DECIPHERING THE TETRAD OF EPIGENETIC CYTOSINE MODIFICATIONS

Monica Yun Liu

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

2016

Supervisor of Dissertation

Signature: _____

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

Graduate Group Chairperson

Signature: _____

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

Dissertation Committee

Benjamin A. Garcia, Ph.D., Presidential Professor of Biochemistry and Biophysics

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

Zhaolan Zhou, Ph.D., Associate Professor of Genetics

Roberto Bonasio, Ph.D., Assistant Professor of Cell and Developmental Biology

Ya-Ming Hou, Ph.D., Professor of Biochemistry and Molecular Biology (Thomas Jefferson University)

To family and friends, sine qua non

ACKNOWLEDGMENTS

sine qua non—"without which, none." This thesis is dedicated to my family and friends, without whom nothing would be possible, or meaningful. Foremost, I thank my parents, whose love and support are the foundations of my life. Whatever else may change, those are constants that I know we will always share. I thank my extended family as well, particularly my grandparents, whose belief in me is astounding and humbling, and whose life stories serve as a source of inspiration. Furthermore, my friends, near and far, are so often like a family to me. Many have made lasting impressions on me, so that even if we are no longer in contact, much of what is good about me has come in part from them. Notably, a few ties have been uniquely close and enduring: Kayla Satmaria, Sophia Li, Melanie Rice, Lynn Wang, Akilah Graham, Abigail Cheung, Sophia Chen, Eugene Serebryany, Tina Ho, Alexandra Adegoke, among others. Alongside these wonderful people, one best friend is all the more extraordinary: Charlie Mo—my partner and confidant, a steady source of joy and support who has enriched my life in countless ways. All these people have my deepest gratitude, now and always.

This thesis specifically has been made possible by many remarkable individuals. First among them is Rahul Kohli, my mentor, friend, and greatest role model. He daily exemplifies not only what I hope to do in my career but also, in many ways, the kind of person I would like to be—unfailingly kind, with a way of making others feel valued. I also thank all the past and present members of the Kohli lab, who have made each day of my Ph.D. years incredibly fun and fulfilling. Among this group, I am especially grateful to Danny Crawford and Jamie DeNizio, who have been the best possible teammates in exploring TET enzyme mechanisms. I also had many wonderful collaborators, especially Hedi Torabifard and Andres Cisneros at Wayne State, who showed me how rewarding the sharing of science can be. Along these lines, my work has connected me to labs across Penn's campus, where I have had the pleasure of getting to know the

labs of Marisa Bartolomei, Ben Garcia, and Ian Blair. I thank all these groups for adopting me and making my time in their labs equally enjoyable. Notably, Joanne Thorvaldsen in the Bartolomei lab has been an exceptionally helpful and kind teacher and collaborator. Finally, I am indebted to my thesis committee members for their encouragement and thoughtful feedback, and to everyone from BMB and the Penn MSTP for a truly outstanding graduate experience.

ABSTRACT

DECIPHERING THE TETRAD OF EPIGENETIC CYTOSINE MODIFICATIONS

Monica Yun Liu

Rahul M. Kohli, M.D., Ph.D.

A tetrad of epigenetic cytosine modifications imbues the DNA code with complex, dynamic meaning. DNA methyltransferase enzymes deposit methyl marks on the 5-carbon of cytosine, forming 5-methylcytosine (mC), which generally mediates long-term, locus-specific transcriptional repression during development and reprogramming. Ten-eleven translocation (TET) family enzymes oxidize the methyl group in three steps, forming predominantly 5-hydroxymethylcytosine (hmC) but also low levels of 5-formylcytosine (fC) and 5-carboxylcytosine (caC). These additional bases likely provide pathways for erasing methylation, but they may also harbor epigenetic functions in their own right. Questions regarding how each base forms and functions drive at the fundamental biology of the epigenome. In this thesis, I chronicle our lab's efforts to probe the epigenome at its source—by deciphering and manipulating TET enzyme mechanisms. In particular, we aim to understand how and why TET enzymes generate rare fC and caC bases rather than hmC alone. Following a review of the field, I describe in Chapter 2 the methods that we developed to study rare cytosine modifications with high sensitivity. In Chapter 3, we applied these techniques to a rigorous kinetic study of how mouse Tet2 establishes and maintains oxidized cytosines. We found that Tet2 is capable of iterative oxidation, staying on a DNA strand to catalyze multiple rounds of oxidation and thereby enabling efficient generation of fC and caC under certain circumstances. In Chapter 4, we asked what structure-function determinants could allow for the generation of fC and caC. We discovered a conserved active site scaffold in human TET2 that specifically supports the formation of all three

oxidized bases, not just hmC. By mutating the active site, we could alter the interactions between key residues to achieve stalling of oxidation at hmC. These mutants have now paved the way for applications in model systems to examine the function of hmC independently of fC and caC, which will allow us to dissect whether the rare, highly oxidized bases are truly critical for epigenetic processes. I describe our progress to date in Chapter 5, along with further mechanistic explorations of the dynamic epigenome.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1: Introduction

This chapter has been substantially expanded from the following publication:

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.[*]

1.1: The dynamic epigenome

1.1.1: Methylation of cytosine in DNA

Epigenetic writers, readers, and erasers transform the simple genetic sequence of A's, C's, G's, and T's into an astoundingly rich text, an interpretive guide for nearly all life's processes. In eukaryotes, one layer of epigenetic regulation involves the covalent modification of cytosine in DNA. For decades, one modified base, 5-methylcytosine (mC), stood out as the predominant epigenetic mark on DNA, the product not of DNA-damaging agents but of specialized DNA methyltransferase enzymes (DNMTs) acting with physiological purpose. Primarily, this purpose is long-term repression of transcription, which contributes to determining cell fates, establishing genomic imprinting, and silencing retrotransposons (Bird, 2011). Genome-wide methylation patterns are generally established by DNMT3a/b enzymes and maintained by DNMT1. The key to maintenance methylation lies in the symmetry of typical modifications, which occur at self-complementary cytosine-guanine dinucleotide sequences (CpGs). DNMT1 specifically recognizes hemi-methylated CpGs and copies the methyl mark from one strand to the other, ensuring stable methylation at any given CpG across cell divisions (Jurkowska et al., 2011). Of the approximately 3 billion bases in the human genome, there are ~28 million CpGs

[*] Author contributions: I was the primary writer of this review. R.M.K. and I conceived the ideas, J.E.D. and E.K.S. made the figures, and everyone edited.

dispersed at low density across the sequence, and ~80% of CpGs are typically methylated (making mC ~0.7% of total bases). By contrast, most of the unmethylated CpGs are highly clustered in ~200-2,000 bp regions called CpG islands. These are often located close to transcription start sites, in prime position to regulate transcription via interaction with RNA polymerases, transcription factors, and epigenetic effectors (Bird, 2011; Smith and Meissner, 2013). Importantly, regulation implies dynamics, and while DNA demethylation has long been recognized as integral to resetting epigenetic marks during embryonic development and certain cellular responses, the mechanisms of mC erasure are less well understood.

1.1.2: Demethylation pathways in an mC-centric epigenome

In the absence of maintenance methylation, DNA replication can lead to a passive reduction in methylation, but active enzymatic mechanisms are necessary for rapid, regulated demethylation. The simplest imaginable mechanism for reversible DNA methylation would be a "toggle switch" at carbon 5 of cytosine. However, the difficulty of breaking carbon-carbon bonds essentially bars this possibility. As one efficient solution, flowering plants such as *Arabidopsis thaliana* have evolved DEMETER family glycosylases that can directly excise the mC base, leaving an abasic site that undergoes base excision repair (BER) to restore unmodified cytosine (Gehring et al., 2006; Morales-Ruiz et al., 2006). Perhaps illustrating the vagaries of evolution, animals lack a DEMETER homologue and therefore require indirect alternatives to solve the problem of DNA demethylation.

These indirect pathways draw upon a well-stocked toolkit of pyrimidine-modifying enzymes. Deamination of mC gained attention, since the resulting thymine, mispaired with guanine, is a substrate for excision by thymine DNA glycosylase (TDG) or methyl-binding domain protein 4 (MBD4), initiating BER to restore C:G (Hendrich et al., 1999; Neddermann and Jiricny, 1993; Rai et al., 2008). However, the leading candidates for catalyzing cytosine

deamination, the AID/APOBEC family enzymes, prefer unmodified C over mC by ~10-fold (Nabel et al., 2012) and are selective for particular sequence contexts in single-stranded DNA (Bransteitter et al., 2003; Kohli et al., 2009; Kohli et al., 2010). These features limit the likelihood that AID/APOBEC enzymes could demethylate any given CpG or could be responsible for genome-wide demethylation. Nevertheless, deamination-driven BER remains a plausible accessory pathway in select circumstances. Of further note, a minor role could be found for deamination of unmodified C to uracil, which can initiate long-patch BER or non-canonical mismatch repair (Franchini et al., 2014; Grin and Ishchenko, 2016). Any mC that falls within the excised patch could therefore be replaced with unmodified C upon repair, without requiring specificity for a single CpG.

1.1.3: New bases, new pathways in the extended epigenome

The predominant mode of active demethylation remained elusive so long as the view of the epigenome centered only on mC. The search broke new ground in 2009 with the discovery that ten-eleven translocation (TET) enzymes oxidize mC to an additional base, 5-hydroxymethylcytosine (hmC), which makes up nearly 40% of mC modifications in the brain (Kriaucionis and Heintz, 2009; Tahiliani et al., 2009). This discovery was essentially the confluence of three known elements, which together transformed our view of the epigenome. First, even before DNA was ascertained to be the hereditary material, hmC had been known to replace cytosine in T-even phage DNA (Wyatt and Cohen, 1953), but it was largely overlooked in eukaryotes. Second, the TET1 isoform had been identified previously from the namesake chromosomal translocation, which results in an oncogenic fusion with mixed lineage leukemia protein (MLL), but its function was not documented (Lorsbach et al., 2003; Ono et al., 2002). Third, J-binding proteins (JBP1 and 2) in trypanosomes were known to oxidize thymine by an Fe(II)/α-ketoglutarate (αKG)-dependent mechanism, as part of the parasite's epigenetic strategy

for immune evasion (Bullard et al., 2014; Cliffe et al., 2009). A computational search for JBP homologues in mammals revealed the missing links between these elements: TET1, 2, and 3 enzymes that preferentially oxidize mC over T to yield hmC as a sixth base of DNA (Figure 1-1) (Tahiliani et al., 2009). Improved assays soon revealed that TET enzymes do not stop at hmC but in rare instances can catalyze stepwise oxidation to 5-formylcytosine (fC) and 5-carboxylcytosine (caC) (He et al., 2011; Ito et al., 2011). These three oxidized mC (ox-mC) bases stably populate the genome in highly diverse cell types and are increasingly thought to carry out independent epigenetic functions.

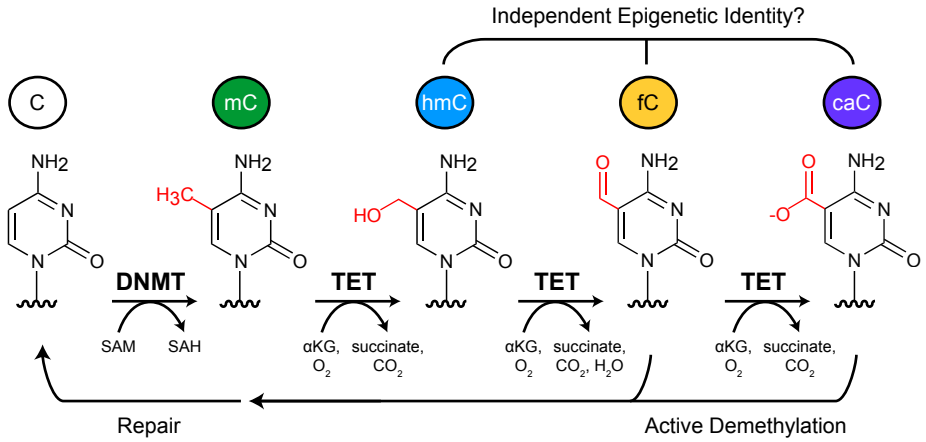


Figure 1-1. Tet enzymes act stepwise on mC to generate the extended epigenome.

DNMT enzymes transfer a methyl group from *S*-adenosyl-L-methionine (SAM) to the 5-carbon of cytosine, leaving *S*-adenosyl-L-homocysteine (SAH) as a byproduct. TET family dioxygenases can then oxidize mC in a three-step, Fe(II)/αKG-dependent mechanism. The oxidized mC bases—hmC, fC and caC—could each play independent roles in epigenetic regulation. However, only fC and caC are thought to be substrates for active DNA demethylation via base excision repair.

With the recognition that the epigenome extends well beyond mC, new pathways for demethylation opened. Passive, replication-dependent loss remained prominent, but the expanded repertoire of purposeful cytosine modifications fueled the pursuit of active, enzymatic mechanisms. DNMTs—classically the writers of methylation—were shown to be capable of direct decarboxylation of caC (Liutkeviciute et al., 2014) or dehydroxymethylation (Chen et al.,

2012; Liutkeviciute et al., 2009). However, these activities are contingent on low levels of *S*-adenosyl-L-methionine (SAM), which is considered improbable under physiological conditions; indeed, these reactions have yet to be observed *in vivo*. Other studies pointed to a mixed TET-AID/APOBEC-BER pathway, whereby mC is oxidized to hmC, deaminated to 5-hydroxymethyluracil (hmU), excised by hmU-specific glycosylases such as TDG or SMUG1, and repaired to unmodified C (Guo et al., 2011). However, AID/APOBEC enzymes have no detectable activity on hmC *in vitro* or in cells (Nabel et al., 2012; Rangam et al., 2012), again making the deamination-mediated pathways less likely.

In the end, one pathway involving TET, TDG, and BER has emerged as the most likely to meet the needs of genome-wide active demethylation (Figure 1-1). TDG was found to excise fC and caC from CpGs while leaving mC and hmC untouched (He et al., 2011; Maiti and Drohat, 2011). Among the many mammalian glycosylases (12 known in humans), TDG is unique in its activity on properly paired, oxidized cytosine bases. It is also the only glycosylase whose deletion is embryonic lethal, suggesting that TDG has a specialized, essential role in mammalian development (Cortazar et al., 2011; Cortellino et al., 2011). TDG depletion results in global accumulation of fC and caC in embryonic stem cells (ESCs), especially at proximal and distal gene regulatory elements (Raiber et al., 2012; Shen et al., 2013; Song et al., 2013). TDG is also required for reprogramming of mouse embryonic fibroblasts (MEFs), a process that involves demethylation both at key loci and genome-wide (Hu et al., 2014). Further supporting the validity of this pathway, demethylation has recently been reconstituted *in vitro* using purified TET, TDG, and BER proteins (Weber et al., 2016).

It is important to note that many studies have added nuances and highlighted additional questions regarding the role of TDG in active demethylation. Several partner proteins, including growth arrest and DNA-damage-inducible protein (Gadd45a) and apurinic/apyrimidinic endonuclease (APE1), have been proposed to facilitate TDG recruitment to sites of demethylation

and promote TDG release from abasic sites, respectively (Fitzgerald and Drohat, 2008; Li et al., 2015b). TET enzymes were also reported to interact with TDG and most of the BER machinery, suggesting potentially synchronous activity (Muller et al., 2014). Still, significant questions remain about how TET enzymes generate fC and caC modifications, how local genomic regions come to be enriched in these rare bases, and how BER can guard against damaging double-stranded breaks when fC/caC occur on both DNA strands. Furthermore, there is no clear consensus on the contribution of passive vs. active mechanisms during the waves of demethylation most prominent in early zygotes and primordial germ cells (Amouroux et al., 2016; Guo et al., 2014; Hajkova et al., 2010; Shen et al., 2014). One plausible model is that the TET-TDG-BER pathway is essential for active demethylation at key loci, whereas passive and TDG-independent mechanisms are likely at play elsewhere (Guo et al., 2014; Hu et al., 2014).

Altogether, the dynamic epigenome represents the coordinated, manifest activities of DNMTs, TETs, deaminases, BER enzymes, and many other proteins and cofactors, which together make each generation of life possible. Boundless open questions will undoubtedly propel the field for the foreseeable future. This thesis focuses on the extension of the epigenome beyond mC, particularly on the enzymology of TET family dioxygenases and the mechanisms by which they generate ox-mC bases for potential epigenetic functions. In the following sections, I review TET biochemistry, properties of the ox-mC bases, and the emerging roles of TET enzyme activity in epigenetic regulation as well as pathology.

1.2: The mechanisms and scope of TET reactivity

1.2.1: Canonical activity in writing the extended epigenome

A mechanistic understanding of TET enzymes begins with their intrinsic properties. TET enzymes belong to the superfamily of Fe(II)/α-ketoglutarate (αKG)-dependent dioxygenases. Accordingly, they are thought to act by harnessing electrons from Fe(II) and αKG to split

molecular oxygen (O_2), resulting in oxidative decarboxylation of αKG to give succinate and a high-energy Fe(IV)-oxo intermediate. The Fe(IV)-oxo species activates the target C–H bond on mC for oxidation, yielding hmC and regenerating Fe(II) in the active site (Lu et al., 2015). Thus, the net reaction on DNA can be expressed as:

$$\text{mC} + \alpha\text{KG} + O_2 \rightarrow \text{hmC} + \text{succinate} + CO_2$$

This mechanism explains how vitamin C, an antioxidant, enhances TET's (and related proteins') activity, likely by maintaining the Fe(II) reduced state (Blaschke et al., 2013; Minor et al., 2013; Yin et al., 2013). Conversely, αKG analogues including succinate, fumarate, and 2-hydroxyglurate can competitively inhibit overall activity (Carey et al., 2015; Laukka et al., 2016; Xiao et al., 2012; Xu et al., 2011), though no TET-specific inhibitors are yet known.

Importantly, while this general mechanism likely underlies each oxidation reaction, TET enzymes are capable of stepwise oxidation of three different substrates (mC, hmC, and fC). Therefore, it becomes critical to understand the mechanisms that are pertinent to each step and that govern progression through the steps. In particular, explanations are needed for the gulf between relatively high levels of hmC and very rare fC and caC: in general, genomic levels of hmC are approximately 10-fold lower than mC, while levels of fC and caC are at least 100-fold lower than hmC—approximately 1 in 10^6 nucleotides (Bachman et al., 2015; Pfaffeneder et al., 2014). This question has implications not only for DNA demethylation but also for the potential epigenetic information encoded by stable fC and caC marks.

A crystal structure of human TET2 bound to mC (Hu et al., 2013b)—along with a structure of a distant TET1 homologue in *Naegleria gruberi* (Hashimoto et al., 2014b)—promised deeper insights into the stepwise oxidation mechanism. The structures showed the substrate mC flipped out of the DNA duplex and positioned in the active site with the target methyl group oriented toward αKG and Fe(II) (see Chapter 4 for details). The cytosine base and cofactors are coordinated by key, conserved residues, which when mutated largely abolish enzymatic activity

(Hu et al., 2013b). The unpaired G is displaced from the helix by Tyr1294, which base stacks with the neighboring C on the opposite strand, effectively explaining TET's specificity for CpGs. No base-specific interactions were observed outside of the central CpG, consistent with a general lack of broader sequence preference. Critically, however, the structures did not readily reveal a mechanism for specific recognition of mC vs. hmC or fC substrates (or products); subsequent crystal structures with hmC and fC further underscored the lack of specific interaction with the 5-modified group (Hu et al., 2015). Thus, based on the crystal structures, TET enzymes merely appeared capable of accommodating the various cytosine derivatives in the active site, but there was little to explain the skewed distribution of oxidation products that are observed in the genome.

Nevertheless, the structures of TET2 enabled computational modeling studies, which suggested that conformational restraints on hmC and fC disfavor hydrogen abstraction from the 5-modified group, resulting in lower reactivity compared to the less restrained mC substrate (Hu et al., 2015; Lu et al., 2016). In line with these models, biochemical studies on several TET homologues reported decreasing reactivity from mC to hmC to fC (Hashimoto et al., 2015; Ito et al., 2011; Pfaffeneder et al., 2011), including ~2- to 5-fold differences in k_{cat} and K_M (but comparable K_d) for human TET2 (Hu et al., 2015).

Decreasing reactivity of the oxidation products, however, raises the question of how fC and particularly caC are observed in any appreciable quantity. Adding TET's substrate preferences to the substantial excess of mC in the genome, higher-order oxidation products should, by probability, almost never come into existence. However, this presents a discrepancy with the proposed functions of fC and caC in active DNA demethylation and other processes—if these rare bases are indeed important, then the enzyme should have mechanisms that enable their formation, despite the odds favoring mC-to-hmC conversion. Therefore, critical questions

surround how TET enzymes establish hmC vs. fC and caC bases, both at a single CpG site and across multiple sites.

Once established, epigenetic marks often must be maintained by copying parental CpG modifications onto newly synthesized DNA. For mammalian DNA methylation, the basis of this epigenetic "memory" is well understood: *de novo* DNMT3a/b enzymes establish new mC marks irrespective of opposite strand methylation, while maintenance DNMT1 enzymes specifically copy mC onto hemi-methylated DNA after replication (Jurkowska et al., 2011). For ox-mCs, epigenetic memory is being actively explored. Isotopic labeling experiments support potentially long-lived modifications (Bachman et al., 2014; Bachman et al., 2015), and sequencing studies suggest that these bases can be stably mapped in many cell lines (Booth et al., 2015; Wu and Zhang, 2015). Stable modifications imply some form of maintenance across cellular generations, though, strictly defined, the copying of ox-mCs onto the opposite strand has yet to be demonstrated. Importantly, in order for TET to act on newly synthesized DNA, DNMTs would first need to generate an mC across from the ox-mC on the parent strand. Biochemical studies have indicated that DNMT1 has poorer activity at CpGs containing an ox-mC mark on one strand; however, DNMT3a/b are minimally affected, suggesting that ox-mCs could indeed coexist at a CpG site (Hashimoto et al., 2012; Ji et al., 2014).

Various studies have attempted to identify and quantify the occupancy of ox-mCs on either strand of a CpG. A base-resolution map of hmC in the genome reported that, while mC nearly always occurs symmetrically at CpGs, hmC is more asymmetric (Yu et al., 2012). However, this result comes with multiple caveats, including limited coverage and sensitivity for low-level hmC modifications, as well as leaving fC and caC unaccounted for. A subsequent sequencing analysis of mC, hmC, and fC showed moderate asymmetry for hmC and fC (~45% difference in levels between strands vs. 18% for mC), though these sequencing methods only cover a subset of genomic CpGs (Booth et al., 2014). Perhaps most notably, all these techniques

report on a pooled population of cells; ideally, readout of strand-specific CpG modifications at the single-cell level would be needed. A significant advance came with single-molecule FRET detection of mC and hmC, which showed that genomic DNA from various mouse tissues contains consistently high levels of dual-modified mCpG/hmCpG units (Song et al., 2016). This asymmetry, with TET substrates on both DNA strands, highlights the need to understand maintenance vs. *de novo* oxidation activity and any differences between the three TET isoforms.

1.2.2: Non-canonical activities suggest additional biological roles

Beyond their canonical role in mC oxidation, TET enzymes are capable of broader reactivity that hints at other potential cellular functions (Figure 1-2). Though initial studies focused on double-stranded DNA, single-stranded 4- to 6-mers are also viable substrates (Kizaki and Sugiyama, 2014). Furthermore, *in vitro* studies showed activity on mC in RNA, and all three oxidized bases have been detected mostly in mRNA from various tissues and species (Fu et al., 2014; Huang et al., 2016; Huber et al., 2015). In a particularly interesting example, hmC was discovered in actively transcribing mRNA in *Drosophila melanogaster*, and levels of hmC decreased upon depletion of the organism's TET homologue (Delatte et al., 2016). As Drosophila lack significant mC in genomic DNA, this opens the door for some TET enzymes to play an added role in RNA biology. At the same time, confirming a physiological role for TET-mediated RNA modification will be important, since the detection of oxidized RNA bases in Tet-null embryonic stem cells (Fu et al., 2014), as well as in organisms that lack TET homologues (Huber et al., 2015), indicates that TET-independent mechanisms could contribute.

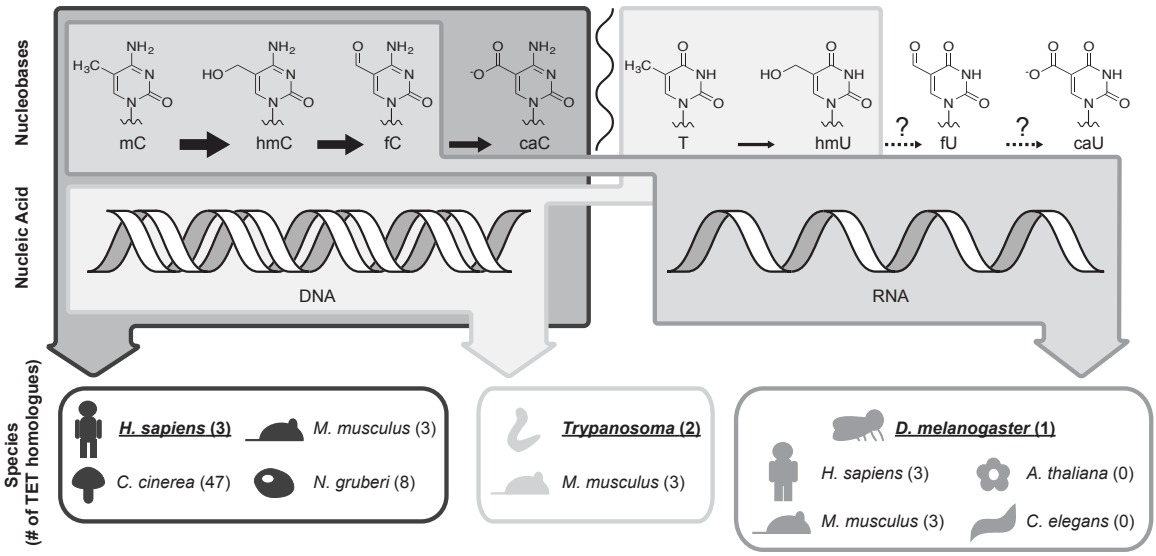


Figure 1-2. The broad scope of TET reactivity.

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 sequentially 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.

Furthermore, although cytosine analogues are the predominant substrate, TET enzymes were shown to have limited activity on thymine in cells (Figure 1-2) (Pfaffeneder et al., 2014). The resulting 5-hydroxymethyluracil was proposed to trigger DNA repair pathways or serve as another epigenetic base, and novel methods for its localization have been developed (Yu et al., 2015). Notably, some dioxygenases in the Fe(II)/αKG-dependent family, such as the trypanosomal JBP1 and JBP2, naturally hydroxylate thymine rather than cytosine (Cliffe et al., 2009). By comparing structure-function relationships across these enzymes, future studies could aim to modulate TET activity on various substrates, possibly including unnatural modifications or nucleic acid analogues.

Understanding the scope of reactivity goes hand-in-hand with the biological contexts where TET homologues are found (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 et al., 2014a). As noted, TET enzymes are also present in organisms such as *D. melanogaster* that lack genomic mC, suggesting potential activity on RNA bases or non-cytosine bases such as N6-methyladenosine, or non-catalytic functions (Delatte et al., 2016; Fu et al., 2014; Zhang et al., 2015). Thus, while TET enzymes have garnered the most attention for their role in DNA demethylation and epigenetic cytosine modification, non-canonical activities suggest possible involvement in even more diverse processes.

1.3: Properties and localization of ox-mC bases

1.3.1: Biophysical and chemical characteristics of ox-mC bases

To bridge from TET enzyme mechanisms to potential epigenetic functions, it is important to understand the biophysical and chemical properties of the ox-mC bases. This has implications for the direct impact of these bases in DNA, but unique properties can also be exploited to detect or localize these rare modifications, offering insight into their potential functions. Indeed, numerous methods to map genomic ox-mCs have been developed (Booth et al., 2015; Wu and Zhang, 2015) and are discussed in the next section.

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 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). Repeats of fC-containing CpGs, which can occur in the genome, were also found to underwind the DNA helix, resulting in a distinct F-form conformation that could influence chromatin packaging or interaction with fC-specific protein readers (Raiber et al., 2015). Moreover, DNA containing at least one fC displayed greater flexibility in a single-molecule cyclization assay, and the increased DNA flexibility correlated with enhanced nucleosome stability (Ngo et al., 2016). Overall, however, the evidence suggests that 5-modification of cytosine largely maintains the structural and sequence integrity of DNA while providing an accessible handle for epigenetic readouts.

The 5-modified groups also offer opportunity for chemical or enzymatic manipulation, which is particularly important for designing assays to distinguish the identity of ox-mC bases (Figure 1-3). The hydroxymethyl group can be glucosylated using UDP-glucose, a reaction typically catalyzed by T4 β-glucosyltransferase (βGT) (Josse and Kornberg, 1962). The glucose moiety can itself be modified, such as with an azide group for downstream biotinylation (Song et al., 2011b), or oxidized with sodium periodate to create an aldehyde handle for pulldown of hmC-enriched DNA (Pastor et al., 2011). In this way, the first genome-wide maps for hmC were generated, though base-resolution sequencing would later provide finer detail. The aldehyde of fC also enables versatile reactions with probes such as primary amines (Hu et al., 2013a), hydroxylamines (Pfaffeneder et al., 2011; Raiber et al., 2012; Song et al., 2013), and hydrazines (Xu et al., 2014), while caC can react specifically with carbodiimides (Ito et al., 2011; Lu et al., 2013). Although signal-to-noise often complicates readout of rare modifications, such chemical tools have made great strides toward sensitive quantification and localization of ox-mCs in various genomes.

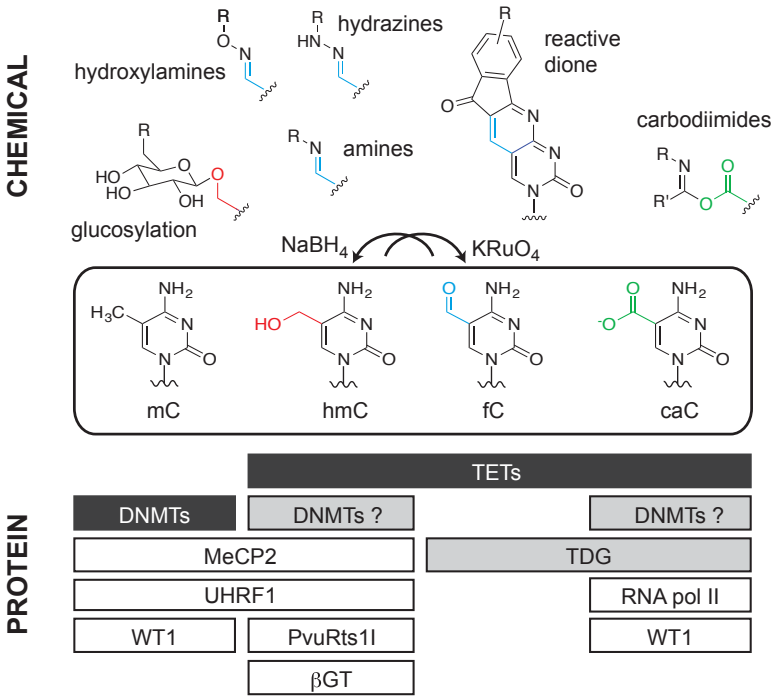


Figure 1-3. Ox-mCs as targets for both chemical manipulation and protein interactions.

Top: Summary of reaction products after treating hmC (red), fC (blue), or caC (green) with classes of chemical modifiers. These modifiers often contain fluorescent labels or provide handles for purification, detection, or localization of ox-mCs in the genome. Bottom: Exogenous and endogenous proteins—some well-characterized and others only recently proposed—can serve as writers (black), erasers (gray), and readers (white) of epigenetic cytosine modifications.

1.3.2: Leveraging the characteristics of ox-mCs for genomic sequencing technologies

Base-resolution sequencing has arguably been the most common application for chemoenzymatic manipulation of cytosine modifications. Most techniques have built on bisulfite sequencing (BS-seq) as a foundation. BS-seq was considered the "gold standard" for reading out C vs. mC: it uses sodium bisulfite to induce chemical deamination of cytosine to uracil, which is read as T upon sequencing, while mC is refractory to deamination and is read as C (Frommer et al., 1992). However, with the expansion of the epigenome beyond mC, it soon became apparent that BS-seq does not distinguish between all the cytosine forms, since C/fC/caC are deaminated while mC/hmC are not (Huang et al., 2010). This necessitated a more nuanced and cautious

interpretation of BS-seq results and drove the field to develop variants of BS-seq that were specific for each modification.

The first milestones came with oxidative bisulfite sequencing (oxBS-seq) and Tet-assisted bisulfite sequencing (TAB-seq) for hmC. In oxBS-seq, hmC is selectively oxidized to fC using potassium perruthenate, allowing for subsequent deamination via bisulfite treatment (Figure 1-3) (Booth et al., 2012). As a result, only mC remains resistant to deamination and is read as C, whereas C/hmC/fC/caC are read as T. By subtracting from regular BS-seq results, the original hmC marks can be identified. TAB-seq avoids the subtraction step, allowing direct readout of hmC. By this method, genomic hmC is first protected by βGT-catalyzed glucosylation; then, the DNA is treated with Tet1 under conditions that oxidize all the mC to fC and caC, allowing for deamination by bisulfite (Yu et al., 2012). Since hmC was protected, it alone is read as C. Together, oxBS-seq and TAB-seq indicated that hmC is relatively enriched at promoter-distal gene-regulatory regions, such as enhancers and insulators, as well as immediately adjacent to transcription factor binding sites and at transcriptionally poised promoters (Booth et al., 2012; Yu et al., 2012). This non-random localization to key regulatory elements hints at fundamental epigenetic functions.

Similar trends were observed with bisulfite-based techniques for fC and caC sequencing, which again leverage the chemical properties of these bases to alter their capacity for deamination. In fC chemically assisted bisulfite sequencing (fCAB-seq), *O*-ethylhydroxylamine protects fC from deamination (Song et al., 2013); for caC (caCAB-seq), reaction with a carbodiimide protects caC (Figure 1-3) (Lu et al., 2013). In reduced bisulfite sequencing (redBS-seq), sodium borohydride selectively reduces fC to hmC (Booth et al., 2014). In all these cases, fC or caC sites are determined by subtracting the results from regular BS-seq. For a more direct readout of fC and caC together, methylase-assisted bisulfite sequencing (MAB-seq) uses the bacterial DNA methyltransferase M.SssI to methylate genomic DNA at CpGs, so that fC and caC

but not unmodified C are left susceptible to deamination (Neri et al., 2015b; Wu et al., 2014). Pretreatment of DNA with sodium borohydride, followed by MAB-seq, results in direct readout of caC at CpGs (Wu et al., 2014). Notably, these forms of analysis are complicated by the rarity of fC and caC, and many low-level modifications (as well as non-CpG modifications) are likely being overlooked due to methodological limitations. *Tdg* knockout in ESCs is usually employed to bolster genomic levels of fC/caC and to identify potential sites of active demethylation. These sites have so far included DNaseI-hypersensitive sites, poised enhancers, insulators, exons, and exon-intron boundaries (Neri et al., 2015b; Wu et al., 2014).

Importantly, bisulfite treatment carries several drawbacks, most notably DNA degradation (up to ~90%) via depyrimidination (Tanaka and Okamoto, 2007). Techniques are increasingly sought that avoid harsh chemical processing, since this could allow for analysis of much smaller quantities of DNA. Enzymes that recognize specific modified cytosines naturally offer some of the best tools. The PvuRts1I family of restriction endonucleases, originally found to restrict T-even phage, selectively cleave DNA containing hmC or glucosyl-hmC but not C or mC (Figure 1-3) (Szwagierczak et al., 2011; Wang et al., 2011). These enzymes can digest genomic DNA in a non-sequence-specific manner, resulting in fragments that contain hmC. Affinity enrichment for hmC-containing digestion products helps to counter the lack of specificity for fC and caC, as well as any residual activity on C or mC (Sun et al., 2013; Sun et al., 2015). By sequencing the fragments, this "Pvu-Seal-seq" approach mapped genomic hmC with such high sensitivity that it found ~10-fold more hmC sites compared to whole-genome TAB-seq analysis (Sun et al., 2015). In particular, Pvu-Seal-seq detected low-level modifications at many non-CpG sites, which cannot readily be found by TAB-seq or oxBS-seq. A variant of Pvu-Seal-seq has also provided direct readout of fC (Sun et al., 2015). Additional bisulfite-free approaches are actively being pursued, including single-molecule real-time (SMRT) sequencing (Song et al., 2011a) and nanopore-based technologies (Wescoe et al., 2014) that do not require PCR amplification. All

told, the genomic profiles of ox-mC bases continue to resolve in finer detail, revealing patterns that invite a broader understanding of the functions played by TET enzymes and cytosine modifications.

1.4: Potential functions of TET and ox-mC bases

1.4.1: General insights from TET deletions and systems-level analyses

While biochemistry and chemical biology continue to delve into the complex mechanisms and properties of TET enzymes and ox-mC bases, cell biology approaches are drawing connections to wide-ranging functions. In general, many critical questions remain open, including the specific roles of each TET isoform and each oxidized modification. Nevertheless, TET enzyme activity has well-established contributions to zygotic development and hematological and neurological health. In addition, fundamental roles in transcription regulation, splicing, genome stability, and metabolism increasingly seem more plausible.

TET knockout models provided a starting point for functional studies. Tet3 deletion leads to neonatal lethality in mice (Gu et al., 2011). Although the precise mechanisms are still debated, Tet3 is the only isoform expressed in zygotes and is likely to mediate essential demethylation processes in early development. By contrast, mice carrying Tet1 or Tet2 single gene deletions are viable and fertile, though they display abnormal methylation patterns (Dawlaty et al., 2011; Li et al., 2011; Moran-Crusio et al., 2011). Tet1 knockout mice tend to have decreased body mass and smaller litter size (Dawlaty et al., 2011), along with deficits in learning and memory (Rudenko et al., 2013; Zhang et al., 2013). Tet2 knockout mice are predisposed to hemapoietic malignancies such as chronic myelomonocytic leukemia (Li et al., 2011; Moran-Crusio et al., 2011), consistent with human TET2 loss-of-function mutations being common in many hematological disorders (Abdel-Wahab et al., 2009; Langemeijer et al., 2009). Since Tet1 and Tet2 are often co-expressed, notably in ESCs, double knockout (DKO) mice were generated to test the extent of

compensatory functions (Dawlaty et al., 2013). Approximately half of these DKO mice were viable and fertile; the other half displayed severe embryonic and neonatal abnormalities, including exencephaly and cerebral hemorrhage, associated with perinatal lethality (Dawlaty et al., 2013). Surviving Tet1/2 DKO mice were prone to developing B cell malignancies (Zhao et al., 2015). Related phenotypes have been observed in other organisms, such as in zebrafish, which require Tet2 and Tet3 homologues for the production of hematopoietic stem cells (Li et al., 2015a).

The challenge now lies in dissecting the exact contributions of each TET isoform and ox-mC base. This is made more difficult by the tendency for two or three TET isoforms to be co-expressed, as well as by the stepwise nature of oxidation. Since any hmC (or fC) that is generated is potentially a substrate for further oxidation, it has so far not been possible to isolate the potential functions of any single modification. However, important strides have been made toward identifying reader proteins for these bases. The transcription factor methyl-CpG-binding protein 2 (MeCP2) was validated as a major hmC-binding protein in the brain, with similar affinity for mC and hmC (Mellen et al., 2012), while Wilms tumor protein 1 (WT1) can recognize caC as well as C and mC (Figure 1-3) (Hashimoto et al., 2014a). Proteomics analyses found numerous hits specific for hmC, fC, or caC (Iurlaro et al., 2013; Spruijt et al., 2013). Interestingly, in both reports, the proteins associated with fC (and caC) greatly outnumber those for hmC, despite hmC being far more prevalent in the genome. On the other hand, a systems analysis of 77 genomic co-localization studies suggested that hmC is the busiest "hub" in the epigenetic communication network within ESCs, linking and even influencing co-evolution of many chromatin-related proteins (Juan et al., 2016). Validation of these interaction and/or signaling networks now becomes a priority, along with understanding the effects of these interactions.

1.4.2: Impacts on transcription

One key question is the effect of ox-mCs on gene expression, either by way of DNA demethylation or as epigenetic marks in their own right. TET enzyme activity is perhaps best known for modulating the expression of pluripotency markers (Oct4, Nanog, etc.) during development and reprogramming (Gu et al., 2011). MicroRNA such as the miR-200 family (Hu et al., 2014), various transcription factors such as Runx1 (Li et al., 2015a), and regulators of meiosis (Yamaguchi et al., 2012) are among other proposed targets. It remains unclear whether mC oxidation alone is sufficient for reversing transcriptional silencing or whether demethylation is required. Toward this end, fusions of TET enzymes to zinc fingers (Chen et al., 2014), transcription activator-like effectors (TALEs) (Maeder et al., 2013), or catalytically inactive Cas9 (Amabile et al., 2016; Liu et al., 2016c) have been used generate ox-mCs at targeted genomic locations. When targeted to a methylated, silenced promoter, these methods can activate transcription, though the potential interplay of stepwise oxidation and base excision repair obscures the mechanisms responsible for reactivation.

A more detailed study used reporter plasmids that were CpG-methylated and TET-oxidized *in vitro* (Muller et al., 2014). When these plasmids were transfected into ESCs, methylation silenced the reporter gene, while the presence of ox-mCs restored expression. Using Tdg-deleted ESCs and complementation with catalytically altered TDG mutants, the study found that TDG-dependent demethylation via both fC and caC was primarily responsible for reactivation of gene expression (Muller et al., 2014). Two important caveats, though, are the inability to restrict cytosine modifications to the most relevant locations on the plasmid (e.g. the promoter), since all CpGs would be modified by this method, and the inability to control which ox-mC marks are made. The ideal experiment would directly demonstrate modification-specific effects on gene expression, but technical challenges must be overcome.

Transcription elongation is also potentially impacted, considering the localization of cytosine modifications to gene bodies and exon-intron junctions (Huang et al., 2014; Iurlaro et al., 2016; Raiber et al., 2012; Wen et al., 2014). Yeast and mammalian RNA polymerase II were found to have reduced elongation rates on fC and caC templates, as well as higher tendencies to backtrack and poorer substrate specificity compared with C/mC/hmC templates (Kellinger et al., 2012). A structural study corroborated that the yeast RNA pol II elongation complex can form hydrogen bonds specifically with caC, resulting in transient transcriptional pausing (Wang et al., 2015). Stalling of transcription opens the door for ox-mCs to play roles in transcriptional fidelity, mRNA processing, chromatin remodeling, and recruitment of additional proteins, especially TDG (Kellinger et al., 2012; Oberdoerffer, 2012; Wang et al., 2015). One study so far has found a partnership between cytosine modifications and CTCF in the regulation of alternative splicing (Marina et al., 2016). CTCF binding sites are present within gene bodies and are enriched for ox-mCs (Sun et al., 2013; Wu et al., 2014; Yu et al., 2012), and CTCF protein preferentially interacts with unmodified C as well as caC (Marina et al., 2016). Intragenic binding of CTCF induces transcriptional stalling, which favors exon inclusion by a kinetic effect (Shukla et al., 2011; Wada et al., 2009). The proposed model suggests that mC evicts CTCF, promoting exon exclusion, while hmC and caC restore CTCF binding and exon inclusion (Marina et al., 2016). Further intersections between transcription and ox-mCs will likely come to light; one particularly interesting angle is whether TET enzymes themselves (or TDG) could act co-transcriptionally.

1.4.3: Connecting TET activity to other fundamental cellular processes

Beyond transcription, TET enzymes and ox-mCs have been implicated in a host of fundamental processes, most notably genome stability. On the one hand, a few reports have suggested that ox-mCs can compromise genome integrity: hmC can promote aneuploidy in cultured cells (Mahfoudhi et al., 2016), or mutagenesis can occur at CpGs containing both caC

and an mC-to-T deamination event (Weber et al., 2016). However, these scenarios are largely artificial, and the majority of evidence favors a role in the maintenance of genome integrity. Two studies described a requirement for TET enzymes in telomere stability (Lu et al., 2014; Yang et al., 2016). Another proposed that hmC is a marker for DNA damage induced by the DNA polymerase inhibitor aphidicolin and is also required for repair (Kafer et al., 2016).

Perhaps most of all, TET activity is thought to protect CpG-dense regions from aberrant methylation, which can lead to C-to-T transition mutations via spontaneous deamination (Bellacosa and Drohat, 2015; Jin et al., 2014; Wiehle et al., 2015). Genome sequencing has mapped ox-mCs to the edges of DNA methylation "canyons," which are generally defined to be hypomethylated regions >3.5 kb containing numerous CpGs (Wiehle et al., 2015). Loss of Tet1 or Tet2 leads to "erosion" of these canyons, with increased methylation spreading into the region (Jin et al., 2014; Wiehle et al., 2015). The proposed mechanism relies on targeting of TET enzymes to unmodified CpGs, with presumed searching for nearby mC substrates. TET1 and TET3 contain CXXC domains that preferentially bind unmodified cytosine, while the *TET2* gene underwent a chromosomal inversion such that its CXXC domain separated into the protein IDAX (Ko et al., 2013). This provides one plausible explanation for why TET enzymes should contain a domain that recognizes an unreactive CpG. It also highlights the necessity for biochemical studies on full-length TET enzymes, since most experiments so far have been limited to the better-behaved catalytic domain, which lacks the N-terminal CXXC region.

Finally, from a holistic point of view, TET enzymes are potentially positioned at the crossroads of not only epigenetic regulation but also metabolism. The dependence on αKG as a co-substrate, coupled with complex interplay with chromatin-related proteins, invites speculation that TET enzymes could form an axis linking metabolic status with epigenetic DNA and chromatin modifications. Indeed, the intersections of epigenetics and metabolism are increasingly appreciated, not only for TET enzymes but also for Fe(II)/αKG-dependent histone demethylases

and SAM-dependent DNMTs and histone methyltransferases (Janke et al., 2015). In summary, myriad connections can be drawn between TET enzymes, ox-mC bases, and biological functions. Given this intricate level of regulation, it is not surprising that perturbations to TET activity manifest in numerous pathologies, as discussed below.

1.5: Pathological perturbations to TET enzyme activity

TET1 was originally identified as a fusion protein to MLL in acute myeloid leukemia (Lorsbach et al., 2003; Ono et al., 2002), and cancers continue to encompass the bulk of TET-related pathologies (Huang and Rao, 2014; Scourzic et al., 2015). TET2 is mutated in approximately 15-40% of various myeloid disorders (Abdel-Wahab et al., 2009; Delhommeau et al., 2009; Langemeijer et al., 2009; Scourzic et al., 2015). These mutations are generally thought to be loss-of-function, with patient samples showing uniformly decreased hmC levels (Ko et al., 2010). Numerous solid tumors, including colon, lung, and endometrial cancers, also harbor mutations in TET1, 2, or 3 (Scourzic et al., 2015). In the absence of TET mutations, TET activity is still often downregulated in cancers, though the mechanisms are unclear. Global reduction in hmC is proposed as a prognostic biomarker, since low hmC correlates with increased tumor growth and metastasis in breast, liver, lung, pancreatic, and prostate cancers (Yang et al., 2013), as well as in melanoma (Lian et al., 2012). Remarkably, re-introducing active TET catalytic domain into colon and melanoma cell lines was shown to slow cancer cell proliferation and reduce invasiveness, leading to prolonged survival in mouse xenograft models (Lian et al., 2012; Neri et al., 2015a).

In addition, metabolic dysregulation can subvert TET function in cancers. One proposed mechanism involves post-translational modification of TET enzymes by *O*-linked β-*N*-acetylglucosamine transferase (OGT). Cancer cells increase glucose utilization (the Warburg effect), which leads to increased production of UDP-*N*-acetylglucosamine (UDP-GlcNAc) (Love

and Hanover, 2005). OGT enzymes use UDP-GlcNAc to modify target proteins at serine and threonine residues. One study found that all three TET isoforms can be *O*-GlcNAcylated, which can stabilize TET protein levels and promote TET3 export from the nucleus, but the effects on TET1 and TET2, as well as effects on catalytic activity, remain unclear (Zhang et al., 2014b). Interaction with TET enzymes may also serve to recruit OGT to histones, especially at H3K4me3-positive CpG-rich promoters of actively transcribed genes (Chen et al., 2013b; Deplus et al., 2013; Vella et al., 2013). Thus, disrupted glucose metabolism can impact TET protein stability and potentially activity, and loss of TET in cancers may perturb the interaction of OGT with other epigenetic targets.

Metabolites such as fumarate and succinate can further disrupt the activity of TET and other Fe(II)/αKG-dependent dioxygenases by competing with αKG (Laukka et al., 2016; Xiao et al., 2012). The Krebs cycle enzymes fumarate hydratase and succinate dehydrogenase are tumor suppressors that are frequently mutated in cancers, leading to accumulation of fumarate and succinate, respectively, up to millimolar levels (Pollard et al., 2005). This can also trigger dysregulation of hypoxia-inducible factor and its downstream targets (Laukka et al., 2016; MacKenzie et al., 2007; Thienpont et al., 2016)—another growing link between TET enzymes and key intracellular processes.

Another important metabolic correlate involves the products of isocitrate dehydrogenase (IDH1 and IDH2) enzymes, which normally convert isocitrate to αKG (Figure 1-4). Cancer-associated, gain-of-function mutations in IDH result in aberrant generation of the oncometabolite 2-hydroxyglutarate (2HG) (Dang et al., 2009; Ward et al., 2010), which competitively inhibits αKG-dependent dioxygenases, including TET (Xu et al., 2011). A photocaged variant of the mutant IDH, generated using an expanded genetic code, has been used to show rapid metabolic perturbations and changes to ox-mC levels upon activation of the neomorphic enzyme (Walker et al., 2016). In strong support of their overlapping pathway, TET and IDH mutations—both

independently common in acute myeloid leukemia—are never seen to co-exist (Figueroa et al., 2010; Gaidzik et al., 2012; Losman and Kaelin, 2013).

Finally, a recent study shows how cytosine modifications and cancer can intersect in nucleotide salvage pathways. Upregulation of the salvage enzyme cytidine deaminase (CDA) in some cancer cell lines resulted in high levels of the deamination products hmU and fU, which could promote cell death when incorporated into DNA (Figure 1-4) (Zauri et al., 2015). All these advances point to novel angles on the biology of ox-mCs and potential therapeutic strategies that merit further exploration.

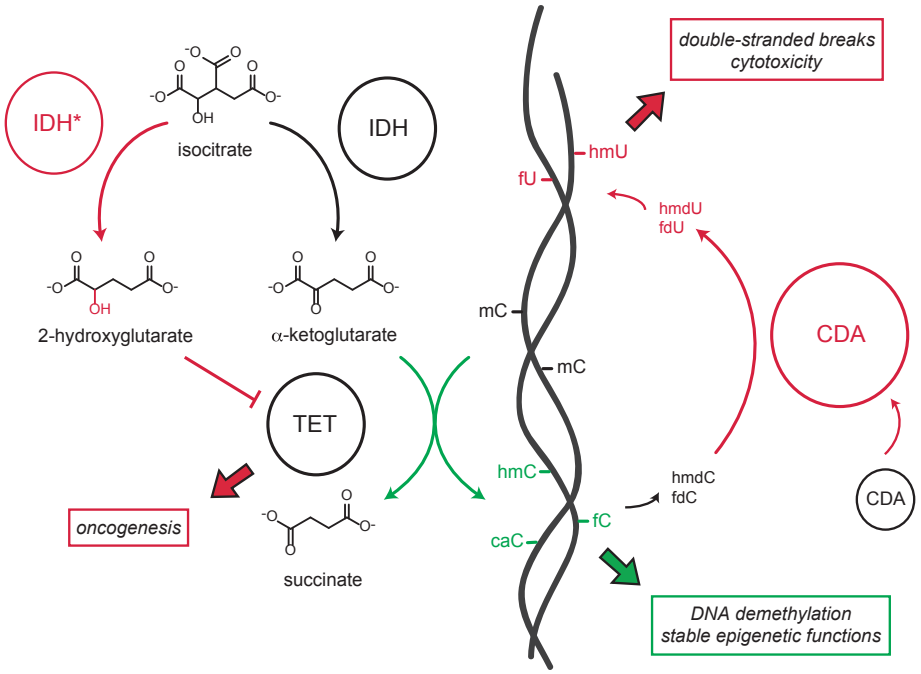


Figure 1-4. Intersection of ox-mCs, metabolism, and pathology.

Physiologically (green), TET enzymes utilize the cofactor αKG supplied by isocitrate dehydrogenase (IDH) to generate ox-mCs, which may contribute to DNA demethylation and other epigenetic functions. Pathologically (red), in some hematologic cancers, gain-of-function IDH mutants (IDH*) instead produce the oncometabolite 2-hydroxyglutarate (2HG), which acts as a competitive inhibitor of TET. As another example, several cancer cell lines upregulate cytidine deaminase (CDA), which can deaminate hmdC and fdC nucleosides. Incorporation of the resultant hmdU and fdU into DNA can lead to DNA breaks and potential cytotoxicity.

1.6: Thesis objectives

Thus, the study of TET enzymes continues to probe new and exciting directions in biology, underscoring the interconnectedness within not only the epigenome but also the life of a cell. This thesis aims to elucidate the biochemical and structural mechanisms that connect the steps of TET-catalyzed oxidation. By understanding in particular how TET enzymes can generate rare fC and caC bases seemingly against the odds, we endeavor to disrupt these steps in order to probe whether fC and caC are indeed essential epigenetic bases, or whether hmC alone may be sufficient for function. In Chapter 2, I describe the novel cellular and *in vitro* assays that our lab developed to address these mechanistic questions in quantitative detail. In Chapter 3, my co-authors and I apply these methods to a rigorous kinetic analysis of stepwise oxidation. We show that mouse Tet2 is capable of iterative oxidation, defined specifically as the ability to catalyze multiple steps of oxidation without fully releasing the DNA strand, and that Tet2 has *de novo* activity, which is minimally impacted by modifications to the opposite strand CpG. The results imply that Tet2 can establish fC and caC marks efficiently under certain circumstances, potentially accessing the functions of these bases without first accumulating hmC. In Chapter 4, we take a structure-function approach to asking how the active site of human TET2 is shaped to influence the progression of stepwise oxidation. Combining biochemistry with computational modeling, we demonstrate that a conserved active site scaffold is required for WT stepwise oxidation. These results suggest that the TET2 active site is specifically shaped to enable formation of fC and caC bases, not just hmC. Furthermore, mutations along the scaffold can reconfigure active site interactions to effectively stall oxidation at hmC, providing the first enzyme variants that break the link between hmC and fC/caC. In Chapter 5, I discuss progress toward introducing these mutants into cellular and *in vivo* models to explore whether hmC is sufficient for various functions or whether fC and caC are required. All together, this work reveals key, intrinsic properties of TET enzymes and lays the groundwork for translating these

properties into a better understanding of how specific ox-mC bases and specific TET isoforms contribute to biological functions.

CHAPTER 2: Quantification of oxidized 5-methylcytosine bases and TET enzyme activity

This chapter has been adapted from the following publication:

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.[*]

2.1: Abstract

To gain insight into the mechanisms and functions of TET family enzymes, rigorous approaches are needed to quantify TET activity in cells and *in vitro*. When we began this work, the tools to study TET activity were relatively rudimentary: alongside immunofluorescence, TLC was initially used to quantify mC and hmC, until it became clear that this technique masked fC and caC products, which were discovered later by 2D-TLC. However, a significant limitation of TLC was its blindness to sequence contexts outside of a few compatible restriction sites. This drove the development of improved assays, such as HPLC and LC-MS/MS, though these early advances still lacked the necessary sensitivity to address questions such as enzyme kinetics and the natural occurrence of cytosine modifications in various genomes. Here, we present the tools developed by our lab, some new and some optimized from existing techniques, to report on each of the five forms of cytosine (unmodified, mC, hmC, fC, and caC) with high specificity and sensitivity. We provide detailed protocols for dot blotting and LC-MS/MS analysis of cytosine

[*] Author contributions: I was the primary writer of this methods article. R.M.K., J.E.D., and I conceived the outline, J.E.D. and I made the figures, and everyone edited. Importantly, Danny Crawford led the development of the *in vitro* reaction conditions, isotopically-labeled substrates, and analysis by liquid scintillation counting and nanoLC-MS/MS. I later retooled and further optimized these approaches, especially LC-MS/MS with help from the Garcia and Blair labs. Chris Nabel also contributed to early dot blotting and chemoenzymatic assays, which I expanded. J.E.D. optimized the βGT-ARP-MspI analysis method.

modifications in genomic DNA. We then describe generation of synthetic oligonucleotide substrates for *in vitro* studies, provide optimized reaction conditions, and introduce several chemoenzymatic and isotope-based assays. These approaches enable mechanistic studies of TET activity, which are key to understanding the role of these enzymes in epigenetic regulation.

2.2: Introduction

Ten-eleven translocation (TET) enzymes are Fe(II)/α-ketoglutarate-dependent dioxygenases that are increasingly tied to diverse biological and pathological processes, including cellular differentiation, reprogramming, and malignancy (Tahiliani et al., 2009; Kohli and Zhang, 2013). There are three mammalian TET isoforms (TET1, 2, and 3) that are all capable of sequentially oxidizing 5-methylcytosine (mC) to 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), and 5-carboxylcytosine (caC) (He et al., 2011; Ito et al., 2011). Stepwise oxidation provides a pathway for DNA demethylation, as the highly oxidized bases fC and caC can be selectively removed by thymine DNA glycosylase (TDG), resulting in an abasic site that can be repaired to regenerate unmodified C (He et al., 2011; Maiti and Drohat, 2011). Emerging evidence also indicates that all three oxidized mC bases (ox-mCs) can exist as stable epigenetic marks with potentially independent functions (Bachman et al., 2014; Bachman et al., 2015; Iurlaro et al., 2013; Spruijt et al., 2013). As TET biology has expanded, there has been a need for robust assays to detect, localize, and quantify these rare genomic modifications (Figure 2-1). Numerous chemical methods now exist for quantification and base-resolution sequencing of ox-mC bases in a variety of cell types (Booth et al., 2015; Song et al., 2012; Wu et al., 2016). At the same time, rigorous biochemical assays are needed to address open mechanistic questions. Early approaches to the study of TET enzymes have been reviewed previously (Shen and Zhang, 2012). Here, we present the most current *in vivo* and *in vitro* methods developed by our lab and others to

distinguish between the modified forms of cytosine and measure TET enzyme activity with high sensitivity.

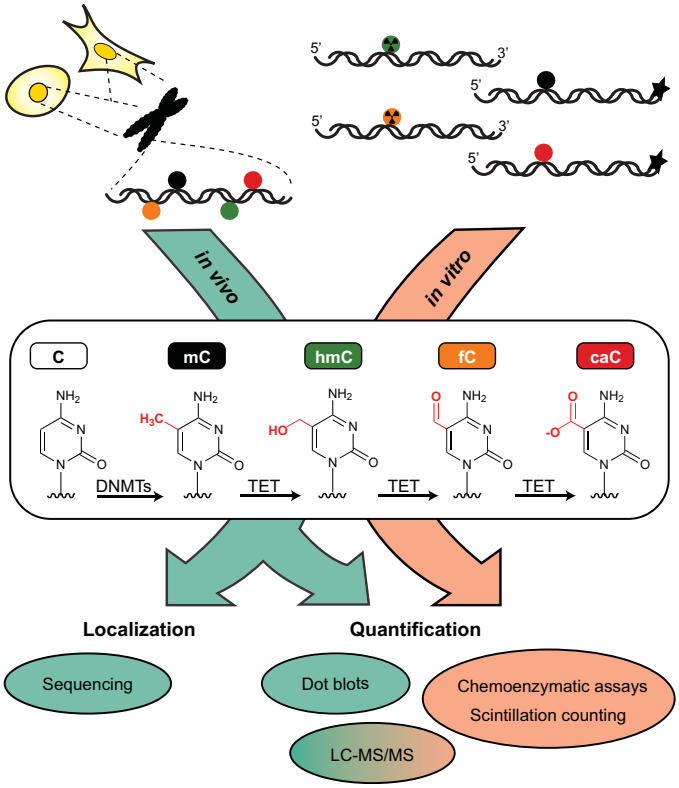


Figure 2-1. Current methods for localization and quantification of TET activity.

These methods provide a snapshot of TET activity, as reflected by modification levels *in vivo* and *in vitro*. Important progress has been made in pushing the limits of detection continually lower, but real-time assays have yet to be developed.

2.3: Analysis of cytosine modifications in cellular DNA

2.3.1: Preparation of genomic DNA from TET-transfected cells

One major area of study centers on the comparative roles of the three TET isoforms and their mutants, many of which have been catalogued in human diseases and may provide key insights into enzyme mechanisms. HEK293T cells provide a convenient overexpression system for assessing the relative activities of TET constructs. HEK293T cells have low levels of

endogenous TET proteins and support efficient transfection and high levels of overexpression. Available constructs include full-length (FL) and catalytic domains (CD) of mouse and human TET1, 2, and 3 cloned into standard mammalian overexpression vectors. Our lab and others have demonstrated activity from the following: hTET1-FL (residues 1-2136), hTET1-CD (1418-2136), hTET2-FL (1-2002), hTET2-CD (1129-2002), hTET3-FL (1-1660), mTet1-FL (1-2007), mTet1-CD (1367-2039), mTet2-FL (1-1912), mTet2-CD (1042-1921), mTet3-FL (1-1668), and mTet3-CD (697-1668) (He et al., 2011; Ito et al., 2010; Tahiliani et al., 2009). In addition, Hu *et al.* crystallized a truncated form of the hTET2 catalytic domain (hTET2-CS, 1129-1936 Δ1481-1843) and demonstrated activity similar to the full-length construct (Hu et al., 2013b). Common negative controls include the corresponding empty expression vector or mutation of the iron-binding HxD motif in the catalytic domain that renders TET inactive.

We use the following protocol to overexpress TET constructs. First, culture HEK293T cells in Dulbecco's Modified Eagle Medium with GlutaMAX (Thermo Fisher Scientific) and 10% fetal bovine serum (Sigma). When cells are 70-90% confluent, transfect using Lipofectamine 2000 (Thermo) according to the manufacturer's protocol. Change media 24 h post-transfection, harvest cells by trypsinization at 48 h, and resuspend pellets in phosphate buffered saline. A sample of transfected cells can be set aside for Western blotting to evaluate protein expression (see Section 2.3.2). Purify genomic DNA (gDNA) using the DNeasy Blood & Tissue Kit (Qiagen), including addition of RNase A where specified in the manufacturer's protocol. Note that this kit is also adaptable for extracting gDNA from tissue specimens and cells that express endogenous TET enzymes.

2.3.2: Qualitative analysis by dot blotting

Dot blots are commonly used to probe for modified bases in gDNA. DNA is denatured to expose the bases, spotted onto an absorbent membrane, and probed with antibodies against each

of the four cytosine modifications. Dot blots offer a clear visual result and can be performed using either serial dilutions or single concentrations of DNA. We consider the former to be semi-quantitative, while the latter is only qualitative but still particularly useful for screening a large number of samples. Dot blotting also works for plasmids but is generally not well suited for short oligonucleotides, likely because these do not adhere consistently to membranes.

The first step is to determine the appropriate amount of DNA for blotting, considering the amount of expected modifications. For gDNA from HEK293T cells overexpressing TET, load 400 ng of gDNA into each well of a Bio-Dot microfiltration apparatus (Bio-Rad). Calculate the total amount of DNA needed (based on number of blots and number of serial dilutions) and dilute to 10 ng/µL in TE buffer (10 mM Tris-Cl, pH 8.0, 1 mM EDTA). Add ¼ volume of 2 M NaOH/50 mM EDTA. Denature the DNA at 95 °C for 10 min, transfer quickly to ice, and add 1 volume of ice-cold 2 M ammonium acetate to stabilize single strands. Serial dilutions may be performed at this point into TE buffer. Meanwhile, prepare membranes for blotting; we have found that Sequi-Blot PVDF membranes (Bio-Rad) give cleaner results than nitrocellulose. Wet membranes in methanol and equilibrate in TE buffer; then, assemble the dot blotting apparatus, taping off any unused wells. Wash each well with 400 µL TE and draw through with gentle vacuum. Purge any air bubbles in the wells, as these can interfere with washing and spotting DNA, and release the vacuum gently to avoid regurgitation that can cross-contaminate wells. Apply 100 µL of DNA samples at the desired dilutions and wash with another 400 µL of TE. Carefully place the membranes into 50 mL conical tubes for blotting. Note that replicate membranes are needed for each separate mC, hmC, fC, and caC blot.

The blotting procedure begins with blocking for 2 h in TBST buffer (50 mM Tris-Cl, pH 7.6, 150 mM NaCl, 0.5% Tween 20) with 5% milk at room temperature. Then, wash 3 times with TBST and incubate at 4 °C overnight with primary antibodies against each modified cytosine (Active Motif offers mouse monoclonal mC and rabbit polyclonal hmC, fC, and caC antibodies).

We use the following antibody dilutions in 5% milk/TBST: 1:5,000 mC; 1:10,000 hmC; 1:5,000 fC; 1:10,000 caC. Volumes should be enough to cover the membrane evenly, and solutions should be poured off cleanly between steps. Wash the blots 3 times with TBST for 5 min each and incubate with secondary 1:2,000 goat anti-mouse-HRP or 1:5,000 goat anti-rabbit-HRP (Santa Cruz Biotechnology) at room temperature for 2 h. Wash 3 times again and, just before imaging, apply Immobilon Western Chemiluminescent HRP Substrate (Millipore) evenly over the entire blot. Expose on an imager with chemiluminescent detection capabilities (we use a Fujifilm LAS-1000), taking care to smooth the blot over the imaging surface and remove air bubbles and excess HRP substrate. As positive and negative controls for this optimized protocol, we typically use gDNA from cells transfected with WT hTET2-CD or empty vector, respectively. Figure 2-2 shows an example of dot blotting results for select TET constructs.

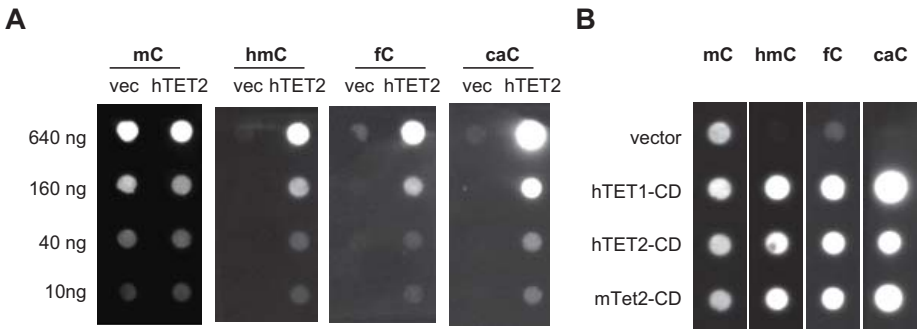


Figure 2-2. Representative dot blots of gDNA from transfected HEK293T cells.

Shown are **(A)** serial dilutions of gDNA from cells transfected with either empty expression vectors or hTET2-CD, and **(B)** dot blots on 400 ng of gDNA from HEK293T cells transfected with empty expression vector, hTET1-CD, hTET2-CD, or mTet2-CD.

2.3.3: Quantitative analysis by LC-MS/MS

To quantify genomic levels of cytosine modifications more rigorously, we favor liquid chromatography tandem mass spectrometry (LC-MS/MS). Several alternatives exist to detect global cytosine modifications, including antibody-based or chemoenzymatic assays (Booth et al.,

2015; Song et al., 2012), but we focus on LC-MS/MS as a direct, reliable, and flexible assay that can quantify diverse modifications simultaneously. A host of sequencing methods have also been developed to localize specific modifications at base resolution; these methods are not covered in this article, as they have been thoroughly reviewed elsewhere (Booth et al., 2015; Wu et al., 2016) (see also Section 1.3.2). Here, we discuss the protocols we developed for micro- and nano-scale analysis of nucleosides by LC-MS/MS, though we note that multiple methods and instruments have been described (Bachman et al., 2014; Pfaffeneder et al., 2014; Tsuji et al., 2014), and optimization for each system will be unique.

We use an Agilent 6460 triple quadrupole mass spectrometer with Agilent 1200 Series HPLC and Supelcosil LC-18-S reverse phase analytical column (5 μm particle size, 2.1 mm x 25 cm, Sigma) (Figure 2-3A). This system offers a good starting platform for nucleoside analysis, is well suited for most applications, and is likely more accessible for most researchers than the nano-scale setup described below. To prepare samples, concentrate 10 μg of purified gDNA by ethanol precipitation and, in a total of 20 μL, degrade the DNA to component nucleosides with 10 U DNA Degradase Plus (phosphodiesterase and phosphatase cocktail available from Zymo Research) in 1X DNA Degradase Buffer (Zymo). Incubate this mixture at 37 °C overnight, then dilute 10-fold into 0.1% formic acid, since on-column retention of caC in particular requires pH adjustment.

Equilibrate the HPLC column to 50 °C in Buffer A1 (5 mM ammonium formate, pH 4.0). Load 2 μL (up to 0.1 μg) of the nucleoside mixture and separate by gradient elution at a flow rate of 0.5 mL/min: 0-5 min, 0% Buffer B1 (4 mM ammonium formate, pH 4.0, 20% (v/v) methanol); 5-12 min, 0-10% B1; 12-12.5 min, 10-100% B1; 12.5-20 min, 100% B1; 20-20.5 min, 100-0% B1; 20.5-28 min, 0% B1. Note that this gradient is optimized for efficient separation of all modified cytosine nucleosides, but the most hydrophobic bases, such as adenosine, elute at >10% B1. Set downstream positive ion mode electrospray ionization for gas temperature of 175 °C, gas

flow of 10 L/min, nebulizer at 35 psi, sheath gas temperature of 300 °C, sheath gas flow of 11 L/min, capillary voltage of 2,000 V, and fragmentor voltage of 70. Collect MS2 scans on the following mass transitions: mC 242.11→126.066 *m/z*; hmC 258.11→142.061 and 124.051; fC 256.09→140.046; caC 272.09→156.041 (A, C, G, and T are optional but can be useful as a loading control: A 252.11→136.062; C 228.10→112.051; G 268.10→152.057; T 243.10→127.050). Optimized collision energies are 10 for mC, fC, and T; 15 for caC; and 25 for hmC. For quantification, generate standard curves from nucleosides (Berry & Associates) ranging from approximately 5 μM to 10 pM; many isotopically-labeled nucleosides are available as internal standards. Fit all peak areas to the standard curve to determine amounts of each modified cytosine in the gDNA sample.

This method attains low femtomolar- and high attomolar-range detection limits, allowing for quantification of 1 modification in 10^5-10^6 of all cytosines. This provides excellent detection of rare oxidized bases, especially in HEK293T cells overexpressing TET, though quantifying fC and caC in physiological samples remains a challenge. However, the triple quadrupole system has two notable limitations: the need for a large electron multiplier voltage and low mass resolution. At best, the Agilent 6460 differentiates only 0.7 amu, which can present a challenge if analytes are 1 amu or less apart, as might occur in some stable isotope labeling studies.

For analysis of isotopically-labeled cytosines, we use nano-LC in tandem with a Q Exactive hybrid quadrupole-orbitrap mass spectrometer (Thermo Scientific) (Figure 2-3B). A key advantage to this system is the very high mass resolution, which improves signal-to-noise and clearly distinguishes between isotopes with 1 amu mass difference. Detection limits for modified nucleosides are generally in the low femtomole range. However, the system is prone to clogging and inconsistent electrospray, so samples and column fittings should be prepared with care.

We make columns from fused-silica tubing (New Objective) with a frit at one end: Dip the column into a 400 μL mixture of 1:3 formamide:KASIL 1624 potassium silicate solution (PQ

Corporation), let polymerize at 100 °C overnight, and trim to ~3 mm. Using a pressure injection cell, pack a 150 µm x ~15 cm pre-column and 100 µm x ~25 cm analytical column with Supelcosil LC-18-S resin (Sigma). Connect columns to an Easy-nLC 1000 (Thermo) with a two-column setup, and add a 10 µm SilicaTip emitter (New Objective). (Alternatively, nanospray tips can be generated with a laser-based micropipette puller [Sutter Instrument], which eliminates the need for frits and significantly reduces dead volume, but this process is delicate and the setup can clog more easily). Equilibrate the pre-column and analytical column in 5-10 column-volumes of Buffer A2 (0.1% formic acid in H_2O) at a constant pressure of 275 bar.

Prepare samples as described above, and inject 1 µL (0.05 µg) onto the LC-MS system. Set a sample loading step to send 5 µL of Buffer A2 through the sample loop to the pre-column at a constant pressure of 275 bar; this way, the sample is bound to the pre-column and desalted by sending the flow-through to waste. Next, run the gradient at a flow rate of 300 nL/min: 0-2 min, 0% Buffer B2 (0.1% formic acid in acetonitrile); 2-7 min, 0-10% B2; 7-37 min, 10-40% B2; 37-39 min, 40-70% B2; 39-45 min, 70% B2. Perform tandem mass spectrometry in positive ion mode nanospray ionization with spray voltage of 2.9 kV, capillary temperature of 275 °C, and normalized collision energy of 35%. Mass transitions and data analysis are the same as described above for the micro-scale set-up. These LC-MS/MS methods, while technically challenging, are broadly applicable for sensitive quantification of cytosine modifications in gDNA. Moreover, in the next Section, we will return to these methods as a powerful tool for quantifying the results of TET reactions on oligonucleotides *in vitro*.

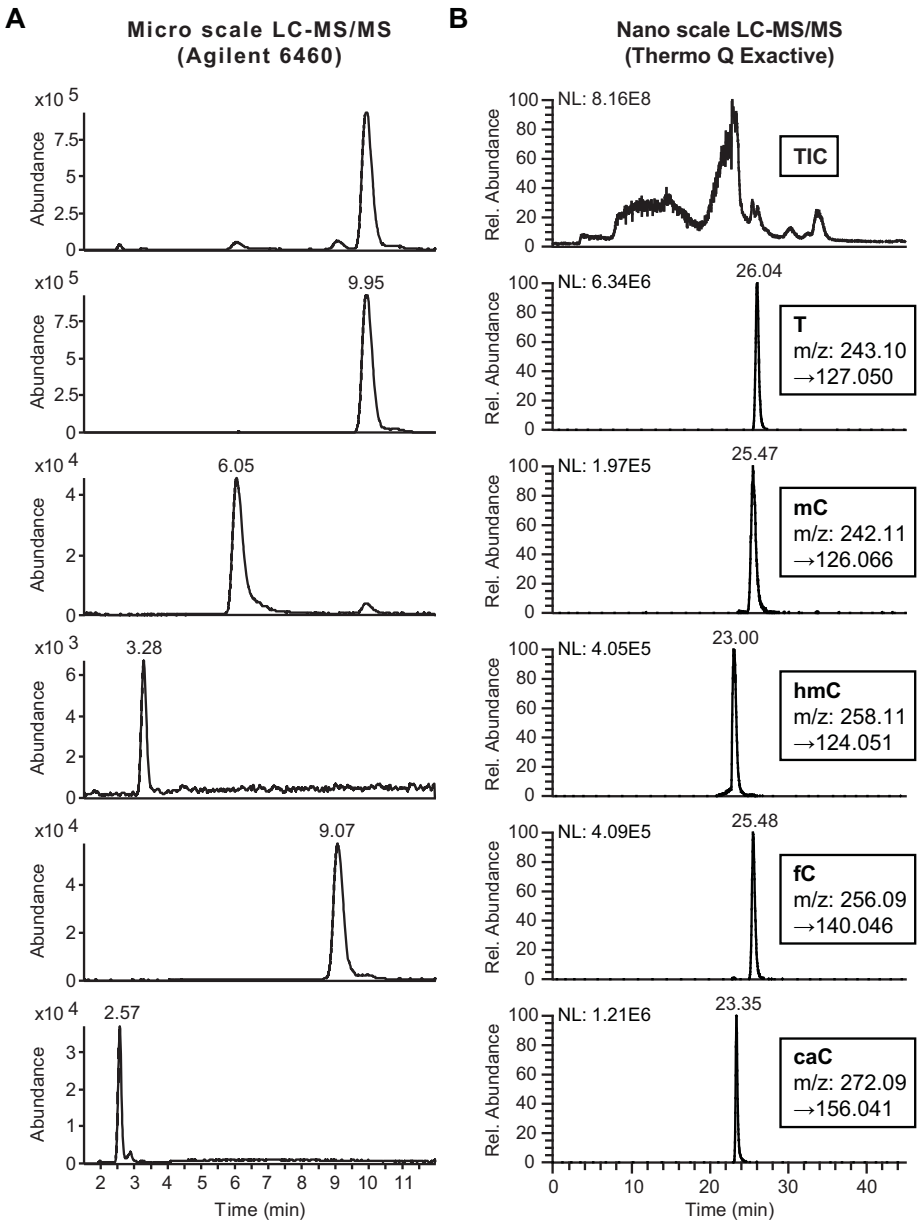


Figure 2-3. Examples of LC-MS/MS analysis of DNA nucleosides.

Protocols can be applied to genomic DNA or oligonucleotide substrates. Shown are the traces from oligonucleotide substrates containing a single mC, treated with hTET2-CS *in vitro*, degraded to nucleosides, and analyzed by **(A)** HPLC with tandem Agilent 6460 triple quadrupole mass spectrometer or **(B)** nano-LC with tandem Thermo Q Exactive hybrid quadrupole-orbitrap mass spectrometer. From top to bottom are the total ion current (TIC) and MS2 chromatograms for T, mC, hmC, fC, and caC deoxyribonucleosides, along with retention times and mass transitions.

2.4: Analysis of cytosine modifications *in vitro*

2.4.1: Purification of TET enzymes from Sf9 insect cells

To prepare recombinant TET enzymes for biochemical analysis, we and most other groups use a baculovirus expression system in Sf9 insect cells. TET constructs are typically cloned into a pFastBac1 vector containing a FLAG-tag at the N terminus. All biochemical studies to date have relied on the catalytic domain constructs, as these retain the activity of the full-length protein but are generally obtained in higher purity and are more stable *in vitro*. Bacmid preparation and transfection were previously described in detail (Shen and Zhang, 2012). We express proteins for only 24 h, which we find minimizes the formation of truncation products. Cells are collected by centrifugation and the pellet is stored at -80 °C until ready for purification. Alternative strategies have been described for expressing and purifying TET proteins from HEK293T cells and, in the crystal structure studies, from *E. coli* (Hu et al., 2013b; Hu et al., 2015); however, we favor Sf9 cells as a relatively straightforward way to obtain pure and active enzyme at intermediate yield.

We adapted a simple protocol for FLAG-based affinity purification of TET enzymes from Sf9 cells; the volumes below are for purification from a 500 mL culture. Resuspend the cell pellet in 10 mL of lysis buffer (50 mM HEPES, pH 7.5, 300 mM NaCl, 0.2% (v/v) NP-40) with cOmplete, EDTA-free Protease Inhibitor Cocktail (Roche, 1 tablet/10 mL) and 10 U/mL of Benzonase Nuclease (Millipore). To lyse the cells, freeze the cell suspension overnight at -20 °C, thaw the following morning on ice, and pass through both a 20-gauge then a 25-gauge needle. For larger volumes, a microfluidizer or sonicator is a good alternative. Clear the lysate by centrifugation at 20,000x*g* for 30 min, collect the supernatant, and pass it through a 0.2 μm syringe filter. Next, prepare a 500 μL column of anti-FLAG M2 affinity gel (Sigma) per manufacturer instructions and equilibrate in lysis buffer. Apply the filtered lysate twice over the column under gravity flow. Wash the protein-bound column with 3 x 10 mL of wash buffer (50

mM HEPES, pH 7.5, 150 mM NaCl, 15% (v/v) glycerol). To elute the protein, apply 500 µL of wash buffer containing 100 µg/mL 3X FLAG peptide (Sigma) to the column, incubating for 5-10 min before collecting the fraction. Collect until no protein is detected by Bio-Rad Protein Assay and SDS-PAGE. Pool fractions, add DTT to 1 mM, aliquot and store at -80 °C. By this method, we are able to obtain highly pure, active TET proteins with minimal steps.

2.4.2: Synthesis and isotopic labeling of TET substrates

Our lab employs a variety of techniques to synthesize substrates for *in vitro* TET reactions and to label the substrates with fluorophores and/or heavy isotopes to enable sensitive detection and quantification. In general, we use DNA oligonucleotides 12-35 nt in length, containing a single TET substrate (mC, hmC, or fC) in a CpG context (although useful HPLC-based assays for substrates as short as 4-6 nt have also been developed (Kizaki and Sugiyama, 2014)). TET enzymes exhibit a strong preference for CpGs (Hu et al., 2013b) but are thought to be less sensitive to surrounding sequences (Yu et al., 2012). Importantly, the self-complementarity of CpGs requires special considerations, since one or both strands of a DNA duplex can contain a TET substrate. This consideration makes it important to control the identity of the top and bottom strands independently, and if both strands are set up to contain a substrate, assays must be designed to distinguish reactivity on each strand.

For most assays, we synthesize oligonucleotides in-house on an Applied Biosystems 394 DNA/RNA synthesizer. All four modified cytosines are now available (Glen Research) and compatible with standard phosphoramidite synthesis protocol, though fC requires post-synthetic processing to obtain the final formyl group from the precursor. Custom-made oligonucleotides are also available from Integrated DNA Technologies (IDT) and the W. M. Keck Biotechnology Resource Laboratory at Yale University, among other facilities. To facilitate tracking of the designated “top” strand, we typically attach 6-carboxyfluorescein (6-FAM) or alternative

fluorophores to either the 5'- or 3'-end of the modified oligonucleotide. This can be done during synthesis or, for 3'-end labeling, can also be done after purification using terminal transferase enzymes and fluorophore-conjugated ddUTP analogs. Importantly, neither the identity nor location of the fluorescent tag alters the reactivity of TET on our substrates. As the first step toward differentiating duplexed strands, we leave the bottom strand unlabeled.

Starting with single strands, the top and bottom strands can be selected to address the experimental goals. In most cases, we use the FAM-labeled top strand containing a single TET substrate and, to ensure that all substrates are double-stranded, anneal a 1.1-1.5-fold excess of unlabeled bottom strand containing unmodified C, so that the resulting DNA duplex contains only one reactive site. Typically, the CpG is embedded in a restriction site for MspI to allow for downstream analysis (see Section 2.4.3). Annealing is performed in a thermocycler: mix the top and bottom strands, heat at 95 °C for 5 min, then cool slowly by decreasing steps of 5 °C for 30 s per step. These duplexes can be added directly to a TET reaction, and reaction products can be visualized by fluorescent detection after separation by denaturing polyacrylamide gel electrophoresis (PAGE), as described in the next Sections.

Fluorescence-based assays are well suited for addressing many important mechanistic questions but have limits of detection in the high femtomole range. To increase the sensitivity of detection for cases of low product formation (e.g. in kinetic studies), one previous strategy has been to radiolabel the 5'-end of full-length substrates with T4 polynucleotide kinase (NEB); however, similar to fluorescent labeling, this method only reports on the activity of a single strand. To measure activity on both strands independently, we developed enzymatic methods to generate substrates where the 5-methyl group of mC is labeled with either $^{13}C^{2}H_3$ or $^{14}CH_3$, which permits several avenues for sensitive and strand-specific detection of product formation (see Section 2.4.4).

The isotopically-labeled substrates are generated enzymatically using the CpG methyltransferase (Figure 2-4A). For $^{13}C^{2}H_3$ labeling, we start with [$^{13}C^{2}H_3$]-*L*-methionine (Sigma) and, in a single reaction mixture, enzymatically generate *S*-[$^{13}C^{2}H_3$-Me]-adenosyl-*L*-methionine ([$^{13}C^{2}H_3$]-SAM) *in situ* and simultaneously transfer the methyl moiety from SAM to DNA. For ^{14}C labeling, we simply start with *S*-[^{14}C-Me]-adenosyl-*L*-methionine ([^{14}C]-SAM) from Perkin Elmer and transfer the methyl group to a CpG-containing oligonucleotide using CpG methyltransferase (New England Biolabs). Notably, CpG methyltransferase acts on double-stranded DNA, so we first anneal complementary oligonucleotides, both containing an unmodified CpG, and purify methylated top and bottom strands separately by HPLC.

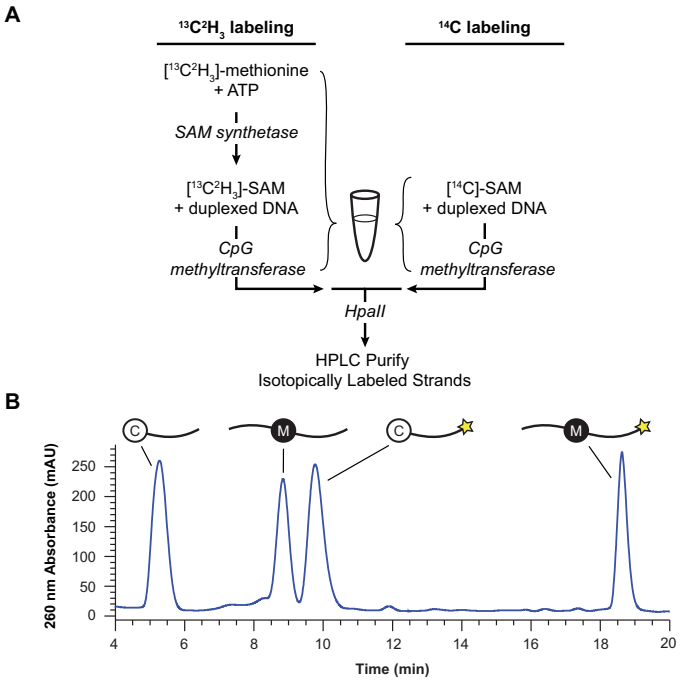


Figure 2-4. Synthesis of isotopically-labeled substrates.

(A) Preparation of oligonucleotide substrates containing isotopically-labeled mC. For $^{13}C^{2}H_3$ labeling (left), [$^{13}C^{2}H_3$]-*S*-adenosylmethionine is enzymatically generated and simultaneously consumed in a single reaction mixture, yielding oligonucleotide duplexes containing [$^{13}C^{2}H_3$]-mC. For $^{14}CH_3$ labeling (right), [^{14}C]-SAM is commercially available, requiring only CpG methyltransferase M.SssI to generate DNA duplexes containing [^{14}C]-mC. Unmethylated strands are digested by the methylation-sensitive restriction enzyme HpaII, and the full-length methylated oligonucleotides are purified by HPLC. **(B)** Representative HPLC chromatogram showing the HpaII-digested products of $^{13}C^{2}H_3$ or $^{14}CH_3$ labeling. Complementary strands are distinguished by a

3'-FAM label (star) on the top strand, which increases on-column retention and is detectable by 490 nm absorbance (not shown).

For the generation of substrates with a $^{13}C^{2}H_3$ modified mC, first purify recombinant *E. coli* SAM synthetase as described (Ottink et al., 2010). The *in situ* generation of [$^{13}C^{2}H_3$]-SAM is initiated by reacting 1 mg/mL SAM synthetase with 1 mM [$^{13}C^{2}H_3$]-*L*-methionine and 2 mM ATP in 1X NEB Buffer 2 (10 mM Tris-HCl, pH 7.9, 50 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT). Incubate this mixture at 37 °C for 15 min, then add 4x volume of duplexed DNA (final concentration 10 μM) and the CpG methyltransferase M.SssI (1:20 total reaction volume, 1000 U/mL) (NEB) in 1X NEB Buffer 2. The simultaneous SAM synthesis and CpG [$^{13}C^{2}H_3$]-methylation is carried out for an additional 5 h at 37° C.

For the generation of substrates with a [^{14}C]-mC label, with [^{14}C]-SAM supplied at ~350 μM in ~500 μL, prepare a reaction containing 35 μM [^{14}C]-SAM, 5.8 μM duplexed DNA, and 224 U/mL CpG methyltransferase (M.SssI), incubating at 37 °C for 4 h. Set aside a small sample of the reaction to determine specific radioactivity for liquid scintillation counting (LSC). Extract DNA from the $^{13}C^{2}H_3$- and ^{14}C-labeling reactions by ethanol precipitation, resuspend the dried pellet in H_2O, and desalt with illustra MicroSpin G-25 Columns (GE Healthcare Life Sciences) equilibrated in water. The resulting oligonucleotide mixture contains methylated products and unmethylated substrate, with and without FAM labels. To digest residual unmethylated substrates, treat with 5,000 U/mL of the methylation-resistant HpaII (NEB) in 1X CutSmart Buffer (NEB) at 37 °C overnight. Ethanol precipitate again, and dissolve the dried pellet in Buffer A3 (100 mM triethylamine acetate (TEAA), pH 7). Purify by ion-pairing HPLC over an Agilent Zorbax Eclipse Plus C18 reverse phase column (3.5 μm particle size, 4.6 mm x 10 cm); we use an Agilent Infinity 1260 Quaternary Pump VL with 1260 FC-AS fraction collector. Equilibrate the column to 65 °C in 65% Buffer A3 and 35% Buffer B3 (50% Methanol/100 mM TEAA, pH 7), and separate over a 20 min gradient from 35% to 45% Buffer B3 at a flow rate of 1 mL/min. Four

major peaks should be observed (along with smaller digestion products): (1) HpaII-digested, unmethylated bottom strand without FAM; (2) full-length, methylated bottom strand without FAM; (3) HpaII-digested, unmethylated top strand with FAM; and (4) full-length, methylated top strand with FAM (Figure 2-4B). Collect fractions (2) and (4), the methylated products +/- FAM, and lyophilize. These purified [$^{13}C^{2}H_3$]-mC or [^{14}C]-mC oligonucleotides can then be duplexed to complementary strands for use in various TET activity assays. For basic analysis of reactivity at one CpG site, we use a complementary strand with an unmodified CpG. Alternatively, the complementary strand can contain mC or ox-mC bases with natural isotope composition, allowing for strand-specific measurements of activity when both strands contain TET substrates.

Finally, some experiments may benefit from longer substrates with multiple targets, rather than simple oligonucleotides. In this case, we use PCR to generate amplicons containing mC, hmC, or fC at all cytosines using ox-mC dNTPs that are commercially available (Trilink). Set up a 50 µL reaction under the following conditions, which have worked well for the majority of our substrates: 1X PCR buffer (20 mM Tris, pH 8.4, 50 mM KCl), 1.5 mM $MgCl_2$, 200 µM of each dNTP (including the desired modified cytosine and no natural cytosine), 1 µM of forward and reverse amplification primers, 1 ng of template DNA (containing unmodified cytosines), and 5 U of Taq polymerase (Invitrogen). When generating mC-containing substrates, add 5% DMSO to aid in denaturation. The thermocycler settings will depend upon the polymerase, nature of the substrate, primer length, etc., but most standard settings translate to these conditions. Amplification of DNA containing caC is not efficient under these conditions, but an alternative strategy using Phusion polymerase has been reported to work (Neri et al., 2015b). Run the entire reaction on an agarose gel and excise the desired band. Use the ZymoClean Gel DNA Recovery Kit (Zymo) to purify the amplicons.

2.4.3: Chemoenzymatic activity assays on full-length oligonucleotides

Despite the diversity of substrates, TET reaction conditions are largely the same for all applications. Our optimized conditions are as follows: 50 mM HEPES, pH 6.5, 100 mM NaCl, 1 mM α-ketoglutarate, 75 μM ammonium iron(II) sulfate (Sigma), 1 mM DTT, and 2 mM sodium ascorbate. The concentration of DNA substrate and TET protein varies based on the experimental goals (see examples below). Importantly, both the α-ketoglutarate and Fe(II) must be fresh, and Fe(II) should be added immediately prior to the start of the reaction to minimize oxidation to Fe(III). Sodium ascorbate is not essential but increases activity by helping to keep Fe(II) in the reduced state (Blaschke et al., 2013; Yin et al., 2013). We note minor differences between our methods and those used successfully by other groups. First, we performed a pH titration for our enzymes and found pH 6.5 to be optimal, with decreasing activity at higher pH; however, other groups routinely perform reactions at pH 8.0 (Hu et al., 2015). ATP has also been reported to stimulate TET activity (He et al., 2011), but it made no detectable difference when evaluated under our reaction conditions.

Incubate the reaction at 37 °C for the desired time. We specifically tested our purified mTet2-CD protein under low turnover conditions (e.g. 500 nM DNA, 10 μg/mL enzyme) and showed linear activity through at least 20 min. Under higher turnover conditions (e.g. 25 nM DNA, 30 μg/mL enzyme), we typically achieve complete conversion of mC and hmC to fC and caC by 30 min; longer reactions are also possible with activity detectable out to 3 h, although turnover slows past ~30 min. Quench the reaction by adding 8x volume of 100% ethanol with 2x volume of Oligo Binding Buffer (Zymo). Purify reactions using the Zymo Oligo Clean & Concentrator kit.

The purified products consist of a mixture of DNA duplexes containing mC, hmC, fC, and/or caC. A number of assays are available to probe for specific bases qualitatively and quantitatively. As noted earlier, Kizaki & Sugiyama reported using 4-6 nt substrates and direct

resolution of the reaction products by HPLC (Kizaki and Sugiyama, 2014). However, longer substrates are needed to understand how strand specificity, sequence context, and a host of other mechanistic questions impact TET activity. Assays on these longer substrates rely on indirect methods involving either enzymatic processing or degradation to nucleosides to deliver rigorous, reliable quantification of TET reaction products *in vitro*.

There are several bacterial restriction enzymes that display variable capacity to cleave the different cytosine modifications. MspI, the most well-known of these enzymes, cleaves the sequence C***X***GG completely when ***X*** is C, mC, or hmC; cleaves partially when ***X*** is fC; and cannot cleave when ***X*** is caC (Figure 2-5A and 2-5B). There is some evidence that MspI cleavage behavior is affected by the identity of the opposite strand CpG (Pais et al., 2015), but the majority of our experiments use substrates that have unmodified CpG on the complement. We have also observed a similar pattern of discrimination with HaeIII on oligonucleotides containing a GGC***X*** sequence. We exploit these restriction enzymes' behavior further by enzymatically modifying hmC and/or chemically modifying fC to prevent cleavage. The difference in cleavage patterns between treated and untreated samples reflects the fraction of these bases in the total reaction mixture. Notably, by selectively excluding either/both chemical modification steps, this method allows the levels of specific bases to be probed.

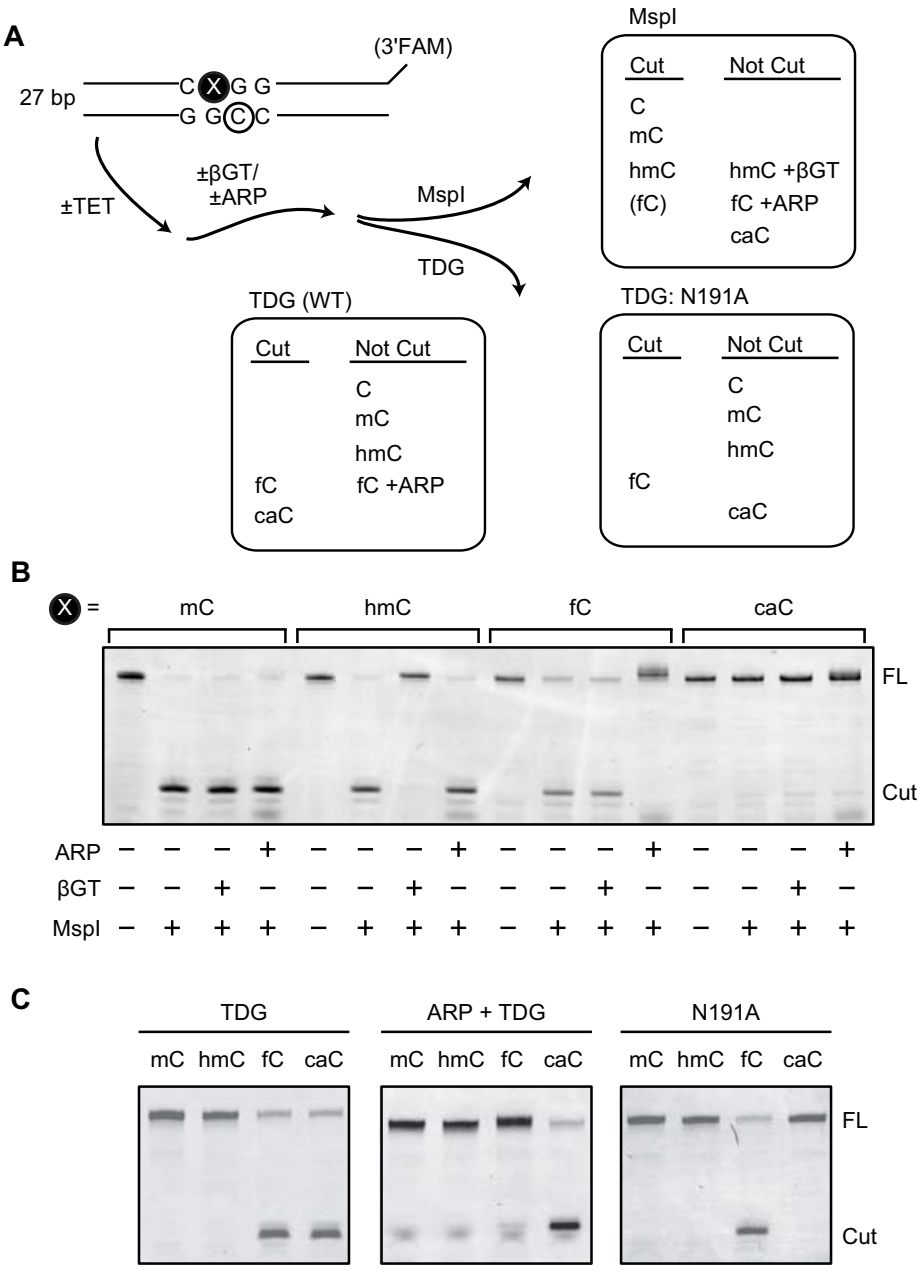


Figure 2-5. Enzyme-coupled assays for TET activity.

(A) Chemoenzymatic assays indirectly measure TET activity on full-length oligonucleotides. Typically, duplexed DNA containing a FAM-labeled reactive strand, which has a mCpG in the MspI cleavage-sequence context, is treated with recombinant TET protein according to optimized assay conditions. Any resulting hmC and fC among reaction products are chemically protected by glucose (via β-glucosyltransferase) or ARP, respectively. Then, the full-length, duplexed DNA is treated with either MspI or TDG (WT or mutant N191A), which each have a unique cleavage capacity against natural and chemically-protected cytosine derivatives. **(B and C)** Denaturing polyacrylamide gels displaying chemoenzymatic activity assay products. Oligonucleotides containing each cytosine modification at the reactive site (X) were employed in lieu of a TET reaction to display MspI **(B)** and TDG **(C)** cleavage signatures, in the presence or absence of various chemical protection steps.

To protect hmC from cleavage, treat the reaction mixture with 2 mM UDP-glucose and 1:25 by volume of T4 β-glucosyltransferase (βGT) (NEB) in 1X CutSmart buffer at 37 °C to transfer the glucose moiety from UDP-glucose to hmC (Terragni et al., 2012). All hmC bases are glucosylated in 30 min, and MspI/HaeIII do not cleave the resulting glucosyl-hmC. To protect fC, mix DNA substrate with at least 35,000-fold molar excess of aldehyde reactive probe (ARP, Dojindo) in a reaction with 6 mM HEPES at pH 5.0 (the lower pH improves the efficiency of the reaction). Incubate at 37 °C overnight, or for at least 3 h. MspI/HaeIII do not cleave fC when ARP is covalently attached. Note that ARP and βGT can be used sequentially to protect both hmC and fC in a reaction mixture: Perform the ARP reaction first and then dilute into 1X CutSmart buffer for the βGT reaction, since βGT is more tolerant of buffer conditions. Proceed directly to MspI digestion in 1X CutSmart buffer at 37 °C for at least 2 h, without need for further purification.

In addition to restriction enzymes, we also utilize thymine DNA glycosylase (TDG) to recognize fC and caC in TET reaction mixtures (Figure 2-5A and 2-5C). TDG is expressed and purified as described previously (Morgan et al., 2007). Treat TET reaction products with 25-fold molar excess of TDG (e.g. 125 nM substrate and 3.125 μM TDG) in TDG buffer (20 mM HEPES, pH 7.5, 100 mM NaCl, 0.2 mM EDTA, 2.5 mM $MgCl_2$) for 4 h at 37 °C. TDG excises fC and caC, leaving abasic sites, but does not react with mC and hmC. To cleave oligonucleotides that now contain abasic sites, add 1:1 volume of 0.3 M NaOH/0.03 M EDTA and incubate at 85 °C for 15 min. As an added means for discrimination of TET products, we use a mutant form of TDG, N191A, that has been shown to preferentially excise fC but not caC(Maiti et al., 2013); this mutant can be purified and used in the same manner to identify fC specifically. Finally, it is possible to selectively cleave caC-containing DNA by treating with ARP (as described above), which protects fC from excision by WT TDG, leaving only caC susceptible. A variant of TDG

that shows preferential excision of caC has also been reported and may provide a complementary approach to quantifying caC in reaction products (Hashimoto et al., 2013).

As the final step of all chemoenzymatic assays, mix samples 1:1 with formamide containing bromophenol blue, separate full-length from cleaved oligonucleotides on a 7 M urea/20% acrylamide/1X TBE gel pre-warmed to 50 °C, and image for FAM fluorescence. Altogether, these chemoenzymatic activity assays provide a complete toolbox to probe for specific ox-mC base modifications in a TET reaction mixture.

2.4.4: Quantitative activity assays on nucleosides

Our chemoenzymatic assays with fluorescent oligonucleotides are both convenient and quantifiable, but digesting the reaction products to nucleosides greatly enhances the accuracy and sensitivity of quantification. This is particularly true of ^{14}C-labeled reactions, for which we can use HPLC to separate modified nucleosides, collect fractions, and perform liquid scintillation counting (LSC) (Figure 2-6A). After reacting [^{14}C]-mC substrate with TET enzymes and purifying as described above, degrade the products to nucleosides using DNA Degradase Plus (Zymo). Note that we typically use radioactivity for reactions where very low turnover is desired and product formation would be difficult to detect by any other means. Since these product peaks would be invisible to UV detection during HPLC, spike the samples with 10 μM each of non-radioactive mC, hmC, fC and caC nucleosides (Berry & Associates), which act as chromatographic markers. Separate the samples by gradient elution over the Supelcosil LC-18-S column, as described in Section 2.3.3. In lieu of mass spectrometry, collect the peaks into 0.25 mL fractions, mix with Opti-Fluor liquid scintillant (Perkin Elmer), and subject to LSC (we use a Tri-Carb 2910 TR (Perkin Elmer)), counting each vial for 10 min using the ^{14}C setting (Figure 2-6B). Measure specific radioactivity from the small sample that was set aside during preparation of radiolabeled substrate (Section 2.4.2), and correct for background radiation using fractions

collected from a non-radioactive control reaction. The resulting measurements of disintegrations per minute (DPM) should be normalized to the input volume and known input concentration to yield molar concentrations of each modified cytosine base.

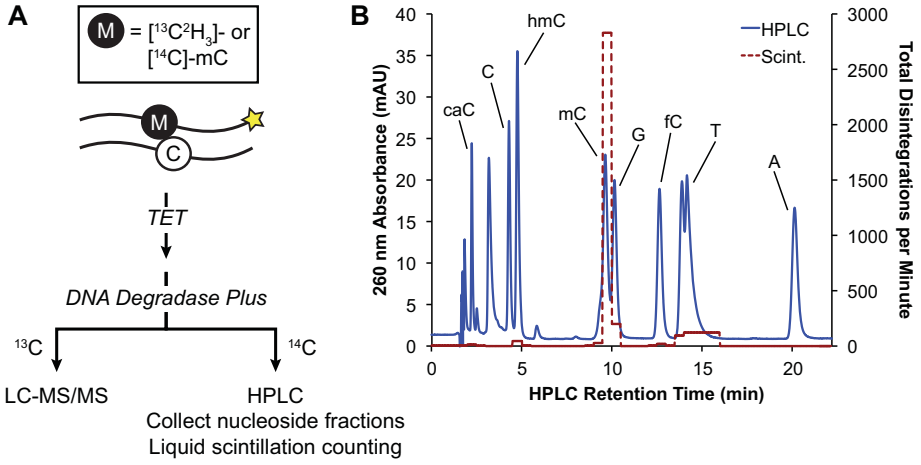


Figure 2-6. High-sensitivity *in vitro* assays.

(A) Isotopically-labeled oligonucleotides duplexed with an unlabeled complementary strand can be reacted with purified TET enzymes *in vitro*. Depending on the nature of the label, the reaction products are degraded to nucleosides and analyzed by either LC-MS/MS or HPLC with liquid scintillation counting to quantify oxidized bases with high sensitivity. **(B)** Representative HPLC chromatogram (solid line) of nucleosides from a degraded TET reaction on ^{14}C-labeled DNA. The samples were spiked with unlabeled mC, hmC, fC, and caC to mark these peak positions for fraction collection, and the fractions were subjected to liquid scintillation counting. The disintegrations per minute (dashed line) for each fraction are overlaid with the HPLC chromatogram. Very low (<1%) product formation can be reliably detected by this approach.

For non-radioactive reactions, LC-MS/MS offers the best method of quantification (Figure 2-3). The application of this protocol using one strand that is selectively $^{13}C^{2}H_3$-labeled can also facilitate tracking of oxidation of two strands of DNA simultaneously. For a typical analysis, degrade at least 5 pmol of TET reaction products to component nucleosides with 1 U DNA Degradase Plus (Zymo) in 10 µL at 37 °C for at least 2 h. Dilute this mixture 10-fold into 0.1% formic acid, and inject at least 20 µL onto the Agilent LC-MS/MS system as described in Section 2.3.3. Alternatively, for better resolution of isotopically-labeled bases, we start with ~20 pmol of reaction products and inject 1 µL onto the nano LC-MS/MS system.

2.5: Acknowledgments

We are grateful to our colleagues who have previously shared reagents or protocols, including Drs. Guoliang Xu (Shanghai Institutes for Biological Sciences), Yi Zhang (Harvard, HHMI), and Alex Drohat (University of Maryland). We thank Dr. Benjamin Garcia for assistance in development of LC-MS/MS protocols. This work was supported by the Rita Allen Foundation.

CHAPTER 3: Tet2 catalyzes stepwise 5-methylcytosine oxidation by an iterative and *de novo* mechanism

This chapter has been adapted from the following publication:

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.*

3.1: Abstract

Modification of cytosine-guanine dinucleotides (CpGs) is a key part of mammalian epigenetic regulation and helps shape cellular identity. Tet enzymes catalyze stepwise oxidation of 5-methylcytosine (mC) in CpGs to 5-hydroxymethylcytosine (hmC), or onward to 5-formylcytosine (fC) or 5-carboxylcytosine (caC). The multiple mC oxidation products, while intricately linked, are postulated to play independent epigenetic roles, making it critical to understand how the products of stepwise oxidation are established and maintained. Using highly sensitive isotope-based studies, we newly show that Tet2 can yield fC and caC by iteratively acting in a single encounter with mC-containing DNA, without release of the hmC intermediate, and that the modification state of the complementary CpG has little impact on Tet2 activity. By revealing

* Author contributions: As a co-author, my specific contribution was gathering and analyzing the data for Figures 3-2 and S3-3; making Figures S3-1, S3-2B, and S3-4; and significantly editing the other figures and text. D.J.C. and R.M.K. primarily conceived the original ideas, and D.J.C. collected and analyzed the data for Figures 3-1 and 3-3. C.S.N., X.J.C., and B.A.G. contributed to assay design and optimization. D.J.C. wrote the first draft of the paper, I wrote the second, and R.M.K. the next, followed by additional editing by all authors. As an interesting follow-up to this paper, another group published contradicting results, and I represented our lab in a long discussion with the primary author and PI of that work.

Tet2 as an iterative, *de novo* mC oxygenase, our study provides insight into how features intrinsic to Tet2 shape the epigenetic landscape.

3.2: Introduction

In mammalian genomes, cytosine base modifications provide an epigenetic information layer that can impact development, differentiation and pluripotency. While 5-methylcytosine (mC) was long considered the predominant modification (Klose and Bird, 2006; Schubeler, 2015), the discovery of Tet family enzymes opened a new and expanded view of the epigenome (Tahiliani et al., 2009). Tet enzymes are α-ketoglutarate-, Fe^{2+}-dependent dioxygenases that can act on mC to generate 5-hydroxymethylcytosine (hmC) in genomic DNA (Figure 1-1), a modification readily detected in many cell types (Kohli and Zhang, 2013). Further, although hmC predominates, Tet enzymes can also catalyze stepwise oxidation of hmC to 5-formylcytosine (fC) and fC to 5-carboxylcytosine (caC), for a total of three oxidized mC (ox-mC) derivatives (He et al., 2011; Ito et al., 2011; Pfaffeneder et al., 2011).

Now viewed as part of the extended epigenome, ox-mC bases appear to have distinct functions. Like mC, they could impact gene expression: ox-mCs interact with different sets of proteins, including transcription factors and RNA polymerase (Spruijt et al., 2013; Wang et al., 2015). They also have distinct genomic profiles, which can persist stably over time and differ by cell type (Bachman et al., 2014; Bachman et al., 2015; Booth et al., 2014; Pfaffeneder et al., 2014; Wu et al., 2014; Yu et al., 2012). Additionally, ox-mCs can play different roles in the dynamic process of DNA demethylation. While all the ox-mCs could potentially facilitate passive, replication-dependent DNA demethylation, fC and caC, but not hmC, are specifically implicated in some proposed pathways for active DNA demethylation, such as base excision repair mediated by thymine DNA glycosylase (He et al., 2011; Kohli and Zhang, 2013; Maiti and Drohat, 2011).

In light of ox-mC modifications, CpGs can be considered complex units of epigenetic information, in which either DNA strand can contain unmodified cytosine or one of its four derivatives. A major question is how these marks are established and maintained. For the methylation code, this task is attributed to the coordinated action of DNA methyltransferases (DNMTs) (Goll and Bestor, 2005; Hashimoto et al., 2012). *De novo* DNMTs largely establish methylation patterns, showing similar preference for both unmodified (C/C) and hemi-methylated (mC/C) CpGs. In contrast, maintenance DNMTs show a strong preference for hemi-methylated CpGs and thereby function to maintain the CpG methylation code after genomic replication.

By contrast with methylation, the mechanisms involved in the generation and maintenance of specific ox-mCs remain unknown. These questions present a particular challenge since Tet enzymes catalyze not one but three reactions at CpGs. While it is now established that mC can be oxidized in a stepwise manner, it remains unknown if these events require multiple encounters between the enzyme and DNA (sequential model) or if caC can be generated in a single encounter with mC-containing DNA (iterative model). This issue is critical to resolve, as the prevalence of mC over hmC raises the question of how highly oxidized bases could be established at a given CpG (Booth et al., 2014; Neri et al., 2015b; Wu et al., 2014). Similarly unexplored is the question of whether ox-mC marks, once established on one strand, can influence the activity of TET on the opposite CpG. This is important because propagation of epigenetic identity depends on maintenance and it is unknown whether, akin to DNMTs, TET enzymes can maintain ox-mC marks across cellular generations.

3.3: Results and Discussion

Here, we focused on understanding how ox-mCs are established and maintained by Tet2. Sequential vs. iterative and maintenance vs. *de novo* models for Tet activity have not yet been resolved in part because prior assays have involved significant substrate depletion and were not

designed to report on strand specific modifications (He et al., 2011; Ito et al., 2011; Tahiliani et al., 2009; Zhang et al., 2014a). To overcome these limitations, we devised highly sensitive, strand-specific, isotopologue-based assays.

Starting from a 27-nt oligonucleotide containing a central CpG moiety, we enzymatically introduced a single isotopically modified methyl group on the substrate (Figure S3-1 and S3-2). Initially, a [^{14}C]-mCpG-containing strand was hybridized to a complementary strand containing an unmodified CpG. After reacting this duplex with mouse Tet2 catalytic domain (mTet2-CD, hereafter referred to as Tet2; Figure S3-3A), the DNA was enzymatically digested to component nucleosides. The nucleoside mixture was subjected to HPLC fractionation and liquid scintillation counting (LSC) to track the kinetics and distribution of [^{14}C]-mC oxidation with high sensitivity (Figure 3-1A). To describe the enzymatic total specific activity (TSA), we accounted for stepwise oxidation. As each detected fC product requires an undetected intermediate hmC and caC requires intermediate hmC and fC, the observed specific activity (SA) for fC and caC generation were multiplied to calculate TSA (Eq. 1).

$$\text{Eq. 1: TSA} = (\text{hmC SA}) + 2\cdot(\text{fC SA}) + 3\cdot(\text{caC SA})$$

We determined optimized enzyme conditions (Figure S3-3) and in an initial assay observed a TSA of 1.3 nmol·min^{-1}·mg^{-1} (Figure 3-1B), which puts an approximate lower limit for turnover at 0.13 min^{-1}. Notably, we detected minimal loss in activity over 20 min, in contrast to prior reports of time-dependent loss of activity (Ito et al., 2011; Zhang et al., 2014a). The sensitivity achieved by our assays also revealed two additional relevant features. First, even at the earliest time points with <1% product formation, we can readily detect the formation of caC. Second, the distribution of the products between hmC, fC and caC were virtually unchanged at early versus late time points. These factors suggest a proficiency for Tet2-catalyzed stepwise oxidation under our reaction conditions.

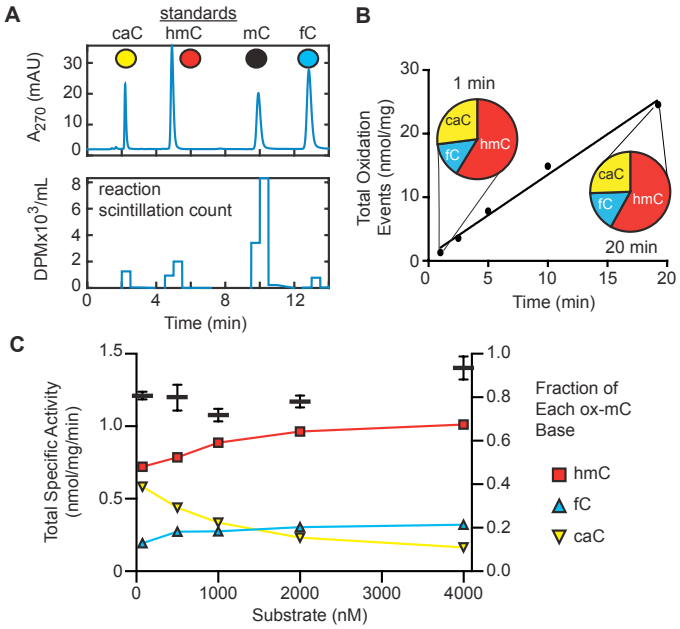


Figure 3-1. Tet2 generates fC and caC early and without a requirement for hmC accumulation.

(A) Example traces for [^{14}C]-mC Tet activity assay. DNA duplexes from Tet reactions were degraded and spiked with standards to delineate fractions containing each base. The fractions were then subjected to scintillation counting. Top: Chromatogram of nucleoside standards (10 µl of 10 µM each). Bottom: Corresponding LSC trace. **(B)** Time course of Tet2 (10 µg/mL) turnover of 500 nM [^{14}C]-mCpG/CpG duplexed DNA showing total oxidation events and fractions of each ox-mC at 1 min and 20 min. **(C)** Titration of 75 nM-4000 nM [^{14}C]-mCpG/CpG with 5 µg/mL Tet2, reacted for 10 min. Total specific activity is plotted on the left y-axis (black bars) as mean and s.d. of duplicate experiments, along with fraction of each ox-mC base on the right y-axis.

A closer examination revealed interesting features at both lower and higher substrate concentrations (Figure 3-1C). On the low end, when reacting 5 µg/mL (maximally 50 nM) Tet2 with as low as 75 nM substrate, activity was near maximal levels. The result suggests a $K_{M,\ DNA}$ which is in the low nM range; otherwise, a greater substrate dependence would be expected. Consistent with this observation, the TSA plateaus as the substrate concentration increases further. Notably the increase in hmC at higher substrate concentrations appears limited relative to fC and caC. Thus, large amounts of fC and caC are formed even when mC is in vast and increasing excess of hmC and C, respectively. For example, under these reaction conditions with 2000 nM substrate, Tet2 generates approximately 25 nM hmC, 8 nM fC and 6 nM caC. These

observations suggest one of two (non-exclusive) possibilities: a sequential oxidation model where mC, hmC and fC substrates have substantially different k_{cat} and K_M values, or an iterative model where Tet2 remains bound to DNA in proximity to the reactive site to establish more highly oxidized bases.

To differentiate between models for how Tet2 establishes ox-mCs, we drew on techniques used previously to examine substrate channeling of metabolites between enzymes (Spivey and Ovadi, 1999). In several metabolic pathways, the product of one reaction is directly fed to the next enzyme without diffusion into bulk solution. Distinct isotopic labels on substrates and products can be used to confirm this molecular hand-off. Given the analogy to the possible models for Tet activity, we set up an isotope-based competition assay that relies upon measuring the isotopic composition of fC and caC produced from [$^{13}C^{2}H_3$]-mC (heavy; *mC) substrates mixed with natural-isotope hmC-containing substrates (light; Figure 3-2A). In the sequential oxidation model, if Tet2 releases the heavy hmC-containing duplex (*hmC) formed from *mC, the isotope will be diluted by the light hmC-containing substrate. Thus, the fraction of downstream heavy fC and caC (*fC and *caC, respectively) products can be no greater than the simultaneous fraction of *hmC/hmC as measured by LC-MS/MS. However, with the iterative oxidation model, if the *hmC product frequently remains bound to Tet2, then the downstream heavy products could be at a higher ratio. At an extreme, if iterative oxidation is highly efficient, we would expect the ratio of *fC/fC and *caC/caC to reflect the initial *mC/hmC ratios.

For our isotope dilution experiment, we enzymatically generated heavy *S*-adenosyl-*L*-methionine (SAM) (Ottink et al., 2010). Following our protocol for generating radiolabeled substrate, we analogously prepared a 27-nt duplex containing a single heavy-mC opposite an unreactive CpG. We also prepared duplexes containing light CpG, mCpG, or hmCpG opposite an unreactive CpG. We reacted 5 µg/mL Tet2 with 100 nM of total duplex DNA containing various mixtures of the *mC-duplex with either light C, mC or hmC-duplex for 10 min. Reaction

products were purified, degraded to nucleosides, and the heavy/light nucleoside ratios were determined with high-precision nano-LC-MS/MS (Figure S3-4).

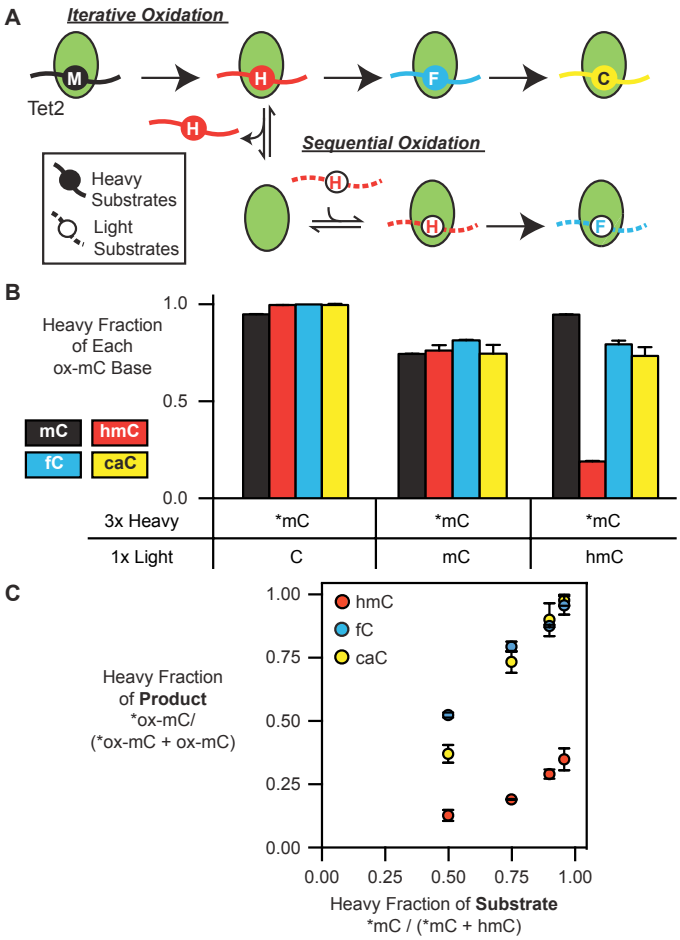


Figure 3-2. fC and caC are formed from iterative oxidation of mC without release of hmC.

(A) Iterative versus sequential oxidation models. Tet2 is shown as green ovals, heavy substrates as filled circles and light as open circles with dotted lines. Tet2 complexed with heavy *mC-containing DNA can directly proceed to heavy fC and caC in iterative oxidation. In sequential oxidation, release of heavy hmC would result in mixing in solution with light hmC, generating predominantly light fC and caC. **(B)** 5 µg/mL Tet2 was incubated for 10 min with heavy *mCpG-containing duplexes mixed in a 3:1 ratio with light CpG-, mCpG- or hmCpG-containing duplexes (100 nM total). Shown is the fraction of heavy isotope for each modification, as analyzed by LC-MS/MS. **(C)** Under similar conditions, Tet2 was incubated with varying ratios of *mC-containing DNA to light hmC-containing DNA. Shown is the heavy fraction of products, compared to the heavy fraction of substrate. The mean and s.d. are shown from triplicate experiments.

As a control, reacting *mC-containing substrate in a 3:1 ratio with non-reactive, unmodified CpG duplex gave >95% yield of heavy *hmC, *fC and *caC, consistent with the isotopic labeling ratio of the *mC (Figure 3-2B), and, as expected, no light or heavy ox-mC products were detected in the absence of Tet2. When Tet2 was reacted with a 3:1 mixture of the heavy *mC and light mC-containing substrates, the heavy:light ratio of hmC, fC and caC were all approximately 3:1, suggesting the absence of any dominant isotope effects. Strikingly, when the *mC-containing substrate was mixed 3:1 with light hmC-containing duplex, the ratio of heavy:light fC and caC were both ~3:1. The accumulation of heavy *fC and *caC is most consistent with the iterative oxidation model where *mC can be converted to higher ox-mCs without obligate release and dilution of the *hmC intermediate.

To determine the extent to which iterative oxidation was occurring, we varied the ratio of heavy *mC to light hmC substrates and quantified the isotopic composition of the resulting products (Figure 3-2C). Across the 1:1, 3:1, 9:1 and 23:1 ratios evaluated, heavy *hmC is generated. However, while *hmC never exceeded light hmC, *fC/fC and *caC/caC ratios were always in great excess of *hmC/hmC. Indeed, these ratios increasingly approach the initial *mC/hmC mixture, consistent with a dominant role for iterative oxidation in establishing the higher ox-mC products. When viewed alongside the [^{14}C] experiments, these results also suggest that fC and caC generation under low turnover conditions was a consequence of iterative activity, as opposed to significantly increased catalytic activity of Tet2 on hmC- or fC-containing duplexes.

In all of the above experiments, we examined duplexes with a single reactive substrate and a non-reactive opposite strand (unmodified CpG). Given the implications for maintenance of the extended epigenome, we next exploited our assays to examine the effects of opposite strand modifications on Tet2 reactivity. Utilizing our [^{14}C]-mCpG assay, which cleanly reports on oxidation of the labeled strand only, we hybridized our original 27-mer [^{14}C]-mCpG strand with

complementary strands containing non-radioactive CpG, mCpG, hmCpG, fCpG or caCpG. We then reacted 75 nM of each duplex with 5 μg/mL Tet2 for 10 min, purified, digested and analyzed using HPLC and LSC as before. Across the substrates, we found that the TSAs for each reaction were very similar, with the largest difference occurring between the opposite strand fC and C, which differ only by a factor of 2 (Figure 3-3). Given the analogy to *de novo* DNMTs, which show minimal differences based on the methylation status of the opposite strand CpG, we suggest that Tet2 is therefore best classified as a *de novo* mC dioxygenase. Moreover, we note that the relative amounts of hmC, fC and caC formed were very similar regardless of the identity of the opposite strand CpG (Figure 3-3). Thus, not only is overall activity largely unaffected, but stalling or iterative oxidation to generate the various ox-mCs is not dictated by the opposite strand CpG.

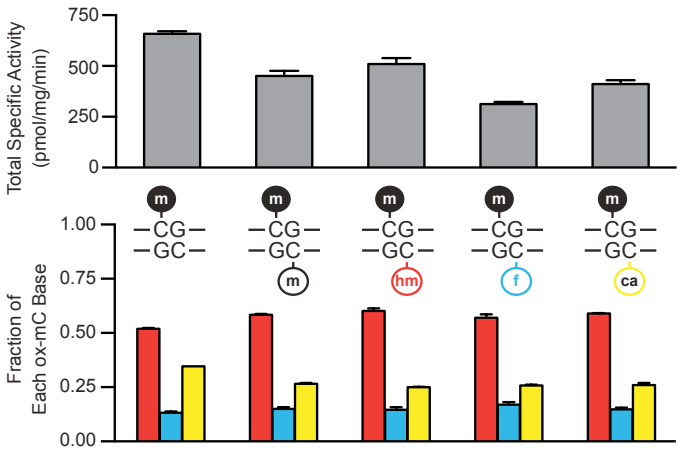


Figure 3-3. Tet2 is a *de novo* 5-methylcytosine dioxygenase.

75 nM [^{14}C]-mCpG was duplexed to an unlabeled strand containing either CpG, mCpG, hmCpG, fCpG or caCpG, then reacted with 5 μg/mL Tet2 for 10 min. TSA and the fraction of each ox-mC were measured by HPLC-LSC. Shown are the mean and SD from duplicate experiments.

Our results provide insight into how Tet2 helps *establish* the extended epigenome. We show that Tet2 can catalyze multiple ox-mC modifications in a single enzyme-substrate

encounter, following the iterative oxidation model. Iterative oxidations helps to explain how fC and caC can be established in genomic DNA despite the relative abundance of mC over hmC or fC. Further, it has implications for the role of ox-mCs as either independent marks or in active DNA demethylation. The fact that Tet2 can iteratively convert a single CpG to fC or caC means that these independent roles can be accessed without an obligatory stable, functional existence as hmC. Additionally, mC bases can be primed for demethylation by iterative oxidation to fC/caC and this need not occur only at sites that first stably exist as hmC.

As noted earlier, in genomic DNA, hmC is far more prevalent than fC and caC (He et al., 2011; Ito et al., 2011; Kohli and Zhang, 2013; Pfaffeneder et al., 2011). Our results indicate that fC and caC, when generated, can derive from a single encounter between Tet2 and mC; however, they do not resolve the question of how or why stalling at hmC is frequently seen *in vivo*, rather than progression to fC and caC. It is feasible that altering the Tet2-DNA encounter lifetime or chromatin accessibility mediate the accumulation of hmC in cells. The extent of iterative oxidation could also be influenced by levels of metabolites such as α-ketoglutarate, interactions with partner proteins, or the non-catalytic domains of Tet2, which could modulate or inhibit activity.

Our results also shed light on possible mechanisms by which ox-mCs are *maintained* in the extended epigenome. We show that CpG modifications on one strand neither impact overall Tet2 activity on the opposite strand nor skew the progression through stepwise oxidation. Thus, Tet2 appears capable of establishing oxidative marks wherever substrates are available, implying that all the permutations of CpG states are biochemically feasible members of the epigenetic repertoire. Notably, while the stable mapping of ox-mCs in various cell types implies maintenance (Bachman et al., 2014; Bachman et al., 2015; Booth et al., 2014; Wu et al., 2014; Yu et al., 2012), our data suggest that substrate preferences intrinsic to Tet2 do not offer a mechanism for such maintenance. Our results imply that alternative cellular factors or the

coordinated activity of different Tet isoforms are more likely to be involved in restoring specific ox-mC marks at a given CpG after cellular division. Indeed, we anticipate that further studies focused on the mechanisms by which Tet enzymes target specific CpGs, regulate iterative oxidation, and coordinate with each other will shed additional light into the generation, maintenance and functional roles of the extended epigenome.

3.4: Acknowledgments

This work was supported by the Rita Allen Foundation Scholar Award (RMK) and NIH grants (R01GM110174, BAG; F30CA196097, MYL).

CHAPTER 4: Mutations along a conserved active site scaffold in TET2 stall oxidation at 5-hydroxymethylcytosine

This chapter has been adapted from the following manuscript:

Liu, M.Y., Torabifard, H., Crawford, D.J., DeNizio, J.E., Cao, X.J., Garcia, B.A., Cisneros, G.A., and Kohli, R.M. (2016). Mutations along a TET2 active site scaffold stall oxidation at 5-hydroxymethylcytosine. Nat. Chem. Biol. *Accepted.*[*]

4.1: Abstract

Ten-eleven translocation (TET) enzymes catalyze stepwise oxidation of 5-methylcytosine (mC) to yield 5-hydroxymethylcytosine (hmC) and the rarer bases 5-formylcytosine (fC) and 5-carboxylcytosine (caC). Stepwise oxidation obscures how each individual base forms and functions in epigenetic regulation and prompts the question of whether TET enzymes primarily serve to generate hmC, or whether they are adapted to produce fC and caC as well. By mutating a single, conserved active site residue in human TET2, Thr1372, we uncovered enzyme variants that permit oxidation to hmC but largely eliminate fC/caC. Biochemical analyses, combined with molecular dynamics simulations, elucidated an active site scaffold that is required for WT stepwise oxidation and that, when perturbed, explains the mutants' hmC-stalling phenotype. Our results suggest that the TET2 active site is shaped to enable higher-order oxidation and provide

[*] Author contributions: This formed the bulk of my thesis work. R.M.K. and I conceived the ideas and managed the decision-making. I performed and analyzed the cellular and biochemical experiments. D.J.C., J.E.D., X.J.C., and B.A.G. helped with assay design and optimization, and J.E.D. in particular helped to collect data for Table 1. For MD simulations, G.A.C., H.T., R.M.K., and I designed the experiments. H.T. performed and analyzed the simulations and provided the associated figures, tables, and methods, which I then edited. I was the primary author of the complete manuscript, and all authors edited.

the first TET variants that could be used to probe the biological functions of hmC separately from fC and caC.

4.2: Introduction

The discovery of ten-eleven translocation (TET) enzymes transformed the known repertoire of epigenetic DNA modifications (Tahiliani et al., 2009). TET enzymes catalyze the oxidation of 5-methylcytosine (mC), the mainstay of the epigenome, into three additional bases: 5-hydroxymethylcytosine (hmC), 5-formylcytosine (fC), and 5-carboxylcytosine (caC) (He et al., 2011; Ito et al., 2010; Ito et al., 2011; Kriaucionis and Heintz, 2009; Pfaffeneder et al., 2011; Tahiliani et al., 2009). Mounting evidence suggests that these oxidized mC (ox-mC) bases stably populate mammalian genomes, aid in DNA demethylation, and potentially encode unique epigenetic information (Bachman et al., 2014; Bachman et al., 2015; Kohli and Zhang, 2013; Liu et al., 2016b; Wu and Zhang, 2015). The central questions now facing the field involve the functions of each individual base and the mechanisms governing their formation.

The overall catalytic mechanism of TET enzymes (TET1–3 in mammals) has been largely inferred from related proteins in the Fe(II)/α-ketoglutarate (α-KG)-dependent family of dioxygenases, such as AlkB (Zheng et al., 2014). Enzymes in this family couple decarboxylation of α-KG with substrate oxidation via a transient Fe(IV)-oxo intermediate, with succinate and CO_2 as byproducts. TET enzymes apply this general mechanism to not one but three stepwise reactions, raising the question of whether these enzymes are specialized for one particular step of oxidation, or for three-step oxidation as a whole. Moreover, stepwise oxidation obscures the function of individual ox-mCs, creating a need to break the linkage between steps in order to study each base in isolation.

The first step of oxidation, conversion of mC to hmC, has so far drawn the most attention, as it best explains the physiological levels of cytosine modifications: in the human

genome, mC accounts for approximately 0.6–1% of all bases, hmC is typically 1–5% of mC, and fC and caC are at least 1–2 orders of magnitude rarer than hmC (Wu and Zhang, 2015). Consistent with these observations, biochemical studies have shown that mC substrate is preferred over hmC and fC, with 2- to 5-fold differences in K_M and k_{cat} reported for human TET2 (Hu et al., 2015). Crystal structures did not reveal substrate-specific interactions that could explain these differences (Hu et al., 2013b; Hu et al., 2015), but computational modeling suggested that hydrogen abstraction is more efficient on mC than on hmC and fC, which adopt unfavorable conformations (Hu et al., 2015; Lu et al., 2016). Together, these studies portray TET enzymes as predominantly serving to generate hmC; in this case, decreased capacity for further oxidation would help to maintain stable levels of hmC for epigenetic functions. Indeed, most functional studies on ox-mC bases have focused on hmC in health and disease, with fC/caC considered as fairly negligible.

However, this view does not explain why fC and caC are present at all, and it contrasts with evidence for the importance of higher-order oxidation. Most notably, fC and caC, but not hmC, are substrates for base excision by thymine DNA glycosylase (TDG); the resulting abasic site can be repaired to regenerate unmodified cytosine (He et al., 2011; Maiti and Drohat, 2011; Weber et al., 2016). This is the leading candidate pathway for active DNA demethylation (Kohli and Zhang, 2013). Apart from being intermediates in demethylation, fC/caC potentially also function as stable epigenetic marks. Genomic sequencing has mapped fC/caC to gene regulatory regions separate from hmC (Wu and Zhang, 2015), and proteomic analysis has described distinct "reader" proteins for each ox-mC base (Iurlaro et al., 2013; Spruijt et al., 2013). Furthermore, mouse Tet2 is capable of iterative oxidation: it can catalyze multiple rounds of oxidation upon a single encounter with mC-containing DNA, without releasing the hmC-containing DNA strand (Crawford et al., 2016). Although the prevalence of genomic hmC implies that most encounters are not iterative, this mechanism could allow TET enzymes to generate fC and caC marks without

first accumulating hmC. Together, these studies encourage the alternate view that TET enzymes are specialized for making not only hmC but fC and caC as well—even that conversion of hmC to fC could be the key "committed" step to DNA demethylation.

To resolve these competing views of TET function, one question comes to the fore: whether TET enzymes are adapted to facilitate higher-order oxidation. The mC-to-hmC step is most favored, but if fC and caC serve important functions, mechanisms should be in place to permit their formation, yet these mechanisms remain largely unknown. They could be extrinsic to TET—e.g. other proteins could recruit TET enzymes or regulate their activity. However, intrinsic features, especially structure-function support for higher-order oxidation, would suggest an enzyme specifically shaped to generate not one but three epigenetic bases.

We examined the active site of human TET2 for potential structure-function determinants of stepwise oxidation. In the crystal structures of TET2 bound to DNA, the enzyme is truncated to the minimal regions necessary for catalytic activity (hTET2-CS, residues 1129–1936 Δ1481–1843) (Figure 4-1a) (Hu et al., 2013b; Hu et al., 2015). The target nucleobase is everted out of the DNA duplex and occupies a tunnel-like space in the active site, with the 5-modified group pointing toward the α-KG analogue and Fe(II) (Figure 4-1b). Although the residues that form this tunnel have no obvious interaction with the 5-modified groups (Hu et al., 2013b; Hu et al., 2015), we hypothesized that they could impact the progress of stepwise oxidation by hydrogen bonding or steric interactions. We therefore targeted two conserved residues located close to the 5-methyl group (Figure 4-1). By substituting all 20 amino acids at these positions, notably Thr1372, we uncovered a relationship between the side chain properties and stepwise oxidation activity, including variants that stall oxidation at hmC, with little to no fC/caC formed. Molecular dynamics simulations, coupled with biochemical analyses, revealed that a conserved Thr1372-Tyr1902 active site scaffold is required for efficient fC/caC formation, providing the first evidence that wild-type TET2 is specifically shaped to enable higher-order oxidation. We further

show that mutations along this core scaffold can reconfigure active site interactions to stall oxidation at hmC, which opens opportunities to test the importance of hmC versus fC/caC in biological and pathological systems.

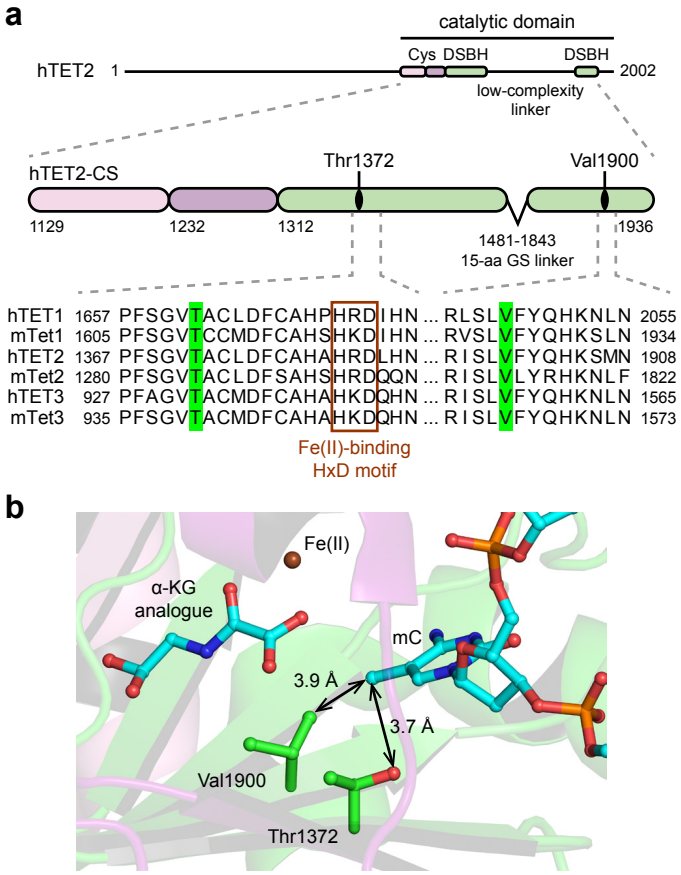


Figure 4-1. Thr1372 and Val1900 were targeted for their potential role in TET2-catalyzed cytosine oxidation.

(a) Schematic of the hTET2-CS construct (drawn to scale, adapted from Hu et al., 2013). The two Cys-rich domains are shown in pink and purple, and the double-stranded β-helix (DSBH) domain is in green; residues are numbered as in the complete hTET2 protein. Both Thr1372 and Val1900 are conserved across mouse and human TET proteins. **(b)** Structure of the hTET2-CS active site (PDB 4NM6) highlighting the targets for mutagenesis, Thr1372 and Val1900. The mC base flips into the active site pocket, pointing toward Fe(II) and the α-KG analogue N-oxalylglycine. Shown are the nearest distances between the residues and the 5-methyl carbon.

4.3: Results

4.3.1: Saturation mutagenesis at Thr1372

We interrogated the active site of human TET2 by performing saturation mutagenesis, which can comprehensively capture structure-function relationships at a particular residue. Using the hTET2-CS construct, we generated plasmids encoding all 20 natural amino acids at either the Thr1372 or Val1900 positions. The plasmids were transiently transfected into HEK293T cells, and genomic DNA (gDNA) was purified from the cells after 48 hr. Using dot blotting to assess the qualitative pattern of genomic cytosine modifications, we found that the Val1900 position is fairly tolerant to mutation, with a variety of mutants showing WT-like stepwise oxidation or reduced overall activity, while bulky and charged residues largely inactivate the enzyme (Figure S4-1a).

We focused our attention on the Thr1372 mutants. TET2 overexpression was confirmed to be uniform by Western blot of cell lysates, with only T1372P having slightly reduced expression (Figure S4-1b). Dot blotting showed that, more so than for Val1900, mutations at Thr1372 produced distinctive patterns of cytosine oxidation, which cluster based on the biochemical properties of the side chain (Figure 4-2a). Replacing Thr1372 with a proline, positively charged (H, K, R), or bulkier hydrophobic residue (I, F, L, M, W, Y) renders TET2 inactive. Only the T1372S mutant, which preserves the side chain hydroxyl group, exhibits WT-like activity. Smaller residues (A, C, G) are proficient at oxidation to fC and caC, but at reduced levels compared to WT. Most remarkably, the acidic or related polar residues (D, E, N, Q) and the nearly isosteric valine permit WT-like formation of hmC but no fC or caC, as detected by dot blot. Given this stalling of oxidation at hmC, Thr1372 appeared to play a unique role in stepwise oxidation.

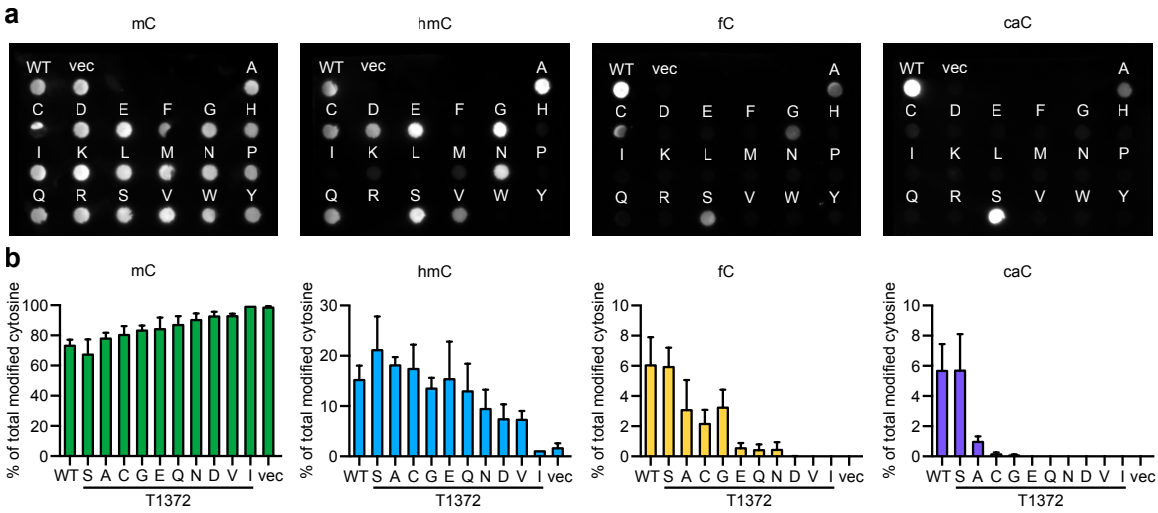


Figure 4-2. Screen for mutant activity.

(a) Dot blots for mC, hmC, fC, and caC in 400 ng of genomic DNA isolated from transfected HEK293T cells. DNA from cells transfected with WT hTET2-CS or empty vector was spotted first, followed by the Thr1372 mutants in alphabetical order (uncropped image in Figure S4-10a). Further analysis of mutant phenotypes focused on variants that were capable of oxidation at least to hmC. **(b)** Genomic levels of mC, hmC, fC, and caC modifications produced by catalytically active Thr1372 mutants, quantified by LC-MS/MS as the percent of total C modifications. Mutants are approximately presented in decreasing order of activity, from WT-like T1372S, to A/C/G that form highly oxidized bases at reduced levels, to E/Q/N/D/V that largely stall at hmC. Shown are the mean and s.d. from independent experiments (WT $n = 7$, vec $n = 6$, mutants $n = 3$, T1372I $n = 2$).

4.3.2: Nucleoside LC-MS/MS quantifies range of mutant activity

We quantified the cellular activity of all Thr1372 mutants capable of oxidizing at least to hmC. The gDNA was degraded to component nucleosides and analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS) (Figure S4-2). In 0.1 µg of HEK293T gDNA, limits of detection in the low femtomole range enabled reliable quantification of 1 in 10^3–10^4 of all cytosines. While the total modified cytosine bases (mC + ox-mCs) were similar across all conditions, the distribution of specific modifications differed significantly. In vector-transfected cells, ox-mC products are minimal: 1.6 ± 1.0% of total cytosine modifications are hmC, with no fC or caC detected (Figure 4-2b). Cells overexpressing WT hTET2-CS contain 15.2 ± 2.8% hmC, 6.0 ± 1.9% fC, and 5.7 ± 1.8% caC, demonstrating robust TET-dependent oxidation at a genomic level.

The mutants exhibit a gradient of activity reflected in the fraction of genomic ox-mC bases (Figure 4-2b). T1372S is the only mutant with WT-like levels of fC and caC, and hmC levels slightly higher than WT. T1372A/C/G mutants generate WT-like levels of hmC but only one-third to one-half as much fC and barely detectable caC. Further down the activity gradient, the E/Q/N/D/V mutants produce hmC at levels at least half that of WT, but fC and caC are near or below detection limits, consistent with the dot blotting results. Among this group, T1372E appears to have the highest activity with WT-like hmC levels and <1% fC, while T1372V is lowest, generating half as much hmC but no fC. Finally, the slightly bulkier T1372I mutant resembles the vector control, underscoring the steric constraints at this position. Thus, the LC-MS/MS results more clearly elucidated the patterns seen on dot blot, showing a spectrum of activity among the Thr1372 mutants correlating with the side chain properties, with E/Q/N/D/V mutants stalling oxidation at hmC.

4.3.3: Computational modeling reveals Thr1372-Tyr1902 scaffold

To probe potential mechanisms behind the mutants' effects, we turned to classical molecular dynamics (MD) simulations of all the active Thr1372 variants. We drew from our experience with AlkB (Fang et al., 2013; Fang and Cisneros, 2014) to model WT hTET2-CS and the Thr1372 mutants bound to each of the four cytosine derivatives (see Figures S4-5 through S4-9 and Tables S4-2 through S4-8 for details). Our simulations were based on the crystal structure of TET2 in complex with DNA containing mC (PDB 4NM6) (Hu et al., 2013b), using α-KG and an Fe(II) surrogate (Mg(II)). Our WT models with hmC and fC proved mostly consistent with the more recently published structures of TET2 with these bases (Hu et al., 2015); we observe all the key interactions between the enzyme, α-KG, active site metal ion, and DNA substrate for varying durations across our simulations. Furthermore, energy decomposition analysis (EDA) and the

root-mean-square deviation (RMSD) comparing the simulations to the reference crystal structure show that the cytosine bases stably occupy the active site across time in all our models.

The hmC models in particular revealed distinct patterns of active site interactions in WT, A/C/G, and E/Q/N/D/V mutants, consistent with hmC being the fulcrum of the observed stalling effect. These patterns helped us to define a key structural scaffold in the WT enzyme that is required for efficient stepwise oxidation. This WT active site scaffold consists of a Thr1372-Tyr1902 hydrogen bond that critically supports optimal non-bonded interactions between Tyr1902 and the substrate cytosine base (Figure 4-3a). The Thr1372-Tyr1902 hydrogen bond is observed in 65% of the simulation time (average over five runs of 50 ns each), and the total non-bonded interaction energy between these residues is –3.37 kcal/mol (Figure 4-3b). Tyr1902, thus oriented by Thr1372, shows significant non-bonded interaction with the hmC base (–6.10 kcal/mol). This core scaffold is present across all WT models bound to mC/hmC/fC/caC and remains fully intact in the T1372S mutant, consistent with this mutant's WT-like activity in cells.

All the other mutants eliminate the Thr1372-Tyr1902 hydrogen bond, perturbing the interaction between Y1902 and the substrate base, with a corresponding loss of enzymatic activity. For the A/C/G mutants, loss of the Thr1372-Tyr1902 scaffold appears to weaken interactions between misaligned active site components, as exemplified by T1372A (Figure 4-3). Combined with the gDNA results, we term the A/C/G phenotype "low-efficiency," since these mutants permit higher-order oxidation but at reduced levels compared to WT.

In our modeling, the E/Q/N/D/V mutants go a step further: they not only eliminate the Thr1372-Tyr1902 scaffold but also elicit new hydrogen bonds specifically with hmC. These new interactions, not present in WT models, position hmC in a different orientation relative to Tyr1902 (Figure 4-3). For instance, in T1372E, the Glu1372 hydrogen bonds directly with the 5-hydroxymethyl group for 88% of the simulation time (average over two runs of 50 ns each). Direct hydrogen bonding to hmC is also observed in T1372D and Q, whereas in T1372N and V,

the new hydrogen bond is between hmC and other nearby residues (Figure S4-5, Tables S4-3b and S4-4b). For example, T1372V elicits an hmC-Asp1384 hydrogen bond (38% of simulation time, average over two runs of 50 ns each). We suggest that the loss of the Thr1372-Tyr1902 scaffold, together with new interactions specific to hmC, could contribute to the unique stalling phenotype of T1372E/Q/N/D/V mutants, which we term "hmC-dominant."

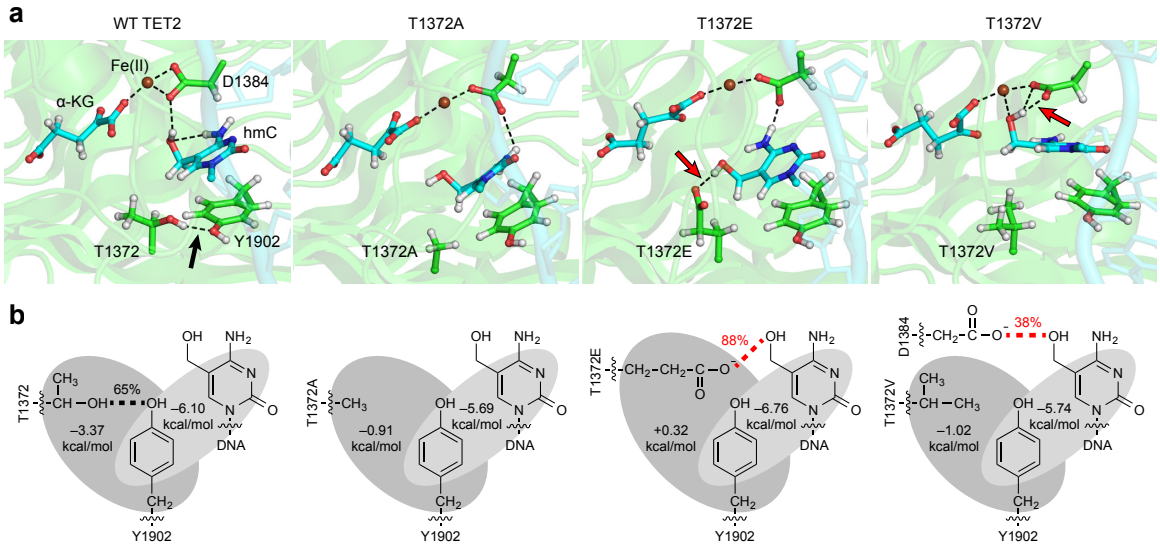


Figure 4-3. Molecular dynamics modeling reveals a critical Thr1372-Tyr1902 scaffold that is disrupted in the low-efficiency and hmC-dominant mutants.

(a) Selected snapshots from MD simulations highlighting key active site components and hydrogen bonds. In WT enzyme (and T1372S), Thr1372 forms a hydrogen bond (black arrow) with Tyr1902, which orients Tyr1902 for optimal non-bonded interactions with the substrate. Low-efficiency mutants such as T1372A disrupt this scaffold, while hmC-dominant mutants such as T1372E/V not only disrupt the scaffold but also elicit new hydrogen bonds (red arrows) with the 5-hydroxymethyl group of hmC. **(b)** Simplified scheme of interactions between key residues and hmC, as determined by MD. Hydrogen bonds (dashed lines) are quantified as percentage of simulation time observed. The values are an average over 2–5 simulation runs of 50 ns each (see Methods). Non-bonded interactions are indicated in gray and total energies of interaction are given in kcal/mol. (Additional modeling data in Figures S4-5 through S4-9 and Tables S4-2 through S4-8.)

4.3.4: Biochemical characterization of TET2 variants

With results from cells and MD showing that side chain properties can define WT, low-efficiency, and hmC-dominant phenotypes, we subjected the TET variants to rigorous comparison

in vitro. We first used driving conditions to compare the maximum extent of the variants' activity and then used limiting conditions to compare the reactivity on mC versus hmC. Representative hTET2-CS variants—WT and T1372S, A, E, and V—were expressed and purified from Sf9 insect cells (Figure S4-3a). To drive oxidation forward, we reacted excess enzyme with limiting substrate: 27-bp oligonucleotides containing a single reactive mC, hmC, or fC duplexed to an unmodified complementary strand. The reaction products were quantified by LC-MS/MS and the results corroborated by three complementary, chemoenzymatic assays (Figure S4-3b and 3c).

In reactions with 20 nM mC-containing duplexes, 30 μg/mL (maximally 0.57 μM) of WT, T1372S, and T1372A convert nearly all substrate to oxidized products in 30 min (Figure 4-4a). However, while WT and T1372S advance efficiently through stepwise oxidation, turning over ~93% of substrate to fC and caC, T1372A lags behind, forming predominantly hmC (30%) and fC (54%) and only 13% caC. This aligns with the gDNA and modeling results, indicating that low-efficiency mutants are capable of oxidation to caC but at reduced levels compared to WT.

The hmC-dominant T1372E and V mutants show noticeably reduced activity on mC (54% and 76% of mC substrate remaining), and oxidation products are strongly restricted to hmC, with 4% and 1% conversion to fC, respectively (Figure 4-4a). Compared to the gDNA results, where the levels of hmC produced by the E/V mutants are within 2-fold of WT (Figure 4-2b), this indicates that other factors can likely tune the activity of TET2 and/or the levels of hmC in cells. Importantly, the patterns of oxidation and hmC stalling hold true in cells and *in vitro*. T1372E is observed to be slightly more active than T1372V, consistent with the gDNA results and suggesting a trade-off between more hmC production and better stringency of stalling. Time course analysis further demonstrates that overall reactivity on mC decreases from WT to the low-efficiency T1372A, and the hmC-dominant E/V mutants fail to produce significant fC even after 3 hours (Figures 4-4b and S4-3d). To validate that the hmC-dominant phenotype is not restricted to the truncated CS form of the protein, we also generated the T1372E mutation in the full

catalytic domain of TET2 (hTET2-FCD, residues 1129–2002) and noted similar results (Figure 4-4a).

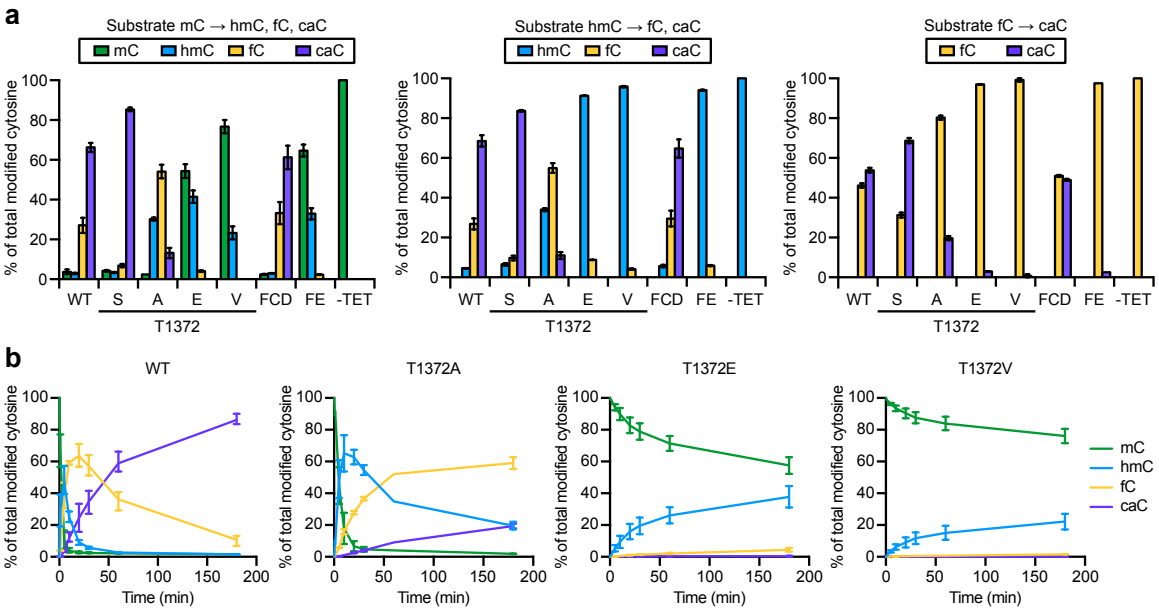


Figure 4-4. Biochemical characterization of purified hTET2 mutants.

(a) TET2 variants (30 µg/mL) were reacted with 20 nM dsDNA substrates containing mC, hmC, or fC for 30 min. The reaction products were purified, degraded to nucleosides, and quantified by LC-MS/MS. WT and T1372E were also generated in the full catalytic domain of TET2 (FCD and FE, respectively) to confirm that the phenotypes are the same as in the hTET2-CS constructs. Mean values are plotted (n = 2), and error bars represent the range. **(b)** Time course for reactions of 30 µg/mL purified TET2 on 25 nM mC substrates. Mean values are plotted (WT n = 3, mutants n = 2), and error bars represent the range.

When all available substrate is hmC, WT and T1372S again convert >93% of substrate to fC and caC. T1372A produces 65% fC/caC, while T1372E and T1372V are able to produce only 8% and 3% fC, respectively (Figure 4-4a). When starting with fC substrate under the same conditions, WT enzymes convert about half of fC to caC, corroborating that the final step of oxidation is the least efficient (Hashimoto et al., 2015; Hu et al., 2015). T1372A generates 19% caC, ~1/3 of the WT level, while E/V mutants make <3% caC, near or below the detection limits of our assays. These results strongly support our model that the Thr1372-Tyr1902 scaffold is

required for WT TET2 activity. Loss of the active site scaffold decreases the activity of low-efficiency mutants and has a more severe effect on hmC-dominant mutants, which do not make significant fC/caC even under driving reaction conditions.

Since TET2 is known to prefer mC over hmC, we next turned to enzyme-limiting conditions to distinguish whether the decrease in overall activity alone was sufficient to explain the restriction of oxidation products to hmC. We compared the reactivity of WT, T1372A, and T1372E mutants on mC versus hmC by titrating enzyme against 745-bp substrates fully modified with mC or hmC. We chose to simplify our kinetic analysis to measure total oxidation products (i.e. substrate consumed), since iterative oxidation links the kinetics of each oxidation step in ways not easily dissected (Crawford et al., 2016). By this analysis, WT TET2 consumes 2.9 ± 0.2 nmol of mC substrate per mg enzyme per minute, while activity on hmC decreases 2.6-fold to 1.1 ± 0.1 nmol/mg/min (Table 4-1, Figure S4-4). This mild decrease in activity on hmC is consistent with previously published observations (Hu et al., 2015). The T1372A mutant displays similar activity on mC and is only 5.5-fold slower in hmC-to-fC conversion, in line with this mutant's capacity for less efficient higher-order oxidation. By contrast, relative to the most proficient WT reaction, the T1372E mutant is 5.9-fold slower in mC-to-hmC conversion but 48-fold slower in hmC-to-fC conversion. Thus, the hmC-dominant mutant exhibits decreased activity overall, but the usual mild preference for mC substrate is not sufficient to explain the larger loss of activity on hmC, which underlies the stalling effect.

4.3.5: Tyr1902 mutagenesis strongly supports our model

Our MD simulations suggested that active site scaffold mutations could introduce aberrant interactions that contribute to hmC stalling. We were cognizant of the challenges to modeling new interactions with classical MD and therefore subjected this model to an independent test: mutating the other scaffold residue, Tyr1902, to Phe. Our modeling predicts that

Y1902F would liberate Thr1372 to form a hydrogen bond directly with hmC (18% of simulation time, average over two runs of 50 ns each), potentially favoring an hmC-dominant phenotype (Figure 4-5a). Taking the hypothesis one step further, by adding a T1372A mutation to Y1902F, our modeling predicts that the T1372A/Y1902F double mutant could rescue activity by alleviating the aberrant hydrogen bonding interaction.

To test these predictions, we compared the activities of purified T1372A, Y1902F, and T1372A/Y1902F enzymes *in vitro*. The results strikingly confirmed our predictions. Compared to the WT mC-to-hmC reaction, the Y1902F single mutant is 9.9-fold slower in mC-to-hmC conversion and 36-fold slower in hmC-to-fC conversion (Table 4-1, Figure S4-4). Addition of the second T1372A mutation partially restores activity, so that the double mutant is only 2.8-fold slower in mC-to-hmC conversion and 14-fold slower in hmC-to-fC conversion. Under driving conditions, the Y1902F mutant leaves 38% of mC substrate unreacted, with products consisting of 49% hmC, 13% fC, and no caC (Figure 4-5b)—similar to T1372E/V but with less stringent stalling at hmC. The introduction of a second mutation in the T1372A/Y1902F double mutant rescues activity, such that 97% of mC substrate is consumed, like the T1372A single mutant.

Substrate consumed (nmol/mg/min)	WT	T1372A	T1372E	Y1902F	T1372A/Y1902F
mC	2.9 ± 0.2	2.9 ± 0.1	0.48 ± 0.02	0.29 ± 0.03	1.0 ± 0.1
hmC	1.1 ± 0.1	0.51 ± 0.03	0.059 ± 0.006	0.079 ± 0.025	0.20 ± 0.02

Table 4-1. Activity of representative TET2 variants on mC and hmC.

Values are mean ± s.e.m. from three independent experiments.

To complement these LC-MS/MS results, rather than digesting the reaction products to nucleosides, we treated the intact oligonucleotides with purified TDG followed by DNA gel electrophoresis to differentiate strands containing mC or hmC from strands containing fC or caC

(Figure 4-5c). While Y1902F shows only trace generation of fC/caC, the addition of the second mutation in T1372A/Y1902F restores stepwise oxidation and mirrors the results for T1372A. Thus, our structural modeling correctly predicts the biochemical behavior of the Y1902F and T1372A/Y1902F mutants, strongly supporting both the requirement of the Thr1372-Tyr1902 scaffold for WT stepwise oxidation and the contribution of aberrant active site interactions to the hmC-dominant phenotype.

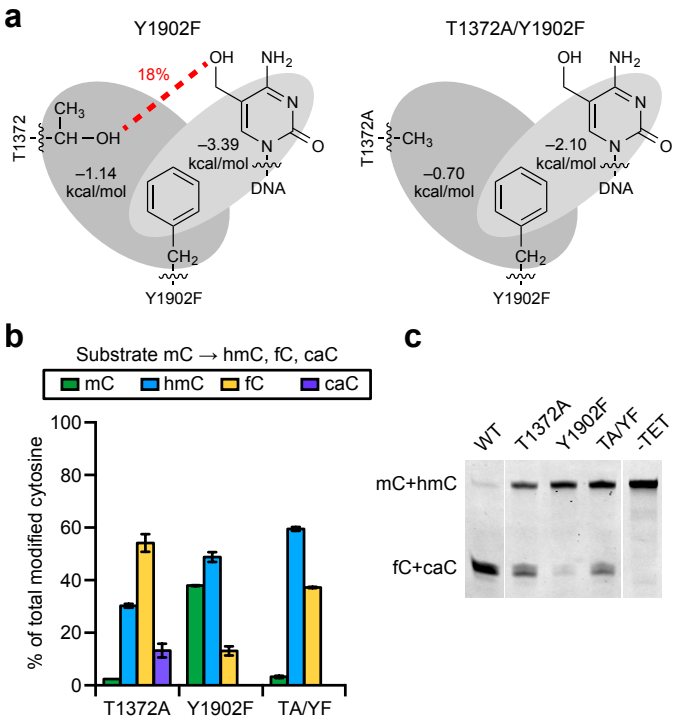


Figure 4-5. T1372A/Y1902F double mutant rescues the hmC-dominant phenotype by configuring active site interactions.

(a) Our modeling predicts that in the Y1902F single mutant, Thr1372 would hydrogen bond instead with hmC, producing an hmC-dominant phenotype. Addition of a T1372A mutation to Y1902F would remove hydrogen bonding, which is predicted to restore activity. The values shown are an average over 2–3 simulation runs of 50 ns each (see Methods). **(b)** Reaction of 30 μg/mL purified mutants on 20 nM mC substrate, analyzed by LC-MS/MS. Mean values are plotted (n = 2), and error bars represent the range. As predicted by our model, Y1902F mimics hmC-dominant mutants, with relatively low activity on mC and little fC formed. The double mutant (TA/YF) restores activity to resemble the T1372A single mutant. **(c)** To highlight fC and caC in the reaction products, the purified oligos were treated with recombinant TDG. After alkaline-mediated cleavage at the resulting abasic sites, denaturing PAGE was used to separate intact oligos containing mC and hmC from cleaved oligos that contained fC and caC (uncropped image in Figure S4-10b).

4.4: Discussion

TET-catalyzed stepwise oxidation populates the mammalian epigenome with three ox-mC bases, making it critical to dissect how each individual base forms and functions. Previous studies have elucidated various biases in favor of the first oxidation step, mC-to-hmC conversion, implying that TET enzymes may be primarily adapted for making hmC, with fC/caC as rare oxidative "overflow" products. However, in light of evidence for the importance of fC/caC in active DNA demethylation and as stable epigenetic marks, we asked whether TET enzymes bear structural features that specifically support fC/caC formation. We have now shown that a conserved Thr1372-Tyr1902 active site scaffold is required for efficient higher-order oxidation by human TET2, suggesting that the enzyme is shaped to enable production of not only hmC but fC/caC as well. We further uncover Thr1372 mutations that effectively abrogate higher-order oxidation by disrupting the active site scaffold; these are the first human TET variants that dissociate the steps of oxidation, providing a new tool to directly test the functions of hmC versus fC/caC.

As a structure-function determinant in TET2, the Thr1372-Tyr1902 scaffold invites comparison to known TET homologues. The Thr-Tyr pair is perfectly conserved across mouse and human TET1, 2, and 3 (Figure S4-1c), raising the possibility that corresponding mutations in TET1 and 3 could likewise tune TET activity. Notably, while a large number of TET mutations have been identified in various malignancies (Abdel-Wahab et al., 2009; Scourzic et al., 2015), but we are not aware of any mutations at the Thr1372 or Tyr1902 positions. As an example in more distant homologues, the trypanosomal J-binding protein JBP1 is predicted to have a Thr-Tyr pair while JBP2 has Ser-Tyr (Figure S4-1c). JBP1 and 2 are capable of oxidizing thymine to both 5-hydroxymethyluracil (hmU) and 5-formyluracil, though a glucosyltransferase normally diverts hmU to form base J as part of the trypanosome's mechanism for immune evasion (Bullard et al., 2014; Cliffe et al., 2009).

A particularly intriguing exception to the conserved Thr1372-Tyr1902 scaffold is the *Naegleria* Tet-like protein NgTet1, which is also capable of higher-order oxidation. Using a structure-based algorithm (Pei et al., 2008) recently borne out by crystal structures (Hashimoto et al., 2015), we found that Thr1372 and Tyr1902 align with Ala212 and Phe295 in NgTet1, respectively (Figure S4-1c), making NgTet1 analogous to our T1372A/Y1902F double mutant. In NgTet1, it was proposed that these residues form a hydrophobic pocket in the active site, which accommodates hmC as it rotates from a product to a substrate conformation for further oxidation (Hashimoto et al., 2015). The active site mutations A212V/N, which could sterically hinder hmC binding within this pocket, were found to partially stall hmC oxidation. In human TET2, the Ala-Phe double mutant only permits low-efficiency stepwise oxidation, suggesting that the Thr-Tyr dyad may have evolved to fine-tune efficient fC/caC generation. By leveraging this scaffold, our results offer the first variants that produce distinct stepwise oxidation patterns in human TET enzymes.

Our combined computational and biochemical approach shows how T1372E/Q/N/D/V mutants could reconfigure active site interactions to produce the hmC-dominant phenotype, characterized by moderate loss of overall catalytic activity as well as a specific decrement at the hmC-to-fC step. To account for the additional loss of activity on hmC, our modeling most prominently implicates new hydrogen bonding to hmC in these mutants. Our calculations correctly predict the hmC-dominant behavior of the Y1902F single mutant, as well as rescued activity in the T1372A/Y1902F double mutant. Indeed, it is quite unusual that the addition of a second mutation rescues activity of the first, helping to bolster our mechanistic model. We note, however, that other related mechanisms could also play a role and are not mutually exclusive with this model. Such mechanisms include restriction of substrate/product rotation (Hashimoto et al., 2015; Lu et al., 2016), changes in protein dynamics (Figure S4-9), and/or altered accessibility of the active site to DNA and α-KG—all of which could occur in concert with aberrant hydrogen

bonding to hmC. These possibilities reflect the complex dynamics of TET-DNA interactions, which remain priorities for future research. Importantly, independent of the mechanism of action, the hmC-dominant Thr1372 mutants fill the need for experimental tools to dissect the individual steps of mC oxidation.

These new TET variants potentially allow for the first direct studies of the epigenetic functions of hmC as distinct from fC and caC. Until now, functional studies have by necessity been all-or-none, showing that loss of one or more TET isozymes can produce diverse phenotypes. Examples range from inability of TET triple-knockout mouse embryonic fibroblasts to undergo reprogramming (Hu et al., 2014), to cancer cell proliferation with loss of TET1 or TET2 (Lian et al., 2012; Neri et al., 2015a), to neonatal lethality in TET3-deleted mice (Gu et al., 2011), among others. In many cases, reintroduction of a single active TET isozyme can fully rescue the phenotype. Such systems provide ideal opportunities to introduce low-efficiency and hmC-dominant TET variants to probe whether hmC alone is sufficient to rescue the defect, whether fC/caC are required, or whether interacting enzymes such as TDG are actually the key players.

These *in vivo* applications will bring new challenges as well, such as examining the mutants' activity under more physiological conditions. Our study illustrates one limit to predicting cellular outcomes based on biochemical properties: although all the mutants perturb the enzyme's reactivity *in vitro*, in HEK293T cells the amount of hmC generated can be close to WT. Many explanations are possible, including that HEK293T cell overexpression likely represents a non-steady state system, in which all cytosine modifications reach unusually high levels with limited means of removing these marks. It will be interesting to see whether normal cells, expressing endogenous TET enzymes, maintain a homeostatic level of ox-mC bases. It will also be important to determine whether the mutant phenotypes in TET2 translate to other TET isoforms, which is needed both for applying the mutants in various biological systems and for

helping to address whether TET1/2/3 have similar or distinct mechanisms of action. Finally, given recently reported structures of TET2 bound to ox-mC bases (Hu et al., 2015), we envision that chemical biology approaches, including additional mutagenesis or unnatural modifications along the Thr1372-Tyr1902-cytosine scaffold, could further hone selectivity for particular bases and potentially uncover TET variants that stringently stall at fC as well or accelerate conversion to caC.

4.5: Methods

4.5.1: Saturation cassette mutagenesis

A codon-optimized hTET2-CS construct (residues 1129–1936 Δ1481–1843) was designed with an N-terminal FLAG tag and unique restriction sites flanking the Thr1372 and Val1900 codons, purchased as a gene block from Integrated DNA Technologies (IDT), and cloned into a pLEXm vector for mammalian expression. Thirty-eight pairs of complementary oligos encoding all amino acid substitutions at both positions (as well as the Y1902F mutation) were ordered, annealed, and cloned by cassette mutagenesis in place of the WT sequence (Table S4-1). Mutations were confirmed by gene sequencing and/or digestion at a unique restriction site within the oligo.

4.5.2: TET2 overexpression in HEK293T cells

HEK293T cells (mycoplasma tested and verified by ATCC) were cultured in Dulbecco's Modified Eagle Medium (DMEM) with GlutaMAX (Thermo Fisher Scientific) and 10% fetal bovine serum (Sigma). Cells were transfected with WT or mutant hTET2-CS, or an empty vector control, using Lipofectamine 2000 (Thermo) according to the manufacturer's protocol. Media was changed 24 h after transfection, cells were harvested by trypsinization 48 h after transfection

and resuspended in phosphate-buffered saline, and genomic DNA (gDNA) was purified from four-fifths of the collected cells using the DNeasy Blood & Tissue Kit (Qiagen).

4.5.3: Western blot for FLAG-tagged hTET2-CS

One-fifth portion of the transfected cells was lysed using CytoBuster Protein Extraction Reagent (EMD Millipore). The clarified lysates were diluted 50-fold into CytoBuster and run on two 8% SDS-PAGE gels, with WT sample as a standard on each gel. To further standardize the blots, the gels were cut at the 70-kDa marker, so that the upper half contained the Hsp90 control band and the bottom half hTET2-CS. The Hsp90 halves of both gels were transferred together onto a single PVDF membrane, and the two TET halves were transferred onto another membrane, using an iBlot Gel Transfer Device (Thermo). Membranes were blocked for 2 h at room temperature with 5% (w/v) milk in Tris-buffered saline with 0.1% (v/v) Tween-20 (TBST), washed 3× with TBST, blotted with primary 1:10,000 anti-FLAG M2 (Sigma, cat. no. F1804) or 1:1,000 anti-Hsp90α/β (Santa Cruz Biotechnology, cat. no. sc-13119) antibodies at 4 °C overnight, washed, blotted with secondary 1:5,000 goat anti-mouse-HRP (Santa Cruz Biotechnology, cat. no. sc-2005) for 2 h, washed, and imaged with Immobilon Western Chemiluminescent HRP Substrate (Millipore) on a Fujifilm LAS-1000 imager with 30-s exposures.

4.5.4: Dot blot for cytosine modifications in gDNA

Purified gDNA from HEK293T cells was diluted to 10 ng/μL in Tris-EDTA (TE) buffer, pH 8.0. To this was added ¼ volume of 2 M NaOH/50 mM EDTA. The DNA was denatured for 10 min at 95 °C and transferred quickly to ice, followed by addition of 1:1 ice cold 2 M ammonium acetate. Sequi-Blot PVDF membranes (Bio-Rad) were cut to size, wet with MeOH and equilibrated in TE buffer, then assembled into a 96-well Bio-Dot microfiltration apparatus

(Bio-Rad). Each well was washed with 400 µL TE drawn through with gentle vacuum, and 400 ng of gDNA was loaded, followed by another TE wash. Membranes were blocked for 2 h in 5% milk/TBST, washed 3× with TBST, and blotted at 4 °C overnight with primary antibodies against each modified cytosine (Active Motif)—1:5,000 mouse anti-mC (cat. no. 39649); 1:10,000 rabbit anti-hmC (cat. no. 39769); 1:5,000 rabbit anti-fC (cat. no. 61223); 1:5,000 rabbit anti-caC (cat. no. 61225). Blots were then washed, incubated with secondary 1:2,000 goat anti-mouse-HRP or 1:5,000 goat anti-rabbit-HRP (Santa Cruz Biotechnology, cat. no. sc-2004) for 2 h, washed, and imaged as described above.

4.5.5: Nano LC-MS/MS analysis of gDNA

Based on published protocols (Liu et al., 2016a), we adapted and optimized LC-MS/MS methods for our systems. To quantify genomic levels of cytosine modifications in HEK293T cells, 20 µg of purified gDNA was concentrated by ethanol precipitation and degraded to component nucleosides with 20 U DNA Degradase Plus (Zymo) in 20 µL at 37 °C overnight. A 150 µm × 17 cm precolumn and 100 µm × 26 cm analytical reverse phase column were made from fused-silica tubing (New Objective) with a Kasil frit: The column was dipped into a 1:3 formamide:Kasil 1624 potassium silicate solution (PQ Corporation), polymerized at 100 °C overnight and trimmed to ~3 mm. Using a pressure injection cell, the columns were packed with Supelcosil LC-18-S resin (Sigma). Using this column setup equilibrated in Buffer A1 (0.1% formic acid in H_2O), the nucleoside mixture was diluted 10-fold into 0.1% formic acid, and 1 µL was injected onto an Easy-nLC 1000 (Thermo) nano LC. The sample was desalted for 5 min over the precolumn, nucleosides resolved using a gradient of 0–30% of Buffer B1 (0.1% formic acid in acetonitrile) over 30 min at a flow rate of 600 nL/min, and tandem MS/MS performed by positive ion mode electrospray ionization on a Q Exactive hybrid quadrupole-orbitrap mass spectrometer (Thermo), with a spray voltage of 2.9 kV, capillary temperature of 275 °C, and normalized

collision energy of 30%. Mass transitions were mC 242.111→26.066 *m/z*, hmC 258.11→124.051, fC 256.09→140.046, caC 272.09→156.041, and T 243.10→127.050. Standard curves were generated from standard nucleosides (Berry & Associates) ranging from 10 μM to 5 nM (10 pmol to 5 fmol total) (Figure S4-2). The sample peak areas were fit to the standard curve to determine amounts of each modified cytosine in the gDNA sample and expressed as the percent of total cytosine modifications in each sample.

4.5.6: Molecular dynamics simulations

Forty-four molecular dynamics (MD) simulations were carried out on WT and all experimentally tested mutants (T1372S/C/A/E/Q/N/D/V, Y1902F, T1372A/Y1902F) with all four cytosine derivatives (mC/hmC/fC/caC), α-KG, and Fe(II)/Mg(II) (see Figures S4-5 through S4-9 and Tables S4-2 through S4-8 for details). All structures were modeled based on WT hTET2-CS bound to mC-containing DNA (PDB 4NM6) (Hu et al., 2013b). Initially, the PDB structure was evaluated with MOLPROBITY (Chen et al., 2010) to check all possible rotamers, followed by hydrogen atom addition to every system with the Leap program (Schafmeister et al., 1995) using the ff99SB parameter set (Case et al., 2005) and solvation in a truncated octahedral box of TIP3P water (Jorgensen et al., 1983). In addition, protonation states of titratable residues were tested with PropKa3.0 (Dolinsky et al., 2004; Dolinsky et al., 2007; Olsson et al., 2011), which confirmed that the default ionization at pH 7 was correct for all residues. Both coordinated histidines are protonated on ND1. All systems were explicitly neutralized with potassium counterions, which were added to the system using the Leap program. The final system size was ~60,000 total atoms with 17–21 counterions. All structures were minimized with 3,000 steps of conjugate gradient, followed by gradual warm-up to 300 K using Langevin dynamics with a collision frequency of 1.0 ps^{-1} in the NVT ensemble for 100 ps. All simulations were performed with the GPU version of the pmemd program in AMBER12 (Case et al., 2005). The force field

parameters for all cytosine derivatives (developed in house), α-KG, Fe(II)/Mg(II), and Zn that are not available in the default ff99SB set are provided in Supplementary Data Set 1 (see online version of publication). The iron cation was approximated by using Mg(II) parameters based on the precedent established by our previous studies on AlkB (Fang et al., 2013; Fang and Cisneros, 2014); this approximation was also validated again for our systems (Figure S4-7 and Tables S4-3, S4-4, S4-8) (Bradbrook et al., 1998; Oda et al., 2005).

Once the systems achieved the target temperature, production MD simulations were performed using Langevin dynamics with a collision frequency of 1.0 ps^{-1} in the NPT (Canonical) ensemble with the Berendsen barostat using a 2-ps relaxation time at 300 K. The production length for each of the simulations was 50 ns, and snapshots were saved every 10 ps, and all snapshots were subjected to subsequent analysis (see below). Values reported are generally a time average over calculations from all snapshots. The most relevant simulations were performed 2–5 times for 50 ns each, with the results averaged across all simulations (the number of simulations for each system is denoted in Table S4-2). All systems were simulated using the Amberff99SB force field with a 1-fs step size and a 9-Å cutoff for non-bonded interactions. SHAKE was used for all the simulations, and the smooth particle mesh Ewald (PME) method (Essmann et al., 1995) was employed to treat long-range Coulomb interactions. Hydrogen bond, root mean square deviation (RMSD), and distance analysis on trajectories were carried out using the CPPTRAJ module (Roe and Cheatham, 2013) available in the AMBER 12 suite, and the trajectories were visualized with the VMD program (Humphrey et al., 1996). Hydrogen bond analysis criteria were 1) angles over 120 degrees and 2) O-H distances less than 3 Å (default cpptraj settings). RMSD and distance analysis are presented in Figures S4-7 and S4-8.

Additional analyses to investigate intermolecular interactions in the active site were carried out by non-covalent interaction analysis (NCI) and energy decomposition analysis (EDA). NCI is a visualization tool to identify non-covalent interactions between molecules (Johnson et

al., 2010). The results obtained from the NCI analysis consist of surfaces between the interacting molecules. These surfaces are assigned specific colors to denote the strength and characteristic of the interactions: green surfaces denote weak interactions (e.g. van der Waals), blue surfaces strong attractive interactions (e.g. hydrogen bonds), and red surfaces strong repulsive interactions. The NCI calculations were performed with the NCI-Plot program (Contreras-Garcia et al., 2011). We focused on the hmC systems, and a representative snapshot from every system was subjected to NCI analysis. In all cases, the hmC substrate was considered as a ligand interacting with a spherical region of 10 Å around the binding site. All calculations were obtained with a step size of 0.2 Å for the cube and a cutoff of 5 Å for the calculation of the interactions between the nucleotides and the active site. The NCI analysis for a selected snapshot of WT and all mutants in the presence of hmC are presented in Figure S4-5. We further examined the WT and T1372A/E/V mutants in the presence of mC and fC; these NCI analyses are presented in Figure S4-6. The snapshots for NCI plots have been selected to highlight the most frequent interactions relevant to the underlying mechanism.

All EDA calculations were carried out with an in-house FORTRAN90 program to determine the non-bonded interactions (Coulomb and VdW interactions) for all the residues (Dewage and Cisneros, 2015; Elias and Cisneros, 2014; Graham et al., 2012). The average non-bonded interaction between a particular cytosine derivative and every other residue, ΔE_{int}, is approximated by $\Delta E_{int}=<\Delta E_i>$, where *i* represents an individual residue, ΔE_i represents the nonbonded interaction (Coulomb or VdW) between residue *i* and the particular cytosine derivative, and the broken brackets represent averages over the complete production ensemble obtained from the MD simulations. This analysis has been previously employed for QM/MM and MD simulations to study a number of protein systems (Cisneros et al., 2009; Cui and Karplus, 2003; Fang et al., 2013; Fang and Cisneros, 2014; Marti et al., 2003; Senn et al., 2005). The EDA results for all protein residues with mC/hmC/fC/caC are presented in Table S4-2, and specific

non-bonded interactions are shown in Table S4-3. Hydrogen bond analyses for WT and all mutants with all cytosine bases are shown in Tables S4-4 through S4-7. As noted, the above-described analyses were performed on each individual snapshot over each individual simulation, and the reported data consist of the averages over all the simulations for each system.

4.5.7: Purification of hTET2 variants from Sf9 insect cells

WT and select hTET2-CS mutants were subcloned into a pFastBac1 vector for expression in Sf9 insect cells as described previously (Liu et al., 2016a). WT and T1372E were also generated in the full catalytic domain (hTET2-FCD, residues 1129–2002). Proteins were expressed for 24 h, and the cell pellet from a 500-mL culture was resuspended in lysis buffer (50 mM HEPES, pH 7.5, 300 mM NaCl, 0.2% (v/v) NP-40) with cOmplete, EDTA-free Protease Inhibitor Cocktail (Roche, 1 tablet/10 mL) and 10 U/mL of Benzonase Nuclease (Millipore). Cells were lysed by one freeze-thaw cycle followed by passage through a 20-gauge and then a 25-gauge needle. The lysate was cleared by centrifugation at 20,000*g* for 30 min, and the supernatant was passed through a 0.2-µm syringe filter. A 250-µL column of anti-FLAG M2 affinity gel (Sigma) was prepared per manufacturer instructions and equilibrated in lysis buffer. The filtered lysate was applied twice to the column under gravity flow, and bound protein was washed with 10 mL then 2 × 5 mL of wash buffer (50 mM HEPES, pH 7.5, 150 mM NaCl, 15% (v/v) glycerol). Elutions of 250 µL were collected in wash buffer containing 100 µg/mL 3× FLAG peptide (Sigma), with each elution incubated on the column for 5 min before collection, until no protein was detected by Bio-Rad Protein Assay and SDS-PAGE. Fractions were pooled, DTT added to 1 mM, and aliquots flash frozen in liquid nitrogen and stored at –80 °C.

4.5.8: TET reactions in vitro

For reactions under "driving" conditions, purified TET2 enzymes were reacted with fluorescein (FAM)-labeled, 27-bp oligonucleotides containing a central reactive site (5'-GTA TCT AGT TCA ATC ***X***GG TTC ATA GCA FAM-3', ***X*** = mC, hmC, or fC), duplexed with a complementary strand containing an unmodified CpG. Protein concentrations were measured by the Bio-Rad Protein Assay and standardized by diluting in elution buffer. A mixture of 20–25 nM duplexed DNA, 50 mM HEPES, pH 6.5, 100 mM NaCl, 1 mM α-ketoglutarate, 1 mM DTT, and 2 mM sodium ascorbate was pre-warmed to 37 °C. Immediately before the reaction, fresh ammonium iron(II) sulfate (Sigma) was added to 75 μM, and at time $t = 0$, TET2 was added to a final concentration of 30 μg/mL (maximally 0.57 μM of hTET2-CS and 0.30 μM of hTET2-FCD). Reaction volumes were typically 200–350 μL. After incubation at 37 °C for 30 min (or at designated time points), the reactions were quenched by addition of 8 volumes of 100% ethanol with 2 volumes of Oligo Binding Buffer (Zymo). Reaction products were purified using the Zymo Oligo Clean & Concentrator kit, eluted in LC-MS grade H_2O, and analyzed by LC-MS/MS and/or enzyme-coupled assays (Liu et al., 2016a).

For enzyme titration experiments, substrates were generated by PCR using 5-methyl- or 5-hydroxymethyl-dCTP and standard protocols for Taq polymerase. Each 745-bp amplicon contained a total of 391 modified cytosines (280 in CpG context) and was purified by gel extraction. Reaction conditions were the same as above, except for using 80 ng of PCR substrates and 1.856–72.5 μg/mL of enzyme in a 25-μL reaction. Following randomized analysis by LC-MS/MS, the percentage of total oxidation products (i.e. substrate consumed) was converted to nanomoles based on the known composition of the substrate. Plots were generated of total oxidation products versus enzyme concentration (Figure S4-4), and the slopes from linear regression were compiled in Table 4-1.

4.5.9: Chemoenzymatic assays of TET activity

We designed three chemoenzymatic assays to probe for specific cytosine modifications (Liu et al., 2016a). Concentrated, purified reaction products representing 50 µL of the TET reaction (up to 1.25 pmol) were used for each assay.

To distinguish mC-containing oligos, the restriction enzyme MspI (NEB) was used, which normally cleaves CCGG sites containing C, mC, or hmC, with partial activity on fC and no activity on caC (Ito et al., 2011). A combination of aldehyde reactive probe (ARP) (Thermo) and T4 β-glucosyltransferase (βGT) (NEB) were used to protect fC and hmC, respectively, from MspI cleavage, leaving only mC susceptible. The reaction products, along with controls, were treated first with 4.4 µM ARP in 6 mM HEPES, pH 5.0 (10 µL total volume), incubated at 37 °C overnight, then diluted into 20 µL with 1× CutSmart Buffer (NEB), 2 mM uridine diphosphoglucose (UDP-Glc) and 1:25 volume of βGT for 30 min at 37 °C. To this mixture was added 50 U MspI in 1× CutSmart Buffer and digestion carried out at 37 °C for >2 h.

To visualize the extent of higher-order oxidation to fC and caC, the reaction products were treated with 25-fold molar excess of thymine DNA glycosylase (TDG) purified as described below, in TDG buffer (20 mM HEPES, pH 7.5, 100 mM NaCl, 0.2 mM EDTA, 2.5 mM $MgCl_2$) for 2–4 h at 37 °C. After the reaction, 1:1 volume of 0.3 M NaOH/0.03 M EDTA was added and the mixture incubated at 85 °C for 15 min to cleave oligos at abasic sites. The TDG mutant N191A, which was previously found to excise fC and not caC (Maiti et al., 2013), was also purified and used in the same manner to identify fC specifically.

As the final step of all three chemoenzymatic processes, the samples were mixed 1:1 with formamide containing bromophenol blue loading dye, loaded onto a 7 M urea/20% acrylamide/1X TBE gel prewarmed to 50 °C, and imaged for FAM fluorescence on a Typhoon 9200 variable mode imager.

4.5.10: LC-MS/MS analysis of reaction products

Concentrated, purified reaction products representing 200 µL of the TET reaction (up to 5 pmol) were degraded to component nucleosides with 1 U DNA Degradase Plus (Zymo) in 10 µL at 37 °C overnight. The nucleoside mixture was diluted 10-fold into 0.1% formic acid, and 20 µL were injected onto an Agilent 1200 Series HPLC with a 5 µm, 2.1 × 250 mm Supelcosil LC-18-S analytical column (Sigma) equilibrated to 50 °C in Buffer A2 (5 mM ammonium formate, pH 4.0). The nucleosides were separated in a gradient of 0–10% Buffer B2 (4 mM ammonium formate, pH 4.0, 20% (v/v) methanol) over 7 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 175 °C, gas flow of 10 L/min, nebulizer at 35 psi, sheath gas temperature of 300 °C, sheath gas flow of 11 L/min, capillary voltage of 2,000 V, fragmentor voltage of 70 V, and delta EMV of +1,000 V. Collision energies were optimized to 10 V for mC, fC, and T; 15 V for caC; and 25 V for hmC. MRM mass transitions and data analysis were as described above.

4.5.11: Purification of hTDG from E. coli

We adapted a published protocol (Morgan et al., 2007) to express and purify WT and N191A TDG from BL21(DE3) cells. 1-L cultures were grown to OD ~0.6, cooled gradually to 16 °C, induced with 0.25 mM IPTG at OD ~0.8, and grown for another 4 h. Cells were collected by centrifugation, resuspended in 20 mL TDG lysis buffer (50 mM NaPhos, pH 8.0, 300 mM NaCl, 25 mM imidazole) with protease inhibitors, and lysed by four passes on a microfluidizer. The lysate was cleared by centrifugation at 20,000*g* for 20 min, then passed through a 0.22-µm syringe filter. A 1-mL column of HisPur cobalt resin (Thermo) was equilibrated in TDG lysis buffer, and the lysate bound by two applications to the column under gravity flow. The column was washed three times with 5 mL of TDG lysis buffer containing 1 M NaCl, then three times

with 5 mL of regular TDG lysis buffer. Elutions of 1 mL each were collected in TDG lysis buffer containing increasing concentrations of imidazole: 50, 100, 150, 200, 250, and 500 mM imidazole. Elutions were evaluated by SDS-PAGE and dialyzed overnight at 4 °C into TDG storage buffer (20 mM HEPES, pH 7.5, 100 mM NaCl, 1 mM DTT, 0.5 mM EDTA, 1% (v/v) glycerol). Final protein concentrations were measured with the Bio-Rad Protein Assay and aliquots stored at –80 °C.

4.6: Acknowledgments

We thank B. Niedziolka and the Wistar Institute Protein Expression Facility for help with protein expression in Sf9 cells and all members of our labs for insightful discussions. Computing time from Wayne State C&IT and additional mass spectrometry resources from I. Blair's lab are gratefully acknowledged. This work was supported by the Rita Allen Foundation Scholar Award to R.M.K. and NIH grants (R01 GM110174 to B.A.G., R01 GM108583 to G.A.C., and F30 CA196097 to M.Y.L.).

CHAPTER 5: Future directions and concluding remarks

In summary, this thesis has described our recent insights into the biochemical and structural mechanisms that govern TET-catalyzed stepwise oxidation. We are broadly interested in the intrinsic properties of the TET1/2/3 family, but hmC-to-fC conversion draws our particular attention. We view this step as a tipping point in stepwise oxidation—a rare event whereby the enzyme can potentially change the epigenetic readout at a given CpG and commit to active demethylation via base excision repair. However, the rarity of fC and caC bases has raised questions about whether they might be accidental rather than purposeful modifications, and whether they are truly needed for epigenetic functions. Using novel, quantitative assays, we showed that Tet2 has the capacity to operate by a *de novo* and iterative mechanism, which can facilitate the generation of fC/caC. We further discovered a conserved structural scaffold in the active site of TET2, which specifically enables fC/caC formation. These results strongly suggest that TET enzymes have evolved in favor of generating all three oxidized bases, not just the most prevalent hmC, and therefore that fC/caC may serve important functions distinct from hmC. Critically, we revealed that mutating a single amino acid of TET2 can disrupt the active site scaffold and largely restrict oxidation to hmC. These TET variants now offer the first tools to directly test the functions of hmC independently of fC/caC and to ascertain the biological significance of the hmC-to-fC step. Below, I outline progress toward introducing these variants into suitable model systems. I also discuss ongoing efforts to uncover the properties of all three TET isoforms and to examine activity on non-canonical substrates. Indeed, we are now positioned to explore many new avenues toward the larger goal of understanding the extended epigenome.

5.1: Determine whether hmC is sufficient for MEF reprogramming to iPSCs, or whether fC and caC are required

5.1.1: Introduction

To test our low-efficiency and hmC-dominant TET variants, we searched for model systems that met key criteria.* First, since TET isoforms have potentially redundant roles, we focused on systems in which deletion of all three TETs produced a clear phenotype. Second, the phenotype had to be reversible by the introduction of a single active TET isoform. We were struck by a report detailing the requirement of TET activity for reprogramming (Hu et al., 2014). This study systematically examined reprogramming of mouse embryonic fibroblasts (MEFs) into induced pluripotent stem cells (iPSCs) using retroviral transduction of *Oct4, Sox2*, and *Klf4* (OSK factors, *c-Myc* optional) (Nakagawa et al., 2008; Takahashi and Yamanaka, 2006). Hu et al. found that TET triple-knockout (TKO) MEFs completely failed to reprogram, but re-introducing any active TET catalytic domain (*Tet1*, *2*, or *3*) along with the OSK factors could fully rescue the block. Interestingly, *Tdg* deletion produced the same reprogramming block and could likewise be rescued by transduction of active *Tdg*.

The defect was traced to the mesenchymal-to-epithelial (MET) transition that occurs early in reprogramming (Hu et al., 2014). The evidence suggested that active demethylation involving TET and TDG is required to reactivate expression of the miR-200 family of microRNAs, which are essential for MET. In support of this model, ectopic expression of any single miR-200 family member could at least partially rescue reprogramming in TET TKO MEFs. Thus, these miRNAs were deemed essential for mediating the activation of downstream

* R.M.K. and I conceived the original ideas, and all subsequent ESC work was done with Joanne Thorvaldsen in the lab of Marisa Bartolomei.

pluripotency genes, while the requirement for TET activity comes earlier in the reprogramming process.

This example illustrates how all-or-none studies of TET activity can reveal striking roles in biological processes, yet the key question remains of which ox-mC base(s) are truly required. The *Tdg* deletion and rescue experiments would seem to suggest that fC and caC, not just hmC, are essential for MET, and indeed that these bases must undergo excision and repair, since catalytically inactive *Tdg* failed to rescue the phenotype. However, TDG has promiscuous roles in epigenetic regulation beyond fC/caC excision, including repair of U:G and T:G mismatches and transcriptional co-activation (Bellacosa and Drohat, 2015). These roles could plausibly contribute to reprogramming independently of TET activity. Therefore, we endeavored to introduce low-efficiency and hmC-dominant variants of mouse Tet2 into the TET TKO model system to directly dissect whether hmC is sufficient to rescue reprogramming, whether fC and caC are required, or whether enzymes such as TDG are actually the key players.

5.1.2: Preliminary results

To best match the materials used in the original study, we obtained pMXs retroviral expression constructs from Guoliang Xu and WT and TET TKO ESCs from Xiajun Li. Sequence alignment showed that the Thr1372 residue in human TET2 corresponds to Thr1285 in mouse Tet2. We therefore generated WT mouse Tet2 and T1285A, E, V, and W mutants in the pMXs vector. Based on our prior work, we expect the T>A mutant to be low-efficiency, capable of making all three ox-mCs at lower levels compared to WT; T>E and T>V should be hmC-dominant, with little to no fC/caC formed; and T>W should be catalytically inactive. Preliminary overexpression of WT, T1285E, and T1285W in HEK293T cells supported these predictions (Figure 5-1), though these experiments need to be repeated with the full set of mutants.

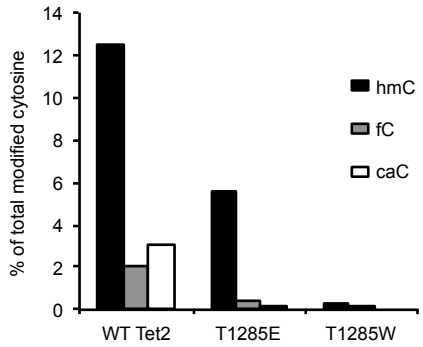


Figure 5-1. Preliminary overexpression of WT Tet2 and T1285E/W mutants.

Based on the hmC-dominant phenotype of the human T1372E mutant, we generated corresponding mutations in mouse Tet2. Quantification of ox-mC bases by LC-MS/MS showed that the mouse Tet2 homologues behave according to expectations, suggesting that these variants can be used to test the specific role of hmC in MEF reprogramming.

We cultured the WT and TET TKO ESCs under the conditions described by Hu et al. In brief, ESCs were maintained on a feeder layer of mitomycin C-treated MEFs. The cell culture media was based on standard E14 media, consisting of DMEM with 15% heat-inactivated FBS, 2 mM GlutaMAX, 0.1 mM nonessential amino acids, 1 mM sodium pyruvate, 0.1 mM β-mercaptoethanol, 1× penicillin/streptomycin, and 1000 U/ml ESGRO leukemia inhibitory factor (LIF); in addition, the media was supplemented with MEK and GSK3 inhibitors (2i), 1 μM PD0325901 and 3 μM CHIR99021. In anticipation of using these ESCs to generate chimeric embryos, we labeled WT and TKO cell lines with GFP by lentiviral transduction, followed by selection for single clones. All cell lines used in the experiment were tested for mycoplasma. Furthermore, the GFP-labeled WT and TKO ESCs were genotyped by PCR and karyotyped to confirm a normal chromosome count of 40.

To establish the experimental protocol, we performed reprogramming trials using WT MEFs containing an Oct4-GFP marker for induction of pluripotency. The MEFs were freshly isolated from embryonic day 12.5 (E12.5) mice and, after one passage, seeded at 5×10^4 cells per well in a 6-well plate the day before transduction. Meanwhile, PLAT-E cells (a derivative of

HEK293T cells specialized for retroviral production) were seeded at 3×10^6 cells per 10-cm dish, and separate dishes were transfected with 8 μg of *Oct4, Sox2*, *Klf4*, and *c-Myc* pMXs constructs using Lipofectamine 2000. For infection of MEFs, virus-containing supernatant was collected from the PLAT-E cultures in two batches, 48 and 72 hr post-transfection, passed through a 45-μm filter, and supplemented with 4 μg/mL polybrene in 10 mL total volume.

To optimize reprogramming efficiency, various conditions were tested. First, MEFs were transduced with either all four Yamanaka factors or OSK only. The supernatants containing each virus were added to the MEFs in 1:1:1:1 ratios, substituting one volume of virus-free media (with polybrene) for *c-Myc* in the OSK conditions. All wells received an additional volume of virus-free media with polybrene, for a total of 2.5 mL per well. After two rounds of infection, on day 0 of reprogramming, the media was changed to fresh ESC media and replenished every day thereafter. Since several studies have suggested that various factors enhance the efficiency of reprogramming (Blaschke et al., 2013; Chen et al., 2011; Chen et al., 2013a), we tested OSK reprogramming in standard E14 media (1,000 U/mL LIF) with various supplements: the GSK3 inhibitor CHIR99021 (3 μM), the MEK inhibitor PD0325901 (1 μM), 50 ug/mL vitamin C (in the more stable form of 2-phospho-L-ascorbic acid), and/or 5 ng/mL basic fibroblast growth factor. Cells were monitored daily for growth of reprogrammed colonies and induction of Oct4-GFP. Furthermore, for each of the conditions tested, one well was left undisturbed to grow in the original 6-well plates, and a duplicate well was replated on day 4 onto a MEF feeder layer in 6-cm dishes.

Morphological changes were evident in all transduced cells within 2 days of changing into ESC media. Four-factor transduction, as expected, drove the most rapid and efficient reprogramming, with rapid cell growth that necessitated additional passaging from 6-cm to 10-cm dishes. Isolated GFP-positive colonies were observed starting on day 8. By comparison, the OSK-transduced conditions underwent more gradual changes, forming fewer but rounder, more

compact colonies. Among this group, cells fed with E14 media plus 2i and vitamin C displayed the most efficient reprogramming, followed by cells supplemented with CHIR99021 alone. In both of these cases, GFP-positive colonies were visible on day 14, while the remaining conditions lagged until day 17 (data not shown). After 17 days (14 for OSKM), the cells were fixed and stained for alkaline phosphatase (AP) as a marker for pluripotency (Figure 5-2).

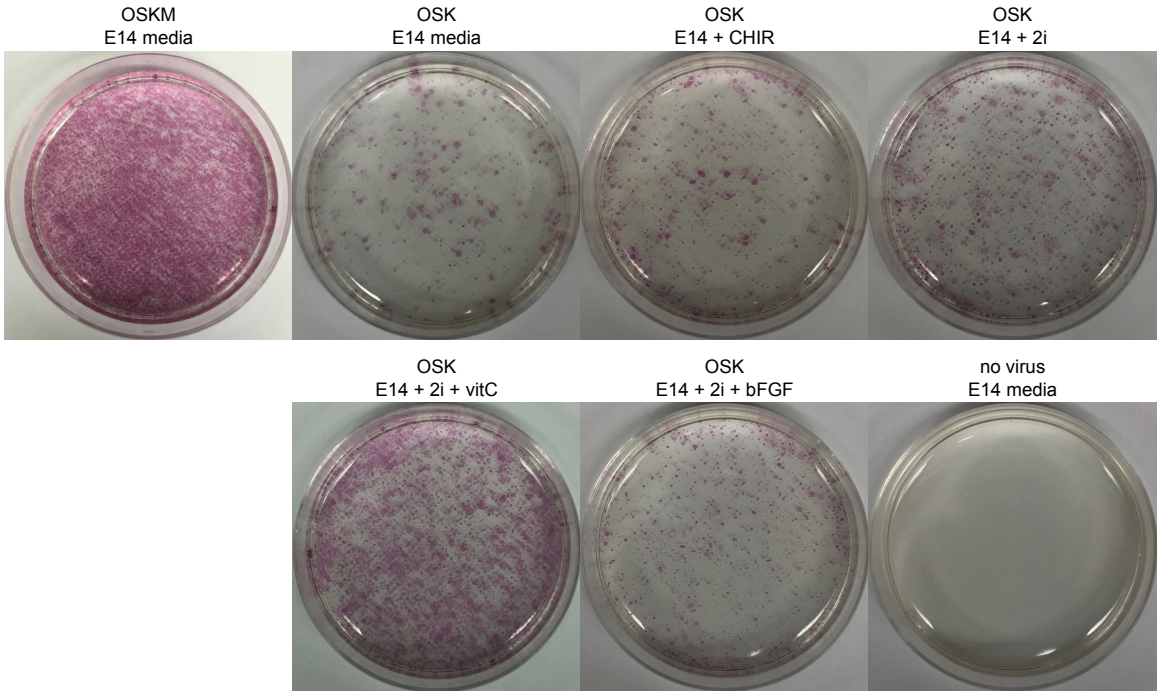


Figure 5-2. Reprogramming trials using Oct4-GFP MEFs.

To optimize the efficiency of reprogramming, conditions were varied to test transduction with OSKM vs. OSK factors, as well as different growth media. Cells were either grown continuously in a 6-well plate or replated onto a feeder layer in 6-cm dishes (shown here). After 17 days (14 days for OSKM condition), cells were fixed and stained for alkaline phosphatase. E14 media indicates standard ESC growth media containing 1000 U/mL LIF; CHIR, the GSK3 inhibitor CHIR99021 (3 μM); 2i, CHIR99021 plus the MEK inhibitor PD0325901 (1 μM); vitC, 50 ug/mL vitamin C (2-phospho-L-ascorbic acid); and bFGF, 5 ng/mL basic fibroblast growth factor.

Overall, AP staining was much more prevalent than GFP fluorescence, perhaps capturing earlier stages of reprogramming. The OSKM plates showed the highest density of AP-positive colonies, followed by OSK with E14/2i/vitamin C media. Notably, vitamin C was specifically

excluded from the Hu et al. study, likely because it has controversial, isoform-specific effects on Tet-mediated reprogramming (Chen et al., 2013a). Finally, media supplemented with CHIR99021 alone appeared to boost reprogramming efficiency as judged by GFP fluorescence, but interestingly these plates did not produce more AP staining compared with the remaining conditions tested. In summary, reprogramming was successful across all conditions, occurring very rapidly and with high efficiency in OSKM-transduced MEFs, and vitamin C was a particularly effective media supplement.

5.1.3: Next steps

These trial experiments pave the way for generating and reprogramming WT and TET TKO MEFs. Essentially all the key elements are now in place: TET mutant constructs made in pMXs, ESC lines confirmed, and reprogramming protocol established. The next major step would be to inject the WT and TKO ESCs into mouse blastocysts to generate chimeric embryos, which we harvest at E12.5 to make MEFs. These MEFs would be FACS sorted for GFP-labeled cells and seeded into 6-well plates for retroviral transduction. The OSKM method (E14 media with early splitting onto 10-cm dishes) is likely preferred, since it produces highly efficient reprogramming. This is important for maximizing the dynamic range within which we may observe differences upon introduction of TET variants.

We expect the results first to recapitulate the Hu et al. study, with WT MEFs reprogramming to AP-positive iPSC-like colonies and TKO MEFs blocked at the early MET step. Re-introduction of WT Tet2 catalytic domain should fully rescue reprogramming. The effects of the low-efficiency and hmC-dominant mutants are difficult to predict but will be informative regardless of the result. The reported reprogramming block with *Tdg* deletion would support a prediction that fC and caC are required. In this case, the hmC-dominant mutants may not be sufficient, whereas the T1285A low-efficiency mutant might fully or partially rescue the defect,

giving insight into the levels of fC and caC needed to restore function. Alternatively, the hmC-dominant T1285E/V mutants, which are expected to produce hmC at different levels, could possibly rescue the phenotype. This would suggest that hmC alone is sufficient for reprogramming. I suspect this result is less likely, especially since we believe TET to be specialized for generating low levels of fC/caC, but it would be very interesting to map and quantify the hmC levels produced by the mutants and to correlate those patterns with, for instance, any changes in gene expression of mesenchymal/epithelial markers and microRNAs. I would speculate that hmC alters transcriptional activity and may recruit protein readers—its prevalence in cells suggests it is not merely a placeholder awaiting active demethylation—but fC/caC may be needed for global changes such as cell fate transitions.

To interpret the effects of the Tet mutants, we anticipate needing to overcome challenges in downstream analysis. Perhaps most notably, we will need to examine the levels of specific ox-mC modifications, both genome-wide in the reprogrammed cells and at relevant loci. This will require improvement of our current mass spectrometry methods to push the limits of detection even further, perhaps involving enrichment for modified cytosines or chemical labeling to enhance signal intensity (Huang et al., 2016; Tang et al., 2015). We are also becoming more familiar with sequencing methods to localize ox-mC modifications. Indeed, sequencing may provide insight into whether modifications need to be targeted to specific loci, since it is surprising that any active Tet catalytic domain can completely rescue reprogramming in TKO MEFs. This raises questions about the roles of the different Tet isoforms and their N-terminal halves, which offer further opportunities for research (see Section 5.2).

5.1.4: Additional applications of low-efficiency and hmC-dominant TET variants

Although we have focused most effort on the MEF reprogramming study, our TET mutants lend themselves to many potential applications. Notably, these are not restricted to TET

triple-knockout models but may also include cases where one or two TET isoforms predominate. We recently initiated a collaboration to introduce our mutants into zebrafish lacking Tet2 and Tet3. In this model, hemapoietic stem cells fail to develop normally, but the defect can be rescued by injecting mRNA that encodes human TET2 (or TET3) into one-cell-stage embryos (Li et al., 2015a). We have cloned the TET2 mutant constructs for *in vitro* transcription and are awaiting the results from pilot experiments being performed in the Goll lab. Further collaborative work on mouse Tet1 mutants is being led by Blake Caldwell in the Bartolomei lab.

As additional examples, two studies demonstrated that downregulation of TET1 in colon cancer (Neri et al., 2015a) and TET2 in melanoma (Lian et al., 2012) drives cancer cell proliferation both in culture and in mouse xenograft models. In both studies, inducible overexpression of the corresponding TET isoform dramatically slowed proliferation, even in actively growing tumors. Recalling also the isolated role of TET2 in hematological malignancies and TET3 in zygotic development, we propose that introducing mutants into these settings can help to broaden understanding of ox-mCs in health and disease.

5.2: Directly compare the activities of TET1, 2, and 3 *in vitro* and in doxycycline-inducible cell lines

5.2.1: Introduction

Another major effort has been aimed at comparing the activities of the three TET isoforms in cells and *in vitro*.[*] This has been a key question in the field, but the idea also sprang from our observation that the TET2 hmC-dominant mutants produced WT-like levels of hmC when overexpressed in HEK293T cells, yet displayed significant loss of activity *in vitro*. We

[*] R.M.K., J.E.D., and I conceived the ideas. J.E.D. is leading the biochemical experiments, while I have focused on cells.

wondered how the cellular environment regulates levels of genomic ox-mCs and therefore set out to generate HEK293T stable cell lines in which we could titrate levels of TET protein expression. Using these cell lines, we could express any TET construct in a controlled manner, with expression levels closer to a physiological range, and directly compare enzyme activities in cells. We therefore included not only TET2 variants but also WT TET1 and TET3. Combined with our growing suite of biochemical assays, we are now poised to compare the activities of all three TET isoforms for the first time.

5.2.2: Preliminary results

To generate the stable cell lines, we obtained HEK293T cells modified for high-efficiency, low-background (HILO) recombination-mediated cassette exchange (RMCE)—i.e. site-specific incorporation of a gene into a tetracycline-controlled module (Khandelia et al., 2011). We first cloned WT TET2 catalytic domain into the recombination cassette plasmid and co-transfected it into HILO-HEK293T cells along with a nuclear-localized Cre recombinase. Following puromycin selection for approximately two weeks, the surviving colonies were pooled into a polyclonal cell line. Pilot doxycycline induction experiments demonstrated a titratable range of ox-mC production in the genomic DNA (Figure 5-3).

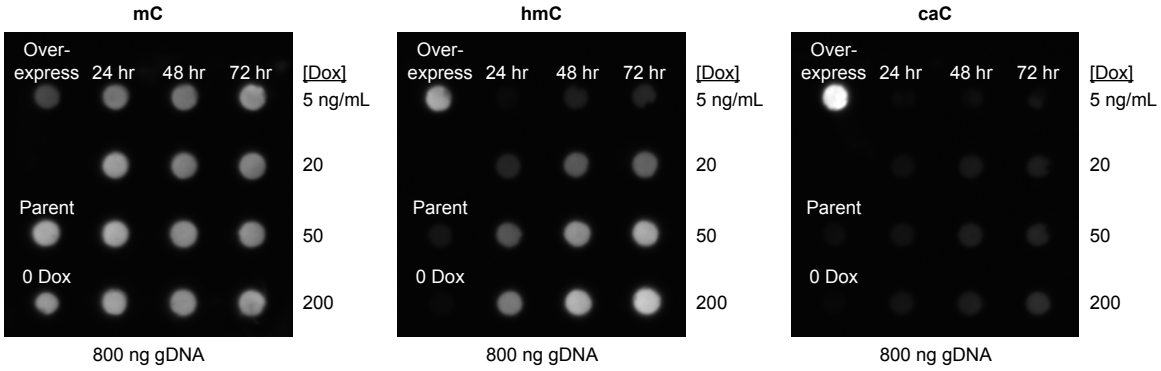


Figure 5-3. Doxycycline induction of WT TET2-CD polyclonal cell lines.

Cells were harvested 24, 48, or 72 hr after addition of doxycycline and genomic DNA was isolated for dot blotting. Controls include gDNA from cells overexpressing WT TET2-CD and the parent HILO-HEK293T cells before the *Tet2* gene was introduced.

From this polyclonal line, we performed limiting dilutions to generate monoclonal and "oligoclonal" lines usually containing 3-10 pooled colonies. Genotype analysis by PCR was used to confirm homogeneous recombination into the proper locus. Initially, we focused our analysis on WT TET2 and the T1372E mutant. Doxycycline titrations, followed by dot blotting of gDNA, showed that monoclonal lines generated very few ox-mCs, perhaps because the cells had been over-stressed, but oligoclonal lines behaved similarly to the polyclonal controls (data not shown). Comparing WT to T1372E, we found that with lower protein expression, the T1372E mutant produced lower hmC levels in cellular DNA, though the difference may be less in cells than *in vitro* (Figure 5-4). This suggests that the equivalent hmC levels observed previously were likely due to overexpression, though neither overexpression nor stable expression fully reflect the reaction rates measured *in vitro*. Notably, the levels of hmC in WT doxycycline-induced cells are near the limits of detection by our mass spectrometry methods, which has so far prevented accurate quantification of ox-mCs in cells. This will be an important area of development as we push toward quantifying rarer modifications.

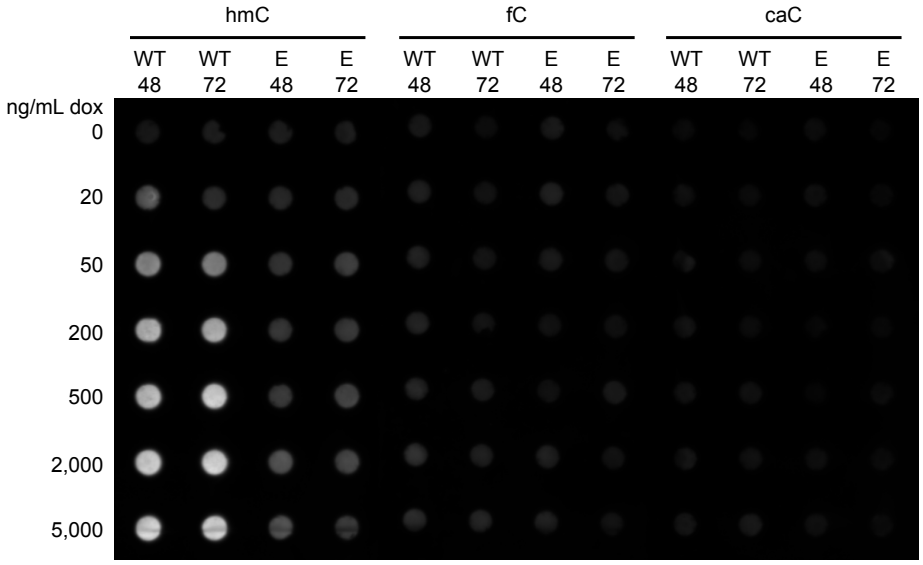


Figure 5-4. Comparison of WT TET2-CD and T1372E mutant activity in cells.

Protein expression was induced with doxycycline for 48 or 72 hr, and genomic ox-mCs were probed by dot blotting.

We have now generated stable, oligoclonal cell lines for WT human TET1, TET2, and TET3 catalytic domains along with several mutants: TET2-T1372A/E/V/W and TET1-T1662E/V/W. Most lines have been confirmed by PCR amplification and sequencing of the TET gene locus. Further validation of each cell line is underway. In addition, we have cloned, expressed, and purified human TET1, 2, and 3 catalytic domains for parallel biochemical analyses.

5.2.3: Next steps

We envision studying the stable cell lines alongside a rigorous biochemical comparison of the three TET isoforms. Focusing first on the WT enzymes, this will enable us to address important questions about the mechanisms of the different isoforms and the extent of redundant vs. distinct functions. For instance, we aim to perform our opposite strand activity assays (Chapter 3) on all three TET isoforms *in vitro* to determine whether they differ in *de novo* or

maintenance activity. We can then look for evidence of how biochemical preferences translate to a cellular setting. In this example, we propose to use variations on hairpin bisulfite sequencing to capture the modification status at complementary CpGs. We would predict that a maintenance TET isoform would tend to generate more symmetrically modified CpGs compared to an isoform with purely *de novo* activity. Thus, we now have a unique ability to examine how the behaviors of each TET isoform cross between biochemical assays and a cellular model. We also have the versatility to further manipulate and disrupt these models, not only with the TET mutants but also potentially with siRNAs or co-expressed proteins. Finally, with the stable cell lines, we are not limited to studies of the catalytic domains but can feasibly introduce full-length TET constructs. This would provide significant insight into the functions of the enzymes' N-terminal halves, which remain poorly understood.

5.3: Additional projects and concluding remarks

Indeed, TET enzymes offer ample opportunities for future research. Additional directions being pursued in our lab include examining TET activity on unnatural cytosine substrates, such as 5-vinyl- or 5-ethynylcytosine. We have shown that TET oxidation chemistry can extend to these alternative substrates, which may guide the development of new TET activity assays and/or activity-based probes. I also recently initiated a study of TET activity on DNA/RNA hybrids, aiming to characterize the relative activity on each nucleic acid substrate singly or in duplexes. More adventurous ideas include feeding cells with ^{14}C-labeled methionine and using HPLC-LSC to screen for potentially unknown genomic modifications, or developing a platform in *E. coli* for experiments in directed evolution of TET enzymes. We may also revisit an idea to use TDG base excision as part of a novel sequencing strategy for fC and caC—a project that occupied my first few months in lab, was stalled at the library preparation step, and may now be more within reach, given our lab's growing familiarity with genome sequencing technologies.

It is always a pleasure to witness a field evolve, to see each unknown frontier crossed and reset a little farther away. I could not have asked for a better lab environment from which to experience and contribute to the field of TET enzymes and the extended epigenome. Our understanding has grown exponentially since TET's discovery in 2009, yet this young field is arguably just reaching its adolescent stage. Fundamental questions still remain unsolved—the functions of each individual base, the intrinsic and extrinsic factors that regulate TET activity, etc.—but steady interest and rapid technological advances promise answers soon to come.

APPENDIX

Supplementary Information for Chapter 3

Supplementary Methods

Enzymatic preparation of *S*-[$^{13}C^{2}H_3$-Me]-adenosyl-*L*-methionine ([$^{13}C^{2}H_3$]-SAM). [$^{13}C^{2}H_3$]-SAM was enzymatically synthesized using [$^{13}C^{2}H_3$]-*L*-methionine (Sigma) and recombinant SAM synthetase. The *E. coli* SAM synthetase gene (*metK*) was cloned in a pET41 vector with an N-terminal HIS tag, expressed and purified as described (Ottink et al., 2010). To prepare [$^{13}C^{2}H_3$]-SAM, 0.2 mM [$^{13}C^{2}H_3$]-*L*-methionine was reacted with 2 mM ATP and 0.2 mg/ml SAM synthetase in 1X NEB Buffer 2 (10 mM Tris-Cl, 50 mM NaCl, 10 mM $MgCl_2$, 1 mM DTT, pH 7.9) (Figure S3-1A). The mixture was incubated at 37°C for 15 min, after which duplexed DNA was added with CpG methyltransferase to generate [$^{13}C^{2}H_3$]-labeled mC oligos *in situ*, as described below.

Enzymatic preparation of mC oligonucleotides. 49.2 µM Oligo 1 and 50.8 µM Oligo 2 (Table S3-1) (49.2 µM duplex) were annealed in a thermocycler by incubation for 4 min at 95 °C followed by 30 s stepwise decreases of 5 °C to 40 °C. Similar protocols were utilized to generate substrates with specific introduction of (A) ^{14}C-methylation or (B) $^{13}C^{2}H_3$-methylation (Figure S3-1A). For (A), final concentrations of 5.8 µM Duplex in a reaction containing 1X NEB Buffer 2, 35 µM *S*-[^{14}C-Me]-adenosyl-*L*-methionine ([^{14}C]-SAM; Perkin Elmer) and 224 U/mL CpG Methyltransferase (NEB) was reacted at 37 °C for 4 h. For (B), final concentrations of 10 µM Duplex in a reaction containing 1X NEB Buffer 2, 1:5 [$^{13}C^{2}H_3$]-SAM enzymatic reaction mix (above), and 1000 U/mL CpG Methyltransferase (NEB) was reacted at 37° C for 5 h. For (A), a small sample of the reaction was collected to determine specific radioactivity for liquid scintillation counting (LSC). For either (A) or (B), at the end of methylation, the reaction was precipitated with 0.1 volume 3 M Sodium Acetate, pH 5.2 and 2.5 volumes of 100% ethanol on

ice for 1 h and pelleted by centrifugation at 13,000 rpm for 30 min. The supernatant was then removed by pipetting, the pellet was washed with 70% ethanol/75 mM sodium acetate pH 5.2 stored at -20 °C and briefly re-centrifuged. The wash supernatant was removed and the pellet dried by air. The dried reaction pellet was dissolved in 0.2 initial reaction volume H_2O and desalted with G-25 spin columns equilibrated in water (CS-901, Princeton Separations). The desalted, methylated oligonucleotide mix was digested with 5,000 U/mL HpaII (NEB) in 1X Cutsmart Buffer (NEB), at 37 °C overnight. The digestion reaction was ethanol precipitated and washed as before, and the dried pellet dissolved in Purification Buffer A (100 mM triethylamine acetate (TEAA), pH 7). This sample was purified by ion-pairing HPLC using a 4.6 x 100 mm Zorbax Eclipse Plus C18 (Agilent) column pre-equilibrated to 65°C in 65% Purification Buffer A/35% Purification Buffer B (50% Methanol/100 mM TEAA, pH 7) and separated over a 20 min gradient from 35% to 45% Purification Buffer B at 1 mL/min, collecting 0.5 mL fractions (Figure S3-2). The HPLC conditions separate the 36-FAM-labeled full-length methylated Oligo 1 from Oligo 2 and unmethylated digested strands. Peaks were pooled and lyophilized for further use as [^{14}C]-mC-Oligo 3 or [$^{13}C^{2}H_3$]-mC-Oligo 3 (Table S1) and [^{14}C]-mC-Oligo 4 or [$^{13}C^{2}H_3$]-mC-Oligo 4 (Table S3-1).

Purification of Tet2. The plasmid encoding the catalytic domain of mTet2 (1042-1912) (referred to as Tet2 throughout) cloned into pFastBac1 plasmid with an N-terminal FLAG tag was generously provided by Yi Zhang. Tet2 protein was expressed in Sf9 cells as previously described (Ito et al., 2010). 6 g of cells were resuspended in lysis buffer (50 mM HEPES, 500 mM NaCl, 0.1% NP-40 pH 7.4 at 4 °C) with cOmplete EDTA-free Protease Inhibitor Cocktail (Roche). Cells were lysed by 3 passes through a microfluidizer. The lysate was then clarified by centrifugation and bound to 1 mL α-Flag M2-affinity gel (Sigma) by gentle agitation for 2 h at 4 °C and applied to a PolyPrep column (BioRad). Bound protein was washed with 4 X 1 mL wash buffer (20 mM HEPES, 150 mM NaCl, 15% glycerol pH 7.4 at 4 °C) and eluted with 100 µg/mL

3X FLAG peptide (Sigma) in wash buffer, collecting 500 µL fractions. Fractions were evaluated for protein by BioRad Protein Assay (BioRad) and by SDS-PAGE (Figure S3-3A) for estimation of concentration. Fractions were pooled, DTT added to 1 mM, and aliquots frozen at -80 °C.

Tet2 activity assay. Duplexed oligonucleotide substrates were annealed in a thermocycler as described above. Optimized assay reaction conditions were used which yielded highest activity and linear turnover with time (Figure S3-3B). Notably, the conditions which show optimal activity may promote maintenance of active enzyme as the selected pH is associated with slower Fe(II) oxidation relative to higher pH conditions (Morgan and Lahav, 2007). To a reaction mixture containing 50 mM HEPES pH 6.5, 100 mM NaCl, 1 mM 2-ketoglutarate, 1 mM DTT, 2 mM sodium ascorbate and the pre-annealed oligo, 75 µM freshly prepared ammonium iron(II) sulfate (Sigma) was added, followed by Tet2. The substrate and enzyme concentrations used in each reaction are explicitly noted in each figure legend, with typical reaction volumes of 50-200 µL. The reactions were incubated at 37 °C, quenched by addition of 8X 100% ethanol and 2X Oligo Binding Buffer (Zymo), purified over Oligo Clean and Concentrator columns (Zymo) per manufacturer instructions and eluted in H_2O. For the two highly quantitative HPLC and LC-MS/MS assays described below, the eluted reaction products were then degraded to component nucleosides with DNA Degradase Plus (Zymo) per manufacturer instructions at 37 °C overnight and further analyzed (Figure S3-1B). For semi-quantitative assays (Figure S3-3C), the purified reaction products were used (without degradation to nucleosides). For the MspI coupled assay to detect most ox-mC products (Ito et al., 2011), the purified reaction products were incubated with 2 mM UDP-glucose and 1:25 by volume of T4 β-glucosyltransferase (βGT, New England Biolabs) to glucosylate hmC products and incubated at 37 ºC for 30 min. Subsequently, the products were digested by MspI (3 U/µL) at 37 ºC for 2 hrs, resolved on denaturing polyacrylamide gel and imaged for FAM fluorescence on a Typhoon scanner. For the TDG coupled assay to detect fC and caC products, the reaction products were

incubated with a 10-fold molar excess of TDG (purified as described in Maiti and Drohat, 2011). After incubation at 37 ºC for 2 hrs, an equal volume of 0.3 M NaOH/0.03 M EDTA was added and the reaction incubated at 85 ºC for 15 min to cleave abasic sites. The samples were then processed and imaged as described for the MspI coupled assay.

HPLC analysis of [^{14}C]-labeled nucleosides. 10 μM each mC, hmC, fC and caC nucleosides (Berry and Associates) were added to the degraded Tet reactions (as chromatographic controls), and this sample was injected onto a 2.1 x 250 mm Supelcosil LC-18S analytical column (Sigma) equilibrated to 50°C in 100% Analysis Buffer A (5 mM ammonium formate, pH 6.0). The nucleosides were separated in a gradient of 0-30% Analysis Buffer B (4 mM ammonium formate pH 6.0, 20% methanol) over 20 min at a flow rate of 0.5 mL/min. Fractions (0.25 mL in target areas, 1.0 mL for everywhere else) were collected and mixed with Opti-Fluor liquid scintillant (Perkin Elmer) for liquid scintillation counting (LSC) on a Tri-Carb 2910 TR (Perkin Elmer). Each vial was counted for 10 min using the ^{14}C DPM setting, with an automatic background correction made from the DPM measurement of a vial containing an appropriate liquid scintillant/HPLC buffer mixture. The outputted DPM measurements were corrected for inputted volume to reflect the total DPM of each fraction and plotted against the HPLC UV trace from that run, which shows the nucleoside standards to confirm identities (Figure 3-2A). The total radioactivity of each peak on the chromatogram was then analyzed to calculate percentage of each product, and normalized to the known inputted concentration to convert to molar quantities of products.

LC-MS/MS analysis of [^{13}C^{2}H$_3$]-labeled nucleosides. A 150 μm x 17 cm precolumn and 100 μm x 26 cm analytical reverse phase column were made by first preparing a Kasil frit (a 1:3 Formamide:Kasil 1624 (PQ corporation) mixture was drawn into column using capillarity, polymerized at 100 °C overnight and trimmed to ~5 mm), and packed with Supelcosil LC-18S resin (Sigma). Nano LC-MS chromatography was performed using an Easy-nLC 1000 (Thermo)

with a two-column setup. The sample was bound to the precolumn and desalted by 5 min of isocratic flow of 0.1% formic acid, then separated on a gradient of 0-30% of acetonitrile into 0.1% formic acid over 30 min at a flow rate of 600 nL/min. Nucleosides were subjected to positive ion mode electrospray ionization in a Q Exactive hybrid quadrupole-orbitrap mass spectrometer (Thermo), with a capillary temperature of 275 °C and spray voltage of 2.9 kV. Total ion count and tandem MS transitions were collected, and the ratios of peak areas of heavy and light nucleosides were compared for each individual nucleotide modification (Figure S3-4).

Supplementary Figures and Tables

Name	Sequence (all bases are 2'-deoxynucleotides)	Source/Notes
Oligo 1	5'-GTA TCT AGT TCA ATC CGG TTC ATA GCA-(36-FAM)-3'	Synthesized by Integrated DNA Technologies (IDT)
Oligo 2	5'-TGC TAT GAA CCG GAT TGA ACT AGA TAC-3'	IDT
Oligo 3	5'-GTA TCT AGT TCA ATC mCGG TTC ATA GCA-(36-FAM)-3'	Chemoenzymatic Preparation (isotopically labeled) or IDT
Oligo 4	5'-TGC TAT GAA CmCG GAT TGA ACT AGA TAC-3'	Chemoenzymatic Preparation (isotopically labeled) or IDT
Oligo 5	5'-TGC TAT GAA ChmCG GAT TGA ACT AGA TAC-3'	Synthesized*
Oligo 6	5'-TGC TAT GAA CfCG GAT TGA ACT AGA TAC-3'	Synthesized by TriLink
Oligo 7	5'-TGC TAT GAA CcaCG GAT TGA ACT AGA TAC-3'	Synthesized*

Table S3-1. Oligonucleotides used in Chapter 3.

*Oligos 5 and 7 were prepared in house using standard phosphoramidite chemistry (reagents from Glen Research) on an ABI394 synthesizer (Applied Biosystems).

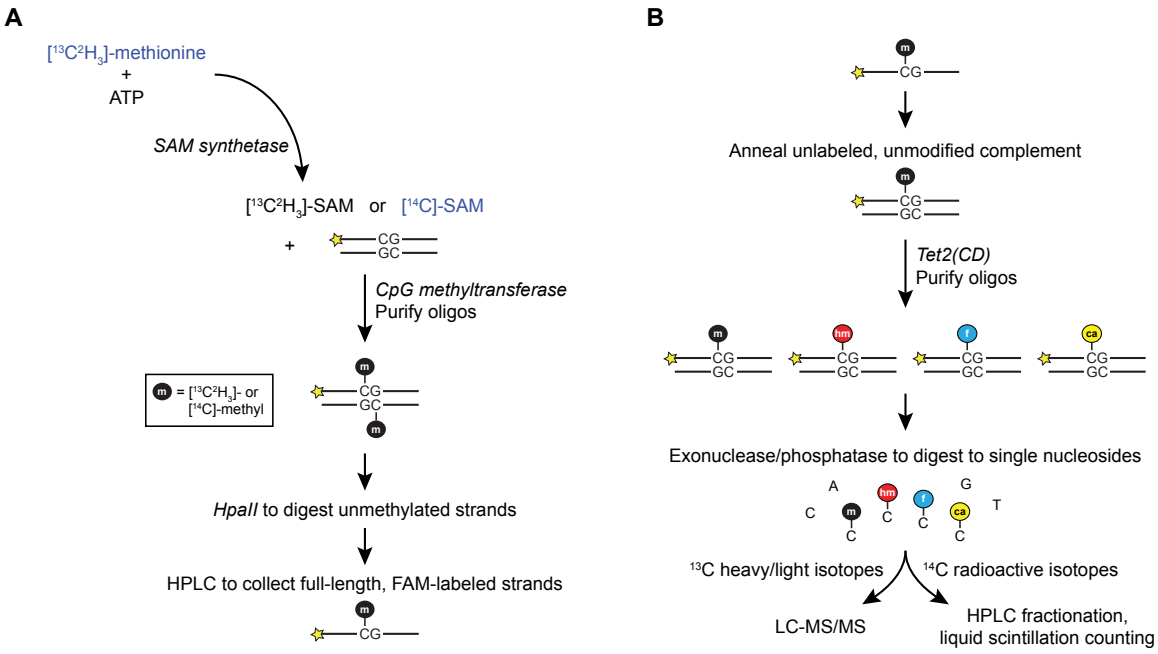


Figure S3-1. Schematic of experimental setup and analysis.

(A) Chemoenzymatic preparation of isotopically-labeled mC-containing substrates, as described in Supplementary Methods. The starting reagents purchased from commercial sources are indicated in blue: [^{14}C]-SAM or [$^{13}C^{2}H_{3}$]-methionine, which was enzymatically converted into [$^{13}C^{2}H_{3}$]-SAM. **(B)** Tet activity assays. Duplexed oligonucleotide substrates were incubated with Tet2. The purified reaction products were degraded to nucleoside mixtures, which were then analyzed by either HPLC-LSC or LC-MS/MS.

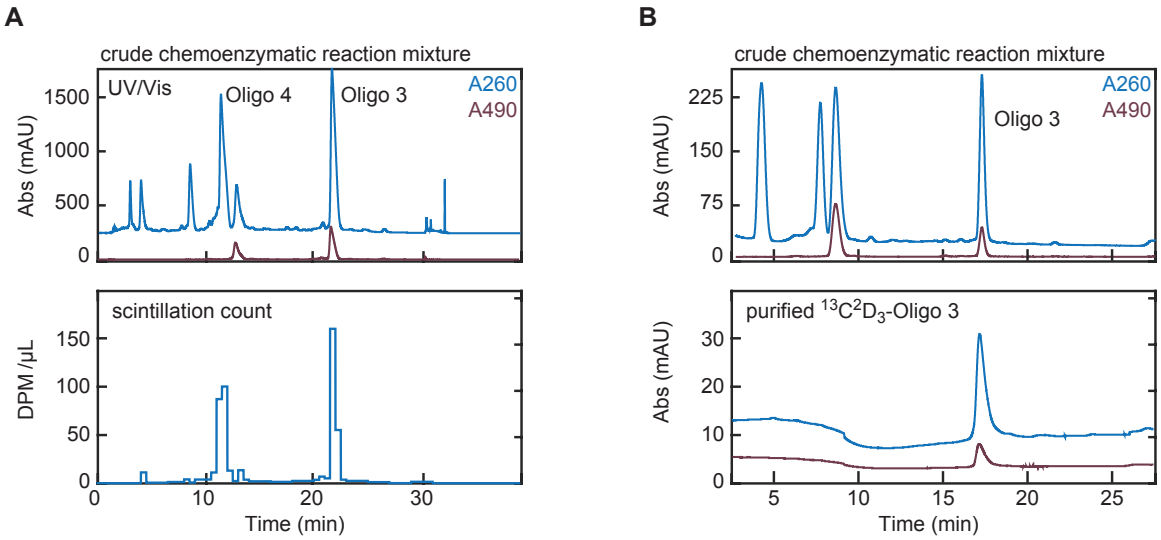


Figure S3-2. Preparation of isotopically labeled substrates.

(A) Chemoenzymatic generation of $^{14}CH_3$-labeled substrate. Shown is an HPLC chromatogram for the crude of the enzymatically-generated substrates, highlighting the A260 (blue) and FAM strand only (purple trace). At bottom is the associated scintillation counts, showing that Oligo 3 is radiolabeled with $^{14}CH_3$. Unlabeled peaks represent digested, unmethylated oligonucleotides. Purified radiolabeled Oligo 3 was used in all experiments involving LSC. **(B)** Chemoenzymatic generation of $^{13}C^{2}D_3$-labeled substrate. Shown at top is the crude chemoenzymatic mixture from generation of the labeled substrate. The peak corresponding to Oligo 3 was purified and used in all experiments involving isotope dilution. The trace at bottom demonstrates the homogeneity of the HPLC purified oligonucleotide.

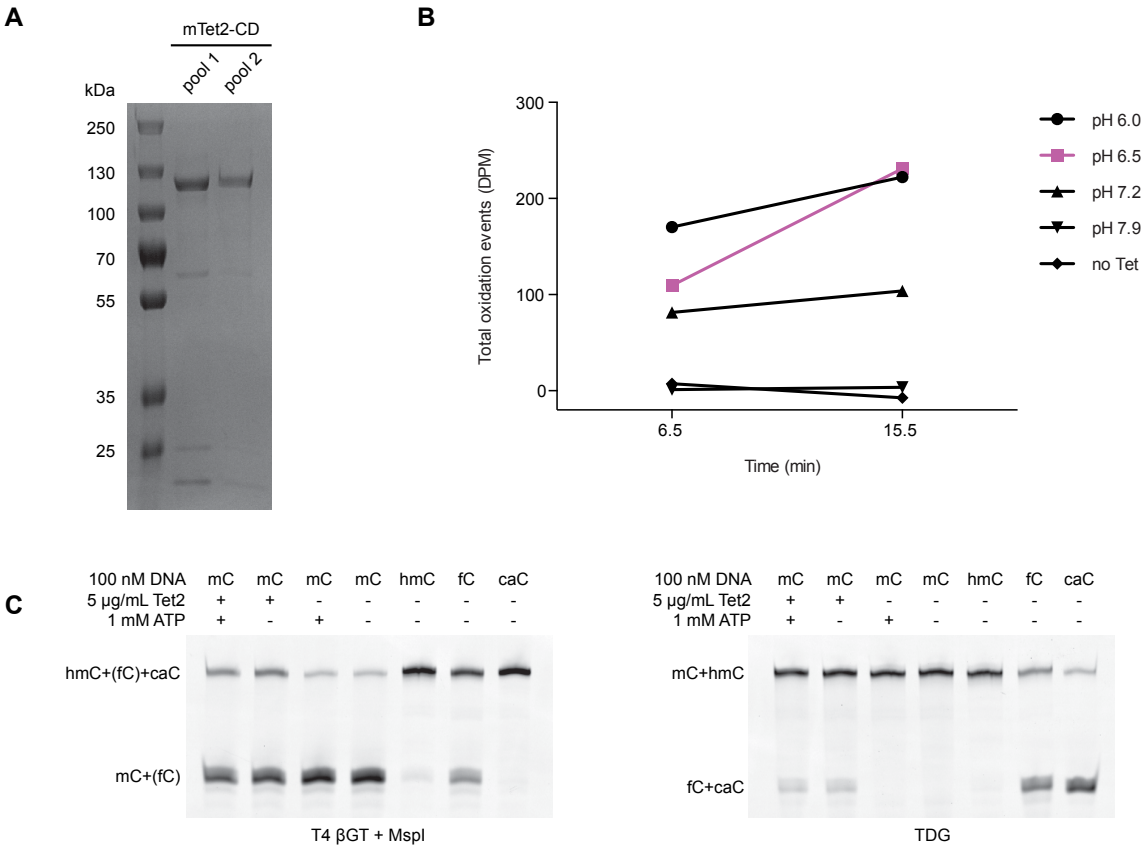


Figure S3-3. Optimized *in vitro* reaction conditions.

(A) Coomassie-stained SDS-PAGE gel showing purified Tet2. Elution fractions from purification were combined into two pools. **(B)** pH-dependence of Tet activity was assessed using 200 nM [^{14}C]-labeled mC substrate (Oligo3/Oligo 4 duplex) reacted with 4 µg/ml of Tet2 for 6.5 or 15.5 min. The total oxidation events are plotted as the disintegrations per minute (DPM) of hmC + 2*DPM of fC + 3*DPM of caC, after background subtraction of the controls without Tet2. At pH 6.5, the enzyme activity was linear with time. **(C)** Under optimal reaction conditions, Tet2 was reacted with DNA in the presence or absence of 1 mM ATP. The products were analyzed by protection from MspI digestion (left gel). T4 βGT treatment protects hmC-containing DNA from cleavage by MspI, while caC is not cleaved and fC is a poor substrate (reaction with control oligonucleotides lanes 5-8). Additionally, fC and caC formation were detected by treatment with TDG (right gel). Both assays indicate that ATP does not significantly enhance activity (comparison of lanes 1 and 2) and additionally demonstrate that no products are detected in the absence of enzyme (lane 4).

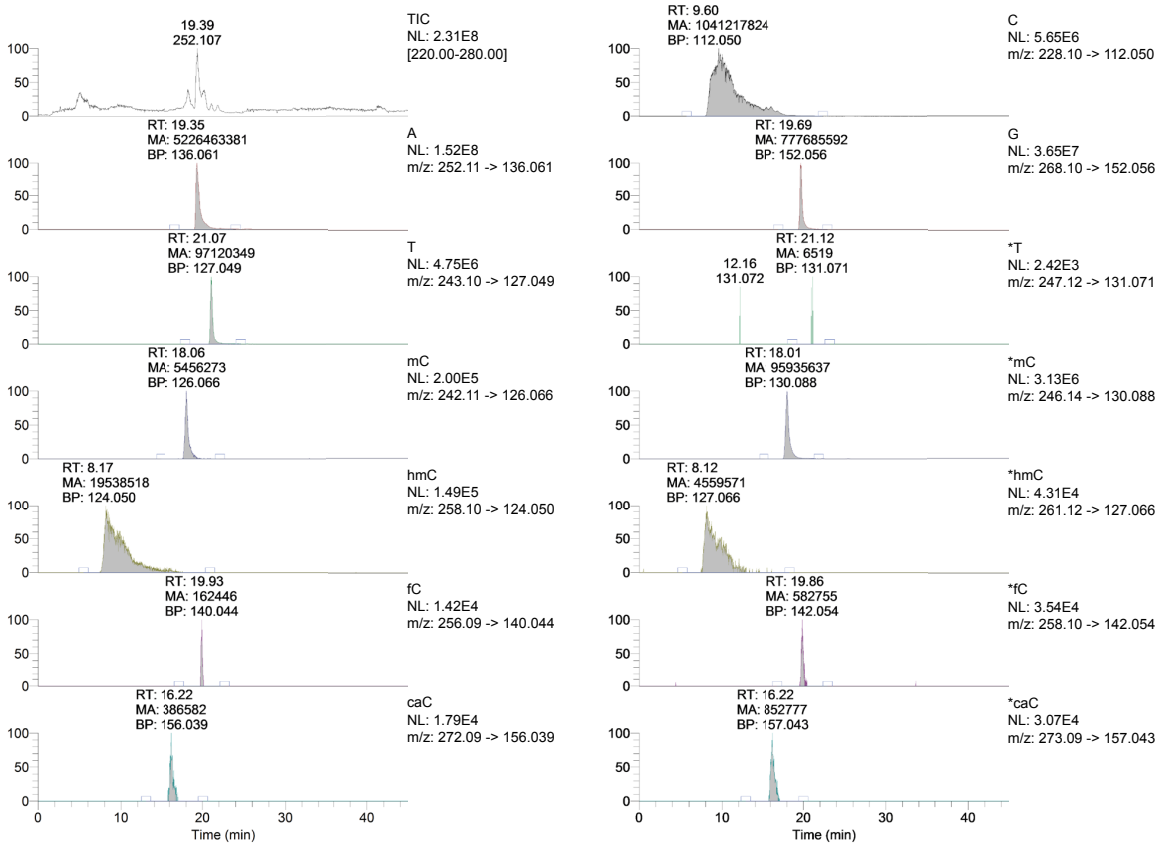


Figure S3-4. Representative mass chromatograms.

Shown are the total ion count and mass transitions for each of the degraded nucleosides in their light and heavy forms (* denotes heavy isotopes).

Supplementary Information for Chapter 4

Supplementary Figures

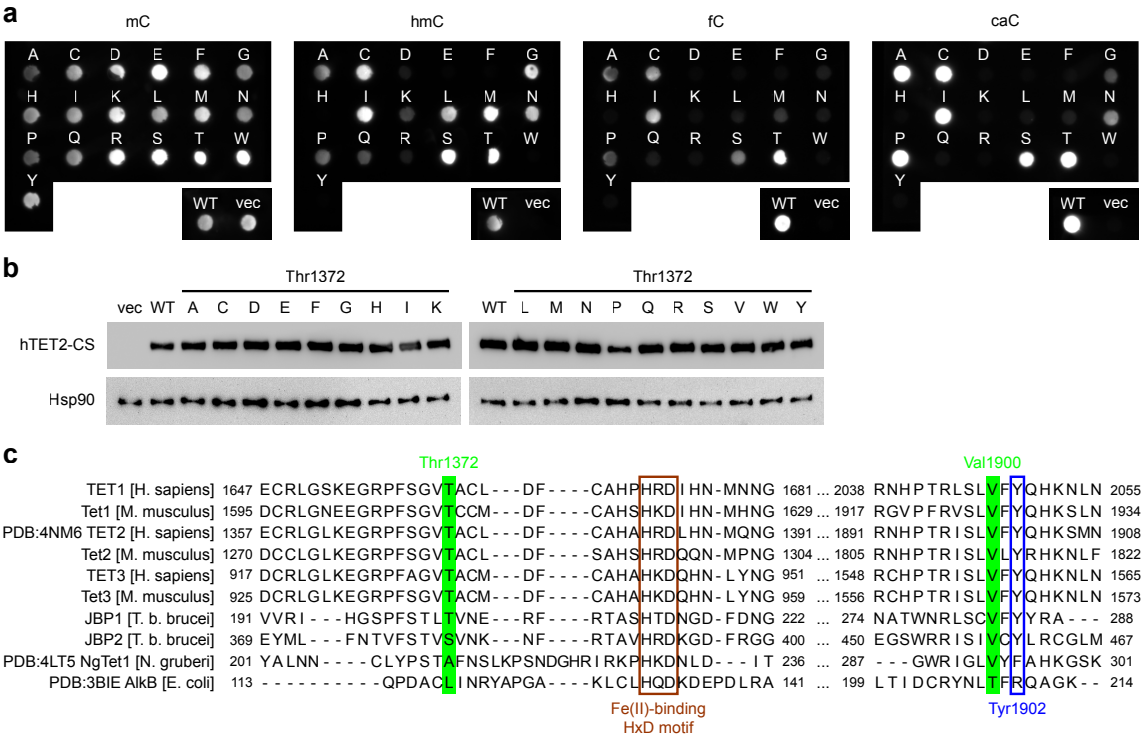
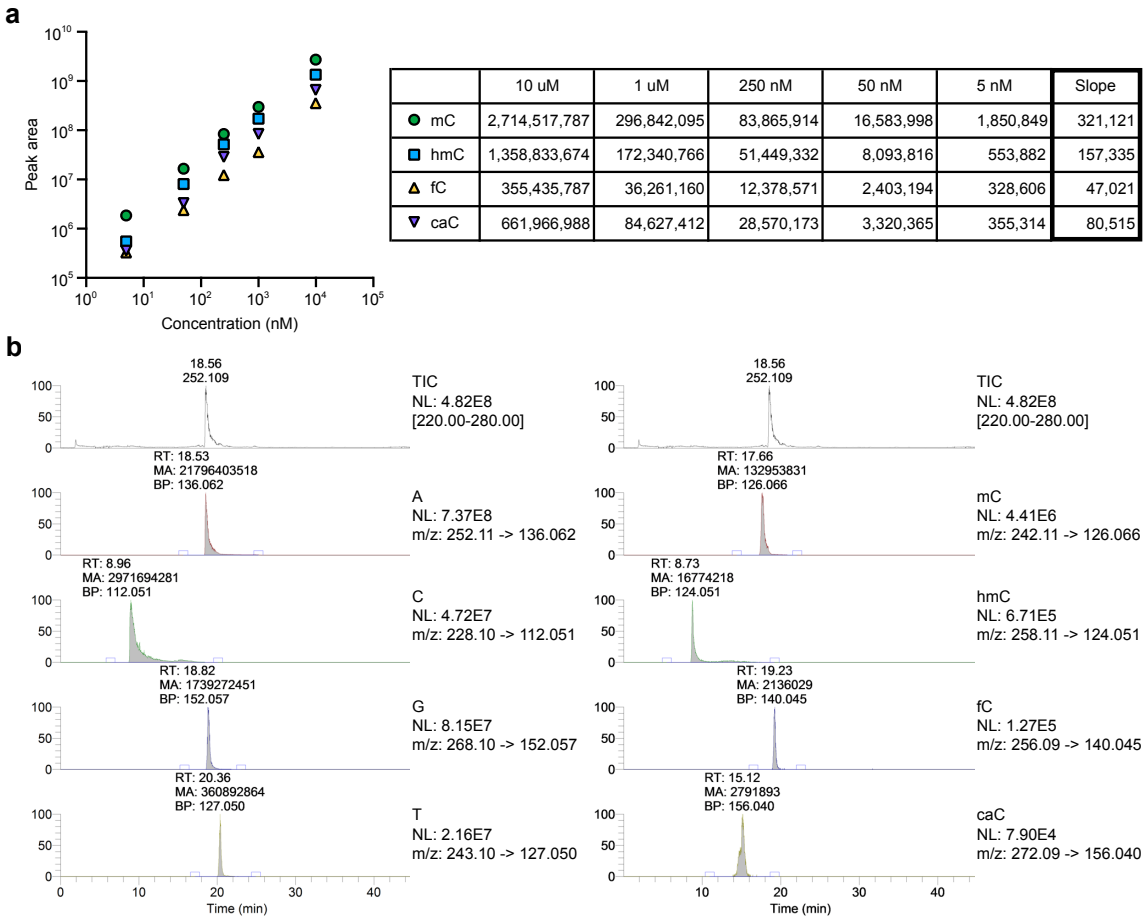


Figure S4-1. Saturation mutagenesis along the conserved active site scaffold.

(a) Dot blots for mC, hmC, fC, and caC in 400 ng of genomic DNA isolated from HEK293T cells transfected with Val1900 mutants. Mutants are in alphabetical order, followed by WT and vector-transfected samples. To maintain consistency, all results shown here and in Figure 4-2a are cropped from the same representative blots (uncropped dot blots in Figure S4-10a). **(b)** Western blot using anti-FLAG antibody to detect hTET2-CS mutants in lysates of transfected HEK293T cells. Hsp90α/β served as a loading control. WT and Thr1372 mutants are shown in alphabetical order, along with an empty vector-transfected control. **(c)** Multiple sequence alignment of human and mouse TET isoforms, the trypanosomal JBP1/2 thymidine hydroxylases, the *Naegleria* Tet-like protein NgTet1, and AlkB of *E. coli*. All these homologues, except AlkB, have been shown to be capable of multistep oxidation on their natural substrates. The residues of interest, Thr1372 and Val1900 in TET2, are highlighted (green), along with the key scaffold residue, Tyr1902 (blue), and HxD motif (red) characteristic of the Fe(II)/α-KG-dependent family of dioxygenases. Alignments were done using the PROMALS3D algorithm, based on the crystal structures of hTET2 (PDB 4NM6), NgTet1 (PDB 4LT5), and AlkB (PDB 3BIE).



	10 uM	1 uM	250 nM	50 nM	5 nM	Slope
mC	2,714,517,787	296,842,095	83,865,914	16,583,998	1,850,849	321,121
hmC	1,358,833,674	172,340,766	51,449,332	8,093,816	553,882	157,335
fC	355,435,787	36,261,160	12,378,571	2,403,194	328,606	47,021
caC	661,966,988	84,627,412	28,570,173	3,320,365	355,314	80,515

Figure S4-2. LC-MS/MS analysis of modified cytosine nucleosides.

(a) Representative standard curves for mC, hmC, fC, and caC nucleosides. So that all data were weighted equally, the slopes were calculated by scaling the data points to 1 nM and taking the average. **(b)** Representative LC-MS/MS spectrum of gDNA from HEK293T cells expressing WT hTET2-CS. Using the slopes of the standard curves, the peak areas were converted into concentrations of each modified cytosine in the gDNA sample. These were then expressed as the percent of total cytosine modifications.

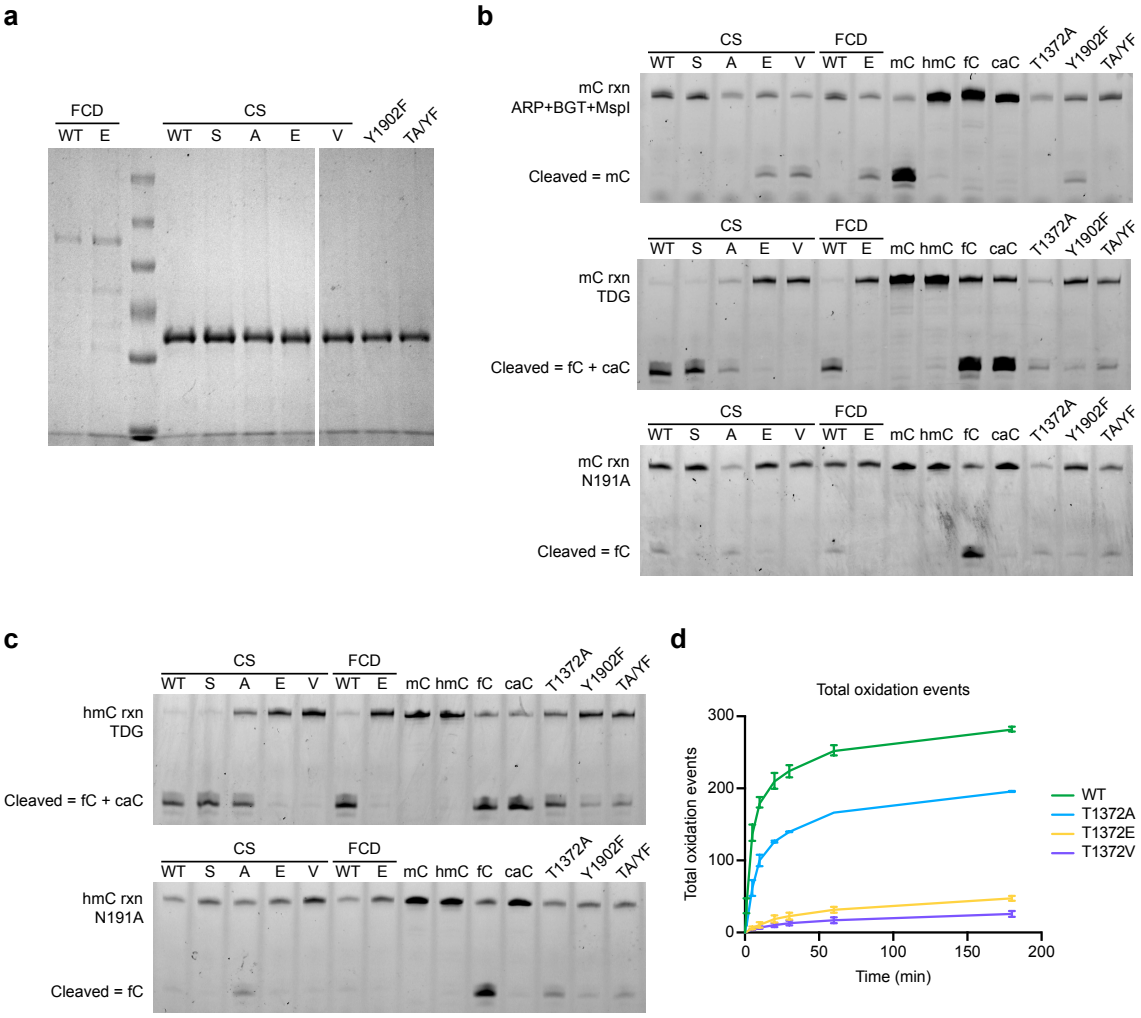


Figure S4-3. Biochemical characterization of select TET2 mutants.

(a) SDS-PAGE of TET2 variants purified from Sf9 insect cells: WT hTET2-FCD and T1372E-FCD, plus WT hTET2-CS and T1372S/A/E/V, Y1902F, and T1372A/Y1902F mutants. **(b,c)** Reactions of 30 μg/mL TET2 with 20 nM dsDNA substrates containing **(b)** mC or **(c)** hmC. The reaction products were purified and subjected to both LC-MS/MS (Fig. 4a) and chemoenzymatic assays, as described in the Methods. Control mC, hmC, fC, and caC substrates without TET were used to illustrate the cleavage patterns in each assay. These orthogonal, complementary assays corroborate the quantitative LC-MS/MS results. **(d)** As another measure of activity, we plotted the total oxidation events over the 3-h time course (Fig. 4b), counting hmC once, fC twice, and caC three times to reflect the number of oxidation steps required to generate each base from mC substrate: Total oxidation events (arbitrary units) = 1 × (% hmC) + 2 × (% fC) + 3 × (% caC). The results further illustrate the distinct WT, low-efficiency, and hmC-dominant phenotypes. Mean values are plotted (WT $n = 3$, mutants $n = 2$), and error bars represent the range.

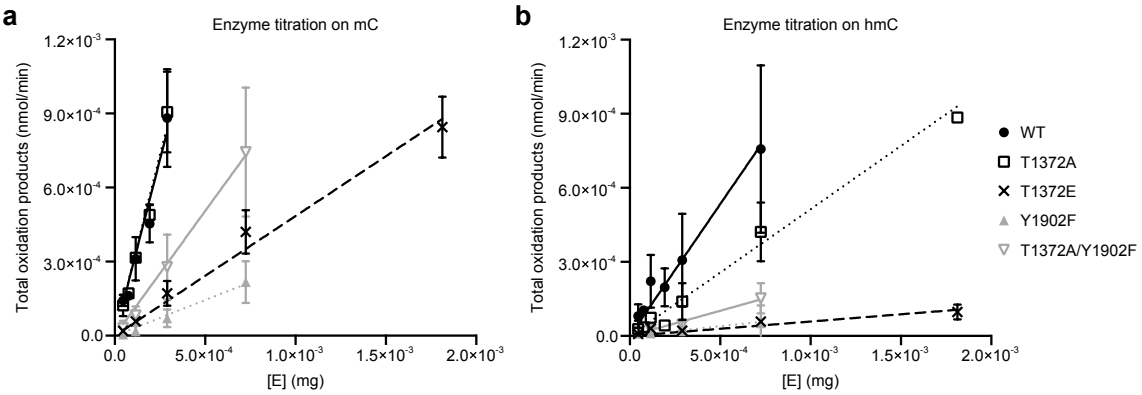


Figure S4-4. Enzyme titrations to compare reactivity of select TET variants.

PCR amplicons fully modified with **(a)** mC or **(b)** hmC were reacted with varying concentrations of enzyme for 30 min, and total oxidation products were quantified by LC-MS/MS. For the mC reaction, total oxidation products are hmC + fC + caC; for the hmC reaction, total oxidation products are fC + caC. Linear dependence of activity with enzyme concentration suggests that the assays are reporting on steady-state consumption of mC or hmC substrate. Under all conditions shown, for determination of the specific activity, <50% of the substrate is consumed. Shown are the mean ± s.d. from three independent experiments. The slopes of the linear regression lines are given in Table 4-1.

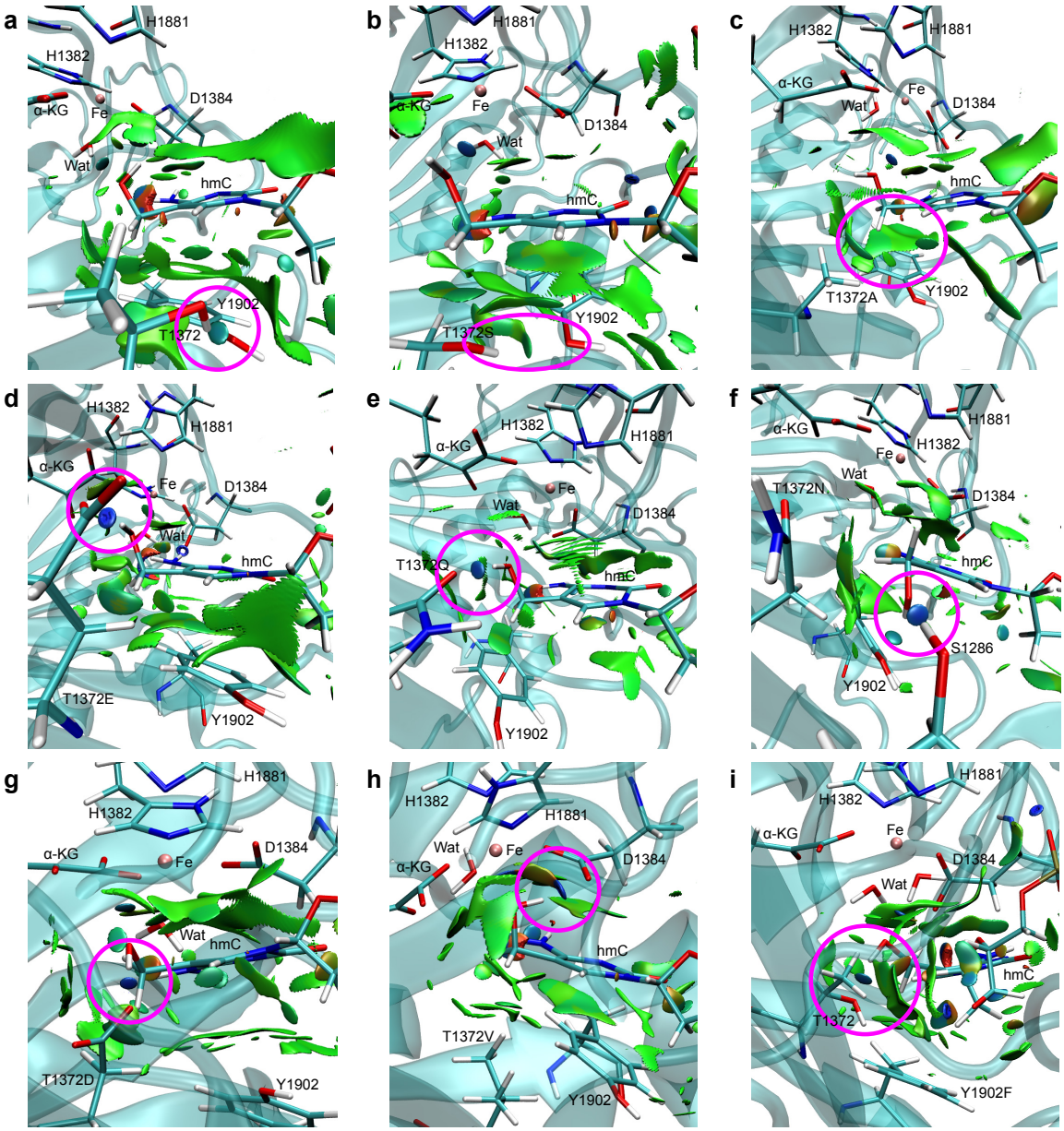


Figure S4-5. Non-covalent interaction (NCI) analysis on a representative snapshot for WT hTET2-CS and mutants in the presence of hmC.

(a) WT **(b)** T1372S **(c)** T1372A **(d)** T1372E **(e)** T1372Q **(f)** T1372N **(g)** T1372D **(h)** T1372V **(i)** Y1902F. Green surfaces denote weak interactions (e.g. van der Waals), blue surfaces are strong attractive interactions (e.g. hydrogen bonds), and red surfaces are strong repulsive interactions. Key interactions are circled. The coordinating water occupying the sixth (equatorial) position is omitted for clarity. The WT Thr1372-Tyr1902 active site scaffold is preserved in T1372S. T1372A removes the hydrogen bonding partner, leaving weakened non-covalent interactions in the active site. The hmC-dominant mutants T1372E/Q/N/D/V elicit a new hydrogen bond directly with the 5-hydroxymethyl moiety; for E/Q/D, the hydrogen bond involves the mutated residue itself, while for N/V the hydrogen bond involves nearby residue(s). The isovalue for NCI is 0.3 au, and -0.2 au $<$ sign(λ_2)ρ $<$ 0.2 au.

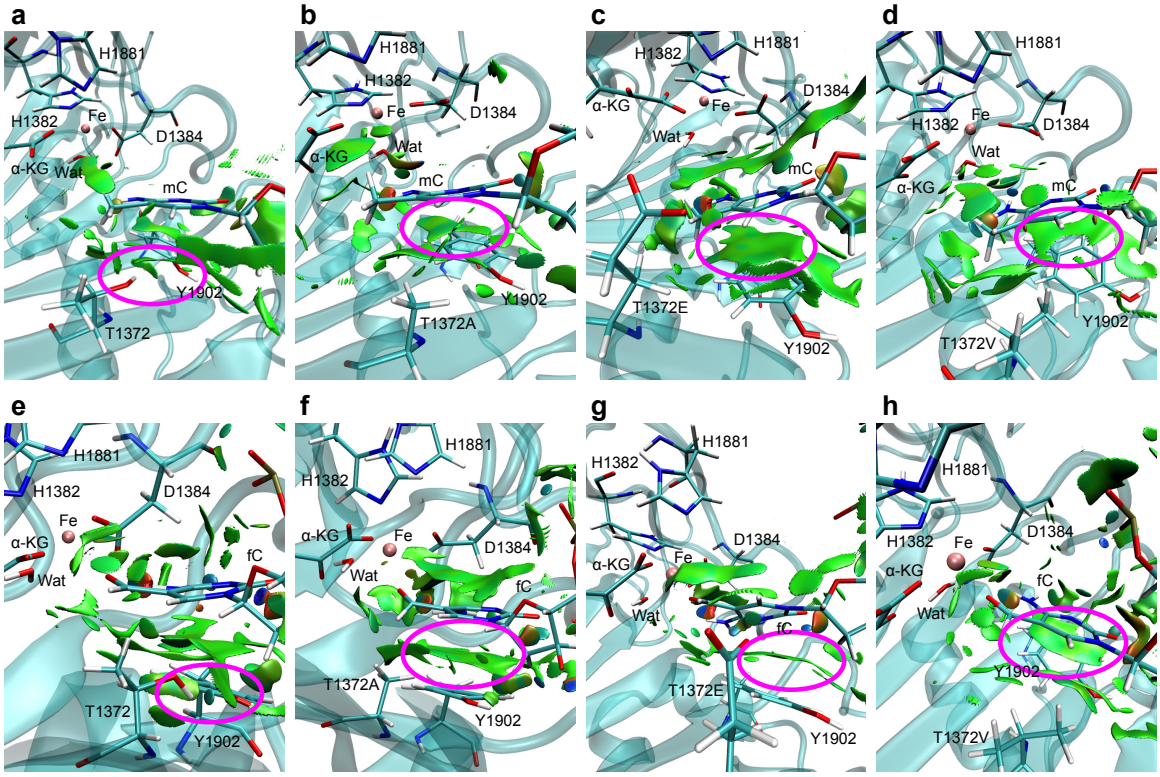


Figure S4-6. NCI analysis on a representative snapshot for WT and T1372A, E, and V mutants in the presence of mC and fC.

(a) WT with mC **(b)** T1372A with mC **(c)** T1372E with mC **(d)** T1372V with mC **(e)** WT with fC **(f)** T1372A with fC **(g)** T1372E with fC **(h)** T1372V with fC. Green surfaces denote weak interactions (e.g. van der Waals), blue surfaces are strong attractive interactions (e.g. hydrogen bonds), and red surfaces are strong repulsive interactions. Key interactions are circled. The coordinating water occupying the sixth (equatorial) position is omitted for clarity. The WT Thr1372-Tyr1902 active site scaffold is present in mC and fC models, as well as hmC (Supplementary Fig. 5a), but the aberrant new hydrogen bonding in hmC-dominant mutants is specific to hmC and is not observed with mC or fC. The isovalue for NCI is 0.3 au, and -0.2 au $<$ sign(λ_2)ρ $<$ 0.2 au.

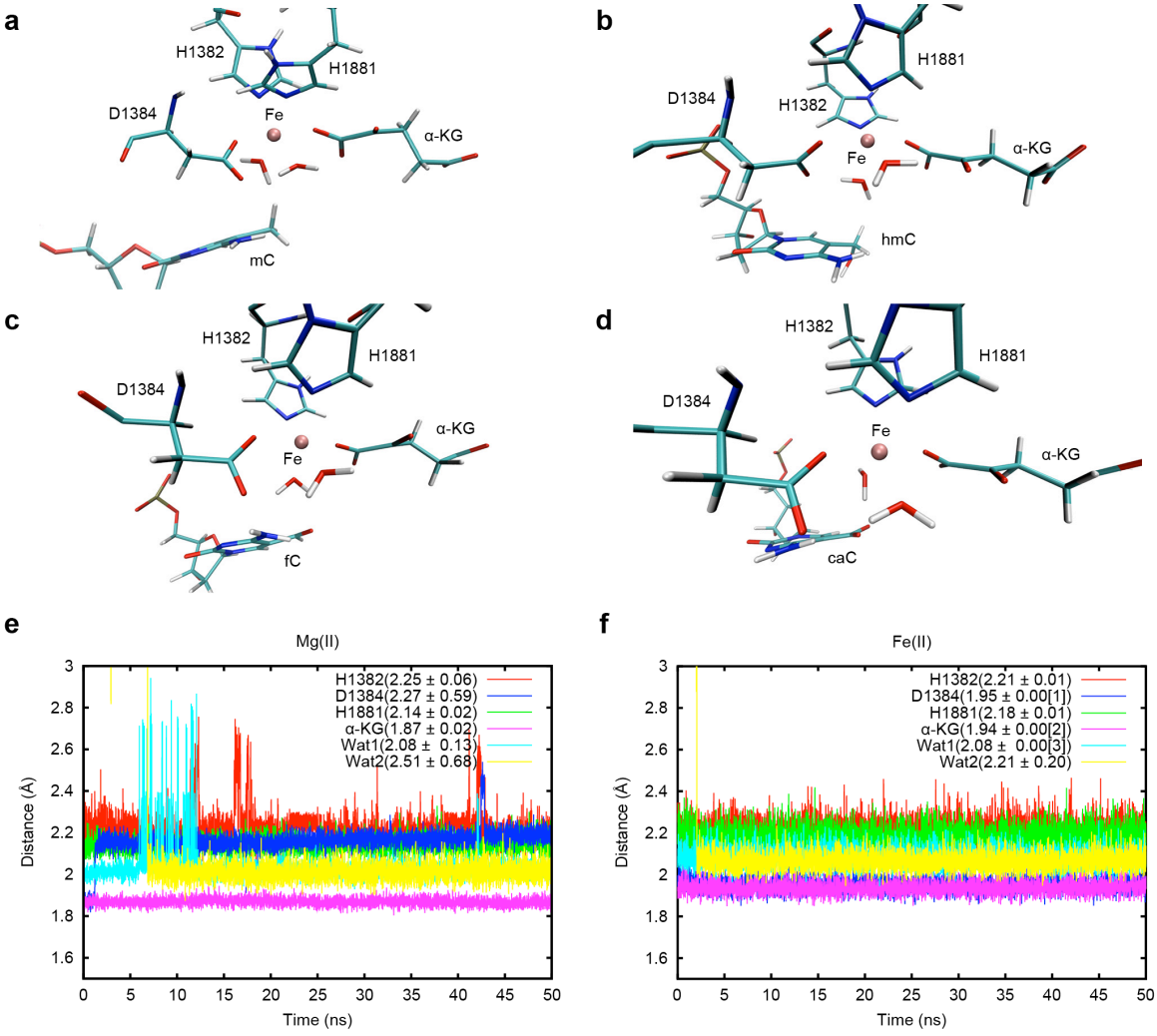
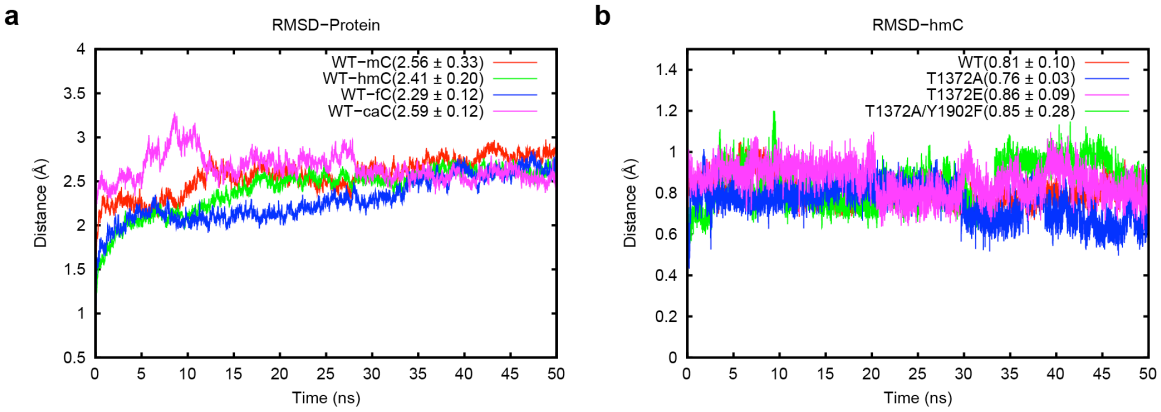


Figure S4-7. Metal ion coordination in MD simulations.

Coordination sphere in WT simulations for **(a)** mC, **(b)** hmC, **(c)** fC, and **(d)** caC. The Fe(II) surrogate (denoted Fe) was simulated by a Mg(II) and is hexa-coordinated in all systems. Note that in the crystal structure and our initial structure for MD simulations, α-KG is coordinated to iron in a bidentate fashion via O2' and O1. However, over the course of the simulation, as shown here, α-KG loses one of its coordination interactions to become a monodentate ligand (via only O1). The sixth (equatorial) position is occupied by a water molecule. This is consistent with our previous QM/MM studies of the reaction mechanism of AlkB, which is used as a prototype to understand TET enzymes. To validate the appropriateness of the surrogate, test simulations were performed for WT with hmC using Mg(II) and Fe(II). Panels **(e)** and **(f)** show the distance of all the ligands in the first coordination shell of the metal, Mg(II) and Fe(II) respectively, for the duration of the trajectory. These results validate our point-charge force field used for modeling. Note that Water 2 comes into the active site and coordinates to the metal cation after α-KG becomes a monodentate ligand (after 7 ns and 2 ns in Mg(II) and Fe(II) simulations, respectively). The mean RMSDs for Wat2 in Mg(II) and Fe(II) simulations decrease to 2.01 ± 0.02 and 2.06 ± 0.01, respectively, after excluding the distances before coordinating to metal. Further analysis is provided in Supplementary Table 8. The values in parentheses are mean ± s.d. The numbers in square brackets are the third significant figure for values <0.005.



c

	RMSD-Protein (hmC)
WT	2.41 ± 0.20
T1372S	2.43
T1372C	2.40
T1372A	2.10 ± 0.37
T1372E	2.85 ± 0.18
T1372Q	2.30 ± 0.20
T1372N	2.29 ± 0.34
T1372D	2.45 ± 0.26
T1372V	2.79 ± 0.06
Y1902F	2.53 ± 0.26
T1372A/Y1902F	2.29 ± 0.23

Figure S4-8. Root mean square deviation (RMSD) analysis with respect to the crystal structure.

(a) RMSD plots for protein backbone in a representative simulation of WT TET2 with mC/hmC/fC/caC show stability across the 50 ns simulation. **(b)** RMSD plots for the hmC base (all atoms) in WT, T1372A, T1372E, and T1372A/Y1902F show small conformational changes for the cytosine base throughout the simulations. **(c)** RMSD values for protein backbone in WT and mutants with hmC-containing DNA are shown. The mean ± s.d in **(a-c)** are calculated based on the mean value from each replicate simulation. No errors are provided with T1372S and T1372C since those simulations were only performed once.

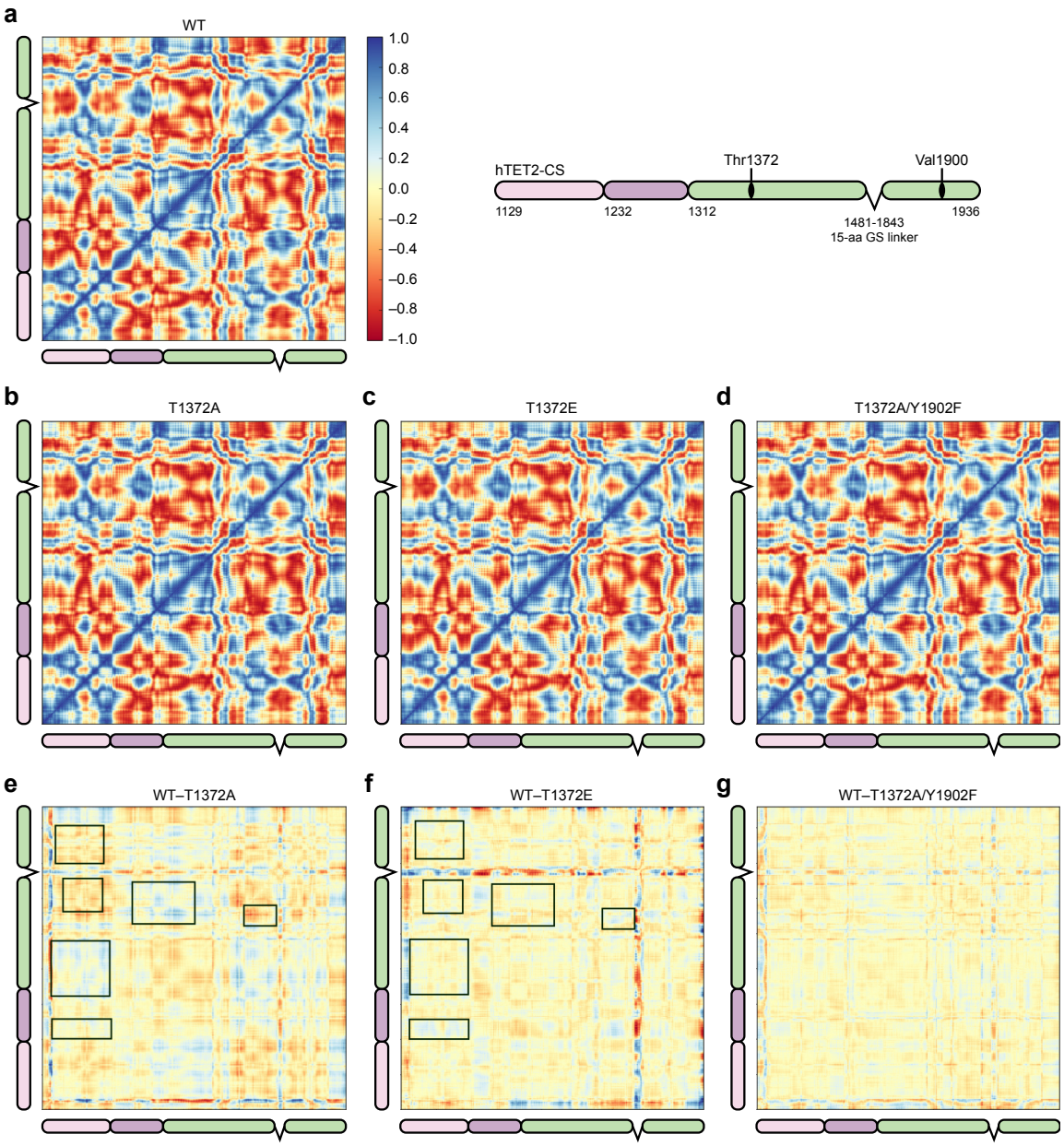


Figure S4-9. Correlation analysis for motions of all protein residues.

Correlation plots for **(a)** WT, **(b)** T1372A, **(c)** T1372E, and **(d)** T1372A/Y1902F. Correlation analysis by residue was carried out using the cpptraj module of Amber14, across the entire simulations. Residue pairs with correlated motions are shown in blue, while anti-correlated motions are shown in red. The correlation difference plots for **(e)** T1372A, **(f)** T1372E and **(g)** T1372A/Y1902F compare the mutant correlation plot to that of the WT and were calculated using an in-house python script. The range in difference plots was narrowed to –0.3 to 0.3 to highlight areas that appear different. Illustrative, regional changes in the single mutants are marked with boxes. For instance, residues 1425–1480 and 1400–1425 in T1372E are correlated but in T1372A are anti-correlated. The double mutant shows a pattern more consistent with the WT, suggesting that protein dynamics could be an added mechanism contributing to the differential reactivity of the variants.

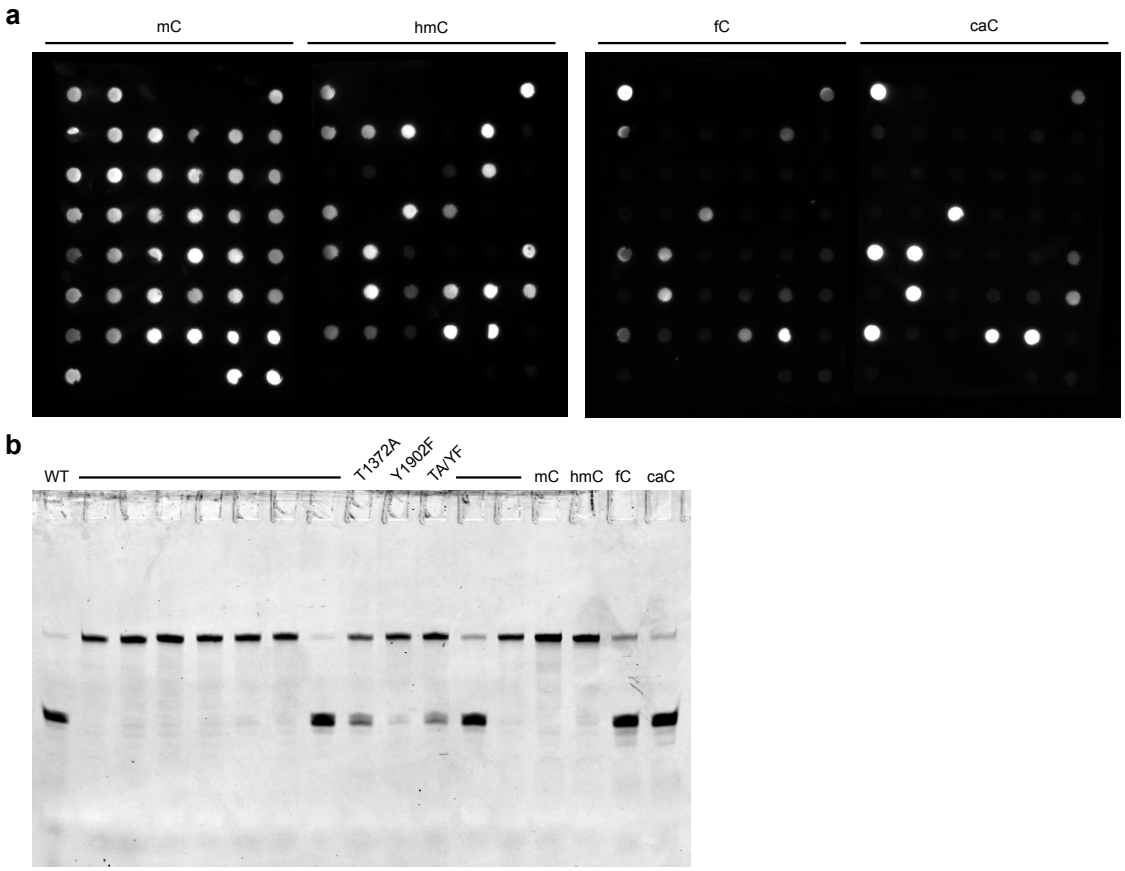


Figure S4-10. Uncropped versions of images used in the main text.

(a) Dot blots of mC and hmC (left) and fC and caC (right) used for Figure 4-2a (and Figure S4-1a). **(b)** For Figure 4-5c, 13 purified TET variants were reacted with oligonucleotides containing mC, and the reaction products were purified and treated with TDG to yield cleavage products at sites of fC and caC formation. These products were analyzed by DNA polyacrylamide gel electrophoresis. Relevant lanes are labeled; the four lanes at right are mC, hmC, fC, and caC oligo controls illustrating the specificity of TDG in this assay.

Name	Sequence (5'-3')
T1372A	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGGCCGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372A-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCGCCACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372C	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGTGCGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372C-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCGCACACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372D	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGGACGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372D-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCGTCCACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372E	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGGAGGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372E-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCCTCCACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372F	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGTTCGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372F-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCGAACACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372G	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGGGCGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372G-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCGCCCACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372H	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGCACGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372H-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCGTGCACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372I	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGATCGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372I-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCGATCACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372K	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGAAGGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372K-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCCTTCACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372L	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGCTGGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372L-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCCACGCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372M	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGATGGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372M-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCCATCACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372N	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGAACGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372N-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCGTTCACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372P	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGCCCGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372P-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCGGGCACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372Q	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGCAGGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372Q-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCCTGCACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372R	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGAGGGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372R-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCCCTCACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372S	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGAGCGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372S-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCGCTCACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372V	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGTGGTGGCCTGCCTGGACTTCTGCGCCCACGCACC
T1372V-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCCACCCCGCTGAAGGGTCGACCCTCCTTCAGGC

Table S4-1. DNA oligonucleotides used for cassette mutagenesis.

(Continued on next page)

T1372W	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGGTGTGGGCCTGCCTGGACTTCTGCGCCCACGCACACC
T1372W-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCCCACACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
T1372Y	CTAGGCCTGAAGGAGGGTCGACCCTTCAGCGGGGGTGTACGCCTGCCTGGACTTCTGCGCCCACGCACACC
T1372Y-r	GTGCGTGGGCGCAGAAGTCCAGGCAGGCGTACACCCCGCTGAAGGGTCGACCCTCCTTCAGGC
V1900A	CGCGTATAAGCTTGGCCTTCTACCAGCACACAAGAGCATGAACGAGGCCTAAACACGGG
V1900A-r	CTAGCCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAAGGCCAAGCTTATA
V1900C	CGCGTATAAGCTTGTGCTTCTACCAGCACAAGAGCATGAACGAGGCCTAAACACGGG
V1900C-r	CTAGCCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAAGCACACAAGCTTATA
V1900D	CGCGTATAAGCTTGGACTTCTACCAGCACACAAGAGCATGAACGAGGCCTAAACACGGG
V1900D-r	CTAGCCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAAGTCCAAGCTTATA
V1900E	CGCGTATAAGCTTGGAGTTCTACCAGCACAAGAGCATGAACGAGGCCTAAACACGGG
V1900E-r	CTAGCCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAACTCCCAAGCTTATA
V1900F	CGCGTATAAGCTTGTTCTTCTACCAGCACACAAGAGCATGAACGAGGCCTAAACACGGG
V1900F-r	CTAGCCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAAGAACAAGCTTATA
V1900G	CGCGTATAAGCTTGGGCTTCTACCAGCACACAAGAGCATGAACGAGGCCTAAACACGGG
V1900G-r	CTAGCCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAAGCCCCAAGCTTATA
V1900H	CGCGTATAAGCTTGCACTTCTACCAGCACACAAGAGCATGAACGAGGCCTAAACACGGG
V1900H-r	CTAGCCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAAGTGCCAAGCTTATA
V1900I	CGCGTATAAGCTTGATCTTCTACCAGCACACAAGAGCATGAACGAGGCCTAAACACGGG
V1900I-r	CTAGCCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAAGATCAAGCTTATA
V1900K	CGCGTATAAGCTTGAAGTTCTACCAGCACACAAGAGCATGAACGAGGCCTAAACACGGG
V1900K-r	CTAGCCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAACTTCAAGCTTATA
V1900L	CGCGTATAAGCTTGCTGTTCTACCAGCACACAAGAGCATGAACGAGGCCTAAACACGGG
V1900L-r	CTAGCCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAACAGCAAGCTTATA
V1900M	CGCGTATAAGCTTGATGTTCTACCAGCACACAAGAGCATGAACGAGGCCTAAACACGGG
V1900M-r	CTAGCCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAACATCAAGCTTATA
V1900N	CGCGTATAAGCTTGAACTTCTACCAGCACACAAGAGCATGAACGAGGCCTAAACACGGG
V1900N-r	CTAGCCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAAGTTCAAGCTTATA
V1900P	CGCGTATAAGCTTGCCCTTCTACCAGCACACAAGAGCATGAACGAGGCCTAAACACGGG
V1900P-r	CTAGCCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAAGGGCAAGCTTATA
V1900Q	CGCGTATAAGCTTGCAGTTCTACCAGCACACAAGAGCATGAACGAGGCCTAAACACGGG
V1900Q-r	CTAGCCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAACTGCAAGCTTATA
V1900R	CGCGTATAAGCTTGAGGTTCTACCAGCACACAAGAGCATGAACGAGGCCTAAACACGGG
V1900R-r	CTAGCCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAACCTCAAGCTTATA
V1900S	CGCGTATAAGCTTGAGCTTCTACCAGCACACAAGAGCATGAACGAGGCCTAAACACGGG

(Table S4-1, continued)

V1900S-r	CTAGCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAAGCTCAAGCTTATA
V1900T	CGCGTATAAGCTTGACCTTCTACCAGCACACAAGAGCATGAACGAGCCTAAACACGGG
V1900T-r	CTAGCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAAGGTCAAGCTTATA
V1900W	CGCGTATAAGCTTGTGGTTCTACCAGCACACAAGAGCATGAACGAGCCTAAACACGGG
V1900W-r	CTAGCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAACCACACAAGCTTATA
V1900Y	CGCGTATAAGCTTGTACTTCTACCAGCACACAAGAGCATGAACGAGCCTAAACACGGG
V1900Y-r	CTAGCCCGTGTTTTAGGCTCGTTCATGCTCTTGTGCTGGTAGAAGTACAAGCTTATA
Y1902F	CGCGTATAAGCTTGGTGTTCTTCCAGCACACAAGAGCATGAACGAGCCTAAACACGGG
Y1902F-r	CTAGCCCGTGTTTAGGCTCGTTCATGCTCTTGTGCTGGAAGAACACCAAGCTTATA

(Table S4-1, continued)

	mC	hmC	fC	caC
WT	–106.58 ± 8.56 (n = 3)	–114.03 ± 10.77 (n = 5)	–101.56 ± 5.51 (n = 3)	–160.05 ± 29.12 (n = 3)
T1372S	–75.83	–83.92	–99.14	–191.59
T1372C	–86.09	–105.90	–88.32	–116.01
T1372A	–69.25	–110.20 ± 15.58 (n = 2)	–110.11	–134.67
T1372E	–92.54	–92.98 ± 1.75 (n = 2)	–68.71	–18.46
T1372Q	–89.49	–114.93 ± 38.25 (n = 2)	–119.55	–92.80
T1372N	–101.45	–113.70 ± 5.25 (n = 2)	–102.18	–134.69
T1372D	–74.64	–63.41 ± 51.68 (n = 2)	–71.82	–10.39
T1372V	–88.26	–76.21 ± 16.39 (n = 2)	–64.76	–61.85
Y1902F	–78.21	–133.45 ± 5.88 (n = 2)	–87.23	–123.85
T1372A/Y1902F	–90.50	–117.76 ± 33.45 (n = 3)	–89.39	–75.13

Table S4-2. Energy decomposition analysis (EDA) analysis for mC/hmC/fC/caC with all protein residues.

WT enzyme with all nucleobases and most mutants with hmC were simulated 2–5 times for 50 ns each; for these systems, the values shown are mean ± s.d. over all simulation runs, with the number of simulations specified in parentheses. For systems simulated one time for 50 ns, the single EDA results are shown. Values are kcal/mol, time averaged over the entire ensemble.

(a) mC	Coul Y1902-mC	vdw Y1902-mC	Total Y1902-mC	Coul X1372-mC	vdw X1372-mC	Total X1372-mC	Coul X1372-Y1902	vdw X1372-Y1902	Total X1372-Y1902
WT	−0.67 ± 0.62	−4.51 ± 0.95	−5.18 ± 0.67	−0.76 ± 0.49	−2.00 ± 0.25	−2.76 ± 0.74	−2.65 ± 0.59	−0.90 ± 0.17	−3.55 ± 0.43
T1372S	−0.09	−3.89	−3.98	−1.30	−1.72	−3.02	−3.60	−0.65	−4.25
T1372C	−0.50	−5.53	−6.03	−1.23	−1.22	−2.45	−0.20	−1.34	−1.54
T1372A	−6.40	−4.44	−10.84	−0.30	−1.20	−1.50	0.12	−1.09	−0.97
T1372E	−1.61	−6.22	−7.83	27.09	−1.60	25.49	1.77	−1.21	0.56
T1372Q	−1.97	−6.39	−8.36	−2.71	−2.21	−4.92	−0.31	−1.17	−1.48
T1372N	−0.31	−5.91	−6.22	−2.28	−2.19	−4.47	−0.42	−1.39	−1.81
T1372D	−0.12	−5.30	−5.42	24.01	−1.42	22.59	0.87	−1.25	−0.38
T1372V	0.04	−4.44	−4.40	−0.55	−2.26	−2.81	0.02	−0.96	−0.94
Y1902F	0.72	−4.77	−4.05	−0.30	−2.05	−2.35	−0.42	−1.08	−1.50
T1372A/Y1902F	−0.08	−1.83	−1.91	−0.46	−1.91	−2.37	0.2	−0.86	−0.66

Table S4-3. EDA analysis for key interactions.

EDA analysis between key residues and **(a)** mC, **(b)** hmC, **(c)** fC, and **(d)** caC for all systems. X1372 denotes the residue at the 1372 position. Energies of Coulombic (Coul) and van der Waals (vdw) interactions are given in kcal/mol. As noted in Supplementary Table 2, some systems were simulated 2–5 times for 50 ns each; for these systems, the values shown are mean ± s.d. over all simulation runs. The values in parentheses are for the model with Fe(II) parameters. The simulations with iron were performed two times, each time for 50 ns. (Continued on next page)

(b) hmC	Coul Y1902-hmC	vdw Y1902-hmC	Total Y1902-hmC	Coul X1372-hmC	vdw X1372-hmC	Total X1372-hmC	Coul X1372-Y1902	vdw X1372-Y1902	Total X1372-Y1902
WT	−0.97 ± 0.70	−5.13 ± 1.26	−6.10 ± 0.78	−1.39 ± 0.44	−2.2 ± 0.64	−3.59 ± 0.89	−2.35 ± 0.54	−1.02 ± 0.05	−3.37 ± 0.50
	(−0.91 ± 0.04)	(−5.07 ± 0.68)	(−5.98 ± 0.71)	(−0.28 ± 0.06)	(−1.81 ± 0.40)	(−2.09 ± 0.46)	(−1.90 ± 0.27)	(−1.13 ± 0.11)	(−3.03 ± 0.16)
T1372S	−1.16	−5.43	−6.59	−1.02	−1.87	−2.89	−2.13	−0.89	−3.02
	(−1.10 ± 0.80)	(−5.08 ± 0.12)	(−6.18 ± 0.92)	(−0.79 ± 0.21)	(−1.67 ± 0.50)	(−2.46 ± 0.29)	(−1.90 ± 1.08)	(−0.93 ± 0.13)	(−2.83 ± 0.95)
T1372C	−0.60	−5.77	−6.37	−1.85	−2.09	−3.94	−0.24	−1.35	−1.59
	(−0.48 ± 0.18)	(−5.75 ± 0.17)	(−6.23 ± 0.35)	(−1.82 ± 0.06)	(−2.19 ± 0.04)	(−4.00 ± 0.10)	(−1.13 ± 1.41)	(−1.20 ± 0.11)	(−2.33 ± 1.30)
T1372A	−0.86 ± 1.36	−4.8 ± 0.81	−5.69 ± 0.55	−0.51 ± 0.47	−0.89 ± 0.30	−1.40 ± 0.76	0.13 ± 0.08	−1.04 ± 0.07	−0.91 ± 0.16
	(−2.08 ± 0.04)	(−4.81 ± 2.26)	(−6.89 ± 2.31)	(−0.30 ± 0.25)	(−0.78 ± 0.31)	(−1.08 ± 0.06)	(0.04 ± 0.14)	(−1.05 ± 0.06)	(−0.01 ± 0.08)
T1372E	−1.98 ± 0.33	−4.78 ± 1.43	−6.76* ± 1.75	19.16 ± 0.03	−0.60 ± 0.23	18.56 ± 0.20	1.69 ± 0.83	−1.37 ± 0.11	0.32 ± 0.72
	(−1.86 ± 0.73)	(−5.46 ± 1.06)	(−7.32 ± 1.77)	(24.06 ± 5.42)	(−1.39 ± 0.7)	(22.67 ± 4.71)	(0.22 ± 2.48)	(−1.41 ± 0.03)	(−1.19 ± 2.51)
T1372Q	−1.26 ± 0.35	−4.43 ± 1.90	−5.69 ± 1.56	−4.34 ± 1.04	−1.69 ± 1.33	−6.03 ± 2.37	0.02 ± 0.24	−1.41 ± 0.14	−1.39 ± 0.38
	(−2.02 ± 0.12)	(−5.83 ± 1.40)	(−7.85 ± 1.52)	(−7.76 ± 2.10)	(−2.08 ± 0.14)	(−9.84 ± 1.96)	(0.05 ± 0.15)	(−1.54 ± 0.04)	(−1.49 ± 0.19)
T1372N	−1.15 ± 1.62	−3.95 ± 1.03	−5.10 ± 0.59	−5.02 ± 3.01	−2.32 ± 1.13	−7.34 ± 4.14	0.50 ± 0.35	−1.27 ± 0.02	−1.77 ± 0.33
	(0.05 ± 0.08)	(−4.79 ± 0.01)	(−4.74 ± 0.08)	(−8.58 ± 0.28)	(−2.78 ± 0.03)	(−11.36 ± 0.25)	(−0.91 ± 0.04)	(−1.25 ± 0.01)	(−2.16 ± 0.04)
T1372D	−0.74 ± 2.75	−5.51 ± 1.11	−6.25* ± 3.86	24.38 ± 1.70	−1.60 ± 0.88	22.78 ± 0.82	−0.30 ± 3.93	−1.06 ± 0.04	−1.36 ± 3.89
	(−0.60 ± 0.64)	(−5.29 ± 0.75)	(−5.89 ± 0.11)	(26.46 ± 7.95)	(−1.97 ± 0.76)	(24.49 ± 7.20)	(2.26 ± 0.36)	(−1.23 ± 0.04)	(1.03 ± 0.16)
T1372V	−0.69 ± 0.62	−5.06 ± 0.35	−5.74 ± 0.26	−0.22 ± 0.11	−2.08 ± 0.18	−2.30 ± 0.29	0.03 ± 0.04	−1.05 ± 0.10	−1.02 ± 0.08
	(−1.11 ± 0.42)	(−3.06 ± 2.02)	(−4.17 ± 2.43)	(−0.64 ± 0.45)	(−1.06 ± 0.57)	(−1.70 ± 0.11)	(0.05 ± 0.04)	(−1.11 ± 0.23)	(−1.06 ± 0.19)
Y1902F	−0.01 ± 0.95	−3.40 ± 0.90	−3.39 ± 0.06	−3.33 ± 2.54	−1.51 ± 0.45	−4.84 ± 2.09	−0.24 ± 0.32	−0.91 ± 0.04	−1.14 ± 0.28
	(−0.57 ± 0.01)	(−2.35 ± 0.18)	(−2.92 ± 0.19)	(−6.65 ± 0.7)	(−0.93 ± 0.10)	(−7.68 ± (0.6)	(−0.04 ± 0.06)	(−0.98 ± 0.05)	(−1.02 ± 0.11)
T1372A/Y1902F	0.06 ± 0.23	−2.16 ± 1.43	−2.10 ± 1.21	−0.77 ± 0.30	−1.08 ± 0.36	−1.85 ± 0.65	0.13 ± 0.08	−0.82 ± 0.05	−0.70 ± 0.11
	(0.08 ± 0.52)	(−2.26 ± 1.61)	(−2.18 ± 1.08)	(−0.71 ± 0.52)	(−1.34 ± 0.44)	(−2.05 ± 0.08)	(0.13 ± 0.08)	(−0.78 ± 0.04)	(−0.65 ± 0.04)

(Table S4-3, continued) *The altered orientation of hmC relative to Tyr1902 tends to destabilize this non-bonded interaction but in T1372E/D can serve as a stabilizing force even though the orientation is disrupted.

(c) fC	Coul Y1902-fC	vdw Y1902-fC	Total Y1902-fC	Coul X1372-fC	vdw X1372-fC	Total X1372-fC	Coul X1372-Y1902	vdw X1372-Y1902	Total X1372-Y1902
WT	−1.76 ± 0.11	−5.43 ± 0.54	−7.19 ± 0.59	−2.47 ± 0.33	−1.83 ± 0.21	−4.30 ± 0.17	−1.72 ± 0.29	−1.06 ± 0.14	−2.78 ± 0.26
T1372S	−1.68	−5.30	−6.98	−2.27	−1.87	−4.14	−2.00	−1.01	−3.01
T1372C	−0.48	−5.87	−6.35	−1.70	−1.99	−3.69	−0.12	−1.18	−1.30
T1372A	−1.66	−5.91	−7.57	−0.44	−0.74	−1.18	0.21	−1.19	−0.98
T1372E	−0.52	−5.08	−5.60	24.25	−1.13	23.12	2.99	−1.74	1.25
T1372Q	−1.06	−5.16	−6.22	−7.99	−2.19	−10.18	0.28	−1.74	−1.46
T1372N	−0.55	−5.60	−6.15	−6.15	−3.14	−9.29	−0.22	−1.04	−1.26
T1372D	0.22	−4.83	−4.61	22.16	−1.58	20.58	−9.28	−0.10	−9.38
T1372V	−1.82	−5.60	−7.42	−0.06	−2.22	−2.28	0.21	−1.16	−0.95
Y1902F	0.22	−4.16	−3.94	−1.60	−1.35	−2.95	−0.49	−0.93	−1.42
T1372A/Y1902F	0.66	−4.46	−3.80	−0.97	−1.40	−2.37	0.22	−0.96	−0.74

(d) caC	Coul Y1902-caC	vdw Y1902-caC	Total Y1902-caC	Coul X1372-caC	vdw X1372-caC	Total X1372-caC	Coul X1372-Y1902	vdw X1372-Y1902	Total X1372-Y1902
WT	−0.66 ± 0.75	−4.83 ± 1.32	−5.49 ± 0.57	−13.63 ± 6.22	0.48 ± 0.69	−13.15 ± 5.56	0.58 ± 1.10	−1.31 ± 0.19	−0.73 ± 0.91
T1372S	0.25	−3.33	−3.08	−10.67	−0.31	−10.98	−1.18	−0.97	−2.15
T1372C	0.44	−5.49	−5.05	−2.98	−2.24	−5.22	−0.49	−1.36	−1.85
T1372A	−0.20	−5.74	−5.94	−1.80	−1.64	−3.44	0.28	−1.02	−0.74
T1372E	−1.38	−4.80	−6.18	88.21	−2.81	85.4	−6.58	−0.88	−7.46
T1372Q	−0.31	−6.23	−6.54	−22.87	−0.55	−23.42	−22.87	−1.22	−24.09
T1372N	−1.58	−4.83	−6.41	−14.17	−1.18	−15.35	−0.02	−1.22	−1.24
T1372D	−0.71	−5.96	−6.67	88.21	−3.11	85.1	−13.79	0.23	−13.56
T1372V	−1.21	−4.21	−5.42	−1.87	−2.45	−4.32	0.1	−1.26	−1.16
Y1902F	−0.52	−5.86	−6.38	−14.04	−0.12	−14.16	−0.62	−0.83	−1.45
T1372A/Y1902F	−0.21	−3.65	−3.86	−1.81	−0.45	−2.26	0.12	−0.78	−0.66

(Table S4-3, continued)

(a) mC	X1372-Y1902
WT	77
T1372S	85
T1372C	-
T1372A	-
T1372E	-
T1372Q	-
T1372N	-
T1372D	-
T1372V	-
Y1902F	-
T1372A/Y1902F	-

(b) hmC	X1372-Y1902	hmC-X1372	hmC-other
WT	65 (57)	- (-)	16-D1384, 16-αKG (-)
T1372S	56 (48)	- (-)	- (-)
T1372C	- (-)	- (-)	- (54-D1384)
T1372A	- (-)	- (-)	17-αKG (-)
T1372E	- (-)	88 (64)	- (18-αKG)
T1372Q	- (-)	21 (56)	42-D1384 (-)
T1372N	- (-)	- (25)	22-S1286, 16-Y1902, 14-αKG (10-αKG, 10-R1261)
T1372D	- (-)	16 (10)	16-αKG (33-αKG)
T1372V	- (-)	- (-)	38-D1384 (11-D1384, 11-αKG)
Y1902F	- (-)	18 (63)	11-D1384 (-)
T1372A/Y1902F	- (-)	- (-)	57-R1261, 43-αKG, 15-T1259, 12-S1286, 11-S1284 (33-R1261, 17-αKG, 10-S1284)

Table S4-4. Major hydrogen bonding interactions observed in simulations.

Hydrogen bond analysis between key residues and **(a)** mC, **(b)** hmC, **(c)** fC, and **(d)** caC. X1372 denotes the residue at the 1372 position. Values are percentage of simulation time in which the hydrogen bond was observed. For systems simulated 2–5 times, the results from each simulation were averaged. The values in parentheses are for the model with Fe(II) parameters. The simulations with iron were performed two times, each time for 50 ns. Only bonds observed in >10% of simulation time are included. (Continued on next page)

(c) fC	X1372-Y1902	fC-X1372	fC-other
WT	55	-	-
T1372S	53	-	-
T1372C	-	-	-
T1372A	-	-	-
T1372E	-	-	-
T1372Q	-	-	-
T1372N	-	-	-
T1372D	59	-	-
T1372V	-	-	-
Y1902F	-	-	-
T1372A/Y1902F	-	-	-

(d) caC	X1372-Y1902	caC-X1372	caC-other
WT	14	78	81-R1261
T1372S	39	49	186-R1261*
T1372C	-	-	-
T1372A	-	-	-
T1372E	27	-	38-R1261
T1372Q	-	98	-
T1372N	-	57	21-R1261
T1372D	80	-	27-R1261
T1372V	-	-	12-R1261
Y1902F	-	87	-
T1372A/Y1902F	-	-	-

(Table S4-4, continued) *This value is the summation of hydrogen bonding between R1261 and caC, which both have two hydrogen bonding groups. Since the hydrogen bonding is simultaneous, the percentage is >100%.

	H-acceptor			H-donor	
	O2	N4[N3]	O5	O5-H5	N4-H41/42*
WT	H1904(HE2-NE2) 10 H1386(HD1-ND1) 22	[H1904(HE2-NE2) 12]	WAT(H-O)* 13	D1384(OD2) 5 D1384(OD1) 11	D1384(OD2) 40 N1387(OD1) 38
T1372S	H1904(HE2-NE2) 65	-	WAT(H-O)* 67	-	-
T1372A	H1386(HD1-ND1) 10	[H1904(HE2-NE2) 12]	WAT(H-O)* 42	α-KG(O)** 17 WAT(H-O)* 38	N1387(OD1) 36
T1372E	H1904(HE2-NE2) 35 H1386(HD1-ND1) 32	WAT(H-O)* 16	WAT(H-O)* 76	E1372(OE1/2)** 88	D1384(OD2) 21 N1387(OD1) 33
T1372Q	H1904(HE2-NE2) 27 H1386(HD1-ND1) 24	-	WAT(H-O)* 40	Q1372(OE1) 21 D1384(OD2) 42	N1387(OD1) 41
T1372N	H1904(HE2-NE2) 31 WAT(H-O)* 14	WAT(H-O)* 14	S1286(HG-OG) 21 WAT(H-O)* 42	Y1902(O) 16 α-KG(O)** 14	D1384(OD2) 13
T1372D	H1904(HE2-NE2) 47	-	WAT(H-O)* 34	D1372(OD1) 16 α-KG(O)** 16	WAT(H-O)* 25
T1372V	H1904(HE2-NE2) 53 H1386(HD1-ND1) 12	-	-	D1384(OD1) 38	D1384(OD2) 19 N1387(OD1) 14
Y1902F	H1386(HD1-ND1) 47	-	WAT(H-O)* 73	T1372(OG1) 18 D1384(OD2) 11	D1384(OD2) 23 N1387(OD1) 42
T1372A/Y1902F	-	WAT(H-O)* 27	R1261(HH11/12-NH1)* 40 R1261(HH21/22-NH2)* 17 S1284(HG-OG) 10	α-KG(O)** 43 T1259(OG1) 12	D1384(OD2) 40 α-KG(O)** 14

Table S4-5. Hydrogen bond analysis for hmC in WT and all mutants.

Atoms are labeled using PDB nomenclature. Values are percentage of simulation time in which the hydrogen bond is observed. For systems simulated 2–5 times, the results from each simulation were averaged. Only bonds observed in >10% of simulation time are included. *The hydrogen bonds percentage is the summation of hydrogen bonds over both hydrogens of the indicated atom. **The percentage of hydrogen bond for α-KG(O) is the summation of hydrogen bonds percentage over O1, O2, and O3 and for E1372(O) is over OE1 and OE2.

	H-acceptor			H-donor	
	O2	N4[N3]	O5	O5-H5	N4-H41/42*
WT-1	H1904(H2-N2) 1 H1386(H1-N1) 66	H1904(H2-N2) 4 [H1904(H2-N2) 4]	WAT(H-O)* 19	D1384(O2) 25	D1384(O2) 18 N1387(O1) 16
WT-2	H1904(H2-N2) 17 H1386(H1-N1) 2	[H1904(H2-N2) 12]	WAT(H-O)* 9	D1384(O1) 13	D1384(O2) 46 N1387(O1) 50
WT-3	H1904(H2-N2) 24	-	WAT(H-O)* 8	D1384(O1) 41	D1384(O2) 24 N1387(O1) 3
WT-4	H1904(H2-N2) 8 H1386(H1-N1) 13	[H1904(H2-N2) 20]	WAT(H-O)* 10	-	D1384(O2) 69 N1387(O1) 59
WT-5	H1904(H2-N2) 2 H1386(H1-N1) 27	[H1904(H2-N2) 26]	WAT(H-O)* 20	D1384(O2) 1	D1384(O2) 44 N1387(O1) 61
WT-avg	H1904(H2-N2) 10±10 H1386(H1-N1) 22±27	[H1904(H2-N2) 12±11]	WAT(H-O)* 13±6	D1384(O2) 5±11 D1384(O1) 11±18	D1384(O2) 40±20 N1387(O1) 38±27

Table S4-6. Comparison of hydrogen bond analysis for hmC in WT across simulations.

The hydrogen bonds are reported for five simulations on WT, each performed for 50 ns. Atoms are labeled using PDB nomenclature. Values are percentage of simulation time in which the hydrogen bond is observed. The mean ± s.d. across the five simulations is reported in the last row. *The hydrogen bonds percentage is the summation of hydrogen bonds over both hydrogens connected to heavy atom. Hydrogen bonds with water molecules are not included.

	H-acceptor			H-donor
	O2	N4[N3]	O5	N4-H41/42
mC	H1904(HE2-NE2) 48 S1290(HG-OG) 35	-	-	D1384(OD2) 10
fC	H1386(HD1-ND1) 17 H1904(HE2-NE2) 43	[H1904(HE2-NE2)* 10]	-	D1384(OD2) 26 N1387(OD1) 29
caC	H1386(HD1-ND1) 55 S1286(HG-OG) 13	[H1904(HE2-NE2)* 25]	R1261(HH11/12-NH1)* 81 T1372(HG-OG) 78	N1387(OD1) 63

Table S4-7. Hydrogen bond analysis for mC/fC/caC in WT.

Atoms are labeled using PDB nomenclature. Values are percentage of simulation time in which the hydrogen bond is observed, averaged over three simulations of 50 ns each. Only bonds observed in >10% of simulation time are included. *The hydrogen bonds percentage is the summation of hydrogen bonds over both hydrogens connected to heavy atom. Hydrogen bonds with water molecules are not included.

(a)		Coulomb Energy	Exchange Energy	Repulsion Energy	Polarization Energy	Total Energy
QM	Wat-Mn(II)	–49.16	–3.78	8.91	–22.1	–66.14
	Wat-Fe(II)	–48.93	–3.83	9.08	–24.3	–68.11
	Wat-Mg(II)	–48.15	–0.81	2.52	–22.26	–68.7
		Coulomb Energy	vdw Energy	Total Energy		
MM	Wat-Mn(II)	–49.11	2.7	–46.41		
	Wat-Fe(II)	–49.11	0.45	–48.66		
	Wat-Mg(II)	–49.11	–0.30	–49.41		

(b)	Total Y1902-hmC	Total T1372-hmC	Total T1372-Y1902	H-bond T1372-Y1902	H-bond hmC-T1372	H-bond hmC-other	Coordination #
Fe(II)	–5.98 ± 0.71	–2.09 ± 0.46	–3.03 ± 0.16	57	-	-	6
Mg(II)	–6.10 ± 0.78	–3.59 ± 0.89	–3.37 ± 0.50	65	-	16-D1384, 16-αKG	6

(c)	H-acceptor			H-donor		
	O2	N4[N3]	O5	O5-H5	N4-H1/42	
Fe(II)	H1904(HE2-NE2) 82	-	-	α-KG(O2) 11		
Mg(II)	H1904(HE2-NE2) 10 H1386(HD1-ND1) 22	[H1904(HE2-NE2) 12]	WAT(H-O) 13	D1384(OD1) 11 D1384(OD2) 5	D1384(OD2) 40 N1387(OD1) 38	

Table S4-8. Comparison of Mg(II) and Fe(II) modeling parameters.

(a) All simulations presented in the main text were carried out using established parameters for Mg(II) serving as an Fe(II) surrogate. This choice was made since there are no Fe(II) parameters defined by default in the ff99SB parameter set. Note that for these classes of force fields, all cations (divalent or otherwise) are represented as a charged sphere (van der Waals radii and positive charge); therefore, the only difference between cations in point-charge force fields is the size of the sphere. Nevertheless, before making this approximation, we performed quantum mechanical energy decomposition analysis calculations on heterodimers comprising Fe(II)/Mg(II)/Mn(II) ions and water (using LMOEDA in GAMESS) and MM (using force fields in Tinker) to make sure that Mg(II) parameters provide a correct description of the non-bonded interactions for our system. The parameters used for Fe(II) are from the Amber parameter database at the University of Manchester (http://research.bmh.manchester.ac.uk/bryce/amber). The energies are in kcal/mol. **(b)** EDA analysis between key residues and hmC for WT in presence of Fe(II) and Mg(II). Energies of Coulombic (Coul) and van der Waals (vdw) interactions are given in kcal/mol. **(c)** Hydrogen bond analysis for hmC in WT in presence of Fe(II) and Mg(II). Atoms are labeled using PDB nomenclature. Values are percentage of simulation time in which the hydrogen bond is observed.

BIBLIOGRAPHY

Abdel-Wahab, O., Mullally, A., Hedvat, C., Garcia-Manero, G., Patel, J., Wadleigh, M., Malinge, S., Yao, J., Kilpivaara, O., Bhat, R.*, et al.* (2009). Genetic characterization of TET1, TET2, and TET3 alterations in myeloid malignancies. Blood *114,* 144-147.

Amabile, A., Migliara, A., Capasso, P., Biffi, M., Cittaro, D., Naldini, L., and Lombardo, A. (2016). Inheritable Silencing of Endogenous Genes by Hit-and-Run Targeted Epigenetic Editing. Cell *167,* 219-232.e14.

Amouroux, R., Nashun, B., Shirane, K., Nakagawa, S., Hill, P.W., D'Souza, Z., Nakayama, M., Matsuda, M., Turp, A., Ndjetehe, E.*, et al.* (2016). De novo DNA methylation drives 5hmC accumulation in mouse zygotes. Nat. Cell Biol. *18,* 225-233.

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.

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.

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

Blaschke, K., Ebata, K.T., Karimi, M.M., Zepeda-Martinez, J.A., Goyal, P., Mahapatra, S., Tam, A., Laird, D.J., Hirst, M., Rao, A., Lorincz, M.C., and Ramalho-Santos, M. (2013). Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. Nature *500,* 222-226.

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., Marsico, G., Bachman, M., Beraldi, D., and Balasubramanian, S. (2014). Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. Nat. Chem. *6,* 435-440.

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

Bradbrook, G.M., Gleichmann, T., Harrop, S.J., Habash, J., Raftery, J., Kalb, J., Yariv, J., Hillier, I.H., and Helliwell, J.R. (1998). X-Ray and molecular dynamics studies of concanavalin-A

glucoside and mannoside complexes Relating structure to thermodynamics of binding. J. Chem. Soc. , Faraday Trans. *94,* 1603-1611.

Bransteitter, R., Pham, P., Scharff, M.D., and Goodman, M.F. (2003). Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. Proc. Natl. Acad. Sci. U. S. A. *100,* 4102-4107.

Bullard, W., Lopes da Rosa-Spiegler, J., Liu, S., Wang, Y., and Sabatini, R. (2014). Identification of the glucosyltransferase that converts hydroxymethyluracil to base J in the trypanosomatid genome. J. Biol. Chem. *289,* 20273-20282.

Carey, B.W., Finley, L.W., Cross, J.R., Allis, C.D., and Thompson, C.B. (2015). Intracellular alpha-ketoglutarate maintains the pluripotency of embryonic stem cells. Nature *518,* 413-416.

Case, D.A., Cheatham, T.E.,3rd, Darden, T., Gohlke, H., Luo, R., Merz, K.M.,Jr, Onufriev, A., Simmerling, C., Wang, B., and Woods, R.J. (2005). The Amber biomolecular simulation programs. J. Comput. Chem. *26,* 1668-1688.

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, J., Guo, L., Zhang, L., Wu, H., Yang, J., Liu, H., Wang, X., Hu, X., Gu, T., Zhou, Z.*, et al.* (2013a). Vitamin C modulates TET1 function during somatic cell reprogramming. Nat. Genet. *45,* 1504-1509.

Chen, J., Liu, J., Chen, Y., Yang, J., Chen, J., Liu, H., Zhao, X., Mo, K., Song, H., Guo, L.*, et al.* (2011). Rational optimization of reprogramming culture conditions for the generation of induced pluripotent stem cells with ultra-high efficiency and fast kinetics. Cell Res. *21,* 884-894.

Chen, Q., Chen, Y., Bian, C., Fujiki, R., and Yu, X. (2013b). TET2 promotes histone O-GlcNAcylation during gene transcription. Nature *493,* 561-564.

Chen, V.B., Arendall, W.B.,3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D Biol. Crystallogr. *66,* 12-21.

Cisneros, G.A., Perera, L., Schaaper, R.M., Pedersen, L.C., London, R.E., Pedersen, L.G., and Darden, T.A. (2009). Reaction mechanism of the epsilon subunit of E. coli DNA polymerase III: insights into active site metal coordination and catalytically significant residues. J. Am. Chem. Soc. *131,* 1550-1556.

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.

Contreras-Garcia, J., Johnson, E.R., Keinan, S., Chaudret, R., Piquemal, J.P., Beratan, D.N., and Yang, W. (2011). NCIPLOT: a program for plotting non-covalent interaction regions. J. Chem. Theory Comput. *7,* 625-632.

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.

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.

Cui, Q., and Karplus, M. (2003). Catalysis and specificity in enzymes: a study of triosephosphate isomerase and comparison with methyl glyoxal synthase. Adv. Protein Chem. *66,* 315-372.

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.

Dang, L., White, D.W., Gross, S., Bennett, B.D., Bittinger, M.A., Driggers, E.M., Fantin, V.R., Jang, H.G., Jin, S., Keenan, M.C.*, et al.* (2009). Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature *462,* 739-744.

Dawlaty, M.M., Breiling, A., Le, T., Raddatz, G., Barrasa, M.I., Cheng, A.W., Gao, Q., Powell, B.E., Li, Z., Xu, M.*, et al.* (2013). Combined deficiency of Tet1 and Tet2 causes epigenetic abnormalities but is compatible with postnatal development. Dev. Cell. *24,* 310-323.

Dawlaty, M.M., Ganz, K., Powell, B.E., Hu, Y.C., Markoulaki, S., Cheng, A.W., Gao, Q., Kim, J., Choi, S.W., Page, D.C., and Jaenisch, R. (2011). Tet1 is dispensable for maintaining pluripotency and its loss is compatible with embryonic and postnatal development. Cell. Stem Cell. *9,* 166-175.

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.

Delhommeau, F., Dupont, S., Della Valle, V., James, C., Trannoy, S., Masse, A., Kosmider, O., Le Couedic, J.P., Robert, F., Alberdi, A.*, et al.* (2009). Mutation in TET2 in myeloid cancers. N. Engl. J. Med. *360,* 2289-2301.

Deplus, R., Delatte, B., Schwinn, M.K., Defrance, M., Mendez, J., Murphy, N., Dawson, M.A., Volkmar, M., Putmans, P., Calonne, E.*, et al.* (2013). TET2 and TET3 regulate GlcNAcylation and H3K4 methylation through OGT and SET1/COMPASS. EMBO J. *32,* 645-655.

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.

Dolinsky, T.J., Czodrowski, P., Li, H., Nielsen, J.E., Jensen, J.H., Klebe, G., and Baker, N.A. (2007). PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. Nucleic Acids Res. *35,* W522-5.

Dolinsky, T.J., Nielsen, J.E., McCammon, J.A., and Baker, N.A. (2004). PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res. *32,* W665-7.

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.

Fang, D., and Cisneros, G.A. (2014). Alternative Pathway for the Reaction Catalyzed by DNA Dealkylase AlkB from Ab Initio QM/MM Calculations. J. Chem. Theory Comput. *10,* 5136-5148.

Fang, D., Lord, R.L., and Cisneros, G.A. (2013). Ab initio QM/MM calculations show an intersystem crossing in the hydrogen abstraction step in dealkylation catalyzed by AlkB. J Phys Chem B *117,* 6410-6420.

Figueroa, M.E., Abdel-Wahab, O., Lu, C., Ward, P.S., Patel, J., Shih, A., Li, Y., Bhagwat, N., Vasanthakumar, A., Fernandez, H.F.*, et al.* (2010). Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation. Cancer Cell *18,* 553-567.

Fitzgerald, M.E., and Drohat, A.C. (2008). Coordinating the initial steps of base excision repair. Apurinic/apyrimidinic endonuclease 1 actively stimulates thymine DNA glycosylase by disrupting the product complex. J. Biol. Chem. *283,* 32680-32690.

Franchini, D.M., Chan, C.F., Morgan, H., Incorvaia, E., Rangam, G., Dean, W., Santos, F., Reik, W., and Petersen-Mahrt, S.K. (2014). Processive DNA demethylation via DNA deaminase-induced lesion resolution. PLoS One *9,* e97754.

Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L., and Paul, C.L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc. Natl. Acad. Sci. U. S. A. *89,* 1827-1831.

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.

Gaidzik, V.I., Paschka, P., Spath, D., Habdank, M., Kohne, C.H., Germing, U., von Lilienfeld-Toal, M., Held, G., Horst, H.A., Haase, D., *et al.* (2012). TET2 mutations in acute myeloid leukemia (AML): results from a comprehensive genetic and clinical analysis of the AML study group. J. Clin. Oncol. *30,* 1350-1357.

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

Goll, M.G., and Bestor, T.H. (2005). Eukaryotic cytosine methyltransferases. Annu. Rev. Biochem. *74,* 481-514.

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.

Grin, I., and Ishchenko, A.A. (2016). An interplay of the base excision repair and mismatch repair pathways in active DNA demethylation. Nucleic Acids Res. *44,* 3713-3727.

Gu, T.P., Guo, F., Yang, H., Wu, H.P., Xu, G.F., Liu, W., Xie, Z.G., Shi, L., He, X., Jin, S.G., *et al.* (2011). The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. Nature *477,* 606-610.

Guo, F., Li, X., Liang, D., Li, T., Zhu, P., Guo, H., Wu, X., Wen, L., Gu, T.P., Hu, B., *et al.* (2014). Active and passive demethylation of male and female pronuclear DNA in the Mammalian zygote. Cell. Stem Cell. *15,* 447-458.

Guo, J.U., Su, Y., Zhong, C., Ming, G.L., and Song, H. (2011). Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. Cell *145,* 423-434.

Hajkova, P., Jeffries, S.J., Lee, C., Miller, N., Jackson, S.P., and Surani, M.A. (2010). Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. Science *329,* 78-82.

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. (2014a). 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. (2014b). Structure of a Naegleria Tet-like dioxygenase in complex with 5-methylcytosine DNA. Nature *506,* 391-395.

Hashimoto, H., Zhang, X., and Cheng, X. (2013). Selective Excision of 5-Carboxylcytosine by a Thymine DNA Glycosylase Mutant. J. Mol. Biol. *425,* 971-976.

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.

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

Hu, J., Xing, X., Xu, X., Wu, F., Guo, P., Yan, S., Xu, Z., Xu, J., Weng, X., and Zhou, X. (2013a). Selective chemical labelling of 5-formylcytosine in DNA by fluorescent dyes. Chemistry *19,* 5836-5840.

Hu, L., Li, Z., Cheng, J., Rao, Q., Gong, W., Liu, M., Shi, Y.G., Zhu, J., Wang, P., and Xu, Y. (2013b). 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.

Hu, X., Zhang, L., Mao, S.Q., Li, Z., Chen, J., Zhang, R.R., Wu, H.P., Gao, J., Guo, F., Liu, W.*, et al.* (2014). Tet and TDG mediate DNA demethylation essential for mesenchymal-to-epithelial transition in somatic cell reprogramming. Cell. Stem Cell. *14,* 512-522.

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., Chavez, L., Chang, X., Wang, X., Pastor, W.A., Kang, J., Zepeda-Martinez, J.A., Pape, U.J., Jacobsen, S.E., Peters, B., and Rao, A. (2014). Distinct roles of the methylcytosine oxidases Tet1 and Tet2 in mouse embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. *111,* 1361-1366.

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., and Rao, A. (2014). Connections between TET proteins and aberrant DNA modification in cancer. Trends Genet. *30,* 464-474.

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, 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.

Iurlaro, M., McInroy, G.R., Burgess, H.E., Dean, W., Raiber, E.A., Bachman, M., Beraldi, D., Balasubramanian, S., and Reik, W. (2016). In vivo genome-wide profiling reveals a tissue-specific role for 5-formylcytosine. Genome Biol. *17,* 141-016-1001-5.

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.

Janke, R., Dodson, A.E., and Rine, J. (2015). Metabolism and Epigenetics. Annu. Rev. Cell Dev. Biol. *31,* 473-496.

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, C., Lu, Y., Jelinek, J., Liang, S., Estecio, M.R., Barton, M.C., and Issa, J.P. (2014). TET1 is a maintenance DNA demethylase that prevents methylation spreading in differentiated cells. Nucleic Acids Res. *42,* 6956-6971.

Johnson, E.R., Keinan, S., Mori-Sanchez, P., Contreras-Garcia, J., Cohen, A.J., and Yang, W. (2010). Revealing noncovalent interactions. J. Am. Chem. Soc. *132,* 6498-6506.

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.

Josse, J., and Kornberg, A. (1962). Glucosylation of deoxyribonucleic acid. III. alpha- and beta-Glucosyl transferases from T4-infected Escherichia coli. J. Biol. Chem. *237,* 1968-1976.

Juan, D., Perner, J., Carrillo de Santa Pau, E., Marsili, S., Ochoa, D., Chung, H.R., Vingron, M., Rico, D., and Valencia, A. (2016). Epigenomic Co-localization and Co-evolution Reveal a Key Role for 5hmC as a Communication Hub in the Chromatin Network of ESCs. Cell. Rep. *14,* 1246-1257.

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

Kafer, G.R., Li, X., Horii, T., Suetake, I., Tajima, S., Hatada, I., and Carlton, P.M. (2016). 5-Hydroxymethylcytosine Marks Sites of DNA Damage and Promotes Genome Stability. Cell. Rep. *14,* 1283-1292.

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.

Khandelia, P., Yap, K., and Makeyev, E.V. (2011). Streamlined platform for short hairpin RNA interference and transgenesis in cultured mammalian cells. Proc. Natl. Acad. Sci. U. S. A. *108,* 12799-12804.

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

Klose, R.J., and Bird, A.P. (2006). Genomic DNA methylation: the mark and its mediators. Trends Biochem. Sci. *31,* 89-97.

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.

Ko, M., Huang, Y., Jankowska, A.M., Pape, U.J., Tahiliani, M., Bandukwala, H.S., An, J., Lamperti, E.D., Koh, K.P., Ganetzky, R.*, et al.* (2010). Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. Nature *468,* 839-843.

Kohli, R.M., Abrams, S.R., Gajula, K.S., Maul, R.W., Gearhart, P.J., and Stivers, J.T. (2009). A portable hotspot recognition loop transfers sequence preferences from APOBEC family members to activation-induced cytidine deaminase. J. Biol. Chem. *284,* 22898-22904.

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, TDG and the dynamics of DNA demethylation. Nature *502,* 472-479.

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

Langemeijer, S.M., Kuiper, R.P., Berends, M., Knops, R., Aslanyan, M.G., Massop, M., Stevens-Linders, E., van Hoogen, P., van Kessel, A.G., Raymakers, R.A.*, et al.* (2009). Acquired mutations in TET2 are common in myelodysplastic syndromes. Nat. Genet. *41,* 838-842.

Laukka, T., Mariani, C.J., Ihantola, T., Cao, J.Z., Hokkanen, J., Kaelin, W.G.,Jr, Godley, L.A., and Koivunen, P. (2016). Fumarate and Succinate Regulate Expression of Hypoxia-inducible Genes via TET Enzymes. J. Biol. Chem. *291,* 4256-4265.

Li, C., Lan, Y., Schwartz-Orbach, L., Korol, E., Tahiliani, M., Evans, T., and Goll, M.G. (2015a). Overlapping Requirements for Tet2 and Tet3 in Normal Development and Hematopoietic Stem Cell Emergence. Cell. Rep. *12,* 1133-1143.

Li, Z., Cai, X., Cai, C.L., Wang, J., Zhang, W., Petersen, B.E., Yang, F.C., and Xu, M. (2011). Deletion of Tet2 in mice leads to dysregulated hematopoietic stem cells and subsequent development of myeloid malignancies. Blood *118,* 4509-4518.

Li, Z., Gu, T.P., Weber, A.R., Shen, J.Z., Li, B.Z., Xie, Z.G., Yin, R., Guo, F., Liu, X., Tang, F.*, et al.* (2015b). Gadd45a promotes DNA demethylation through TDG. Nucleic Acids Res. *43,* 3986-3997.

Lian, C.G., Xu, Y., Ceol, C., Wu, F., Larson, A., Dresser, K., Xu, W., Tan, L., Hu, Y., Zhan, Q.*, et al.* (2012). Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma. Cell *150,* 1135-1146.

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

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

Liutkeviciute, Z., Kriukiene, E., Licyte, J., Rudyte, M., Urbanaviciute, G., and Klimasauskas, S. (2014). Direct decarboxylation of 5-carboxylcytosine by DNA C5-methyltransferases. J. Am. Chem. Soc. *136,* 5884-5887.

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.

Lorsbach, R.B., Moore, J., Mathew, S., Raimondi, S.C., Mukatira, S.T., and Downing, J.R. (2003). TET1, a member of a novel protein family, is fused to MLL in acute myeloid leukemia containing the t(10;11)(q22;q23). Leukemia *17,* 637-641.

Losman, J.A., and Kaelin, W.G.,Jr. (2013). What a difference a hydroxyl makes: mutant IDH, (R)-2-hydroxyglutarate, and cancer. Genes Dev. *27,* 836-852.

Love, D.C., and Hanover, J.A. (2005). The hexosamine signaling pathway: deciphering the "O-GlcNAc code". Sci. STKE *2005,* re13.

Lu, F., Liu, Y., Jiang, L., Yamaguchi, S., and Zhang, Y. (2014). Role of Tet proteins in enhancer activity and telomere elongation. Genes Dev. *28,* 2103-2119.

Lu, J., Hu, L., Cheng, J., Fang, D., Wang, C., Yu, K., Jiang, H., Cui, Q., Xu, Y., and Luo, C. (2016). A computational investigation on the substrate preference of ten-eleven-translocation 2 (TET2). Phys. Chem. Chem. Phys. *18,* 4728-4738.

Lu, X., Song, C.X., Szulwach, K., Wang, Z., Weidenbacher, P., Jin, P., and He, C. (2013). Chemical Modification-Assisted Bisulfite Sequencing (CAB-Seq) for 5-Carboxylcytosine Detection in DNA. J. Am. Chem. Soc. *135,* 9315-9317.

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.

MacKenzie, E.D., Selak, M.A., Tennant, D.A., Payne, L.J., Crosby, S., Frederiksen, C.M., Watson, D.G., and Gottlieb, E. (2007). Cell-permeating alpha-ketoglutarate derivatives alleviate pseudohypoxia in succinate dehydrogenase-deficient cells. Mol. Cell. Biol. *27,* 3282-3289.

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.

Mahfoudhi, E., Talhaoui, I., Cabagnols, X., Della Valle, V., Secardin, L., Rameau, P., Bernard, O.A., Ishchenko, A.A., Abbes, S., Vainchenker, W., Saparbaev, M., and Plo, I. (2016). TET2-mediated 5-hydroxymethylcytosine induces genetic instability and mutagenesis. DNA Repair (Amst) *43,* 78-88.

Maiti, A., and Drohat, A.C. (2011). Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine: Potential Implications for Active Demethylation of CpG Sites. J. Biol. Chem. *286,* 35334-35338.

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.

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.

Marti, S., Andres, J., Moliner, V., Silla, E., Tunon, I., and Bertran, J. (2003). Preorganization and reorganization as related factors in enzyme catalysis: the chorismate mutase case. Chemistry *9,* 984-991.

Mellen, M., Ayata, P., Dewell, S., Kriaucionis, S., and Heintz, N. (2012). MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. Cell *151,* 1417-1430.

Minor, E.A., Court, B.L., Young, J.I., and Wang, G. (2013). Ascorbate induces ten-eleven translocation (Tet) methylcytosine dioxygenase-mediated generation of 5-hydroxymethylcytosine. J. Biol. Chem. *288,* 13669-13674.

Morales-Ruiz, T., Ortega-Galisteo, A.P., Ponferrada-Marin, M.I., Martinez-Macias, M.I., Ariza, R.R., and Roldan-Arjona, T. (2006). DEMETER and REPRESSOR OF SILENCING 1 encode 5-methylcytosine DNA glycosylases. Proc. Natl. Acad. Sci. U. S. A. *103,* 6853-6858.

Moran-Crusio, K., Reavie, L., Shih, A., Abdel-Wahab, O., Ndiaye-Lobry, D., Lobry, C., Figueroa, M.E., Vasanthakumar, A., Patel, J., Zhao, X.*, et al.* (2011). Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. Cancer. Cell. *20,* 11-24.

Morgan, B., and Lahav, O. (2007). The effect of pH on the kinetics of spontaneous Fe(II) oxidation by O2 in aqueous solution--basic principles and a simple heuristic description. Chemosphere *68,* 2080-2084.

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.

Muller, U., Bauer, C., Siegl, M., Rottach, A., and Leonhardt, H. (2014). TET-mediated oxidation of methylcytosine causes TDG or NEIL glycosylase dependent gene reactivation. Nucleic Acids Res. *42,* 8592-8604.

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.

Nakagawa, M., Koyanagi, M., Tanabe, K., Takahashi, K., Ichisaka, T., Aoi, T., Okita, K., Mochiduki, Y., Takizawa, N., and Yamanaka, S. (2008). Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. Nat. Biotechnol. *26,* 101-106.

Neddermann, P., and Jiricny, J. (1993). The purification of a mismatch-specific thymine-DNA glycosylase from HeLa cells. J. Biol. Chem. *268,* 21218-21224.

Neri, F., Dettori, D., Incarnato, D., Krepelova, A., Rapelli, S., Maldotti, M., Parlato, C., Paliogiannis, P., and Oliviero, S. (2015a). TET1 is a tumour suppressor that inhibits colon cancer growth by derepressing inhibitors of the WNT pathway. Oncogene *34,* 4168-4176.

Neri, F., Incarnato, D., Krepelova, A., Rapelli, S., Anselmi, F., Parlato, C., Medana, C., Dal Bello, F., and Oliviero, S. (2015b). 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.

Oberdoerffer, S. (2012). A conserved role for intragenic DNA methylation in alternative pre-mRNA splicing. Transcription *3,* 106-109.

Oda, A., Yamaotsu, N., and Hirono, S. (2005). New AMBER force field parameters of heme iron for cytochrome P450s determined by quantum chemical calculations of simplified models. J. Comput. Chem. *26,* 818-826.

Olsson, M.H., Sondergaard, C.R., Rostkowski, M., and Jensen, J.H. (2011). PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. J. Chem. Theory Comput. *7,* 525-537.

Ono, R., Taki, T., Taketani, T., Taniwaki, M., Kobayashi, H., and Hayashi, Y. (2002). LCX, leukemia-associated protein with a CXXC domain, is fused to MLL in acute myeloid leukemia with trilineage dysplasia having t(10;11)(q22;q23). Cancer Res. *62,* 4075-4080.

Ottink, O.M., Nelissen, F.H., Derks, Y., Wijmenga, S.S., and Heus, H.A. (2010). Enzymatic stereospecific preparation of fluorescent S-adenosyl-L-methionine analogs. Anal. Biochem. *396,* 280-283.

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.

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.

Pei, J., Kim, B.H., and Grishin, N.V. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. *36,* 2295-2300.

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.

Pollard, P.J., Briere, J.J., Alam, N.A., Barwell, J., Barclay, E., Wortham, N.C., Hunt, T., Mitchell, M., Olpin, S., Moat, S.J., *et al.* (2005). Accumulation of Krebs cycle intermediates and over-expression of HIF1alpha in tumours which result from germline FH and SDH mutations. Hum. Mol. Genet. *14,* 2231-2239.

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

Raiber, E.A., Beraldi, D., Ficz, G., Burgess, H.E., Branco, M.R., Murat, P., Oxley, D., Booth, M.J., Reik, W., and Balasubramanian, S. (2012). Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase. Genome Biol. *13,* R69.

Raiber, E.A., Murat, P., Chirgadze, D.Y., Beraldi, D., Luisi, B.F., and Balasubramanian, S. (2015). 5-Formylcytosine alters the structure of the DNA double helix. Nat. Struct. Mol. Biol. *22,* 44-49.

Rangam, G., Schmitz, K.M., Cobb, A.J., and Petersen-Mahrt, S.K. (2012). AID enzymatic activity is inversely proportional to the size of cytosine C5 orbital cloud. PLoS One *7,* e43279.

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.

Rudenko, A., Dawlaty, M.M., Seo, J., Cheng, A.W., Meng, J., Le, T., Faull, K.F., Jaenisch, R., and Tsai, L.H. (2013). Tet1 is critical for neuronal activity-regulated gene expression and memory extinction. Neuron *79,* 1109-1122.

Schafmeister, C.E.A.F., Ross, W.S., and Romanovski, V. (1995). The leap module of AMBER.

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.

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

Scourzic, L., Mouly, E., and Bernard, O.A. (2015). TET proteins and the control of cytosine demethylation in cancer. Genome Med. *7,* 9-015-0134-6. eCollection 2015.

Senn, H.M., O'Hagan, D., and Thiel, W. (2005). Insight into enzymatic C-F bond formation from QM and QM/MM calculations. J. Am. Chem. Soc. *127,* 13643-13655.

Shen, L., Inoue, A., He, J., Liu, Y., Lu, F., and Zhang, Y. (2014). Tet3 and DNA replication mediate demethylation of both the maternal and paternal genomes in mouse zygotes. Cell. Stem Cell. *15,* 459-470.

Shen, L., Wu, H., Diep, D., Yamaguchi, S., D'Alessio, A.C., Fung, H.L., Zhang, K., and Zhang, Y. (2013). Genome-wide Analysis Reveals TET- and TDG-Dependent 5-Methylcytosine Oxidation Dynamics. Cell *153,* 692-706.

Shen, L., and Zhang, Y. (2012). Enzymatic analysis of Tet proteins: key enzymes in the metabolism of DNA methylation. Methods Enzymol. *512,* 93-105.

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., Clark, T.A., Lu, X.Y., Kislyuk, A., Dai, Q., Turner, S.W., He, C., and Korlach, J. (2011a). Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. Nat. Methods *9,* 75-77.

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, C.X., Szulwach, K.E., Dai, Q., Fu, Y., Mao, S.Q., Lin, L., Street, C., Li, Y., Poidevin, M., Wu, H.*, et al.* (2013). Genome-wide Profiling of 5-Formylcytosine Reveals Its Roles in Epigenetic Priming. Cell *153,* 678-691.

Song, C.X., Szulwach, K.E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.H., Zhang, W., Jian, X.*, et al.* (2011b). Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. Nat. Biotechnol. *29,* 68-72.

Song, C.X., Yi, C., and He, C. (2012). Mapping recently identified nucleotide variants in the genome and transcriptome. Nat. Biotechnol. *30,* 1107-1116.

Spivey, H.O., and Ovadi, J. (1999). Substrate channeling. Methods *19,* 306-321.

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.

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.

Sun, Z., Terragni, J., Borgaro, J.G., Liu, Y., Yu, L., Guan, S., Wang, H., Sun, D., Cheng, X., Zhu, Z., Pradhan, S., and Zheng, Y. (2013). High-resolution enzymatic mapping of genomic 5-hydroxymethylcytosine in mouse embryonic stem cells. Cell. Rep. *3,* 567-576.

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.

Szwagierczak, A., Brachmann, A., Schmidt, C.S., Bultmann, S., Leonhardt, H., and Spada, F. (2011). Characterization of PvuRts1I endonuclease as a tool to investigate genomic 5-hydroxymethylcytosine. Nucleic Acids Res. *39,* 5149-5156.

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.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126,* 663-676.

Tanaka, K., and Okamoto, A. (2007). Degradation of DNA by bisulfite treatment. Bioorg. Med. Chem. Lett. *17,* 1912-1915.

Tang, Y., Zheng, S.J., Qi, C.B., Feng, Y.Q., and Yuan, B.F. (2015). Sensitive and simultaneous determination of 5-methylcytosine and its oxidation products in genomic DNA by chemical derivatization coupled with liquid chromatography-tandem mass spectrometry analysis. Anal. Chem. *87,* 3445-3452.

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.

Thienpont, B., Steinbacher, J., Zhao, H., D'Anna, F., Kuchnio, A., Ploumakis, A., Ghesquiere, B., Van Dyck, L., Boeckx, B., Schoonjans, L.*, et al.* (2016). Tumour hypoxia causes DNA hypermethylation by reducing TET activity. Nature *537,* 63-68.

Tsuji, M., Matsunaga, H., Jinno, D., Tsukamoto, H., Suzuki, N., and Tomioka, Y. (2014). A validated quantitative liquid chromatography-tandem quadrupole mass spectrometry method for monitoring isotopologues to evaluate global modified cytosine ratios in genomic DNA. J. Chromatogr. *953-954,* 38-47.

Vella, P., Scelfo, A., Jammula, S., Chiacchiera, F., Williams, K., Cuomo, A., Roberto, A., Christensen, J., Bonaldi, T., Helin, K., and Pasini, D. (2013). Tet proteins connect the O-linked N-acetylglucosamine transferase Ogt to chromatin in embryonic stem cells. Mol. Cell *49,* 645-656.

Wada, Y., Ohta, Y., Xu, M., Tsutsumi, S., Minami, T., Inoue, K., Komura, D., Kitakami, J., Oshida, N., Papantonis, A.*, et al.* (2009). A wave of nascent transcription on activated human genes. Proc. Natl. Acad. Sci. U. S. A. *106,* 18357-18361.

Walker, O.S., Elsasser, S.J., Mahesh, M., Bachman, M., Balasubramanian, S., and Chin, J.W. (2016). Photoactivation of Mutant Isocitrate Dehydrogenase 2 Reveals Rapid Cancer-Associated Metabolic and Epigenetic Changes. J. Am. Chem. Soc. *138,* 718-721.

Wang, H., Guan, S., Quimby, A., Cohen-Karni, D., Pradhan, S., Wilson, G., Roberts, R.J., Zhu, Z., and Zheng, Y. (2011). Comparative characterization of the PvuRts1I family of restriction enzymes and their application in mapping genomic 5-hydroxymethylcytosine. Nucleic Acids Res. *39,* 9294-9305.

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.

Ward, P.S., Patel, J., Wise, D.R., Abdel-Wahab, O., Bennett, B.D., Coller, H.A., Cross, J.R., Fantin, V.R., Hedvat, C.V., Perl, A.E.*, et al.* (2010). The common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting alpha-ketoglutarate to 2-hydroxyglutarate. Cancer. Cell. *17,* 225-234.

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.

Wescoe, Z.L., Schreiber, J., and Akeson, M. (2014). Nanopores discriminate among five C5-cytosine variants in DNA. J. Am. Chem. Soc. *136,* 16582-16587.

Wiehle, L., Raddatz, G., Musch, T., Dawlaty, M.M., Jaenisch, R., Lyko, F., and Breiling, A. (2015). Tet1 and Tet2 Protect DNA Methylation Canyons against Hypermethylation. Mol. Cell. Biol. *36,* 452-461.

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.

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

Xiao, M., Yang, H., Xu, W., Ma, S., Lin, H., Zhu, H., Liu, L., Liu, Y., Yang, C., Xu, Y., *et al.* (2012). Inhibition of alpha-KG-dependent histone and DNA demethylases by fumarate and succinate that are accumulated in mutations of FH and SDH tumor suppressors. Genes Dev. *26,* 1326-1338.

Xu, L., Chen, Y.C., Chong, J., Fin, A., McCoy, L.S., Xu, J., Zhang, C., and Wang, D. (2014). Pyrene-based quantitative detection of the 5-formylcytosine loci symmetry in the CpG duplex content during TET-dependent demethylation. Angew. Chem. Int. Ed Engl. *53,* 11223-11227.

Xu, W., Yang, H., Liu, Y., Yang, Y., Wang, P., Kim, S., Ito, S., Yang, C., Wang, P., Xiao, M., *et al.* (2011). Oncometabolite 2-Hydroxyglutarate Is a Competitive Inhibitor of alpha-Ketoglutarate-Dependent Dioxygenases. Cancer Cell *19,* 17-30.

Yamaguchi, S., Hong, K., Liu, R., Shen, L., Inoue, A., Diep, D., Zhang, K., and Zhang, Y. (2012). Tet1 controls meiosis by regulating meiotic gene expression. Nature *492,* 443-447.

Yang, H., Liu, Y., Bai, F., Zhang, J.Y., Ma, S.H., Liu, J., Xu, Z.D., Zhu, H.G., Ling, Z.Q., Ye, D., Guan, K.L., and Xiong, Y. (2013). Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. Oncogene *32,* 663-669.

Yang, J., Guo, R., Wang, H., Ye, X., Zhou, Z., Dan, J., Wang, H., Gong, P., Deng, W., Yin, Y., *et al.* (2016). Tet Enzymes Regulate Telomere Maintenance and Chromosomal Stability of Mouse ESCs. Cell. Rep. *15,* 1809-1821.

Yin, R., Mao, S.Q., Zhao, B., Chong, Z., Yang, Y., Zhao, C., Zhang, D., Huang, H., Gao, J., Li, Z., *et al.* (2013). Ascorbic acid enhances Tet-mediated 5-methylcytosine oxidation and promotes DNA demethylation in mammals. J. Am. Chem. Soc. *135,* 10396-10403.

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.

Yu, M., Song, C.X., and He, C. (2015). Detection of mismatched 5-hydroxymethyluracil in DNA by selective chemical labeling. Methods *72,* 16-20.

Zauri, M., Berridge, G., Thezenas, M.L., Pugh, K.M., Goldin, R., Kessler, B.M., and Kriaucionis, S. (2015). CDA directs metabolism of epigenetic nucleosides revealing a therapeutic window in cancer. Nature *524,* 114-118.

Zhang, G., Huang, H., Liu, D., Cheng, Y., Liu, X., Zhang, W., Yin, R., Zhang, D., Zhang, P., Liu, J., *et al.* (2015). N6-methyladenine DNA modification in Drosophila. Cell *161,* 893-906.

Zhang, L., Chen, W., Iyer, L.M., Hu, J., Wang, G., Fu, Y., Yu, M., Dai, Q., Aravind, L., and He, C. (2014a). 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, Q., Liu, X., Gao, W., Li, P., Hou, J., Li, J., and Wong, J. (2014b). Differential regulation of the ten-eleven translocation (TET) family of dioxygenases by O-linked beta-N-acetylglucosamine transferase (OGT). J. Biol. Chem. *289,* 5986-5996.

Zhang, R.R., Cui, Q.Y., Murai, K., Lim, Y.C., Smith, Z.D., Jin, S., Ye, P., Rosa, L., Lee, Y.K., Wu, H.P.*, et al.* (2013). Tet1 regulates adult hippocampal neurogenesis and cognition. Cell. Stem Cell. *13,* 237-245.

Zhao, Z., Chen, L., Dawlaty, M.M., Pan, F., Weeks, O., Zhou, Y., Cao, Z., Shi, H., Wang, J., Lin, L.*, et al.* (2015). Combined Loss of Tet1 and Tet2 Promotes B Cell, but Not Myeloid Malignancies, in Mice. Cell. Rep. *13,* 1692-1704.

Zheng, G., Fu, Y., and He, C. (2014). Nucleic acid oxidation in DNA damage repair and epigenetics. Chem. Rev. *114,* 4602-4620.