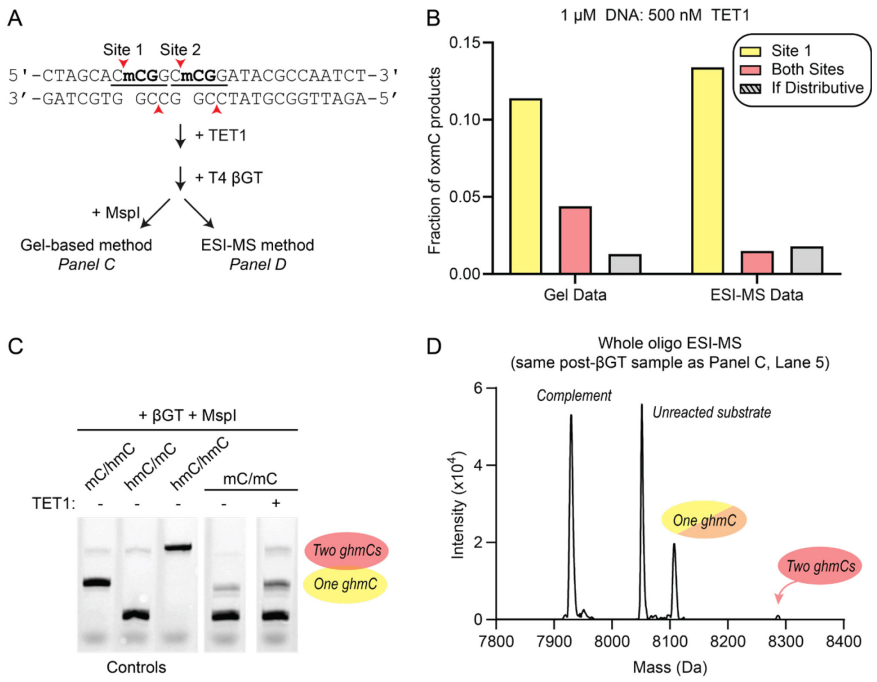


Figure 4-8. Comparison of a modified, gel-based and ESI-MS assays in tandem yield conflicting results regarding TET1 processivity.

A) Experimental steps prior to tandem analyses of the same TET1-treated sample by either a gel-based assay (C) or ESI-MS (D). The gel-based assay is modified (See Figure 4-6B for details) from a similar method used to generate the data in Figure 4-7 and 4-9. **B)** Relative quantification of hmC detected at either a single or both sites via both detection methods. **C-D)** The raw data that is quantified in (B). (C) Notably, this gel-based assay utilizes a single R.E., only allowing for the accurate quantification of hmC at Site 1. To account for incomplete cleavage, the "-TET1" values for fractions of DNA with hmC at Site 1 or Both Sites are subtracted from the "+TET1" values. (D) Prior to collecting this data, we analyzed a control sample with equimolar ratios of mC/mC, ghmC/mC, mC/ghmC and ghmC/ghmC substrates to ensure all were detected equally by ESI-MS (data not shown).

4.2.3: Exploring sequence context preferences of TET enzymes

Despite the inconclusive result regarding processivity, we still wanted to pursue the observation that the two mCpG sites in substrate (1) exhibited different reactivities from both each other and for TET1 and TET2. To our knowledge, that a biochemical difference in substrate preference has been observed for two of the human TET homologs. To explore the mechanistic reason for this discrepancy, we designed two other substrates: In substrate (2), we flipped the sequence in half in order to explore the effect of larger sequence context preferences; in

substrate (3), we only flipped the restriction-enzyme target-sequences (Figure 4-9A). Similar to substrate (1), TET1 and TET2 oxidize each mCpG site to different extents (Figure 4-9B). Surprisingly though, TET1 and TET2 exhibit a similar, roughly 2-fold preference for site 2 in substrate (2). With substrate (3), there are drastically different activity levels on site 2 for TET1 and TET2, with TET1 exhibiting about 2-fold greater activity. Thus, while TET1 appears to consistently prefer site 2, TET2 shows more variability in its preferences.

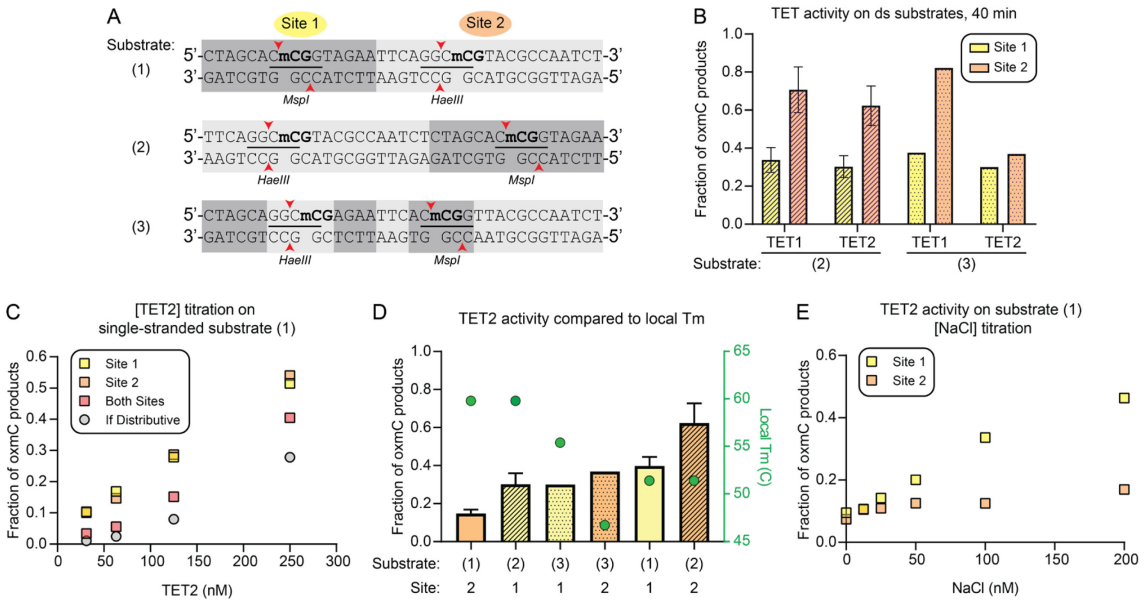


Figure 4-9. Investigating the unequal reactivities of two mCpG sites on the same DNA strand, via the original gel-based assay.

A) Three different substrates are used in the experiments in this figure to test different hypotheses regarding the basis for the differences in oxidation at each site. The color-blocking corresponds substrate (1) and is meant to highlight that substrate (2) is completely flipped, while substrate (3) only switches the local sequence context of the respective R.E. target sequences. **B)** TET1 and TET2 activity on substrates (2) and (3). **C)** TET2 activity on single-stranded substrate (1). **D)** The relative oxidation levels of TET2 on each site measured in (B), as well as Figure 4-7B, and the corresponding local T_m of the 6 surrounding bases. **E)** TET2 activity on substrate (1) in concentrations of NaCl ranging from 0-200 nM.

While it remains unresolved if there are true substrate preference differences for TET1 and TET2, we were intrigued by the variability in site preference exhibited by TET2 and wanted to explore this observation further. First, we incubated TET2 with single-stranded substrate (1) to test the impact of the complement strand (Figure 4-9C). At various concentrations of TET2, the

discrepancy between oxidation levels at the two mCpG sites is lost and each site is reacted to the same extent. While the level of oxidation at both sites is no longer equal to that of the least reacted site, it still occurs at a level higher than we would expect if TET were distributive. Also, this finding suggests that base-pairing has an impact on the relative extent of oxidation that can occur at a particular site. Considering this and that TET2 is a base-flipping enzyme, we hypothesized that the local breathing of duplexed DNA could be a determining factor. To consider this, we ranked the relative oxidation at each site for the three double-stranded substrates and compared the local melting temperatures (T_m) of the +/-3 base regions surrounding each CpG site (Figure 4-9D). Although the dataset is limited, there is a general inverse correlation between oxidation level by TET2 and local T_m of the substrate.

To further explore the role of duplex DNA thermodynamics, we performed incubations of TET2 with substrate (1) at various NaCl concentrations (Figure 4-9E). In all other instances, reactions were performed at 100 nM NaCl, where there is 2.7-fold greater oxidation at site 1 over site 2 for TET2 (Figure 4-7B, 4-8E). In this experiment, the reactivity at site 2 remains relatively consistent for all NaCl concentrations, even with no NaCl in the reaction. Meanwhile, oxidation at site 1 increases linearly with increasing NaCl concentration, effectively decreasing the discrepancy between the two sites at lower NaCl concentrations (Figure 4-9E). Conventionally, increases in salt concentration causes both an increase in DNA melting temperature and decreased protein affinity for DNA. Therefore, we would have expected to see the opposite result if TET's ability to access the Watson-Crick face of the target mC or the strength of the electrostatic interactions with DNA were primary factors regulating TET's substrate preferences.

Despite the lack of clear mechanistic explanation for the variation in reactivity at two mCpGs on the same oligomer, we were still intrigued whether the TET enzymes have a particular sequence context preference, other than the preference of G at the +1 base. To assess this with a greater sequence variation, we re-examined the BS-seq clustering data (Figure 4-3). Although there are only 20 unique sequence contexts with regards to the region +/- 2 bases surrounding each mCpG, I decided to examine the sequence contexts of the most and least reacted sites in these samples (Figure 4-5A,B, grey boxes). For both TET1 and TET2, there is a stark trend: The

most reacted mCpGs have A or T at the -1 base position, while the least reacted mCpGs primarily have G (Table 4-1). This is not likely due to a lack of availability, as there are three times as many mCpGs with G at the -1 position as there are with A (Table 4-2).

TET1				
CpG #	*-2*	*-1*	*+2*	*% hoxmCs per site*
27	T	A	G	62
2	T	A	A	48
28	A	A	C	36
10	A	A	C	33
17	C	T	C	31
20	C	T	G	31
8	G	T	T	29
5	T	G	C	5
12	C	G	C	5
23	T	G	G	5
15	G	G	C	2
19	T	G	C	2
4	T	G	T	0

TET2				
CpG #	*-2*	*-1*	*+2*	*% hoxmCs per site*
2	T	A	A	61
27	T	A	G	53
8	G	T	T	39
28	A	A	C	39
1	T	C	C	35
7	G	T	G	29
9	A	T	G	27
22	T	T	G	12
4	T	G	T	8
5	T	G	C	8
19	T	G	C	8
23	T	G	G	8
12	C	G	C	6
15	G	G	C	6

Table 4-1. Sequence contexts of the most and least reacted mCpG sites in the pUC19 BS-seq samples.

Sequence frequencies			
Base	*-2*	*-1*	*+2*
A	4	4	2
T	10	8	4
C	6	4	12
G	8	12	10

Table 4-2. The base frequency directly surrounding the 28 mCpGs in the 362 bp, pUC19 region.

4.3: Discussion

The original goal of this project was to address both the extent to which TET can elicit clusters of modifications and test for the most likely underlying mechanism, that TET can act processively. Although unable to fully realize our goal of using oxBS-seq to analyze the clustering

of all oxmCs generated on mCpG-containing DNA *in vitro*, we were able to generate data using BS-seq to localize the higher oxidized modifications, fC and caC. The amalgamated datasets consist of TET-treated DNA reads from a variety of reaction conditions with either TET1 or TET2. We observe that both TET1 and TET2 elicit clusters of hoxmCs, in both a qualitative and quantitative manner, despite generally low levels of overall activity. We developed the clustering score to quantify clustering in a manner that takes the oxidation levels of each individual CpG site into account, so that we may assess whether clustering occurs solely due to a greater underlying preference for a particular stretch of CpGs or some additional factor. The finding that both TET1 and TET2 generate an average clustering score greater than 1 for the analyzed region of the pUC19 substrate suggests that an additional factor may be involved. An important caveat of these observations is that the trends may not be relevant when hmC generation is assessed both separately from and together with fC and caC (as all oxmCs). Nevertheless, just as mC clustering must impact hmC clustering so must hmC affect fC/caC clustering, so it is reasonable to predict that hmCpGs will cluster as well.

Structural insight into the network of contacts between TET and DNA initially led us to hypothesize that TET can act processively. We devised an indirect, enzyme- and gel-based assay that allowed us to assess how frequently oxidation of two mCpG sites on the same oligomer occurs in comparison to the likely frequency if the enzyme were only to act distributively. For both TET1 and TET2, we observed about twice as much dual-oxmC product DNA as we would expect to occur by chance. Despite this evidence for processivity, an alternative MS-based assay performed in tandem with a modified, gel-based assay yielded conflicting results: While the observed fraction of DNA with TET-generated hmCs at both CpG sites exceeds the likely fraction if the enzyme were distributive in the gel-based assay, it does not in the MS-based assay.

In the absence of a clear finding regarding processivity, this work has still provided several useful lessons. First, it highlights the importance of assay design. While unlikely that either assay is “wrong,” it is an important reminder that rigorous quantification by multiple methods is required prior to reporting a claim. Second, the gel-based assay provides a relatively facile method for quantifying TET oxidation at two different sites and is the only TET biochemical

experiment to do so, to our knowledge. Tamanaha *et al*. also examined the *in vitro* activity of human TET1 and TET2 on multiple mCpGs but did so as a bulk measurement (Tamanaha et al., 2016), rather than looking at the levels of each individual site.

The gel-based assay also allowed us to explore the unequal reactivities at the two different mCpG sites. Although there was not a consistent pattern for TET1, there was some indication that TET2 activity was impacted by local sequence context. Interestingly, when there is no complement DNA strand and the substrate is single-stranded, TET2 oxidized both sites equally, suggesting local base-pairing surrounding the target mC has a significant impact on relative TET2 activity. In line with this theory, the overall preferences of TET2 for the relative mCpGs in each of the three substrates loosely correlated with T_ms for the region surrounding each mC. However, TET2's preference cannot be solely dictated by the thermodynamics of DNA, as we observe increasing oxidation at site 2 with increasing salt concentrations. Despite this, there is some initial evidence that local sequence context does play a role in TET oxidation preferences, at least when it comes to fC and caC generation. Of the 28 CpGs in the analyzed BS-seq reads, the most consistently oxidized sites are those with an A or T at the -1 position and those that are the least frequently oxidized have G at the -1 position. Notably, however, each of the mCpGs in the short oligomers substrates had C at the -1 position, yet differences in reactivity were still observed.

There are limited previous biochemical studies to test the general sequence preferences of TET, outside of the influence of the +1 base. For example, human TET2-CS has been shown to have more activity on an AT-rich hm/fCpG-containing substrate than a CG-rich one of the same length (Hu, L. et al., 2015). In addition, sequencing of mouse TET1-generated hmC-containing sequences showed 3' sequence specificity at both the +1 and +2 positions, with G and C preferred at the former and a pyrimidine preferred at the latter (Kizaki, Seiichiro et al., 2016). In ESCs, there is TAB-seq data that suggests hmC is enriched in G-rich sequences, but this does not extend to the immediately upstream bases; TCA occur most frequently at the -3 to -1 positions (Yu et al., 2012).

The idea that the general sequence context could be involved in regulating the epigenetic regulation of gene expression would be quite radical; however, it is also clear that GC content varies throughout the genome, which could impact local CpG melting. To that end, it would be interesting to assess both the levels of oxmC clustering in specific genomic regions in various cell types, as well as the relative sequence contexts of the detected oxmCs. Further, the knowledge of both the clustering capacity and sequence preferences will be important for the development of biotechnological tools that utilize TET. For example, if an ideal sequence context were to be identified, it would be useful to embed an unnatural modified cytosine base in that sequence in an activity-based probe (Ghanty et al., 2018). In addition, as epigenetic editing platforms progress (DeNizio et al., 2018), it will be essential to understand the catalytic capacities of TET and other DNA-modifying enzymes to have appropriate expectations for which sequences will be modified and to what extent, once targeted there.

4.4: Methods

The TET proteins used in this chapter were expressed and purified as detailed in Chapter 2 (See section 2.5.2).

4.4.1: Substrate preparation

All DNA oligonucleotides were purchased from Integrated DNA Technologies (IDT), with the exception of the hmC-containing control oligomers (Figure 4-8), which were purchased from Yale Keck Oligonucleotide Synthesis facility. The substrate oligonucleotides, containing modified cytosines, were fluorescently labeled at the 3' end with 6-carboxyfluorescein, with the exception substrate (3), which was 5'-Cy5- and 3'-Cy3-labeled. The complement oligonucleotides did not contain modified cytosines and were unlabeled. All oligonucleotides were HPLC purified and the masses confirmed. All substrates were diluted to 5 or 10 μM (of reactive, methylated substrate, whether single- or double-stranded), prior to use in TET reactions. Duplexed oligos were annealed at ratios of either 1:1.2 or 1:1.5 substrate:complement by heating to 95 °C for 3 min and cooling slowly to 4 °C.

For the clustering experiments, methylated and non-methylated pUC19 DNA was purchased from Zymo Research and is 1 ng/uL. The non-methylated pUC19 DNA was isolated from a methylation-negative strain of bacteria (Dam–, Dcm–). The methylated pUC19 DNA was isolated from the same strain and was then enzymatically methylated at all CpGs by M.SssI methyltransferase. Both pUC19 DNA samples were linearized at position 2177 using ScaI endonuclease.

4.4.2: TET reactions on methylated pUC19 DNA

The mCpG-containing pUC19 substrate was diluted in reaction buffer (50 mM HEPES, pH 6.5, 100 mM NaCl, 1 mM α-KG, 1 mM DTT, and 2 mM sodium ascorbate) so that 100 pg was reacted. Fresh ammonium iron(II) sulfate (Sigma) was added to 75 μM prior to initiation, and purified enzyme, at concentrations ranging from 6-800 nM, was added last to start the reaction (t = 0), bringing the total volume to 25 μL. The reactions were incubated at 37 °C for times ranging from 45 min to 3 hrs. Reactions were quenched by addition of pre-mixed quenching solution (25 μL H_2O, 100 μL Oligo Binding Buffer (Zymo), and 400 μL ethanol) and then purified using the Zymo Oligo Clean & Concentrator kit and eluted in 15 μL of Millipore water.

4.4.3: Bisulfite-sequencing analysis of TET-treated pUC19 DNA

To localize fC and caC in the product pUC19 DNA, bisulfite sequencing was performed using the Epitect Bisulfite Kit (Qiagen). The FFPE protocol was followed, with a longer, 10 hour PCR protocol, in order to minimize mC conversion and maximize fC conversion (Wu, H. et al., 2016).

A 362 base-pair region was amplified from the 2.7 kb pUC19 DNA using bisulfite-specific primers that did not contain any CpGs (forward primer, GGTTATAGTTGTTTTTTGTGTGAAATT GTTATT; reverse primer, CTAACCTTTTACTCACATATTCTTTCCTAC) and allowed for visualization of a single strand of the double-stranded pUC19 substrate. For the PCR reaction, 1 μL of Epimark Hot Start *Taq* DNA Polymerase was mixed with 1X of the supplied reaction buffer and 200 μM of each dNTP, and an optimized PCR method was used (95 °C for 30 s, 40x[95 °C

for 15 s, 58 °C for 30 s, 68 °C for 30 s], 68 °C for 5 min. PCR products were run on a 1.3% agarose, ethidium bromide gel; the 362 bp product band was excised, purified using the Gel DNA Recovery Kit (Zymo), and eluted in 8 µL Millipore water.

TA cloning was performed so that individual DNA molecules could be analyzed. Using the TOPO TA Cloning Kit (Invitrogen), 4 µL of the amplicon was mixed, immediately after gel purification (to maintain 3'-thymine overhangs), with 1 µL salt solution and 1 µL pCR4 vector. The reaction was incubated at room temperature for 30 min-1 hr. Then, 3 µL of the reaction was transformed into Turbo Competent *E. coli* cells (NEB), using a heat shock at 42 °C for 45 s, immediately followed by a 2 min incubation on ice. SOC Broth was added to the cells, and the vial was rotated at 37 °C for 1 hour. The cells were spun down at 9,000 rpm for 5 min, resuspended in 100 µL, and then plated on a KanR/Xgal LB agar plate for incubation at 37 °C for 16 hours and then at room temperature for 8 hours. Individual clones were isolated, cultured, and the plasmid purified using the QIAprep Spin Miniprep kit (Qiagen).

The purified plasmids were sent to GeneWiz (South Plainfield, NJ) for Sanger sequencing analysis. The M13(-21) forward primer was used to visualize the inserted amplicon. Samples in which all CpH were converted to TpH (*i.e.* complete bisulfite conversion) were analyzed for C→T conversion against the amplicon reference sequence.

4.4.4: TET reactions on oligonucleotides

For most of the reactions on the two-mC-containing oligomer substrates, substrates were diluted to 250 nM in the same reaction buffer (50 mM HEPES, pH 6.5, 100 mM NaCl, 1 mM α-KG, 1 mM DTT, and 2 mM sodium ascorbate). Fresh ammonium iron(II) sulfate was added to 75 µM prior to initiation, and 250 nM purified enzyme was added last to start the reaction (t = 0), bringing the total volume to 25 µL. There were three experiments for which these conditions varied slightly: For the experiment in Figure 4-8, 500 nM substrate:1 uM enzyme was used. For the experiment in Figure 4-9C, TET2 was reacted at concentrations ranging from 30-250 nM. For the experiment in Figure 4-9E, NaCl concentrations of the reactions varied from 0-200 nM. All reaction mixtures were incubated at 37 °C for 40 min. Reactions were quenched by addition of

pre-mixed quenching solution (25 µL H_2O, 100 µL Oligo Binding Buffer (Zymo), and 400 µL ethanol). Oligonucleotide products were purified using the Zymo Oligo Clean & Concentrator kit and eluted in 8 µL of Millipore water.

4.4.5: Restriction-enzyme, gel-based analysis of TET-treated oligonucleotide DNA

A graphical representation of the following protocols is shown in Figure 4-6. Prior to downstream analysis, single-stranded DNA was annealed to complement by heating at 95 °C for 3 min then slow-cooling to room temperature (experiment in Figure 4-9C). For the original assay design, in which two different R.E. are utilized (Figure 4-6A), a fC protection step was included in which 4 µL of purified product DNA was diluted into 6 mM HEPES, pH 5 and mixed with 15 mM ARP (Dojindo Molecular Technologies, then Gold Biotechnology, then Cayman Chemical). The 10 µL reaction was incubated at 37 °C for 3 hours. Then, for hmC protection it was diluted directly into CutSmart Buffer, mixed with 2 mM of uridine diphosphoglucose (UDP-Glc) and 1 µL of T4 Phage β-glucosyltransferase (NEB), and incubated at 37 °C for 1-2 hours. The 15 µL reaction was then split in triplicate and incubated with either 1 µL of MspI (NEB), 2 µL of HaeIII (NEB), or a total of 3 µL of both enzymes in 10 µL reactions; these restriction digestions were performed at 37 °C for at least 4 hours. The reaction products were mixed 1:1 with formamide containing bromophenol blue loading dye, heated at 95 °C for 15 minutes, then loaded onto a 7M urea/20% acrylamide/1X TBE gel prewarmed to 50 °C. The gel was imaged for either FAM, Cy3, or Cy5 fluorescence on a Typhoon FLA9500 imaging system (GE Healthcare).

For the modified version of this assay (Figure 4-6B), only the hmC protection step was performed, so the purified reaction products are diluted into Cutsmart and mixed with 2 mM UDP-Glc and 1 µL T4 BGT in a reaction volume of 10 µL; it was incubated at 37 °C for 1 hour. The reaction was then ethanol precipitated and resuspended in 10 µL Millipore water. For restriction enzyme digestion, 1 µL of the sample was diluted into CutSmart Buffer, mixed with 1 µL MspI in a final reaction volume of 10 µL, and incubated at 37 °C for 4 hours. The reaction products were analyzed in the same manner as above. For ESI-MS analysis, the other 9 µL (containing ~2 uM DNA) was sent to Novatia (Newtown, PA).

CHAPTER 5: Conclusions & Future Directions

In summary, this thesis has exposed previously unexplored substrate preferences of human TET enzymes. As biochemists, we are committed to the belief that knowledge of TET's intrinsic reactivities is crucial to contextualizing the biological functions of observed oxmC signatures in the genome. By quantifying the steady-state products generated on different nucleic acid structures via LC-MS/MS, we have identified that TET2 is more promiscuous on different nucleic acid structures than originally thought, and that the nucleic acid identity of the target base itself is the strongest determinant of TET activity (Chapter 2). We also examined the activity of TET on dsDNA with each of the three possible modified cytosine substrates in CpG versus CpH sequence contexts. Despite exhibiting the expected preference for mCpG, TET2 exhibits more tolerance for hmCs when in a CpH context, providing a possible mechanism to explain the depletion of hmCpH observed in cells with relatively high mCpH levels (Chapter 3). Finally, by utilizing creative metrics for quantifying clustering in modification-specific sequencing experiments, we have confirmed that clusters of oxmCs can occur *in vitro*. Although we were unable to definitively confirm whether TET can act processively on multiple mCpGs on a single DNA strand, we obtained evidence to suggest these clusters may be due to local sequence context preferences outside of the +1 base (Chapter 4).

The analysis of steady-state products generated from these different substrates was important, and it sets up crucial downstream experiments that can continue to refine the understanding of TET enzymes. Therefore, in this Chapter, I propose two methods to test the impact of specific reaction steps on the relative substrate preferences of TET. Also, I explore ways in which our current knowledge of the intrinsic properties of TET enzymes can be exploited towards the development of more efficient and specific epigenetic editing complexes.

5.1: Exploring the complexities of TET mechanistic pathways

To date, most of the biochemical experiments with TET enzymes quantify steady-state activity. Single-turnover experiments and other kinetic measurements have likely not been explored due to the complexity of the possible reaction steps of the full TET oxidation pathway (Figure 5-1). As discussed previously, mC, hmC, and fC are each possible substrates and they may be converted to product either through independent, distributive behavior (horizontal pathways) or through an iterative mechanism (highlighted in blue). In this schema, TET is depicted as maintaining the product-substrate flipped into the active site between chemical conversions in the iterative pathway. However, crystal structure data of two different TET homologs (hTET2 and NgTET1) bound to hmC or fC suggest that some conformational change must occur to allow for the exchange of fresh cofactors, Fe(II) and α-ketoglutarate (αKG) (Hashimoto et al., 2015; Hu, L. et al., 2015). Although this exchange may occur while the extrahelical base remains in the active site, it is also possible that it flips back into the duplex while TET maintains non-specific local contacts with DNA, until the target base flips back into the active site for catalysis (Crawford et al., 2016). Even if the latter is not the case the case for catalytic processivity, it remains a possible mechanism through which strand processivity could occur. The nuances and interplay of these mechanisms likely contribute to the difficulty in testing for strand processivity when assessing the steady-state products.

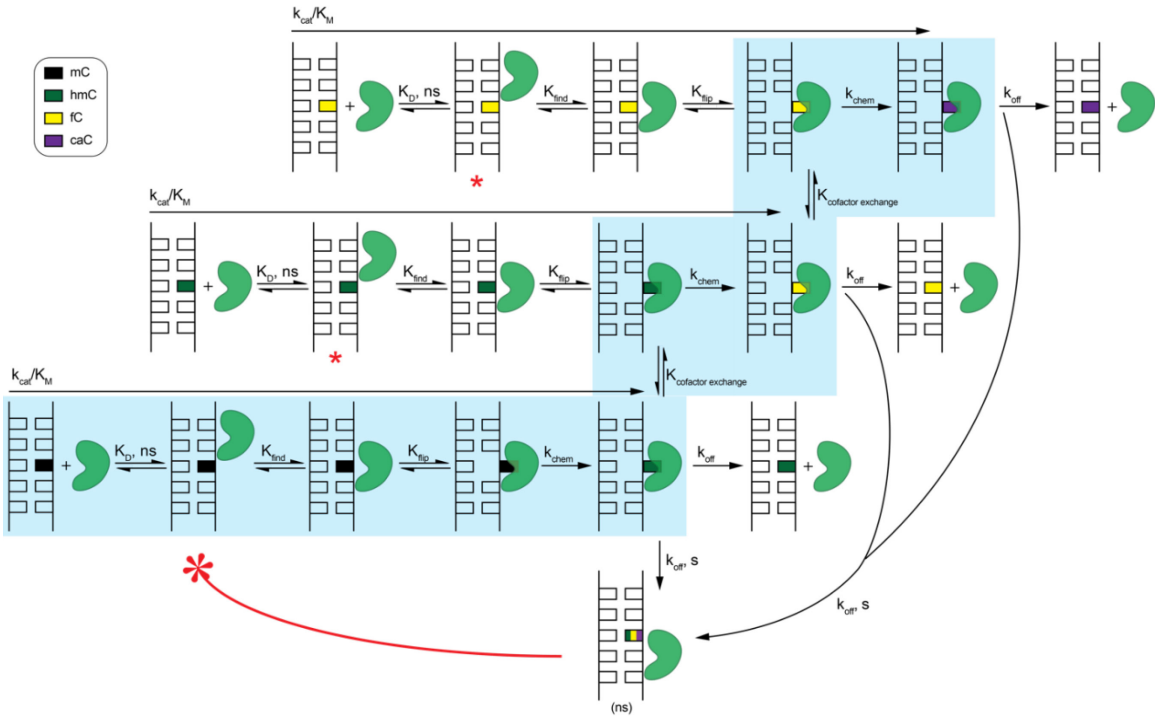


Figure 5-1. Scheme of possible reaction steps for TET enzymes.

A simplified iterative/catalytically processive pathway is highlighted in blue. After specific contact with a target modified cytosine is lost, the enzyme could either dissociate completely (k_{off}), suggesting activity is distributive, or could remain non-specifically bound (k_{off},s). This could allow for strand processivity, mediating TET to another target site of either mC, hmC, or fC, as indicated by red stars. ns, non-specific; s, specific.

In order to fully understand the basis for and potentially exploit TET's inherent substrate preferences, it will be critical for the field to develop methods to test specific reaction steps. Although the chemical reaction steps could be involved, studies of other DNA modifying enzymes implicate mechanisms of substrate engagement in dictating their different substrate preferences. The binding of most DNA-modifying enzymes to their target sites is thought to include a minimum of three binding steps: binding to non-specific DNA, finding the target site, and engaging the nucleotide in the active site (Hendershot and O'Brien, 2017). To achieve the latter, all but one known DNA-modifying enzyme to date (DNA repair glycosylase AlkD) employ nucleotide-flipping. It is less clear, however, how TET finds target sites and positions itself for that reaction step. Nevertheless, I propose that measuring dissociation rates and the

equilibrium constants for nucleotide-flipping of TET enzymes with various substrates will illuminate the mechanisms regulating TET's substrate preferences.

5.1.1: Measuring dissociation rates of TET enzymes from different substrates

Previously, binding affinities (K_D) have been determined for TET2 on DNA. Human TET2-CS exhibited nearly identical affinities for C-, mC-, hmC- and f-CpG-containing dsDNA, with K_Ds of about 0.5-1μM (Hu, L. et al., 2015). The binding affinity of TET2 on mCpH-containing DNA, or even mCpG-containing DNA embedded in a range of local sequence contexts, has not been measured. It is generally assumed that TET enzymes also exhibit relatively similar affinities for non-specific DNA because the majority of the protein-DNA contacts are between TET and the phosphodiester backbone, but this assumption has not been proven. I am also interested in comparing TET enzymes' non-specific and mC-specific affinities with DNA versus RNA. In particular, considering that TET2 exhibited such a strong preference for deoxribomethylcytosine, it would be interesting if enhanced or diminished affinity for dmC over ribomethylcytosine contributed to this mechanism.

Unfortunately, due to the current concentration limits of our recombinant protein preparations, we are unable to generate complete equilibrium binding curves to calculate K_Ds. In addition, the binding, searching, and the conformational rearrangements that ultimately result in flipping are expected to be very rapid, occurring on a millisecond timescale (Hendershot and O'Brien, 2017). Nevertheless, I have begun to address the potential role of substrate engagement mechanisms in the preference for DNA over RNA by assessing the dissociation rates of TET from different nucleic acid-TET complexes. While it is unknown at which point in the possible substrate engagement pathway that the enzyme most likely to falls off DNA in our simplified schema, measuring dissociation rates will allow us to examine nucleic acid preferences broadly across the contributing reaction steps (Figure 5-2A).

To measure dissociation rates, I utilized the fluorescence anisotropy detection of a stopped flow apparatus (Kintek). For these preliminary experiments, I used FAM-labeled versions of the single-stranded 16mer DNA and RNA substrates, each containing a single mCpG, that I

previously described in Chapter 2 (Figure 2-1A). These experiments were performed by first forming TET2-DNA-/RNA-FAM complexes by incubating 1.1 μM TET2 with 100 nM of the FAM-labeled ssDNA or ssRNA probe (for 10-fold excess enzyme) for 1 hour at room temperature. The buffer conditions were similar to those used in the majority of the biochemical experiments in this thesis (50 mM HEPES, pH 6.5, 100 mM NaCl, 1 mM DTT, 2 mM sodium ascorbate, 75 μM ammonium iron(II) sulfate), with the exception that 1 mM N-oxalylglycine (NOG), rather than αKG, was used to inhibit catalysis. The addition of TET2 resulted in an increase in fluorescence anisotropy, relative to that of the probe mixed with only buffer.

To initiate dissociation, the TET2:DNA-/RNA-FAM complexes were each rapidly mixed with an equal volume of unlabeled versions of the DNA/RNA probes at 10 uM (18X excess) (Figure 5-2B,C) and fluorescence anisotropy was measured for 1-3 min. Unfortunately, for both the TET2:DNA-FAM and TET2:RNA-FAM complexes, these conditions were not sufficient to fully dissociate the initial complexes after 60 seconds (Figure 5-2B-C) or up to 3 minutes (data not shown). To calculate accurate k_{off} values (or half-lives, $t_{1/2}$, of the complexes), ideally the dissociation curves would return to baseline. Nevertheless, I performed a follow-up experiment in which I mixed the TET2:RNA-FAM complex with an excess of unlabeled DNA that contained multiple mCpGs, rather than just one (Figure 5-2D). Interestingly, the TET2:RNA-FAM complex exhibits enhanced dissociation in the presence of this multi-mCpG-containing DNA quench in the first 60 seconds. Even after 15 minutes, the TET2:RNA-FAM complex had still not completely dissociated, but it does exhibit a 2-fold change in anisotropy, similar to the scale of the change observed with TET2:DNA-FAM in excess of the single mCG-DNA quench after 60 seconds. Thus, these findings suggest that TET2 has a much slower off-rate from RNA compared to DNA.

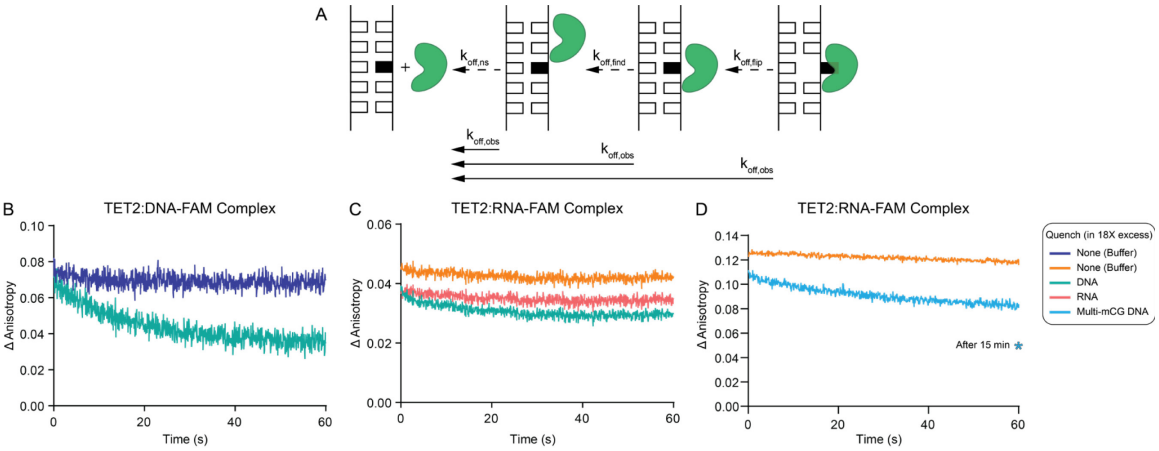


Figure 5-2. Fluorescence polarization can be used to measure the dissociation rates of TET from different substrates.

A) Measured k_{off} rates could reflect dissociation from any of these different levels of substrate engagement (*i.e.*, $k_{off,obs}$). **B-D)** Curves represent the change in anisotropy over time when pre-formed complexes of the catalytic domain of TET2 with either a DNA-FAM (B) or RNA-FAM probe (C,D) are rapidly mixed with either just buffer or an excess of unlabeled quench substrate. The change in anisotropy was calculated by subtracting the relative baseline measurement of non-complexed DNA- or RNA-FAM probes mixed with buffer.

5.1.2: Proposal to investigate the role of nucleotide flipping in TET specificity

To further probe the mechanisms of substrate engagement regulating substrate preference, I propose taking kinetic measurements of nucleotide-flipping. From the crystal structures of TET homologs bound to DNA we know that TET engagement causes DNA kinking and the nearly 180° rotation of the target mC nucleotide out of the DNA duplex and into the active site of TET. As mentioned, this is a common mechanism for nearly all DNA-modifying enzymes. However, the specific steps that proceed this action can vary for different enzymes. For example, there is some evidence to suggest Uracil DNA glycosylase (UDG) engages with extrahelical bases as they spontaneously emerge from duplex DNA via thermal fluctuations during its search for lesions. However, there has also been evidence suggesting that UDG intentionally samples different nucleotides in order to find lesions. Many of the DNA-modifying proteins likely combine these two mechanisms. For example, TDG is thought to precipitate the extrusion of the target base from the duplex and then capture it in the active site;

however, it too is thought to be aided by the fluctuations of G:fC and G:caC base pairs in its ability to fully flip these targets into the active site (Dai et al., 2016; Maiti et al., 2013).

From structural and biochemical data, I have speculated how different substrates can impact the substrate engagement reaction steps. First, because TET has exhibited robust activity in single-stranded DNA, contacts with the opposite strand of duplexed DNA are likely not necessary for the torsion of the DNA required for nucleotide flipping. With regards to both the preference for DNA over RNA and tolerance for hmCpH, there appears to be some relationship between the strength of the interactions with the phosphodiester backbone surrounding the target base and either the strength of base-stacking or helical conformation. Although it would prove informative to test each of the involved reaction steps in substrate engagement if possible, I believe studying the relative occupancy of TET with a flipped-out target nucleotide when bound to substrate DNA will be the most informative to begin with.

Recently, colleagues at the University of Maryland have used ^{19}F 1D NMR to study the role of nucleotide flipping in TDG specificity (Dow et al., 2019). ^{19}F, which is hypersensitive to its local environment, allows for a much greater range of chemical shifts than the corresponding ^{1}H nuclei. By embedding target nucleotides that contain ^{19}F at the 2'-position, they were able to measure the distinct chemical shifts resulting from distinct chemical environments, reflecting occupancies of the nucleotide in different positions; specifically, whether stacked in the duplex or flipped into the TDG active site.

In this study, the authors compared the peak shifts for free versus TDG-bound T(^{19}F)-containing DNA, when the T(^{19}F) is base-paired with either a mismatched G or the canonical A. These experiments were performed 1000-fold above the K_D of TDG with non-specific DNA, thus operating under the assumption that TDG is stably bound to DNA. The T•G-DNA exhibits two NMR peaks when bound to TDG, suggesting that the 2'-^{19}F occupies two distinct chemical environments, one where the base remains stacked in the DNA duplex ($I_{stacked}$) and the other where the base is flipped into the active site of TDG ($I_{flipped}$). In contrast, T•A-DNA exhibits a single NMR peak when bound to TDG, with the same chemical shift as free DNA. Equilibrium

constants for reversible nucleotide flipping (K_{flip}) are calculated for both conditions by taking the ratio of the intensity of the two relative peaks ($I_{flipped}/I_{stacked}$) (Figure 5-3A).

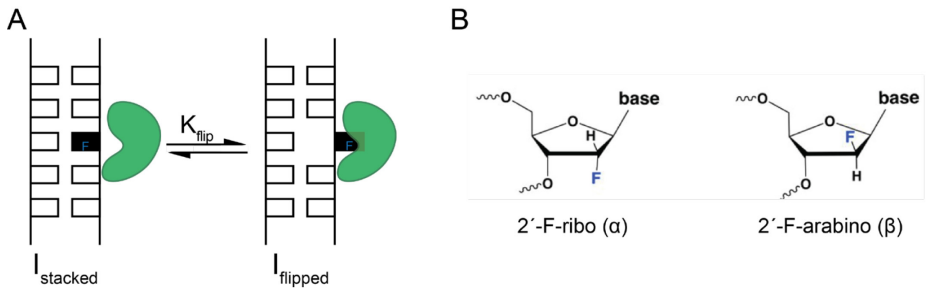


Figure 5-3. ^{19}F NMR can be used to measure the propensity for nucleotide flipping.

A) The peak shifts in ^{19}F 1D NMR correspond to when the ^{19}F-containing nucleotide is either "stacked" in the DNA helix (left) or "flipped" into the active site of the enzyme. The nucleotide flipping equilibrium constant (Kflip) is calculated as the ratio of these two respective peak intensities ($I_{flipped}/I_{stacked}$) **B)** The ^{19}F-containing nucleotide can be synthesized in either the 2'-F-ribo or 2'-F-arabino conformations.

In collaboration with our colleagues, we are thus poised to ask about the relative nucleotide flipping of TDG, but also TET, with fC(^{19}F)-containing DNA which has already been synthesized. Although it would be ideal to also analyze TET nucleotide flipping with mC and hmC, the synthesis can be cumbersome. Nevertheless, using fC(^{19}F)-substrates could still provide insight into the mechanisms dictating the CpG versus CpH preference of TET, since in Chapter 3, we observe a similar difference in the relative ratios of activity for fCpG versus fCpH substrates as we do with those with mCpG versus mCpH. The ^{19}F-containing substrates in the published TDG experiments contain 2'-F-arabino, which is compatible with B-form DNA. However, 2'-F-ribo forms, which are compatible with A-form RNA, of the modified cytosine nucleotides could also be generated (Figure 5-3B). Experiments with these different target nucleotides embedded in either DNA or RNA strands could specifically test the role of nucleotide flipping in our proposed model of DNA versus RNA specificity in Chapter 2.

In summary, I propose that the next logical expansion of the work in this thesis is to test the proposed mechanisms that may be dictating TET substrate specificity. Both of these

approaches, either examining K_Ds/off-rates or nucleotide flipping, require greater concentrations of protein than our current expression and purification methods currently allow. Other members of the lab have been working towards the goal of making more concentrated, active TET proteins in an *E. coli* rather than insect cell expression system. Nevertheless, I anticipate that insights gained in this thesis, as well as those proposed in this section, are critical for towards the overarching goal of being able to interpret the biological purpose of TET's varied activity preferences, as well as for the development of optimal epigenetic editing tools.

5.2: Exploring the future of epigenome editing

This section of the chapter was adapted from parts of the following review article:

DeNizio, J.E., Schutsky, E.K., Berrios, K., Liu, M.Y., Kohli, R.M. Harnessing natural DNA modifying activities for editing of the genome and epigenome. *Curr. Opin. Chem. Biol.* 2018, 45, 10-17.

5.2.1: Introduction to editing complexes

The introduction of site-specific DNA modifications to the genome or epigenome presents great opportunities for manipulating biological systems. Such changes are now possible through the combination of DNA-modifying enzymes with targeting modules, including dCas9, that can localize the enzymes to specific sites, potentially enabling more direct accounting of cause and effect at specific loci. Although the effect of targeted DNA methylation on downregulating gene expression is variable and often modest, there is stronger evidence that the demethylation of a specific loci can result in gene activation (Lei et al., 2018). Nevertheless, there are still many unknowns prohibiting the effective use of these tools, such as: which loci are regulated by DNA methylation? If they are, how many mCpGs in which regulatory regions need to be demethylated to achieve phenotypic changes? And, what factors are regulating the demethylation of these mCpGs? I anticipate that the exploitation of the intrinsic features of TET enzymes, particularly with regards to clustering capacity and substrate specificity, will prove critical towards advancing this technology. In this section, I will review how TET enzymes, as well as other DNA modifying

enzymes, have been utilized in the past and propose new alternatives to how they can be used in the future, highlighting ways in which they can enhance both editing efficiency and/or specificity.

To provide a framework for understanding these methodological developments, we offer a generic definition of the editing complexes (Figure 5-4). The typical formula for an editing complex involves a DNA targeting module (TM) partnered with one or more DNA modifying modules (MM). The TMs can recognize specific nucleotide sequences including certain modified bases within those sequences. MMs can be partnered with accessory modules (MM_X) or can be variants of the natural enzyme with altered biochemical properties (MM*) or with tunable activity (MM†).

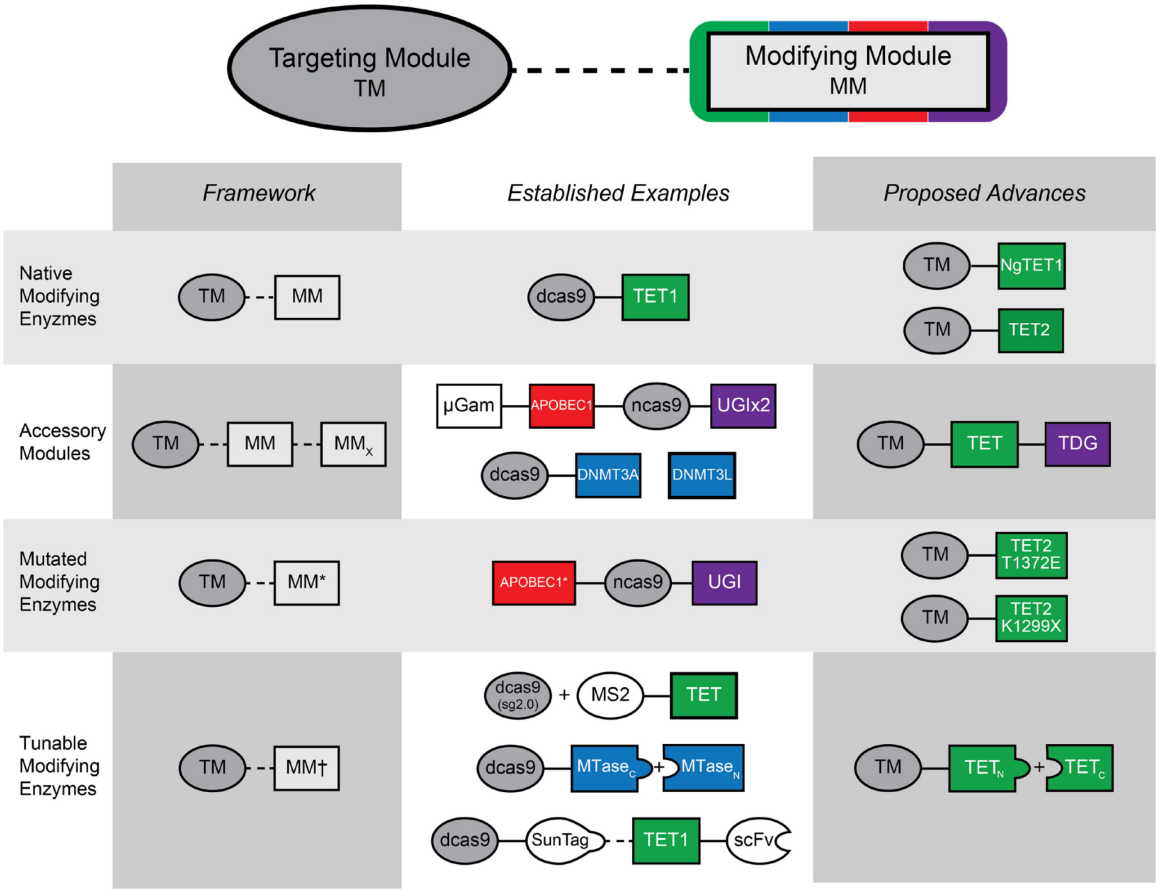


Figure 5-4. A formulaic approach for assessing current epigenetic editing complexes and proposing new advances.

Editing complexes can be described as consisting of a Targeting Module (TM) and Modifying Module (MM), which can be classified based on other distinguishing features, including the addition of accessory modules (MM_X), variants of the native enzymes with altered activities (MM*), or tunable variants

with altered spatiotemporal control (MM†). Established examples of these complexes are shown, as well as anticipated advances based on the knowledge of the structure and function of TET enzymes. Lines connecting the modules represent fusions or linkers, and a plus sign indicates that the two relative entities were co-expressed separately. The dashed line indicates that both fused and unfused/co-expressed versions have been made.

While potential TMs have been well reviewed recently (Hu, J. H. et al., 2016; Nelson and Gersbach, 2016), I will briefly note the available constructs to frame our discussion of DNA-modifying enzymes. The original method for programmable targeting of DNA sequences employed zinc-finger proteins (ZFPs) which recognize a trinucleotide motif and can be concatenated to recognize longer stretches of DNA (Gersbach et al., 2014). The discovery of transcription activator-like effector (TALE) proteins from plant pathogens later offered a means for targeting single bases, and, like with ZFPs, concatenation of TALEs results in selective recognition of longer DNA motifs (Joung and Sander, 2013). Structure-function insights into the TALE residues that contact nucleobases have permitted the development of constructs that differentiate unmodified cytosine from 5mC and ox-mCs (Kubik et al., 2015; Maurer et al., 2016; Rathi et al., 2017; Zhang, Y. et al., 2017). Such modified TMs (TM*) have not yet been employed with different MMs, but the combination would make for a logical and anticipated advance. While ZFPs and TALEs are still being employed, the CRISPR/Cas system offers a more adaptable solution to the challenges of designing TMs, in that different guide RNAs (gRNA) could be used to achieve targeting to multiple locations with a single protein scaffold, rather than having to manipulate individual proteins for each target. Although Cas9 was originally exploited for targeted DNA cleavage, using catalytically inactive Cas9 (dCas9) or nicking variants (nCas9) allows for retention of localization capabilities and partnering with DNA-modifying enzymes as MMs. The discovery of new CRISPR/Cas systems will likely expand the TM choices yet further (Makarova et al., 2015). Across the variety of combinations, TM and MMs can be directly fused, which requires optimization of module ordering and linker lengths, or can be brought together by exploiting protein-protein or protein-RNA interactions (in the case of Cas9) to promote co-localization.

5.2.2: Strategies towards efficient levels of targeted TET oxidation

Due to the varied nature of TET homologs, there are many opportunities to utilize naturally occurring homologs as MMs that could possibly yield efficient editing. With regards to the human homologs, one study utilized the catalytic domains of TET1, TET2, and TET3 and observed the greatest demethylation efficiency with TET2. Despite this, TET1, coupled to ZFPs, TALEs, and more commonly now to dCas9 (Chen, H. et al., 2014; Choudhury et al., 2016; Liu, X. S. et al., 2016; Lo et al., 2017; Maeder et al., 2013; Morita et al., 2016; Xu, X. et al., 2016), is most frequently employed. Intriguingly, plants have a direct demethylation pathway through the protein ROS1, a DNA glycosylase that directly excises 5mC. After fusion with a yeast Gal4-DNA binding domain as a TM, ROS1 has also been employed for direct demethylation and specific reactivation of silenced genes (Parrilla-Doblas et al., 2017). The ROS1 construct represents the concept of using a native MM out of its native context, and such strategies might be more generally employed; for example, the well-behaved NgTET1 (Hashimoto, Pais et al., 2014; Hashimoto et al., 2015) could be employed instead of mammalian TET enzymes.

Additionally, partnering of native proteins to accessory modules (MM_X) can be used to enhance editing. The initial groundbreaking effort for genome editing employed APOBEC1, which is a single-stranded DNA deaminase that catalyzes a C>U transition, as the MM, along with dCas9, which plays a dual role in targeting and in unwinding of the target site so that APOBEC1 may access a ssDNA substrate. With this construct, in *cis* incorporation of the accessory protein UGI – a small phage-derived protein that potently inhibits UDG – prevents downstream repair and greatly increases the efficiency of editing processes (Komor et al., 2016). This idea has also been applied to epigenome editing: fusing DNMT3L as an MM_X to DNMT3A enables multimerization and promotes processive methylation, which potentially helps replicate the natural tendency of methylation clustering (Stepper et al., 2017). It is feasible that the MM_X approach could be extended in the future by combining multiple MMs in *cis*; for example, locus-specific demethylation could be enhanced by fusing TDG to TET-containing constructs.

There are several factors that have been shown to influence the range of activity of a MM surrounding a targeted region. For example, distance from the sgRNA anchoring site when the

TM is dCas9, linker length between the TM and MM, and size of the chosen MM have all been shown to influence the distances over which epigenetic DNA editing is observed (Lei et al., 2018). Similar to the effect seen by fusing DNMT3L as an MM_X (Huang, Y. H. et al., 2017), the SunTag system has been employed to expand the range of activity. This system relies on multiple antibody epitopes that can be recognized by a single-chain variable fragment (scFv). Fusing dCas9 with the epitopes and the MM to the scFv can result in multiple proteins being targeted to the designated locus, amplifying the activity (Huang, Y. H. et al., 2017; Morita et al., 2016). Attaching a TET enzyme to the SunTag system has resulted in several instances in which demethylation was more efficiently spread over a region (Gallego-Bartolome et al., 2018; Morita et al., 2016). Although it will likely be loci specific, the number of SunTags could also manipulated in order to control the number of TET proteins recruited, potentially allowing for control of the relative range that gets oxidized. Theoretically, this level of control could help identify the minimal regions of oxidation necessary for gene activation. To a similar end, if the strand processivity of a particular TET homolog could be enhanced, via mutation enhancement or an interacting protein, we may further amplify the demethylation efficiency or gene activation.

5.2.3: Harnessing intrinsic TET preferences for specific editing

Aside from harnessing native enzymes, existing structure-function knowledge has been exploited to generate variants of DNA-modifying enzymes with increased specificity. Both rational and screening-based approaches have been utilized to assess and/or evolve activities, with the most significant advances coming in genome editing rather than epigenome editing thus far. For example, prior biochemical work isolated sequence determinants of activity for each AID/APOBEC family member (Kohli et al., 2010), and targeted alteration to this region and others in the active site permitted increased precision in base editing (Kim et al., 2017).

Meanwhile, this thesis has demonstrated that TET enzymes, at least *in vitro*, can exhibit relatively promiscuous activity, particularly on different nucleic acid structures. TET has only been targeted to DNA thus far, but it is important to consider that RNA oxidation may be occurring and confounding the phenotypic interpretation of the desired DNA editing. Although not relevant to

dcas9 scaffolds, several other classes of Cas systems have been shown to facilitate RNA-dependent RNA targeting (Cox et al., 2017; Strutt et al., 2018). As we learn more about the different Cas systems and different ones are employed, this may become more of an issue, necessitating the use of a TET mutant with diminished RNA activity. Further, current TET-targeting epigenetic editing systems do not have the specificity to allow us to ascertain the independent roles of oxmCs both *in vitro* and *in vivo*, as all three TET isozymes are known to make all three ox-mCs. Recently discovered hmC-dominant mutants of human TET2 (Liu, M. Y. et al., 2017) or NgTET1 (Hashimoto et al., 2015), or a TET homolog from the mushroom *C. cinerea* that produces mostly fC (Zhang, L. et al., 2014), could be utilized in editing systems to more directly dissect the roles of individual oxmCs.

Lastly, systems within which TET activity can be spatiotemporally regulated to minimize expression could prove advantageous for improving specificity. Splitting Cas9 to create a ligand-dependent activation variant has been an effective means to gain spatiotemporal control over TMs (Oakes et al., 2016), an analogous protein-driven method for controlling MMs is to split the relevant enzyme so that it only reconstitutes at the desired site. A recent study using a split methyltransferase fused to dCas9 exhibited high specificity to target sites and suggested that the strategy limited off-target effects (Xiong et al., 2017). Although not yet attached to a targeting scaffold, TET2 has also been split and placed under the control of a small molecule modulator: Two inactive fragments were individually fused with either FKBP12 or FRB, and upon the addition of rapamycin, TET2 reassembled into a functional enzyme (Lee, M. et al., 2017). Another method for control involves tagging the MM, such as TET1, with the RNA-binding protein MS2, which is directed to dCas9 by incorporation of MS2 hairpin-binding sites into the gRNA (Hess et al., 2016; Xu, X. et al., 2016).

BIBLIOGRAPHY

Arioka, Y., Watanabe, A., Saito, K., and Yamada, Y. (2012). Activation-Induced Cytidine Deaminase Alters the Subcellular Localization of Tet Family Proteins. PLoS One *7,* e45031.

Babcock, M.S., Pednault, E.P., and Olson, W.K. (1994). Nucleic acid structure analysis. Mathematics for local Cartesian and helical structure parameters that are truly comparable between structures. J. Mol. Biol. *237,* 125-156.

Bachman, M., Uribe-Lewis, S., Yang, X., Burgess, H.E., Iurlaro, M., Reik, W., Murrell, A., and Balasubramanian, S. (2015). 5-Formylcytosine can be a stable DNA modification in mammals. Nat. Chem. Biol. *11,* 555-557.

Bachman, M., Uribe-Lewis, S., Yang, X., Williams, M., Murrell, A., and Balasubramanian, S. (2014). 5-Hydroxymethylcytosine is a predominantly stable DNA modification. Nat. Chem. *6,* 1049-1055.

Bakan, A., Meireles, L.M., and Bahar, I. (2011). ProDy: protein dynamics inferred from theory and experiments. Bioinformatics *27,* 1575-1577.

Basanta-Sanchez, M., Wang, R., Liu, Z., Ye, X., Li, M., Shi, X., Agris, P.F., Zhou, Y., Huang, Y., and Sheng, J. (2017). TET1-Mediated Oxidation of 5-Formylcytosine (5fC) to 5-Carboxycytosine (5caC) in RNA. ChemBioChem *18,* 72-76.

Bayly, C.I., Cieplak, P., Cornell, W., and Kollman, P.A. (1993). A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. J. Phys. Chem. *97,* 10269-10280.

Bellacosa, A., and Drohat, A.C. (2015). Role of base excision repair in maintaining the genetic and epigenetic integrity of CpG sites. DNA Repair (Amst) *32,* 33-42.

Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A., and Haak, J.R. (1984). Molecular dynamics with coupling to an external bath. J. Chem. Phys. *81,* 3684-3690.

Bergman, Y., and Cedar, H. (2013). DNA methylation dynamics in health and disease. Nat. Struct. Mol. Biol. *20,* 274-281.

Bird, A. (2011). The dinucleotide CG as a genomic signalling module. J. Mol. Biol. *409,* 47-53.

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. *8,* 1499-1504.

Bohnsack, K.E., Hobartner, C., and Bohnsack, M.T. (2019). Eukaryotic 5-methylcytosine (m(5)C) RNA Methyltransferases: Mechanisms, Cellular Functions, and Links to Disease. Genes (Basel) *10,* 10.3390/genes10020102.

Booth, M.J., Branco, M.R., Ficz, G., Oxley, D., Krueger, F., Reik, W., and Balasubramanian, S. (2012). Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Science *336,* 934-937.

Booth, M.J., Raiber, E.A., and Balasubramanian, S. (2015). Chemical methods for decoding cytosine modifications in DNA. Chem. Rev. *115,* 2240-2254.

Breyer, W.A., and Matthews, B.W. (2001). A structural basis for processivity. Protein Sci. *10,* 1699-1711.

Case, D.A., Cerutti, S.D., Cheatham, T.E.,3rd, Darden, T.A., Duke, R.E., Giese, T.J., Gohlke, H., Goetz, A.W., Greene, D., Homeyer, N.*, et al.* (2017). AMBER 2017.

Chen, C.C., Wang, K.Y., and Shen, C.K. (2012). The mammalian de novo DNA methyltransferases DNMT3A and DNMT3B are also DNA 5-hydroxymethylcytosine dehydroxymethylases. J. Biol. Chem. *287,* 33116-33121.

Chen, H., Kazemier, H.G., de Groote, M.L., Ruiters, M.H., Xu, G.L., and Rots, M.G. (2014). Induced DNA demethylation by targeting Ten-Eleven Translocation 2 to the human ICAM-1 promoter. Nucleic Acids Res. *42,* 1563-1574.

Chen, L., Lin, H., Zhou, W., Xiong, Y., Ye, D., and Guan Correspondence Kun-Liang. (2018). SNIP1 Recruits TET2 to Regulate c-MYC Target Genes and Cellular DNA Damage Response. Cell Reports *25,*

Choi, N.Y., Bang, J.S., Lee, H.J., Park, Y.S., Lee, M., Jeong, D., Ko, K., Han, D.W., Chung, H., Kim, G.J.*, et al.* (2018). Novel imprinted single CpG sites found by global DNA methylation analysis in human parthenogenetic induced pluripotent stem cells. Epigenetics *13,* 343-351.

Choudhury, S.R., Cui, Y., Lubecka, K., Stefanska, B., and Irudayaraj, J. (2016). CRISPR-dCas9 mediated TET1 targeting for selective DNA demethylation at BRCA1 promoter. Oncotarget *7,* 46545-46556.

Ciccarone, F., Valentini, E., Zampieri, M., and Caiafa, P. (2015). 5mC-hydroxylase activity is influenced by the PARylation of TET1 enzyme. Oncotarget *6,* 24333-24347.

Cliffe, L.J., Kieft, R., Southern, T., Birkeland, S.R., Marshall, M., Sweeney, K., and Sabatini, R. (2009). JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J biosynthesis in genomic DNA of African trypanosomes. Nucleic Acids Res. *37,* 1452-1462.

Coey, C.T., Malik, S.S., Pidugu, L.S., Varney, K.M., Pozharski, E., and Drohat, A.C. (2016). Structural basis of damage recognition by thymine DNA glycosylase: Key roles for N-terminal residues. Nucleic Acids Res. *44,* 10248-10258.

Cooper, D.N., and Youssoufian, H. (1988). The CpG dinucleotide and human genetic disease. Hum. Genet. *78,* 151-155.

Cortazar, D., Kunz, C., Selfridge, J., Lettieri, T., Saito, Y., MacDougall, E., Wirz, A., Schuermann, D., Jacobs, A.L., Siegrist, F.*, et al.* (2011). Embryonic lethal phenotype reveals a function of TDG in maintaining epigenetic stability. Nature *470,* 419-423.

Cortellino, S., Xu, J., Sannai, M., Moore, R., Caretti, E., Cigliano, A., Le Coz, M., Devarajan, K., Wessels, A., Soprano, D.*, et al.* (2011). Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. Cell *146,* 67-79.

Cox, D.B.T., Gootenberg, J.S., Abudayyeh, O.O., Franklin, B., Kellner, M.J., Joung, J., and Zhang, F. (2017). RNA editing with CRISPR-Cas13. Science

Crawford, D.J., Liu, M.Y., Nabel, C.S., Cao, X.J., Garcia, B.A., and Kohli, R.M. (2016). Tet2 Catalyzes Stepwise 5-Methylcytosine Oxidation by an Iterative and de novo Mechanism. J. Am. Chem. Soc. *138,* 730-733.

Dai, Q., Sanstead, P.J., Peng, C.S., Han, D., He, C., and Tokmakoff, A. (2016). Weakened N3 Hydrogen Bonding by 5-Formylcytosine and 5-Carboxylcytosine Reduces Their Base-Pairing Stability. ACS Chem. Biol. *11,* 470-477.

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

Delatte, B., Wang, F., Ngoc, L.V., Collignon, E., Bonvin, E., Deplus, R., Calonne, E., Hassabi, B., Putmans, P., Awe, S*., et al.* (2016). Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. Science *351,* 282-285.

DeNizio, J.E., Liu, M.Y., Leddin, E.M., Cisneros, G.A., and Kohli, R.M. (2019). Selectivity and Promiscuity in TET-Mediated Oxidation of 5-Methylcytosine in DNA and RNA. Biochemistry *58,* 411-421.

DeNizio, J.E., Schutsky, E.K., Berrios, K.N., Liu, M.Y., and Kohli, R.M. (2018). Harnessing natural DNA modifying activities for editing of the genome and epigenome. Curr. Opin. Chem. Biol. *45,* 10-17.

Dewage, S.W., and Cisneros, G.A. (2015). Computational analysis of ammonia transfer along two intramolecular tunnels in Staphylococcus aureus glutamine-dependent amidotransferase (GatCAB). J. Phys. Chem. B *119,* 3669-3677.

Di Stefano, B., Sardina, J.L., van Oevelen, C., Collombet, S., Kallin, E.M., Vicent, G.P., Lu, J., Thieffry, D., Beato, M., and Graf, T. (2014). C/EBPalpha poises B cells for rapid reprogramming into induced pluripotent stem cells. Nature *506,* 235-239.

Dor, Y., and Cedar, H. (2018). Principles of DNA methylation and their implications for biology and medicine. Lancet *392,* 777-786.

Dow, B.J., Malik, S.S., and Drohat, A.C. (2019). Defining the Role of Nucleotide Flipping in Enzyme Specificity using 19F NMR. J. Am. Chem. Soc. jas.9b00146.

Dupradeau, F.Y., Pigache, A., Zaffran, T., Savineau, C., Lelong, R., Grivel, N., Lelong, D., Rosanski, W., and Cieplak, P. (2010). The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. Phys. Chem. Chem. Phys. *12,* 7821-7839.

Elias, A.A., and Cisneros, G.A. (2014). Computational study of putative residues involved in DNA synthesis fidelity checking in Thermus aquaticus DNA polymerase I. Adv. Protein Chem. Struct. Biol. *96,* 39-75.

Essmann, U., Perera, L., Berkowitz, M.L., Darden, T., Lee, H., and Pedersen, L.G. (1995). A smooth particle mesh Ewald method. J. Chem. Phys. *103,* 8577-8593.

Falnes, P.O., Bjoras, M., Aas, P.A., Sundheim, O., and Seeberg, E. (2004). Substrate specificities of bacterial and human AlkB proteins. Nucleic Acids Res. *32,* 3456-3461.

Fiser, A., Do, R.K., and Sali, A. (2000). Modeling of loops in protein structures. Protein Sci. *9,* 1753-1773.

Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Scalmani, G., Barone, V., Petersson, G.A., Nakatsuji, H.*, et al.* (2016). Gaussian 09, Revision A.02.

Fu, L., Guerrero, C.R., Zhong, N., Amato, N.J., Liu, Y., Liu, S., Cai, Q., Ji, D., Jin, S.G., Niedernhofer, L.J.*, et al.* (2014). Tet-mediated formation of 5-hydroxymethylcytosine in RNA. J. Am. Chem. Soc. *136,* 11582-11585.

Furst, R.W., Kliem, H., Meyer, H.H., and Ulbrich, S.E. (2012). A differentially methylated single CpG-site is correlated with estrogen receptor alpha transcription. J. Steroid Biochem. Mol. Biol. *130,* 96-104.

Gallego-Bartolome, J., Gardiner, J., Liu, W., Papikian, A., Ghoshal, B., Kuo, H.Y., Zhao, J.M., Segal, D.J., and Jacobsen, S.E. (2018). Targeted DNA demethylation of the Arabidopsis genome using the human TET1 catalytic domain. Proc. Natl. Acad. Sci. U. S. A. *115,* E2125-E2134.

Gersbach, C.A., Gaj, T., and Barbas, C.F.,3rd. (2014). Synthetic zinc finger proteins: the advent of targeted gene regulation and genome modification technologies. Acc. Chem. Res. *47,* 2309-2318.

Ghanty, U., DeNizio, J.E., Liu, M.Y., and Kohli, R.M. (2018). Exploiting Substrate Promiscuity to Develop Activity-Based Probes for TET Family Enzymes. J. Am. Chem. Soc.

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

Graham, S.E., Syeda, F., and Cisneros, G.A. (2012). Computational Prediction of Residues Involved in Fidelity Checking for DNA Synthesis in DNA Polymerase I. Biochemistry *51,* 2569-2578.

Guo, J.U., Su, Y., Shin, J.H., Shin, J., Li, H., Xie, B., Zhong, C., Hu, S., Le, T., Fan, G.*, et al.* (2014). Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. Nat. Neurosci. *17,* 215-222.

Halford, S.E., and Marko, J.F. (2004). How do site-specific DNA-binding proteins find their targets? Nucleic Acids Res. *32,* 3040-3052.

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

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

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

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

He, C., Sidoli, S., Warneford-Thomson, R., Tatomer, D.C., Wilusz, J.E., Garcia, B.A., and Bonasio, R. (2016). High-Resolution Mapping of RNA-Binding Regions in the Nuclear Proteome of Embryonic Stem Cells. Mol. Cell *64,* 416-430.

He, Y.F., Li, B.Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L.*, et al.* (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science *333,* 1303-1307.

Hendershot, J.M., and O'Brien, P.J. (2017). Transient Kinetic Methods for Mechanistic Characterization of DNA Binding and Nucleotide Flipping. Methods Enzymol. *592,* 377-415.

Hess, G.T., Fresard, L., Han, K., Lee, C.H., Li, A., Cimprich, K.A., Montgomery, S.B., and Bassik, M.C. (2016). Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. Nat. Methods *13,* 1036-1042.

Holz-Schietinger, C., and Reich, N.O. (2010). The inherent processivity of the human de novo methyltransferase 3A (DNMT3A) is enhanced by DNMT3L. J. Biol. Chem. *285,* 29091-29100.

Hrit, J., Li, C., Martin, E.A., Goll, M., and Panning, B. (2017). OGT binds a conserved C-terminal domain of TET1 to regulate TET1 activity and function in development. BioRx *doi.org/10.1101/125419,*

Hu, J.H., Davis, K.M., and Liu, D.R. (2016). Chemical Biology Approaches to Genome Editing: Understanding, Controlling, and Delivering Programmable Nucleases. Cell. Chem. Biol. *23,* 57-73.

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

Hu, L., Lu, J., Cheng, J., Rao, Q., Li, Z., Hou, H., Lou, Z., Zhang, L., Li, W., Gong, W.*, et al.* (2015). Structural insight into substrate preference for TET-mediated oxidation. Nature *527,* 118-122.

Huang, W., Lan, M.D., Qi, C.B., Zheng, S.J., Wei, S.Z., Yuan, B.F., and Feng, Y.Q. (2016). Formation and determination of the oxidation products of 5-methylcytosine in RNA. Chem. Sci. *7,* 5495-5502.

Huang, Y., Pastor, W.A., Shen, Y., Tahiliani, M., Liu, D.R., and Rao, A. (2010). The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. PLoS One *5,* e8888.

Huang, Y., Wang, G., Liang, Z., Yang, Y., Cui, L., and Liu, C.Y. (2016). Loss of nuclear localization of TET2 in colorectal cancer. Clin. Epigenetics *8,* 9-eCollection 2016.

Huang, Y.H., Su, J., Lei, Y., Brunetti, L., Gundry, M.C., Zhang, X., Jeong, M., Li, W., and Goodell, M.A. (2017). DNA epigenome editing using CRISPR-Cas SunTag-directed DNMT3A. Genome Biol. *18,* 176-017-1306-z.

Huber, S.M., van Delft, P., Mendil, L., Bachman, M., Smollett, K., Werner, F., Miska, E.A., and Balasubramanian, S. (2015). Formation and abundance of 5-hydroxymethylcytosine in RNA. ChemBioChem *16,* 752-755.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. J. Mol. Graph. *14,* 33-38.

Ito, F., Fu, Y., Kao, S.A., Yang, H., and Chen, X.S. (2017). Family-Wide Comparative Analysis of Cytidine and Methylcytidine Deamination by Eleven Human APOBEC Proteins. J. Mol. Biol. *429,* 1787-1799.

Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466,* 1129-1133.

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

Iurlaro, M., Ficz, G., Oxley, D., Raiber, E.A., Bachman, M., Booth, M.J., Andrews, S., Balasubramanian, S., and Reik, W. (2013). A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. Genome Biol. *14,* R119.

Iyer, L.M., Tahiliani, M., Rao, A., and Aravind, L. (2009). Prediction of novel families of enzymes involved in oxidative and other complex modifications of bases in nucleic acids. Cell Cycle *8,* 1698-1710.

Iyer, L.M., Zhang, D., Burroughs, A.M., and Aravind, L. (2013). Computational identification of novel biochemical systems involved in oxidation, glycosylation and other complex modifications of bases in DNA. Nucleic Acids Res. *41,* 7635-7655.

Iyer, L.M., Zhang, D., de Souza, R.F., Pukkila, P.J., Rao, A., and Aravind, L. (2014). Lineage-specific expansions of TET/JBP genes and a new class of DNA transposons shape fungal genomic and epigenetic landscapes. Proc. Natl. Acad. Sci. U. S. A. *111,* 1676-1683.

Jang, H.S., Shin, W.J., Lee, J.E., and Do, J.T. (2017). CpG and non-CpG methylation in epigenetic gene regulation and brain function. Genes *8,* 2-20.

Jeltsch, A., Wenz, C., Stahl, F., and Pingoud, A. (1996). Linear diffusion of the restriction endonuclease EcoRV on DNA is essential for the in vivo function of the enzyme. Embo J. *15,* 5104-5111.

Ji, D., Lin, K., Song, J., and Wang, Y. (2014). Effects of Tet-induced oxidation products of 5-methylcytosine on Dnmt1- and DNMT3a-mediated cytosine methylation. Mol. Biosyst *10,* 1749-1752.

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

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. (1983). Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. *79,* 926.

Joung, J.K., and Sander, J.D. (2013). TALENs: a widely applicable technology for targeted genome editing. Nat. Rev. Mol. Cell Biol. *14,* 49-55.

Jurkowska, R.Z., Jurkowski, T.P., and Jeltsch, A. (2011). Structure and function of mammalian DNA methyltransferases. Chembiochem *12,* 206-222.

Kaiser, S., Jurkowski, T.P., Kellner, S., Schneider, D., Jeltsch, A., and Helm, M. (2017). The RNA methyltransferase Dnmt2 methylates DNA in the structural context of a tRNA. RNA Biol. *14,* 1241-1251.

Kellinger, M.W., Song, C.X., Chong, J., Lu, X.Y., He, C., and Wang, D. (2012). 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. Nat. Struct. Mol. Biol. *19,* 831-833.

Kim, Y.B., Komor, A.C., Levy, J.M., Packer, M.S., Zhao, K.T., and Liu, D.R. (2017). Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. Nat. Biotechnol. *35,* 371-376.

Kinde, B., Gabel, H.W., Gilbert, C.S., Griffith, E.C., and Greenberg, M.E. (2015). Reading the unique DNA methylation landscape of the brain: Non-CpG methylation, hydroxymethylation, and MeCP2. Proc. Natl. Acad. Sci. U. S. A. *112,* 6800-6806.

Kizaki, S., and Sugiyama, H. (2014). CGmCGCG is a versatile substrate with which to evaluate Tet protein activity. Org. Biomol. Chem. *12,* 104-107.

Kizaki, S., Chandran, A., and Sugiyama, H. (2016). Identification of Sequence Specificity of 5-Methylcytosine Oxidation by Tet1 Protein with High-Throughput Sequencing. ChemBioChem *17,* 403-406.

Kizaki, S., Zou, T., Li, Y., Han, Y.W., Suzuki, Y., Harada, Y., and Sugiyama, H. (2016). Preferential 5-Methylcytosine Oxidation in the Linker Region of Reconstituted Positioned Nucleosomes by Tet1 Protein. Chemistry - A European Journal *22,* 16598-16601.

Ko, M., An, J., Bandukwala, H.S., Chavez, L., Aijo, T., Pastor, W.A., Segal, M.F., Li, H., Koh, K.P., Lahdesmaki, H.*, et al.* (2013). Modulation of TET2 expression and 5-methylcytosine oxidation by the CXXC domain protein IDAX. Nature *497,* 122-126.

Koh, K.P., Yabuuchi, A., Rao, S., Huang, Y., Cunniff, K., Nardone, J., Laiho, A., Tahiliani, M., Sommer, C.A., Mostoslavsky, G.*, et al.* (2011). Tet1 and Tet2 Regulate 5-Hydroxymethylcytosine Production and Cell Lineage Specification in Mouse Embryonic Stem Cells. Stem Cell *8,* 200-213.

Kohli, R.M., Maul, R.W., Guminski, A.F., McClure, R.L., Gajula, K.S., Saribasak, H., McMahon, M.A., Siliciano, R.F., Gearhart, P.J., and Stivers, J.T. (2010). Local sequence targeting in the AID/APOBEC family differentially impacts retroviral restriction and antibody diversification. J. Biol. Chem. *285,* 40956-40964.

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

Komor, A.C., Badran, A.H., and Liu, D.R. (2017). CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes. Cell *169,* 559.

Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature *533,* 420-424.

Kubik, G., Batke, S., and Summerer, D. (2015). Programmable sensors of 5-hydroxymethylcytosine. J. Am. Chem. Soc. *137,* 2-5.

Kulkarni, M., and Mukherjee, A. (2017). Understanding B-DNA to A-DNA transition in the right-handed DNA helix: Perspective from a local to global transition. Prog. Biophys. Mol. Biol. *128,* 63-73.

Lee, J.H., Park, S.J., and Nakai, K. (2017). Differential landscape of non-CpG methylation in embryonic stem cells and neurons caused by DNMT3s. Scientific Reports *7,* 1-11.

Lee, M., Li, J., Liang, Y., Ma, G., Zhang, J., He, L., Liu, Y., Li, Q., Li, M., Sun, D., Zhou, Y., and Huang, Y. (2017). Engineered Split-TET2 Enzyme for Inducible Epigenetic Remodeling. J. Am. Chem. Soc. *139,* 4659-4662.

Lei, Y., Huang, Y.H., and Goodell, M.A. (2018). DNA methylation and de-methylation using hybrid site-targeting proteins. Genome Biol. *19,* 187-018-1566-2.

Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D.*, et al.* (2013). Global epigenomic reconfiguration during mammalian brain development. Science *341,* 1237905.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M.*, et al.* (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462,* 315-322.

Liu, M.Y., DeNizio, J.E., and Kohli, R.M. (2016). Quantification of Oxidized 5-Methylcytosine Bases and TET Enzyme Activity. Methods Enzymol. *573,* 365-385.

Liu, M.Y., DeNizio, J.E., Schutsky, E.K., and Kohli, R.M. (2016). The expanding scope and impact of epigenetic cytosine modifications. Curr. Opin. Chem. Biol. *33,* 67-73.

Liu, M.Y., Torabifard, H., Crawford, D.J., DeNizio, J.E., Cao, X.J., Garcia, B.A., Cisneros, G.A., and Kohli, R.M. (2017). Mutations along a TET2 active site scaffold stall oxidation at 5-hydroxymethylcytosine. Nat. Chem. Biol. *13,* 181-187.

Liu, X.S., Wu, H., Ji, X., Stelzer, Y., Wu, X., Czauderna, S., Shu, J., Dadon, D., Young, R.A., and Jaenisch, R. (2016). Editing DNA Methylation in the Mammalian Genome. Cell *167,* 233-247.e17.

Liutkeviciute, Z., Lukinavicius, G., Masevicius, V., Daujotyte, D., and Klimasauskas, S. (2009). Cytosine-5-methyltransferases add aldehydes to DNA. Nat. Chem. Biol. *5,* 400-402.

Lo, C.L., Choudhury, S.R., Irudayaraj, J., and Zhou, F.C. (2017). Epigenetic Editing of Ascl1 Gene in Neural Stem Cells by Optogenetics. Sci. Rep. *7,* 42047.

Lu, X., Zhao, B.S., and He, C. (2015). TET family proteins: oxidation activity, interacting molecules, and functions in diseases. Chem. Rev. *115,* 2225-2239.

Maeder, M.L., Angstman, J.F., Richardson, M.E., Linder, S.J., Cascio, V.M., Tsai, S.Q., Ho, Q.H., Sander, J.D., Reyon, D., Bernstein, B.E.*, et al.* (2013). Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. Nat. Biotechnol. *31,* 1137-1142.

Mahe, E.A., Madigou, T., Serandour, A.A., Bizot, M., Avner, S., Chalmel, F., Palierne, G., Metivier, R., and Salbert, G. (2017). Cytosine modifications modulate the chromatin architecture of transcriptional enhancers. Genome Res. *27,* 947-958.

Maiti, A., Michelson, A.Z., Armwood, C.J., Lee, J.K., and Drohat, A.C. (2013). Divergent mechanisms for enzymatic excision of 5-formylcytosine and 5-carboxylcytosine from DNA. J. Am. Chem. Soc. *135,* 15813-15822.

Maiti, A., Morgan, M.T., and Drohat, A.C. (2009). Role of two strictly conserved residues in nucleotide flipping and N-glycosylic bond cleavage by human thymine DNA glycosylase. J. Biol. Chem. *284,* 36680-36688.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J., Charpentier, E., Haft, D.H.*, et al.* (2015). An updated evolutionary classification of CRISPR-Cas systems. Nat. Rev. Microbiol. *13,* 722-736.

Marina, R.J., Sturgill, D., Bailly, M.A., Thenoz, M., Varma, G., Prigge, M.F., Nanan, K.K., Shukla, S., Haque, N., and Oberdoerffer, S. (2016). TET-catalyzed oxidation of intragenic 5-methylcytosine regulates CTCF-dependent alternative splicing. Embo J. *35,* 335-355.

Masser, D.R., Hadad, N., Porter, H., Stout, M.B., Unnikrishnan, A., Stanford, D.R., and Freeman, W.M. (2018). Analysis of DNA modifications in aging research. Geroscience *40,* 11-29.

Maurer, S., Giess, M., Koch, O., and Summerer, D. (2016). Interrogating Key Positions of Size-Reduced TALE Repeats Reveals a Programmable Sensor of 5-Carboxylcytosine. ACS Chem. Biol. *11,* 3294-3299.

Melamed, P., Yosefzon, Y., David, C., Tsukerman, A., and Pnueli, L. (2018). Tet Enzymes, Variants, and Differential Effects on Function. Frontiers in Cell and Developmental Biology *6,* 22.

Mellen, M., Ayata, P., and Heintz, N. (2017). 5-Hydroxymethylcytosine Accumulation in Postmitotic Neurons Results in Functional Demethylation of Expressed Genes. Proc. Natl. Acad. Sci. U. S. A. *114,* E7812-E7821.

Mendonca, A., Chang, E.H., Liu, W., and Yuan, C. (2014). Hydroxymethylation of DNA influences nucleosomal conformation and stability in vitro. Biochim. Biophys. Acta *1839,* 1323-1329.

Morgan, M.T., Bennett, M.T., and Drohat, A.C. (2007). Excision of 5-halogenated uracils by human thymine DNA glycosylase. Robust activity for DNA contexts other than CpG. J. Biol. Chem. *282,* 27578-27586.

Morita, S., Noguchi, H., Horii, T., Nakabayashi, K., Kimura, M., Okamura, K., Sakai, A., Nakashima, H., Hata, K., Nakashima, K., and Hatada, I. (2016). Targeted DNA demethylation in

vivo using dCas9-peptide repeat and scFv-TET1 catalytic domain fusions. Nat. Biotechnol. *34,* 1060-1065.

Motorin, Y., Lyko, F., and Helm, M. (2010). 5-methylcytosine in RNA: detection, enzymatic formation and biological functions. Nucleic Acids Res. *38,* 1415-1430.

Muller, T., Gessi, M., Waha, A., Isselstein, L.J., Luxen, D., Freihoff, D., Freihoff, J., Becker, A., Simon, M., Hammes, J.*, et al.* (2012). Nuclear exclusion of TET1 is associated with loss of 5-hydroxymethylcytosine in IDH1 wild-type gliomas. Am. J. Pathol. *181,* 675-683.

Nabel, C.S., Jia, H., Ye, Y., Shen, L., Goldschmidt, H.L., Stivers, J.T., Zhang, Y., and Kohli, R.M. (2012). AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation. Nat. Chem. Biol. *8,* 751-758.

Nabel, C.S., Lee, J.W., Wang, L.C., and Kohli, R.M. (2013). Nucleic acid determinants for selective deamination of DNA over RNA by activation-induced deaminase. Proc. Natl. Acad. Sci. U. S. A. *110,* 14225-14230.

Nelson, C.E., and Gersbach, C.A. (2016). Engineering Delivery Vehicles for Genome Editing. Annu. Rev. Chem. Biomol. Eng. *7,* 637-662.

Neri, F., Incarnato, D., Krepelova, A., Rapelli, S., Anselmi, F., Parlato, C., Medana, C., Dal Bello, F., and Oliviero, S. (2015). Single-Base Resolution Analysis of 5-Formyl and 5-Carboxyl Cytosine Reveals Promoter DNA Methylation Dynamics. Cell. Rep. *10,* 674-683.

Ngo, T.T., Yoo, J., Dai, Q., Zhang, Q., He, C., Aksimentiev, A., and Ha, T. (2016). Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. Nat. Commun. *7,* 10813.

Oakes, B.L., Nadler, D.C., Flamholz, A., Fellmann, C., Staahl, B.T., Doudna, J.A., and Savage, D.F. (2016). Profiling of engineering hotspots identifies an allosteric CRISPR-Cas9 switch. Nat. Biotechnol. *34,* 646-651.

Olson, W.K., Bansal, M., Burley, S.K., Dickerson, R.E., Gerstein, M., Harvey, S.C., Heinemann, U., Lu, X.J., Neidle, S., Shakked, Z.*, et al.* (2001). A standard reference frame for the description of nucleic acid base-pair geometry. J. Mol. Biol. *313,* 229-237.

Pais, J.E., Dai, N., Tamanaha, E., Vaisvila, R., Fomenkov, A.I., Bitinaite, J., Sun, Z., Guan, S., Correa, I.R.,Jr, Noren, C.J.*, et al.* (2015). Biochemical characterization of a Naegleria TET-like oxygenase and its application in single molecule sequencing of 5-methylcytosine. Proc. Natl. Acad. Sci. U. S. A. *112,* 4316-4321.

Parrilla-Doblas, J.T., Ariza, R.R., and Roldan-Arjona, T. (2017). Targeted DNA demethylation in human cells by fusion of a plant 5-methylcytosine DNA glycosylase to a sequence-specific DNA binding domain. Epigenetics *12,* 296-303.

Pastor, W.A., Pape, U.J., Huang, Y., Henderson, H.R., Lister, R., Ko, M., McLoughlin, E.M., Brudno, Y., Mahapatra, S., Kapranov, P.*, et al.* (2011). Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells. Nature *473,* 394-397.

Pearl, L.H. (2000). Structure and function in the uracil-DNA glycosylase superfamily. Mutat. Res. *460,* 165-181.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J. Comput. Chem. *25,* 1605-1612.

Pfaffeneder, T., Hackner, B., Truss, M., Munzel, M., Muller, M., Deiml, C.A., Hagemeier, C., and Carell, T. (2011). The Discovery of 5-Formylcytosine in Embryonic Stem Cell DNA. Angew. Chem. Int. Ed Engl. *50,* 7008-7012.

Pfaffeneder, T., Spada, F., Wagner, M., Brandmayr, C., Laube, S.K., Eisen, D., Truss, M., Steinbacher, J., Hackner, B., Kotljarova, O.*, et al.* (2014). Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA. Nat. Chem. Biol. *10,* 574-581.

Pidugu, L.S., Flowers, J.W., Coey, C.T., Pozharski, E., Greenberg, M.M., and Drohat, A.C. (2016). Structural Basis for Excision of 5-Formylcytosine by Thymine DNA Glycosylase. Biochemistry *55,* 6205-6208.

Raddatz, G., Guzzardo, P.M., Olova, N., Fantappie, M.R., Rampp, M., Schaefer, M., Reik, W., Hannon, G.J., and Lyko, F. (2013). Dnmt2-dependent methylomes lack defined DNA methylation patterns. Proc. Natl. Acad. Sci. U. S. A. *110,* 8627-8631.

Rathi, P., Witte, A., and Summerer, D. (2017). Engineering DNA Backbone Interactions Results in TALE Scaffolds with Enhanced 5-Methylcytosine Selectivity. Sci. Rep. *7,* 15067-017-15361-1.

Renciuk, D., Blacque, O., Vorlickova, M., and Spingler, B. (2013). Crystal structures of B-DNA dodecamer containing the epigenetic modifications 5-hydroxymethylcytosine or 5-methylcytosine. Nucleic Acids Res. *41,* 9891-9900.

Roe, D.R., and Cheatham, T.E.,3rd. (2013). PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J. Chem. Theory Comput. *9,* 3084-3095.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. *234,* 779-815.

Salomon-Ferrer, R., Götz, A.W., Poole, D., Le Grand, S., and Walker, R.C. (2013). Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J. Chem. Theory Comput. *9,* 3878-3888.

Sanchez-Vasquez, E., Alata Jimenez, N., Vazquez, N.A., and Strobl-Mazzulla, P.H. (2018). Emerging role of dynamic RNA modifications during animal development. Mech. Dev. *154,* 24-32.

Schapira, M. (2016). Structural Chemistry of Human RNA Methyltransferases. ACS Chem. Biol. *11,* 575-582.

Schiesser, S., Pfaffeneder, T., Sadeghian, K., Hackner, B., Steigenberger, B., Schroder, A.S., Steinbacher, J., Kashiwazaki, G., Hofner, G., Wanner, K.T., Ochsenfeld, C., and Carell, T. (2013). Deamination, oxidation, and C-C bond cleavage reactivity of 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxycytosine. J. Am. Chem. Soc. *135,* 14593-14599.

Schroder, A.S., Parsa, E., Iwan, K., Traube, F.R., Wallner, M., Serdjukow, S., and Carell, T. (2016). 2'-(R)-Fluorinated mC, hmC, fC and caC triphosphates are substrates for DNA polymerases and TET-enzymes. Chem. Commun. (Camb) *52,* 14361-14364.

Schubeler, D. (2015). Function and information content of DNA methylation. Nature *517,* 321-326.

Schutsky, E.K., DeNizio, J.E., Hu, P., Liu, M.Y., Nabel, C.S., Fabyanic, E.B., Hwang, Y., Bushman, F.D., Wu, H., and Kohli, R.M. (2018). Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. Nat. Biotech. *36,* 1083-1090.

Schutsky, E.K., Nabel, C.S., Davis, A.K.F., DeNizio, J.E., and Kohli, R.M. (2017). APOBEC3A efficiently deaminates methylated, but not TET-oxidized, cytosine bases in DNA. Nucleic Acids Res. *45,* 7655-7665.

Seiler, C.L., Fernandez, J., Koerperich, Z., Andersen, M.P., Kotandeniya, D., Nguyen, M.E., Sham, Y.Y., and Tretyakova, N.Y. (2018). Maintenance DNA Methyltransferase Activity in the Presence of Oxidized Forms of 5-Methylcytosine: Structural Basis for Ten Eleven Translocation-Mediated DNA Demethylation. Biochemistry (N. Y.) *57,* 6061-6069.

Shen, J.C., Rideout, W.M.,3rd, and Jones, P.A. (1994). The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res. *22,* 972-976.

Shen, Q., Zhang, Q., Shi, Y., Shi, Q., Jiang, Y., Gu, Y., Li, Z., Li, X., Zhao, K., Wang, C., Li, N., and Cao, X. (2018). Tet2 promotes pathogen infection-induced myelopoiesis through mRNA oxidation. Nature *554,* 123-127.

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature *479,* 74-79.

Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. Nat. Rev. Genet. *14,* 204-220.

Song, C.X., Diao, J., Brunger, A.T., and Quake, S.R. (2016). Simultaneous single-molecule epigenetic imaging of DNA methylation and hydroxymethylation. Proc. Natl. Acad. Sci. U. S. A. *113,* 4338-4343.

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

Spruijt, C.G., Gnerlich, F., Smits, A.H., Pfaffeneder, T., Jansen, P.W., Bauer, C., Munzel, M., Wagner, M., Muller, M., Khan, F.*, et al.* (2013). Dynamic Readers for 5-(Hydroxy)Methylcytosine and Its Oxidized Derivatives. Cell *152,* 1146-1159.

Squires, J.E., Patel, H.R., Nousch, M., Sibbritt, T., Humphreys, D.T., Parker, B.J., Suter, C.M., and Preiss, T. (2012). Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic Acids Res. *40,* 5023-5033.

Stanford, N.P., Szczelkun, M.D., Marko, J.F., and Halford, S.E. (2000). One- and three-dimensional pathways for proteins to reach specific DNA sites. Embo J. *19,* 6546-6557.

Stepper, P., Kungulovski, G., Jurkowska, R.Z., Chandra, T., Krueger, F., Reinhardt, R., Reik, W., Jeltsch, A., and Jurkowski, T.P. (2017). Efficient targeted DNA methylation with chimeric dCas9-Dnmt3a-Dnmt3L methyltransferase. Nucleic Acids Res. *45,* 1703-1713.

Stivers, J.T., and Jiang, Y.L. (2003). A mechanistic perspective on the chemistry of DNA repair glycosylases. Chem. Rev. *103,* 2729-2759.

Strutt, S.C., Torrez, R.M., Kaya, E., Negrete, O.A., and Doudna, J.A. (2018). RNA-dependent RNA targeting by CRISPR-Cas9. Elife *7,* 10.7554/eLife.32724.

Sudhamalla, B., Wang, S., Snyder, V., Kavoosi, S., Arora, S., and Islam, K. (2018). Complementary Steric Engineering at the Protein-Ligand Interface for Analogue-Sensitive TET Oxygenases. J. Am. Chem. Soc. *140,* 10263-10269.

Sun, Z., Dai, N., Borgaro, J.G., Quimby, A., Sun, D., Correa, I.R.,Jr, Zheng, Y., Zhu, Z., and Guan, S. (2015). A sensitive approach to map genome-wide 5-hydroxymethylcytosine and 5-formylcytosine at single-base resolution. Mol. Cell *57,* 750-761.

Szulik, M.W., Pallan, P.S., Nocek, B., Voehler, M., Banerjee, S., Brooks, S., Joachimiak, A., Egli, M., Eichman, B.F., and Stone, M.P. (2015). Differential stabilities and sequence-dependent base pair opening dynamics of Watson-Crick base pairs with 5-hydroxymethylcytosine, 5-formylcytosine, or 5-carboxylcytosine. Biochemistry *54,* 1294-1305.

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

Tamanaha, E., Guan, S., Marks, K., and Saleh, L. (2016). Distributive Processing by the Iron(II)/Î±-Ketoglutarate-Dependent Catalytic Domains of the TET Enzymes Is Consistent with Epigenetic Roles for Oxidized 5-Methylcytosine Bases. J. Am. Chem. Soc. *138,* 9345-9348.

Tarantino, M.E., Dow, B.J., Drohat, A.C., and Delaney, S. (2018). Nucleosomes and the three glycosylases: High, medium, and low levels of excision by the uracil DNA glycosylase superfamily. DNA Repair *72,* 56-63.

Terragni, J., Bitinaite, J., Zheng, Y., and Pradhan, S. (2012). Biochemical characterization of recombinant beta-glucosyltransferase and analysis of global 5-hydroxymethylcytosine in unique genomes. Biochemistry *51,* 1009-1019.

Van Dongen Stijn F M, Elemans, J.A.A.W., Rowan, A.E., and Nolte, R.J.M. (2014). Processive Catalysis. Angewandte Chemie - International Edition *53,* 11420-11428.

Vanquelef, E., Simon, S., Marquant, G., Garcia, E., Klimerak, G., Delepine, J.C., Cieplak, P., and Dupradeau, F.Y. (2011). R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. Nucleic Acids Res. *39,* W511-7.

Vilkaitis, G., Suetake, I., Klimasauskas, S., and Tajima, S. (2005). Processive methylation of hemimethylated CpG sites by mouse Dnmt1 DNA methyltransferase. J. Biol. Chem. *280,* 64-72.

Wagner, M., Steinbacher, J., Kraus, T.F., Michalakis, S., Hackner, B., Pfaffeneder, T., Perera, A., Muller, M., Giese, A., Kretzschmar, H.A., and Carell, T. (2015). Age-dependent levels of 5-methyl-, 5-hydroxymethyl-, and 5-formylcytosine in human and mouse brain tissues. Angew. Chem. Int. Ed Engl. *54,* 12511-12514.

Wang, F., Becker, J.-., Cieplack, P., and Dupradeau, F.-. (2013). R.E.D. Python: Object oriented programming for Amber force fields
.

Wang, L., Zhang, J., Duan, J., Gao, X., Zhu, W., Lu, X., Yang, L., Zhang, J., Li, G., Ci, W*., et al.* (2014). Programming and inheritance of parental DNA methylomes in mammals. Cell *157,* 979-991.

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

Weber, A.R., Krawczyk, C., Robertson, A.B., Kusnierczyk, A., Vagbo, C.B., Schuermann, D., Klungland, A., and Schar, P. (2016). Biochemical reconstitution of TET1-TDG-BER-dependent active DNA demethylation reveals a highly coordinated mechanism. Nat. Commun. *7,* 10806.

Wen, L., Li, X., Yan, L., Tan, Y., Li, R., Zhao, Y., Wang, Y., Xie, J., Zhang, Y., Song, C*., et al.* (2014). Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. Genome Biol. *15,* R49-2014-15-3-r49.

Wu, D., Hu, D., Chen, H., Shi, G., Fetahu, I.S., Wu, F., Rabidou, K., Fang, R., Tan, L., Xu, S*., et al.* (2018). Glucose-regulated phosphorylation of TET2 by AMPK reveals a pathway linking diabetes to cancer. Nature *559,* 637-641.

Wu, H., Wu, X., Shen, L., and Zhang, Y. (2014). Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. Nat. Biotechnol. *32,* 1231-1240.

Wu, H., Wu, X., and Zhang, Y. (2016). Base-resolution profiling of active DNA demethylation using MAB-seq and caMAB-seq. Nat. Protoc. *11,* 1081-1100.

Wu, H., and Zhang, Y. (2015). Charting oxidized methylcytosines at base resolution. Nat. Struct. Mol. Biol. *22,* 656-661.

Wu, X., and Zhang, Y. (2017). TET-mediated active DNA demethylation: mechanism, function and beyond. Nat. Rev. Genet. *18,* 517-534.

Xiong, T., Meister, G.E., Workman, R.E., Kato, N.C., Spellberg, M.J., Turker, F., Timp, W., Ostermeier, M., and Novina, C.D. (2017). Targeted DNA methylation in human cells using engineered dCas9-methyltransferases. Sci. Rep. *7,* 6732-017-06757-0.

Xu, Q., Wang, K., Wang, L., Zhu, Y., Zhou, G., Xie, D., and Yang, Q. (2016). IDH1/2 Mutants Inhibit TET-Promoted Oxidation of RNA 5mC to 5hmC. PLoS One *11,* e0161261.

Xu, X., Tao, Y., Gao, X., Zhang, L., Li, X., Zou, W., Ruan, K., Wang, F., Xu, G.L., and Hu, R. (2016). A CRISPR-based approach for targeted DNA demethylation. Cell. Discov. *2,* 16009.

Yang, X., Yang, Y., Sun, B.F., Chen, Y.S., Xu, J.W., Lai, W.Y., Li, A., Wang, X., Bhattarai, D.P., Xiao, W*., et al.* (2017). 5-methylcytosine promotes mRNA export - NSUN2 as the methyltransferase and ALYREF as an m(5)C reader. Cell Res. *27,* 606-625.

Yu, M., Hon, G.C., Szulwach, K.E., Song, C.X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B.*, et al.* (2012). Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell *149,* 1368-1380.

Zeng, Y., Yao, B., Shin, J., Lin, L., Kim, N., Song, Q., Liu, S., Su, Y., Guo, J.U., Huang, L.*, et al.* (2016). Lin28A Binds Active Promoters and Recruits Tet1 to Regulate Gene Expression. Mol. Cell *61,* 153-160.

Zhang, L., Chen, W., Iyer, L.M., Hu, J., Wang, G., Fu, Y., Yu, M., Dai, Q., Aravind, L., and He, C. (2014). A TET homologue protein from Coprinopsis cinerea (CcTET) that biochemically converts 5-methylcytosine to 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxylcytosine. J. Am. Chem. Soc. *136,* 4801-4804.

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

Zhang, Y., Liu, L., Guo, S., Song, J., Zhu, C., Yue, Z., Wei, W., and Yi, C. (2017). Deciphering TAL effectors for 5-methylcytosine and 5-hydroxymethylcytosine recognition. Nat. Commun. *8,* 901-017-00860-6.