Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by the MLL Partner TET1

Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by the MLL fusion partner, TET1

Mamta Tahiliani[1], Kian Peng Koh[1], Yinghua Shen[2], William A. Pastor[1], Hozefa Bandukwala[1], Yevgeny Brudno[2], Suneet Agarwal[3], Lakshminarayan M. Iyer[4], David R. Liu[2], L. Aravind[4], Anjana Rao[1]

[1]Department of Pathology, Harvard Medical School and Immune Disease Institute, 200 Longwood Avenue, Boston, Massachusetts 02115, USA.

[2]Department of Chemistry and Chemical Biology and the Howard Hughes Medical Institute, Harvard University, Cambridge, Massachusetts 02138, USA.

[3]Division of Pediatric Hematology/Oncology, Children's Hospital Boston and Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA

[4]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894

*Correspondence should be addressed to

Dr. Anjana Rao
Department of Pathology, Harvard Medical School and Immune Disease Institute
Rm 152, Warren Alpert Bldg, 200 Longwood Avenue, Boston MA 02115
Tel: 617-278-3260; FAX: 617-278-3280, email: arao@idi.harvard.edu

A computational analysis identified TET proteins as mammalian homologs of the trypanosome J-binding proteins, JBP1 and JBP2 that have been proposed to oxidize the 5-methyl group of thymine. We tested the hypothesis that TET proteins modify 5-methylcytosine in DNA. Here we show that TET1, previously characterized as a fusion partner of the MLL gene in acute myeloid leukemia, is a 2-oxoglutarate (2OG)- and Fe(II)-dependent oxygenase that catalyzes the conversion of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (hmC) in cultured cells and *in vitro*. Expression of TET1 in HEK293 cells led to a decrease in 5mC as judged by immunocytochemistry, with concomitant appearance of a novel nucleotide identified by mass spectrometry as hmC. The catalytic domain of TET1 converted 5mC to hmC *in vitro*, whereas a variant with substitutions in residues predicted to bind the catalytic Fe(II) did not generate hmC. hmC can be detected in the genome of mouse ES cells and both TET1 levels and hmC levels decline when ES cells are differentiated. RNA-interference-mediated depletion of TET1 results in a decrease in hmC levels in ES cells. Our results suggest that hmC is a normal constituent of mammalian DNA, and identify TET1 as an enzyme with a potential role in epigenetic regulation through modification of 5mC.

Introduction

5-methylcytosine (5mC) is a minor base in mammalian DNA. It constitutes ~1% of all DNA bases and is found almost exclusively as symmetrical methylation of the dinucleotide CpG (*1*). The majority of methylated CpG is found in repetitive DNA elements, suggesting that cytosine methylation evolved as a defense against transposons and other parasitic elements in DNA (*2, 3*). Indeed, loss of the DNA methyltransferase *Dnmt1* in mice results in reactivation of a class of retrotransposons in somatic cells (*4*). Methylated CpGs are mutagenic; the modified cytosine is prone to deamination to yield thymine, thus generating T:G mismatches which if unrepaired result in C to T transitions. This mechanism has been proposed to effect a slow but progressive inactivation of retrotransposons in the genome (*2*).

Methylation patterns in mammals are remarkably dynamic during early embryogenesis, when CpG methylation is essential for establishing X-inactivation and the asymmetric expression of imprinted genes (*5, 6*). In somatic cells, CpG islands in genes that are not expressed are often highly methylated. DNA methylation enables the recruitment of methyl-CpG binding proteins such as MeCP2 and Kaiso; these in turn recruit histone deacetylases and other chromatin-modifying complexes that create a repressed chromatin environment, ultimately resulting in chromatin compaction and gene silencing (*7, 8*). In addition, CpG methylation may directly interfere with the binding of certain transcriptional regulators to their cognate DNA sequences (*9*), as best shown for the insulator-binding protein CTCF which controls the reciprocal expression of the imprinted genes *Igf2* and *H19* on paternal and maternal chromosomes respectively (*10*). DNA methylation patterns are highly dysregulated in cancer, and changes in methylation status have been postulated to inactivate tumor suppressors and activate oncogenes, thus contributing to tumorigenesis (*11*).

The enzymes that methylate DNA in mammals have been thoroughly characterized. Two de novo methyltransferases, DNMT3A and DNMT3B, symmetrically methylate unmethylated DNA (*12*), whereas the maintenance methyltransferase, DNMT1, maintains the established patterns of DNA methylation by methylating CpGs during DNA replication (*3, 10*). Preventing maintenance methylation through successive replication cycles progressively dilutes the methyl mark and results in “passive” DNA demethylation, a process well-documented to occur in differentiating cells (*13*). An “active” enzymatic mechanism of DNA demethylation has also been postulated; the most convincing

example involves the genome-wide demethylation of the paternal genome that occurs shortly after fertilization, independently of DNA replication (*14-16*). Potential mechanisms for active demethylation include (i) a thermodynamically unfavorable cleavage of the carbon-carbon bond linking the methyl group to the pyrimidine, resulting in release of the methyl moiety; and (ii) a repair-like process in which the methylated base or nucleotide is excised and the lesion is then repaired to replace the original 5mC with an unmethylated C (reviewed in (*5, 17*)). This latter mechanism occurs in plants, where DEMETER, a bifunctional glycosylase restricted to flowering plants, cleaves the glycosidic bond of 5mC and stimulates insertion of an unmethylated C through base-excision repair (*18*). A variety of candidate mammalian demethylase activities have been described, but unfortunately, none of these reports have been confirmed by other laboratories (reviewed in (*19-21*)).

In trypanosomes, a potential analog of 5-methylcytosine is a modified version of thymine called base J (β-D-glucosyl hydroxymethyluracil), which appears to be derived by hydroxylation of thymine followed by glucosylation (*22*). Base J is thought to play an important role in silencing gene transcription. It is present in silenced copies of the genes encoding the variable surface glycoprotein (VSG) responsible for antigenic variation in the host, but is absent from the single expressed copy (*23, 24*). A J-binding protein, JBP1, and a homolog, JBP2, were identified biochemically and through homology searches respectively (*22, 24*). JBP1 and JBP2 are enzymes of the 2OG and Fe(II)-dependent oxygenase superfamily (*22*), and mutations of the predicted Fe(II)- and 2OG-coordinating residues in JBP1 and JBP2 lead to decreased levels of base J in trypanosome and *Leishmania* DNA. Based on these data, the authors proposed that JBP1 and JBP2 oxidize the 5-methyl group of thymine in the first step of J biosynthesis (*22, 25, 26*)

We speculated that homologs of JBP1 and JBP2 might be generally involved in modification of 5-methylpyrimidines, including thymine and 5-methylcytosine. Computational analysis identified the TET family of proteins (TET1, TET2 and TET3) as likely homologs with predicted 2OG and Fe(II)-dependent oxygenase activity. Here we show that TET1 catalyzes the conversion of 5mC to hmC both in cells and *in vitro*. Moreover, we demonstrate that hmC is present in the genome of undifferentiated ES cells, but is not detectable in genomic DNA of two differentiated cell types, previously-activated human T cells and mouse dendritic cells. TET1 and hmC levels diminish in parallel when ES cell differentiation is induced by withdrawal of the cytokine LIF, and RNAi-mediated

depletion of TET1 in ES cells results in decreased hmC. These studies define a new class of enzymes that catalyze a new modification of DNA, and modify our perception of how DNA methylation status may be regulated in mammalian cells.

RESULTS

Identification of a novel family of 2OG-Fe(II) oxygenases with predicted 5-methylpyrimidine oxidase activity

To identify candidate enzymes that catalyze the oxidative modification of 5-methylcytosine, we created a position-specific score matrix (profile) of known 2OG-Fe(II) oxygenases that included the predicted oxygenase domains of JBP1 and JBP2 (Suppl. Fig. S1). We used this profile to conduct a systematic search of the non-redundant database using the PSI-BLAST program (*27*). This search recovered homologous domains in the gp2 proteins from the mycobacteriophages Cooper and Nigel and a related prophage from the *Frankia alni* genome ($e<10^{-4}$). We then generated a new profile which included the gp2 protein sequences, and performed a second search of the protein sequence database of microbes from environmental samples. This search detected numerous homologous proteins potentially derived from uncultured marine phages and prophages. A further search of the non-redundant database, with the newly-detected proteins from the environmental sequences added to the profile, recovered homologous regions in the three paralogous human oncogenes TET1 (CXXC6), TET2 and TET3 and their orthologs found throughout metazoa ($e<10^{-5}$). Homologous domains were also found in fungi and algae. In PSI-BLAST searches, these groups of homologous domains consistently recovered each other prior to recovering any other member of the 2OG-Fe(II) oxygenase superfamily, suggesting that they constitute a distinct family (LM Iyer, L Aravind, *manuscript in preparation*).

To confirm the relationship of the newly-identified proteins (hereafter referred to as the TET/JBP family; see legend to Fig. 1) with classical 2OG-Fe(II) oxygenases, we prepared a multiple alignment of their shared conserved domains and used this to generate a hidden markov model (HMM). A profile-profile comparison of this HMM against a library of HMM's generated for all structurally-characterized domains from the Protein Database (PDB) with the HHpred program (*28*) resulted in recovery of prolyl hydroxylase ($e<10^{-12}$), a canonical member of the 2OG-Fe(II) oxygenase superfamily, as the best hit. Secondary

structure predictions suggested the existence of an N-terminal α-helix followed by a continuous series of β-strands, typical of the double-stranded β-helix (DSBH) fold of the 2OG-Fe(II) oxygenases (*29*) (Fig. 1A, *bottom*). A multiple sequence alignment (Fig. 1A) further showed that the new TET/JBP family displayed all of the typical features of 2OG-Fe(II) oxygenases including: (i) the HxD signature, just downstream of an N terminal β-strand which chelates Fe(II) (x is any amino acid); (ii) a small residue, usually glycine, at the beginning of the strand immediately downstream of the HxD motif, which helps in positioning the active site arginine; (iii) the Hxs motif (where s is a small residue) in the C-terminal part, in which the H chelates the Fe(II) and the small residue helps in binding the 2-oxo acid; (iv) the Rx_5a signature (where a is an aromatic residue) downstream of the above motif – the R in this motif forms a salt bridge with the 2-oxo acid and the aromatic residue helps position the first metal-chelating histidine (the key residues are marked with asterisks in Fig 1A). These observations strongly suggested that members of the TET/JBP family, including TET1, 2 and 3, are catalytically-active 2OG-Fe(II) oxygenases.

Additionally, the metazoan TET proteins contain a unique conserved cysteine-rich region, contiguous with the N-terminus of the DSBH region, which possesses at least eight conserved cysteines and one histidine that are likely to comprise a binuclear metal cluster (Fig. 1A, *top*). Vertebrate TET1 and TET3, and their orthologs from all otheranimals, also possess a CXXC domain, a binuclear Zn-chelating domain with eight conserved cysteines and one histidine, located N-terminal to the 2OG-Fe(II) oxygenase domain (shown schematically for TET1 in Fig. 1B). Analysis of the architectures of CXXC domain proteins suggests that the CXXC domain is an accessory DNA-binding domain, that tends to be combined in the same polypeptide with a variety of domains that possess diverse chromatin-modifying and modification recognition activities. The CXXC domain is found in several chromatin-associated proteins, including the methyl-DNA-binding protein MBD1, the histone methyltransferase MLL, the DNA methyltransferase DNMT1 (*30*) and the lysine demethylases KDM2A (JHDM1A/ FBXL11) and KDM2B (JHDM1B/FBLX10, which also contains a ubiquitin E3 ligase domain) (*31*), and in certain cases has been shown to discriminate between methylated and unmethylated DNA (*32*).

Taken together, the contextual information gleaned from these domain architectures and from gene neighborhoods in the bacteriophage members (Supplementary Information and LM Iyer, L Aravind, *manuscript in preparation*) supported a conserved DNA modification function for the entire TET/ JBP family, namely oxidation of 5-methyl-pyrimidines. In this

study we tested the specific hypothesis that TET proteins might operate on 5mC to catalyze oxidation or oxidative removal of the methyl group, focusing on TET1 as a mammalian example of the TET/JBP family.

Cells overexpressing TET1 show decreased staining for 5-methylcytosine

To determine whether increasing the amount of TET1 expression would alter the global levels of 5mC in cells, we expressed HA-tagged full-length TET1 in HEK293 cells. Two days after transfection, cells were fixed, doubly stained with antibodies specific for 5mC and for the HA epitope, and processed for immunocytochemistry (Fig. 2A). To quantify the relation between HA (TET1) and 5mC staining at a single-cell level in each cell in a given field, the intensities of HA and 5mC staining were measured in each pixel within an area defined as the cell nucleus by DAPI staining (CellProfiler^{TM} image analysis software, Suppl. Fig. S2). The mean pixel intensity for each cell analyzed is represented by a dot in Fig. 2B.

Mock-transfected cells showed substantial variation in 5mC staining intensity (Fig. 2A, *top panel*), either because 5mC levels vary from cell to cell or because the accessibility of 5mC to the antibody differs among cells due to technical considerations (e.g. incomplete denaturation of DNA). This spread in staining intensity is obvious in Fig. 2B, in which mock-transfected cells are represented by blue dots and enclosed in a dotted oval. In cells transfected with TET1, there was a strong visual correlation of HA positivity with decreased staining for 5mC (Fig. 2A, *middle panel*). The population of HA-low cells, presumed to be untransfected, showed a spread of 5mC staining intensity similar to that of the mock-transfected population (Fig. 2B, *left panel*, note overlapping red and blue dots at low HA intensities), whereas the productively transfected (HA-high) cells showed a clear decrease in 5mC staining intensity (Fig. 2B, *left panel*, note uniformly low 5mC staining intensity in the red HA-high dots).

In addition to wild-type TET1, we tested the effect of a mutant TET1 bearing H1671Y, D1673A substitutions predicted to impair Fe(II) binding. Cells expressing this mutant protein did not show decreased staining for 5mC; rather, both visual and quantitative analysis indicated that cells expressing high levels of HA could also display high levels of 5mC staining (Fig. 2A, *bottom panel;* Fig. 2B, *right panel*; see quantification in Fig. 2C).

These results are consistent with the hypothesis that the H1671Y, D1673A mutations compromise a biological function of TET1.

An unidentified nucleotide derived from 5mC is present in HEK293 cells expressing TET1

To determine more quantitatively whether TET1 overexpression affects the intracellular levels of 5mC, we used an approach that allowed us to measure the ratio of 5mC to C at a subset of genomic CpG sites (those present in sequences recognized by the methylation-insensitive restriction enzyme MspI) (*33*). For these experiments we used full-length TET1 (TET1-FL) as well as the predicted catalytic domain of TET1 (TET1-CD, comprising the Cys-rich (C) and DSBH (D) regions; see Fig. 1B). HEK293 cells were transiently transfected with vectors in which expression of wild-type or mutant TET1 proteins was coupled to expression of human CD25, from an internal ribosome entry site (IRES). Control cells (mock) were transiently transfected with a corresponding empty vector that drove expression of CD25 alone. Forty-eight hours after transfection, CD25-expressing cells were purified using magnetic beads, and genomic DNA was isolated and digested with MspI. MspI cleaves DNA at the recognition site C^CGG regardless of whether or not the second C is methylated, producing cleavage products whose 5' ends derive from the dinucleotide CpG and contain either C or 5mC. The 5' phosphates were removed by treatment with calf intestinal phosphatase (CIP) and the fragments were end-labeled with T4 PNK and [γ-^{32}P] ATP, yielding a population of molecules labeled predominantly on C from the dinucleotide CG (the low-level labeling of other nucleotides reflects DNA shearing or contaminating endonucleolytic activity). The samples were digested with snake venom phosphodiesterase and DNase I to generate 5' dNMP's, which were separated by thin-layer chromatography (TLC) on cellulose plates with unlabeled 5'dNMP's included as size standards. The amount of radioactivity deriving from dCMP and 5-methyl-dCMP (5m-dCMP) was then measured by autoradiography and phosphorimager scanning. This method provides a quantitative estimate of the ratio of 5mC to C at all MspI recognition sites in genomic DNA (Fig. 3A).

MspI-digested DNA from cells transfected with the control vector yielded predominantly dCMP and 5m-dCMP as expected (Fig. 3B, lane 1), whereas DNA from cells expressing wild-type TET1-FL or TET1-CD yielded an additional unidentified labelled species (Rf ~0.29) migrating more slowly than dCMP (Fig. 3B, lanes 2 and 4; for solvent composition

see Methods). This new species was not detected in MspI-digested DNA from cells transfected with mutant versions of TET1 (Fig. 3B, lanes 3 and 5). The appearance of the unidentified species was associated with a concomitant decrease in the abundance of 5m-dCMP (Fig. 3B, lanes 2 and 4; quantified in Fig. 3C), suggesting that the unidentified species might be derived through modification of 5m-dCMP. TET1-CD-overexpressing cells had higher levels of the altered species in their genomic DNA than cells expressing TET1-FL (Fig. 3B, lanes 2 and 4); this is likely explained by the higher expression levels of the smaller TET1-CD protein relative to the full-length protein (data not shown). Notably, the unidentified species was not detected when DNA from cells expressing wild-type TET1 proteins was digested with HpaII, the methylation-sensitive isoschizomer of MspI (Fig. 3D). Since HpaII does not cut CCGG sequences if the second C is methylated, this result indicates that the unidentified species may be a derivative of 5mC. Moreover, CCGG sites containing the unidentified species are not cleaved by HpaII.

To confirm that the unidentified species was not an artifact of MspI digestion, we analyzed genomic levels of 5mC using Taq$^{\alpha}$I, a methylation-insensitive enzyme that cuts at T^CGA, and therefore, like MspI, produces DNA fragments whose 5' ends contain C or 5mC from the dinucleotide CpG. Again, the unidentified species was observed by TLC in Taq$^{\alpha}$I-digested DNA from cells expressing wildtype, but not mutant, TET1-CD (Fig. 3E), and the appearance of this species was associated with a corresponding decrease in abundance of 5m-dCMP (Fig. 3F). In all experiments, expression of wild-type TET1-CD correlated with a small but significant increase in the abundance of dCMP (Fig. 3C, F). Together these results demonstrate that expression of wild-type TET1 in HEK293 cells correlates with a decrease in the abundance of 5mC as assayed by immunocytochemistry and TLC, and with the concomitant appearance of a new nucleotide species as shown in a quantitative assay involving restriction enzyme digestion and TLC.

Cells overexpressing TET1 contain increased levels of hmC in their genome

Because many members of the 2OG and Fe(II)-dependent oxygenase superfamily are hydroxylases, we hypothesized that the unidentified species observed in TET1-expressing cells was hmC, and that TET1 hydroxylates the 5-methyl moiety of 5mC, converting it to hmC. These hypotheses are consistent with the slower migration of the unidentified species on cellulose TLC plates relative to 5m-dCMP and dCMP (Figs. 3B, E). Since hmC and its glucosylated derivatives are known to replace cytosine in the genome of T-

even phages (*34*) we used hm-dCMP from unglucosylated T4 phage DNA as an authentic source of hm-dCMP. Unglucosylated T4 phage (T4*) DNA was produced by growing T4 phage in *E.coli* ER1656, a strain deficient in the glucose donor molecule, UDP-glucose. Genomic DNA from T4* phage was digested with Taq$^{\alpha}$I, end-labeled, hydrolyzed and resolved using TLC. These experiments confirmed that the novel nucleotide generated in cells expressing TET1-CD migrated in a TLC assay similarly to authentic hm-dCMP from unglucosylated T4 DNA (Fig. 4A).

We used a mass spectrometry-based approach to identify the novel nucleotide. Genomic DNA was prepared from unsorted HEK293 cells (Fig. 4A) overexpressing wildtype or mutant TET1-CD and digested to 5' dNMPs. Following semi-preparative-scale TLC, the unknown species migrating with an Rf value of ~0.29 was extracted from the TLC plates with water and the extracts were analyzed by high-resolution mass spectrometry (MS). Although several species showed a ~2-fold difference in their abundance in DNA from cells expressing wild-type compared to mutant TET1-CD, most likely because of variation in the amount of resin excised from the respective TLC plates, a singly-charged species with an observed m/z of 336.0582, consistent with a molecular formula of $C_{10}H_{15}NO_8P^-$, was the only species exhibiting a large (~19-fold) difference in abundance between the wild-type and mutant samples (Fig. 4B). To establish the identity of this compound, we performed a series of MS-MS fragmentation experiments at various collision energies (15, 25, 35 and 50V) in both positive and negative ion modes (Fig. 4C, Suppl. Fig. S3 and data not shown). The observed fragmentation pattern of the 336.0582 Da ion (Fig. 4C, *bottom panel*) was consistent with that of hm-dCMP, and was virtually identical to that of authentic hm-dCMP isolated from T4 phage and subjected to MS-MS fragmentation in parallel (Fig. 4C, *top panel*). Taken together with the TLC data, these results identify the new nucleotide species present in the TLC assay of DNA from TET1 expressing cells as hmC, implying that in cells expressing wild-type but not mutant TET1, a significant proportion of 5mC is oxidized to hmC.

TET1 hydroxylates 5mC in vitro

To determine if TET1 was directly responsible for hmC production in transfected cells, we expressed Flag-HA-tagged wildtype and mutant TET1-CD in Sf9 insect cells, and purified the recombinant protein to near homogeneity (Suppl. Fig. S4A). The ability of the purified proteins to catalyze 5mC to hmC conversion was tested by incubating them with double-

stranded DNA oligonucleotides containing a fully methylated Taq$^{\alpha}$I recognition site. Conversion of 5mC to hmC was monitored by TLC (Fig. 5A). Wild-type TET1-CD, but not the H1671Y, D1673A mutant, catalyzed robust conversion of 5mC to hmC (Fig 5A, compare *lanes 2 and 6*). We confirmed that TET1 uses Fe(II) and 2OG as cofactors by independently omitting each factor from the enzymatic reactions. TET1 displayed an absolute requirement for both cofactors (Fig 5A, compare *lane 2 with lanes 3 and 5*). Although some 2OG- and Fe(II)-dependent oxygenases show enhanced activity in the presence of ascorbate (*35*), we saw no decrease in TET1-CD enzymatic activity in the absence of ascorbate. This is likely due to the fact that DTT, which was included in the reaction buffer to counteract the strong tendency of TET1-CD to oxidize (Suppl. Fig S4A-C), is able to play a similar role to ascorbate in reducing inactive Fe(III) to Fe(II) thereby protecting the enzyme from oxidative inactivation (Fig. 5A, compare *lanes 2 and 4*) (*36, 37*). The results are quantified in Fig. 5B.

We used high-resolution mass spectrometry to demonstrate, as before, that a singly-charged species at m/z of 336.0582 was the only species at Rf ~0.29 that differed significantly (35-fold) in abundance when comparing substrates incubated with wild-type and mutant proteins (Fig. 5C). MS-MS experiments at various collision energies showed, as before, that the fragmentation pattern of the species produced by recombinant TET1-CD was identical to that of authentic hm-dCMP derived from unglucosylated T4 phage (Fig. 5D and data not shown). Under the conditions of our assay, recombinant TET1-CD hydroxylated 5mC not only in fully-methylated, but also in hemi-methylated oligonucleotide substrates (Fig. 5E).

Since TET1 was identified as a homolog of the putative thymine hydroxylases JBP1 and JBP2, we tested the ability of TET1-CD to hydroxylate thymine *in vitro*. Recombinant TET1-CD was incubated with double-stranded DNA oligonucleotides containing a SalI recognition site (G^TCGAC). Conversion of thymine to hmU was either very inefficient or did not occur (Suppl. Fig. S5), showing specificity of TET1 for 5mC and not for thymine.

Together these data demonstrate that TET1 is a member of the 2OG and Fe(II)-dependent oxygenase superfamily with the ability to oxidize 5mC to hmC in double-stranded DNA.

hmC is present in ES cell DNA and its abundance decreases upon differentiation or TET1 depletion

We next asked whether hmC was a normal constituent of mammalian DNA. Using the TLC assay, we compared hmC levels in the sequence CCGG (i.e. in MspI cleavage sites) in previously-activated human T cells, mouse dendritic cells, and ES cells. We observed a clear spot corresponding to labelled hmC only in ES cells (Fig. 6A), suggesting an association of hmC with the pluripotent rather than the differentiated state. Quantification of multiple experiments indicates that hmC and 5mC constitute ~4% and 55-60% respectively of all cytosine species in MspI cleavage sites (CCGG) in ES cells (Fig. 6A).

To examine this relation further, we asked whether TET1 and hmC levels were altered when ES cells were differentiated by removal of the cytokine LIF from culture media. TET1 mRNA levels declined by 80% in response to LIF withdrawal for 5 days, compared to the levels observed in undifferentiated ES cells (Fig. 6B). In parallel, hmC levels diminished from 4.4 to 2.6% of total C species (a decline of ~40% from control levels) (Fig. 6C). The discrepancy (~80% decline of TET1 mRNA levels compared to ~40% decline of hmC levels) might be due to the compensatory activity of other TET-family proteins.

Similarly, RNAi-mediated depletion of endogenous TET1 with two different siRNAs resulted in an 83-85% decrease in TET1 mRNA levels and a corresponding ~38% decrease in hmC levels (Fig. 6D, E). Again, the discrepancy is likely due to the presence of TET2 and TET3, which are both expressed in ES cells. These data strongly support the hypothesis that TET1, and potentially other TET family members, are responsible for hmC generation in ES cells under physiological conditions.

Discussion

We have identified TET1, a member of the TET/JBP family of enzymes, as a 2OG and Fe(II)-dependent enzyme capable of oxidizing 5mC to hmC both in cells and *in vitro*. Members of the TET subfamily (TET1, TET2 and TET3) are conserved in jawed vertebrates and are broadly expressed in many cell types (*30*). All three paralogs are likely to possess catalytic activity, as the residues predicted to be necessary for activity are fully conserved.

hmC is a normal constituent of mammalian DNA, at least in certain cell types. Of the limited number of cell types that we analyzed, only ES cells possessed readily detectable levels of hmC in MspI cleavage sites (CCGG). In contrast, hmC was almost undetectable in HEK293 cells and two differentiated primary cell types, dendritic cells and previously-activated T cells. Consistent with our finding that overexpressed and recombinant TET1 catalyzed the conversion of 5mC to hmC in HEK293 cells and *in vitro* respectively, RNAi-mediated depletion of TET1 in ES cells led to a perceptible loss of hmC.

Notably, ES cells induced to differentiate in response to LIF withdrawal showed parallel decreases in TET1 mRNA expression and hmC levels in genomic DNA, suggesting a possible relation between hmC and the pluripotent state of ES cells. To explore this hypothesis in more detail, it will be useful to ask whether hmC and TET proteins localize to specific regions of ES cell DNA – for instance, to genes that are involved in maintaining pluripotency or that are poised to be expressed upon differentiation. hmC may not be confined to CpGs: nearest-neighbour analyses have shown that whereas somatic tissues have negligible non-CpG methylation, 15-20% of total 5mC in ES cells is present in CpT, CpA and (to a minor extent) CpC sequences (*38*).

What are the biological functions and fate of hmC? Figure 7 summarizes some speculations. An attractive hypothesis is that hmC (which is a stable base in the genome of T-even phages (*34*) influences chromatin structure and local transcriptional activity by recruiting proteins that selectively bind to hmC (Fig. 7B). Conversely, the hmC modification might exclude methyl-CpG-binding proteins (MBPs) that normally recognize 5mC, thus displacing chromatin-modifying complexes that are recruited by MBPs (*7, 8*). In support of this latter scenario, it has already been demonstrated that the methyl-binding protein MeCP2 does not recognize hmC (*39*).

The stability of hmC contrasts with the well-documented instability of N-linked hydroxymethyl adducts generated by the DNA repair enzymes AlkB and the JmjC domain-containing histone demethylases, which are also members of the 2OG and Fe(II)-dependent oxygenase superfamily. These nitrogen-linked adducts spontaneously resolve, yielding the unmethylated amino group and formaldehyde (*36*); in contrast, oxidation of carbon-linked methyl groups by JBP1/2 and thymine hydroxylase does not result in removal of the methyl adduct via oxidation (*22, 25, 40*). This difference, which is due to the fact that carbon is a much poorer leaving group than nitrogen, explains our ability to detect hmC in genomic DNA. Nevertheless, hmC has been shown to convert to

cytosine through loss of formaldehyde in photo-oxidation experiments (*41*) and at high pH (*42, 43*), leaving open the possibility that hmC could convert to cytosine under certain conditions in cells (Fig. 7A).

Another possibility is that hmC is an intermediate in a pathway of passive (replication-dependent) DNA demethylation (Fig. 7A, B). The presence of hmC on one strand of the dinucleotide CpG has been shown to prevent DNMT1 from methylating its target C in the opposite strand (*44*). Thus the small but reproducible increase in the relative abundance of unmethylated cytosine at C^CGG and T^CGA sites, observed within two days of expressing TET1-CD in HEK293 cells (Fig. 3C, F), is consistent with a model in which demethylation of 5mC proceeds via hmC. Even a minor reduction in the fidelity of maintenance methylation would be expected to result in an exponential decrease in CpG methylation over the course of many cell cycles. It will be of interest to test if the SAD (SRA) domain protein UHRF1, the DNMT1 partner protein that is responsible for its high selectivity for hemimethylated CpGs (*45*), is also less capable of recognizing hemi-modified hmCpG.

Alternatively, the presence of hmC might be sensed by specific DNA repair mechanisms that replace hmC with C, thus leading to "active" demethylation of DNA (Fig. 7). In fact, bovine thymus extracts have been reported to contain a glycosylase activity specific for hmC (*46*), but this activity has not been further characterized. Moreover, several DNA glycosylases, including TDG and MBD4, have been implicated in DNA demethylation, although none of them has shown ~~a~~ convincing activity on 5mC using in vitro enzymatic assays (*47-50*). Finally, cytosine deamination has also been implicated in demethylation of DNA (*49-51*) . In this context, it is conceivable that deamination of hmC, which would result in its conversion to hmU, might be involved in the conversion of hmC into a substrate for repair enzymes. In support of this hypothesis, high levels of HmU:G glycosylase activity have been reported in fibroblast extracts (*52*). With the description of this new derivative of 5mC, it will be important to reexamine the specificity of known DNA glycosylases and deaminases for DNA containing hmC and its potential derivatives.

Why has hmC been overlooked as a normal constituent of mammalian DNA? With the exception of two papers that reported high levels of hmC in genomic DNA isolated using unconventional but not standard methods (*53*), most previous studies describe hmC as a rare base that is a probable oxidation product of 5mC (*44, 54*). One reason that hmC may have been missed is that it might be present at detectable levels only in specific cell types

(ES cells and not differentiated cells). Another factor may be the relatively low abundance of hmC. In ES cells, hmC is ~4% of all cytosine species in CpG dinucleotides located in MspI cleavage sites (CCGG) (Fig. 6C, E). CpG is ~0.8% of all dinucleotides in the mouse genome (*55*), thus hmC constitutes ~0.032% of all bases or ~1 in every 3000 nucleotides. For comparison, 5mC is 55-60% of all cytosines in CpG dinucleotides in MspI cleavage sites (Fig. 6A), about 14-fold higher than hmC. A more trivial explanation is that some TLC running buffers do not resolve hmC from C (*53*). We were fortunate that our experimental design led us to focus on areas of the genome containing CpG dinucleotides (and hence enriched for 5mC), and that our TLC running conditions could distinguish hmC.

A full appreciation of the biological significance of hmC will depend heavily on the development of tools that allow hmC, 5mC and C to be distinguished unequivocally. We show here that two of the three most commonly used techniques do not meet this criterion. A widely-used mouse monoclonal antibody to 5mC apparently does not recognize hmC by immunocytochemistry (Fig. 2A), thus it will be important to reevaluate previous reports of DNA demethylation based solely on the use of this antibody. Similarly, the methylation-sensitive restriction enzyme, HpaII, fails to cut hmC (Fig. 3D) as previously reported (*56*), raising the possibility that in some instances hmC-modified DNA was incorrectly judged to be methylated. Another methylation-sensitive restriction enzyme, McrBC, is already known to cleave 5mC- and hmC-containing DNA equivalently (*57*), and therefore also does not allow these two nucleotides to be distinguished. It is yet to be determined how bisulfite modification analysis interprets the presence of hmC in DNA. Treatment of DNA with sodium bisulfite promotes the spontaneous deamination of cytosine to uracil, while leaving 5mC unaffected; amplification of the sequence of interest followed by sequencing allows the precise methylation patterns at a given sequence to be determined (*58*). It is known that bisulfite reacts rapidly with hmC at the C5 to form a stable cytosine 5-methylenesulfonate adduct, which is not readily deaminated (*59*). This substituted species, which is expected to form base pairs similar to those formed by cytosine, could be read by polymerases as C during the amplification steps, resulting in the sequence being interpreted as containing 5mC. Alternatively, polymerases may not copy cytosine 5-methylenesulfonate efficiently, in which case the DNA containing this adduct would not be amplified effectively and the sequence containing the original hmC modification would be underrepresented in the amplified DNA.

Notably, disruptions of the *TET1* and *TET2* genetic loci have been reported in association with hematologic malignancies. A fusion of *TET1* with the histone methyltransferase, *MLL*, has been identified in at least two cases of acute myeloid leukemia (AML) associated with t(10;11)(q22;q23) translocation (*30, 60*). Homozygous null mutations and chromosomal deletions involving the *TET2* locus have been found in AML and myeloproliferative disorders, suggesting a tumor suppressor function for TET2 (*61, 62*). It will be interesting to test the involvement of TET proteins and hmC in oncogenic transformation and malignant progression.

References

1. M. Ehrlich, R. Y. Wang, *Science* **212**, 1350 (Jun 19, 1981).
2. J. A. Yoder, C. P. Walsh, T. H. Bestor, *Trends Genet* **13**, 335 (Aug, 1997).
3. M. G. Goll, T. H. Bestor, *Annu Rev Biochem* **74**, 481 (2005).
4. C. P. Walsh, J. R. Chaillet, T. H. Bestor, *Nat Genet* **20**, 116 (Oct, 1998).
5. H. D. Morgan, F. Santos, K. Green, W. Dean, W. Reik, *Hum Mol Genet* **14 Spec No 1**, R47 (Apr 15, 2005).
6. W. Reik, W. Dean, J. Walter, *Science* **293**, 1089 (Aug 10, 2001).
7. X. Nan *et al.*, *Nature* **393**, 386 (May 28, 1998).
8. H. G. Yoon, D. W. Chan, A. B. Reynolds, J. Qin, J. Wong, *Mol Cell* **12**, 723 (Sep, 2003).
9. J. D. Lewis *et al.*, *Cell* **69**, 905 (Jun 12, 1992).
10. A. Bird, *Genes Dev* **16**, 6 (Jan 1, 2002).
11. E. N. Gal-Yam, Y. Saito, G. Egger, P. A. Jones, *Annu Rev Med* **59**, 267 (2008).
12. M. Okano, S. Xie, E. Li, *Nat Genet* **19**, 219 (Jul, 1998).
13. D. U. Lee, S. Agarwal, A. Rao, *Immunity* **16**, 649 (May, 2002).
14. W. Mayer, A. Niveleau, J. Walter, R. Fundele, T. Haaf, *Nature* **403**, 501 (Feb 3, 2000).
15. J. Oswald *et al.*, *Curr Biol* **10**, 475 (Apr 20, 2000).

16. F. Santos, B. Hendrich, W. Reik, W. Dean, *Dev Biol* **241**, 172 (Jan 1, 2002).

17. H. Cedar, G. L. Verdine, *Nature* **397**, 568 (Feb 18, 1999).

18. M. Gehring *et al.*, *Cell* **124**, 495 (Feb 10, 2006).

19. A. P. Wolffe, P. L. Jones, P. A. Wade, *Proc Natl Acad Sci U S A* **96**, 5894 (May 25, 1999).

20. A. P. Bird, *Nature* **321**, 209 (May 15-21, 1986).

21. S. K. Ooi, T. H. Bestor, *Cell* **133**, 1145 (Jun 27, 2008).

22. Z. Yu *et al.*, *Nucleic Acids Res* **35**, 2107 (2007).

23. P. Borst, R. Sabatini, *Annu Rev Microbiol* **62**, 235 (2008).

24. C. DiPaolo, R. Kieft, M. Cross, R. Sabatini, *Mol Cell* **17**, 441 (Feb 4, 2005).

25. S. Vainio, P. A. Genest, B. Ter Riet, H. van Luenen, P. Borst, *Mol Biochem Parasitol* **164**, 157 (Apr, 2009).

26. L. J. Cliffe *et al.*, *Nucleic Acids Res* (Jan 9, 2009).

27. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389 (Sep 1, 1997).

28. J. Soding, A. Biegert, A. N. Lupas, *Nucleic Acids Res* **33**, W244 (Jul 1, 2005).

29. L. Aravind, E. V. Koonin, *Genome Biol* **2**, RESEARCH0007 (2001).

30. R. Ono *et al.*, *Cancer Res* **62**, 4075 (Jul 15, 2002).

31. Y. Tsukada *et al.*, *Nature* **439**, 811 (Feb 16, 2006).

32. M. D. Allen *et al.*, *Embo J* **25**, 4503 (Oct 4, 2006).

33. H. Cedar, A. Solage, G. Glaser, A. Razin, *Nucleic Acids Res* **6**, 2125 (1979).

34. G. R. Wyatt, S. S. Cohen, *Biochem J* **55**, 774 (Dec, 1953).

35. L. Que, Jr., R. Y. Ho, *Chem Rev* **96**, 2607 (Nov 7, 1996).

36. C. Loenarz, C. J. Schofield, *Nat Chem Biol* **4**, 152 (Mar, 2008).

37. L. E. Netto, E. R. Stadtman, *Arch Biochem Biophys* **333**, 233 (Sep 1, 1996).

38. B. H. Ramsahoye *et al.*, *Proc Natl Acad Sci U S A* **97**, 5237 (May 9, 2000).

39. V. Valinluck *et al.*, *Nucleic Acids Res* **32**, 4100 (2004).

40. J. A. Smiley, M. Kundracik, D. A. Landfried, V. R. Barnes, Sr., A. A. Axhemi, *Biochim Biophys Acta* **1723**, 256 (May 25, 2005).

41. E. Privat, L. C. Sowers, *Chem Res Toxicol* **9**, 745 (Jun, 1996).

42. J. G. Flaks, S. S. Cohen, *J Biol Chem* **234**, 1501 (Jun, 1959).

43. A. H. Alegria, *Biochim Biophys Acta* **149**, 317 (Dec 19, 1967).

44. V. Valinluck, L. C. Sowers, *Cancer Res* **67**, 946 (Feb 1, 2007).

45. M. Bostick *et al.*, *Science* **317**, 1760 (Sep 21, 2007).

46. S. V. Cannon, A. Cummings, G. W. Teebor, *Biochem Biophys Res Commun* **151**, 1173 (Mar 30, 1988).

47. B. Zhu *et al.*, *Nucleic Acids Res* **28**, 4157 (Nov 1, 2000).

48. B. Zhu *et al.*, *Proc Natl Acad Sci U S A* **97**, 5135 (May 9, 2000).

49. R. Metivier *et al.*, *Nature* **452**, 45 (Mar 6, 2008).

50. S. Kangaspeska *et al.*, *Nature* **452**, 112 (Mar 6, 2008).

51. K. Rai *et al.*, *Cell* **135**, 1201 (Dec 26, 2008).

52. V. Rusmintratip, L. C. Sowers, *Proc Natl Acad Sci U S A* **97**, 14183 (Dec 19, 2000).

53. N. W. Penn, R. Suwalski, C. O'Riley, K. Bojanowski, R. Yura, *Biochem J* **126**, 781 (Feb, 1972).

54. J. R. Pratt *et al.*, *Transplant Proc* **38**, 3344 (Dec, 2006).

55. M. W. Simmen, *Genomics* **92**, 33 (Jul, 2008).

56. S. Tardy-Planechaud, J. Fujimoto, S. S. Lin, L. C. Sowers, *Nucleic Acids Res* **25**, 553 (Feb 1, 1997).

57. E. Sutherland, L. Coe, E. A. Raleigh, *J Mol Biol* **225**, 327 (May 20, 1992).

58. T. Rein, M. L. DePamphilis, H. Zorbas, *Nucleic Acids Res* **26**, 2255 (May 15, 1998).

59. H. Hayatsu, M. Shiragami, *Biochemistry* **18**, 632 (Feb 20, 1979).

60. R. B. Lorsbach *et al.*, *Leukemia* **17**, 637 (Mar, 2003).

61. F. Viguie *et al.*, *Leukemia* **19**, 1411 (Aug, 2005).

62. F. Delhommeau *et al.*, paper presented at the ASH Annual Meeting and Exposition, San Francisco, CA, December 9, 2008 2008.

63. M. B. Lobocka *et al.*, *J Bacteriol* **186**, 7032 (Nov, 2004).

64. M. Hori *et al.*, *Nucleic Acids Res* **31**, 1191 (Feb 15, 2003).

Acknowledgements

We thank K. Kreuzer for the gift of the T4 phage and *E.coli* ER1656 and CR63, Charles Richardson, Udi Qimron and Ben Beauchamp for advice and assistance in culturing T4 phage, and Melissa Call for advice on production of recombinant proteins in insect cells, and Patrick Hogan for many helpful discussions. This work was supported by NIH grant AI44432, a Scholar Award from the Juvenile Diabetes Research Foundation, and a Seed grant from the Harvard Stem Cell Institute (to A.R).

Figure Legends

Figure 1. Multiple sequence alignment of the TET/JBP family and representative 2OG-Fe(II) dependent dioxygenase domains. **(A)** The cysteine-rich region (C) and the core catalytic domain (D) are aligned separately. The family is unified by the presence of a distinctive proline which might result in a kink in the N-terminal conserved helix, and a conserved aromatic residue (typically F) in the strand following the first conserved helix. This aromatic residue is part of an sx_2a signature that is boxed in the alignment (where s is small, a is an aromatic residue, and x is any amino acid). The alignment also contains the structurally characterized representatives of the dioxygenase superfamily, the *Chlamydomonas* prolyl hydroxylase (P4H) and the *E. coli* AlkB along with their PDB codes. The proteins are labeled by their gene name and species, separated by underscores. The 95% consensus was calculated from a larger alignment of the TET/JBP proteins. The consensus for the TET-specific C-terminal strands was calculated separately from a larger alignment specifically of the metazoan TET homologs. **(B)** A schematic diagram of the predicted domain architecture of TET1 which includes the CXXC-type zinc-binding domain (CXXC, amino acids (aa) 584-624); the cysteine-rich region (Cys-rich, aa1418-1610); and the double stranded beta helix domain predicted to have 2OG-Fe(II) oxygenase activity (DSBH (D), aa 1611-2074). The positions of three bipartite nuclear localization sequences (NLS) are shown.

Figure 2. Expression of TET1 in HEK293 cells results in decreased 5mC staining intensity. (A) Immunocytochemistry of HEK293 cells transiently transfected for two days with empty vector (mock) or plasmids expressing wild-type or mutant HA-TET1 and co-stained with antibodies specific for the HA epitope (green) or 5mC (red). Nuclei were counterstained with DAPI (blue). Selected HA-expressing cells are circled for convenient comparison. Cells expressing wild-type TET1 (green) show decreased 5mC staining, whereas cells expressing mutant TET1 do not show an overall decrease. Scale bars, 10 mm. **(B)** Quantification of staining intensities using the CellProfiler™ image analysis program. Staining intensities of HA (green) and 5mC (red) were measured in individual nuclei (defined by DAPI staining) of all cells in a given field using low magnification images (20x). Data from transfected cell populations (600-1200 cells, pooled from 3

microscope fields) are presented as dot plots (red), superimposed on the corresponding dot plots from mock-transfected cells (blue), with each dot representing each individual cell. The same set of mock-transfected cells is shown in the two panels. The plots show data from one experiment, representative of 3-4 independent experiments. **(C)** The population average of 5mC staining intensities of HA-expressing cells (arbitrarily defined by HA pixel intensities above the highest intensity observed in mock-transfection) in each group of transfected cells are compared with that of mock-transfected cells (set as 1) in each experiment. Values are background-subtracted before normalization and are mean ± SEM from 3-4 experiments; statistical comparisons are based on ANOVA and Bonferroni's post-hoc test.

Figure 3. The genomic DNA of TET1-expressing HEK293 cells contains a modified nucleotide within the dinucleotide CG. (A) Schematic diagram describing the experimental design. **(B-F)** Genomic DNA was purified from transfected HEK293 cells overexpressing TET1 and cleaved with **(B, C)** MspI, **(D)** HpaII or **(E, F)** Taq$^{\alpha}$I, and the end-labeled fragments were digested to 5' dNMPs and resolved by TLC. Cells expressing wild-type but not mutant versions of TET1-FL and TET1-CD show altered relative abundance of unmethylated and methylated cytosine, as well as a modified nucleotide indicated by **?**. **(E, F)** The intensities of the spots corresponding to dCMP, 5m-dCMP and the unidentified nucleotide (**?**, subsequently identified as hm-dCMP) from **(B)** and **(E)** were quantified from three independent transfections and the average amounts are shown expressed as a fraction of all cytosine-derived nucleotides. Error bars represent the s.d. of triplicates. **(C)** Neither 5m-dCMP nor the modified nucleotide are observed when the DNA is digested with HpaII. The data shown are representative of at least three experiments.

Figure 4. The modified nucleotide is identified as 5-hydroxymethylcytosine. **(A)** DNA from T4 phage grown in GalU-deficient *E.coli* ER1656 (T4*), and genomic DNA from unsorted populations of HEK293 cells expressing wild-type or mutant TET1-CD, were digested with Taq$^{\alpha}$I. In the case of T4* DNA, this results in the production of fragments with hmC at the 5' end. The end-labeled fragments were hydrolyzed and the resulting dNMP's were separated by TLC. The modified nucleotide produced in TET1-CD-

expressing HEK293 cells migrates with an Rf of ~0.29 in isobutyric acid:water:ammonia (66:20:1), similarly to authentic hmC isolated from T4* phage. **(B)** Comparison of LC/ESI-MS ions present at an Rf of 0.29 in samples derived from genomic DNA of HEK293 cells expressing wild-type (wt) or mutant (mut) TET1-CD. A species with an observed m/z = 336.06 was significantly more abundant (18.5-fold) in the wildtype sample compared with the mutant sample. **(C)** Mass spectrometry fragmentation (MS/MS) analysis of authentic hm-dCMP (*top*), and the m/z = 336.06 species isolated from HEK293 cells expressing TET1-CD (*bottom*). MS/MS analysis was performed in negative ion mode with a collision energy of 15 V. Expected m/z values are shown in red; observed m/z values are shown in black (anticipated mass accuracy is within 0.002 Da). The fragmentation patterns are identical, identifying the unknown nucleotide as hm-dCMP.

Figure 5. Recombinant Flag-HA-TET1-CD purified from Sf9 cells converts 5mC to hmC in methylated DNA oligonucleotides *in vitro*. **(A)** 2 μg of double-stranded DNA oligonucleotides containing a fully-methylated Taq$^{\alpha}$I site were incubated with 3 μg of purified Flag-HA-TET1-CD or mutant Flag-HA-TET1-CD in a buffer containing 1 mM 2OG, 2 mM ascorbic acid, 75 μM Fe(II) for 3 hours at 37 C (1:10 enzyme to substrate ratio). Recovered oligonucleotides were digested with Taq$^{\alpha}$I, end-labeled, hydrolyzed to dNMP's and resolved using TLC. The faint dCMP spot in each lane is derived from end-labelling of the C at the 5' end of each strand of the oligonucleotide substrate. (T4 PNK is not able to phophorylate blunt ends as efficiently as the 5' overhangs generated by restriction enzyme cleavage.) **(B)** The intensity of dNMP's was quantified using PhosphorImager and the extent of conversion of 5mC to hmC is shown as the ratio (hmC/(hmC+ 5mC)). Error bars represent the standard deviation of triplicates. **(C)** Comparison of species within the TLC spot (Rf = 0.29) of products resulting from synthetic fully-methylated double-stranded oligonucleotides incubated with wild-type (wt) or mutant (mut) Flag-HA-TET1-CD. **(D)** Mass spectrometry fragmentation analysis (MS/MS) of authentic hm-dCMP isolated from mutant T4 phage (top) and the unidentified nucleotide derived from synthetic fully-methylated double-stranded oligonucleotides incubated with wild-type Flag-HA-TET1-CD (bottom). MS/MS analysis was performed in negative ion mode with a collision energy of 15 V. Observed masses are shown in black (mass accuracy was within 0.002 Da). The fragmentation patterns are identical, identifying the unknown nucleotide as hm-dCMP. **(E)** Recombinant Flag-HA-TET1-CD is able to hydroxylate 5mC in fully-methylated (full, *lanes*

5, 6) and hemi-methylated (hemi, *lanes 3, 4*) DNA oligonucleotide substrates. Unme, unmethylated DNA oligonucleotide (*lanes 1, 2*).

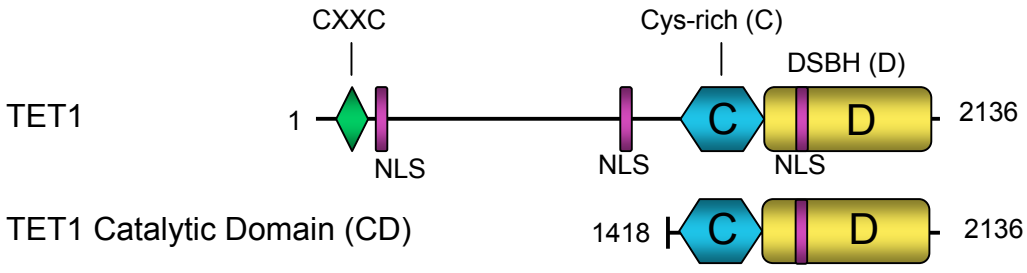
Figure 6. hmC is present in ES cell DNA and its abundance decreases upon differentiation or TET1 depletion

(A) *Left*, TLC showing that hmC is detected in the genome of ES cells, but not dendritic cells or T cells. *Right*, graph quantifying the relative abundance of 5mC, C and hmC in the genomic DNA of ES cells cultured under standard conditions in media containing the cytokine LIF. **(B)** TET1 mRNA levels decline by ~80% in ES cells induced to differentiate by withdrawal of LIF. **(C)** The same differentiated ES cells show a corresponding ~40% decrease in hmC levels. **(D)** Transfection of ES cells using two different RNAi duplexes directed against TET1 decreases TET1 mRNA levels by ~75%. (E) The same TET1-depleted ES cells show a corresponding ~40% decline in hmC levels.

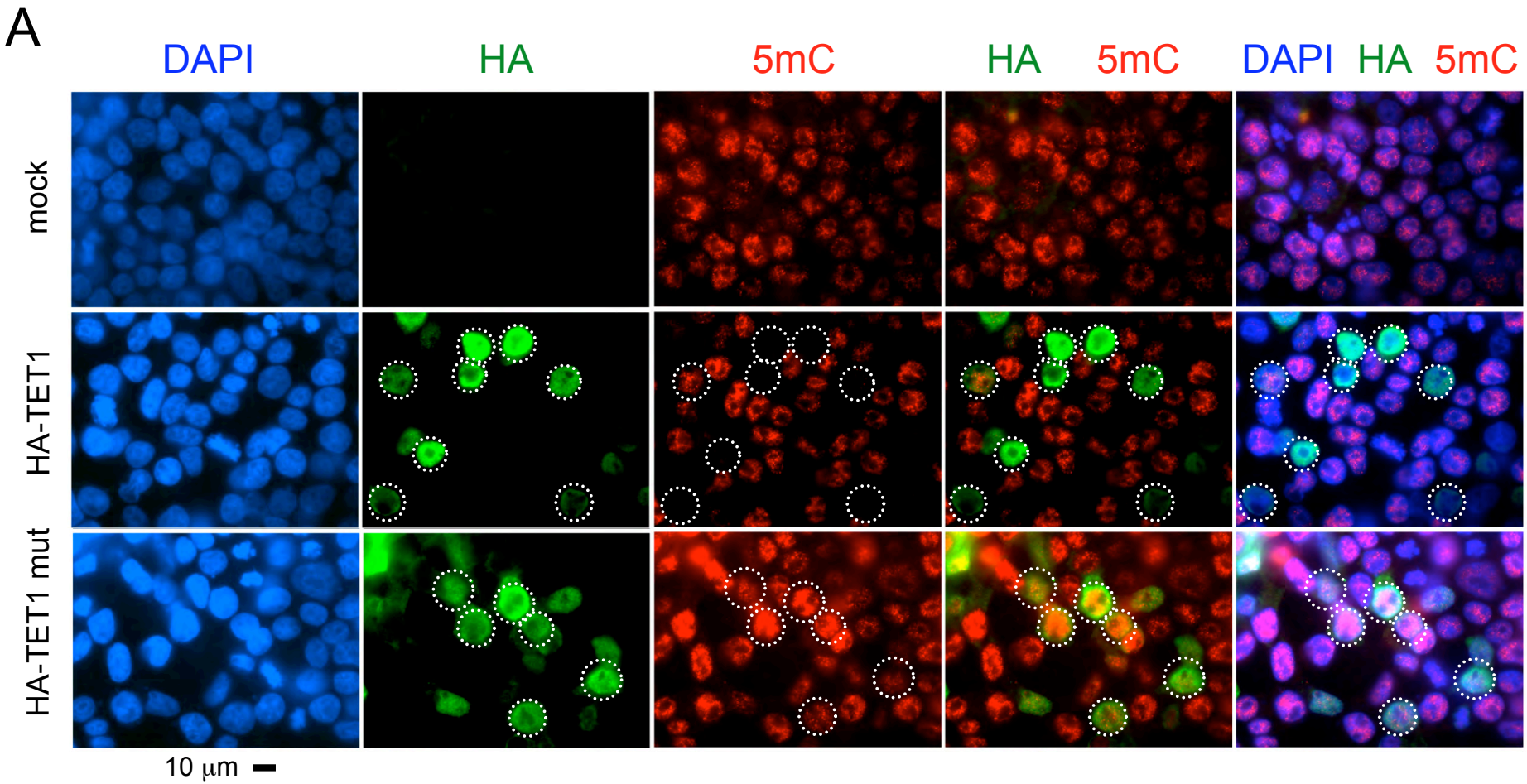
Figure 7. Model speculating on the biological role of hmC and its potential as an intermediate in DNA demethylation. **(A)** Integration of hmC into the known pathways of DNA methylation and passive replication-dependent demethylation. Known pathways are shown in black and the new findings in this study in red. Dashed lines and question marks are used to indicate pathways that may exist but have not been experimentally established. **(B)** Potential biological mechanisms involving hmC. hmC may recruit specialized binding proteins; and conversion of 5mC to hmC may displace methyl-binding proteins from DNA. hmC may also be an intermediate that facilitates passive DNA demethylation: conversion of 5mC to hmC in hemi-methylated DNA may interfere with recognition by DNMT1 and associated SRA-domain proteins. Finally, hmC may be recognised by specialized DNA repair proteins in specific cell types, and so function as an intermediate in “active” (repair-mediated) DNA demethylation.

A

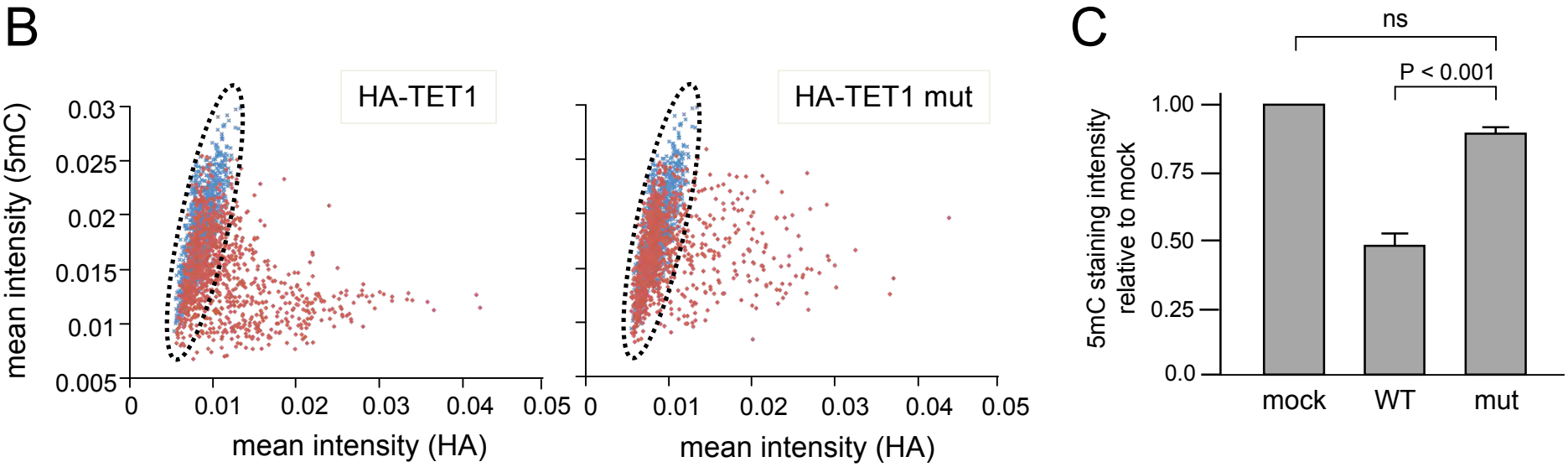
B



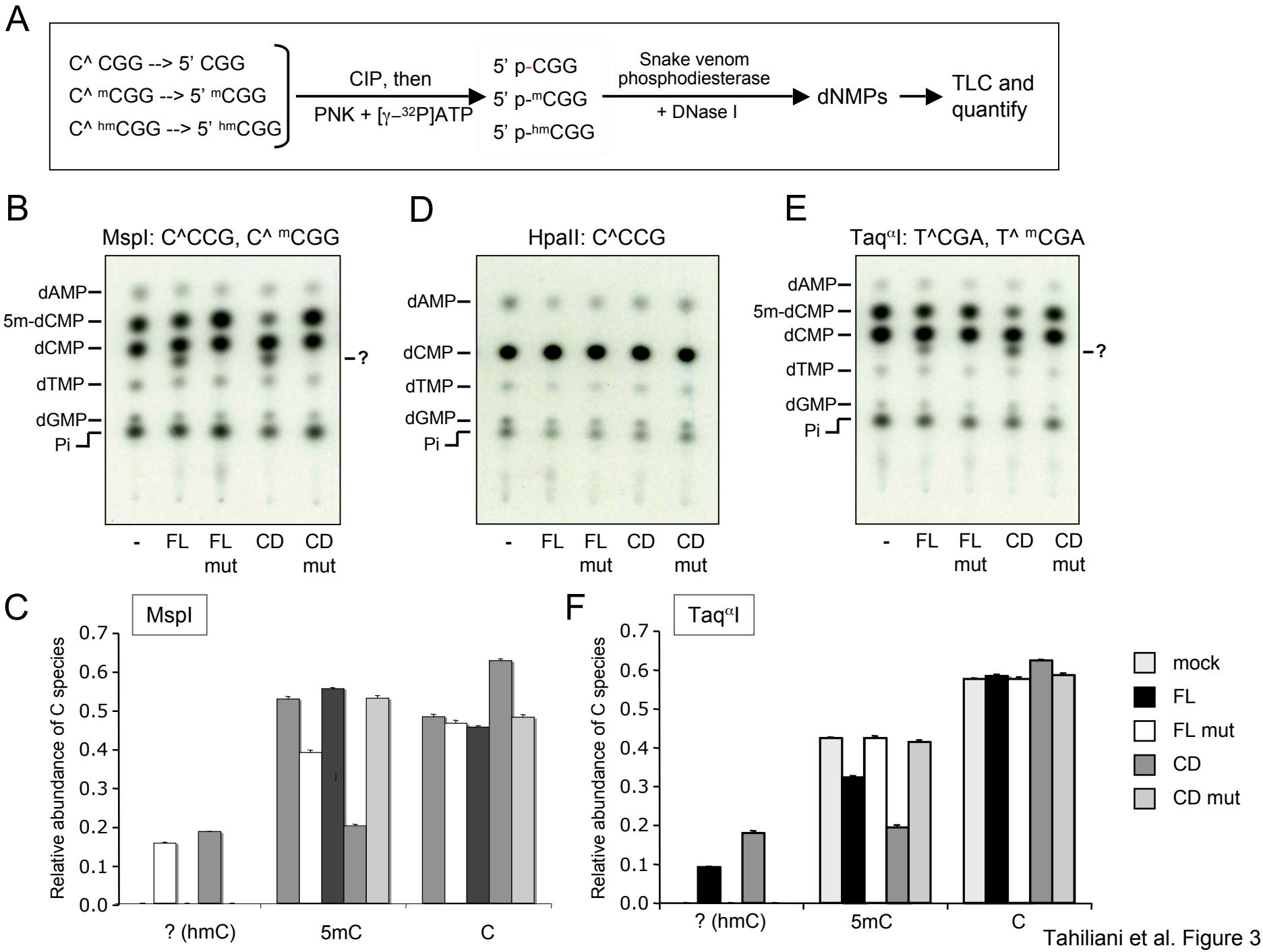
Tahiliani et al. Figure 1



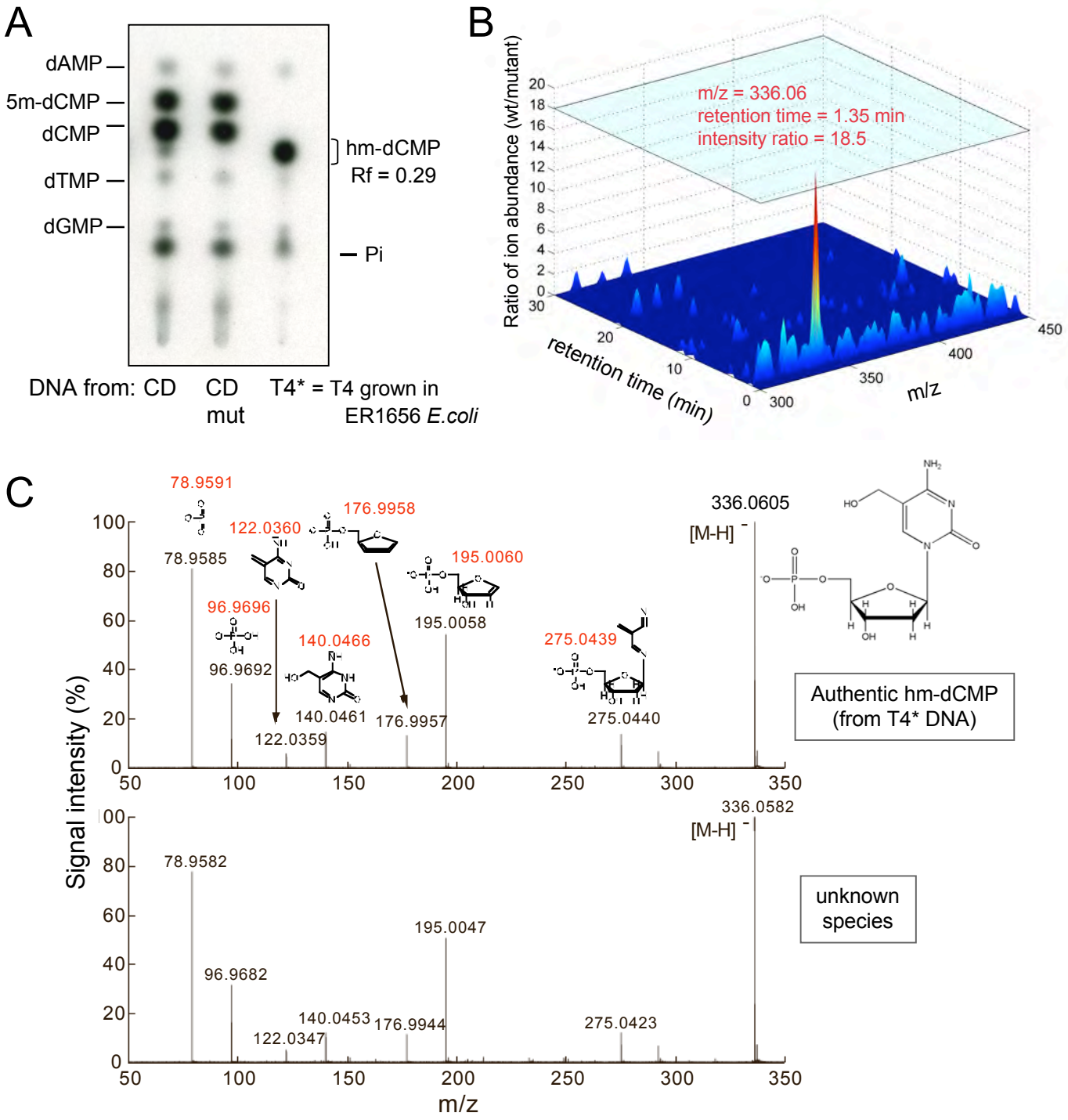
Tahiliani et al. Figure 2A



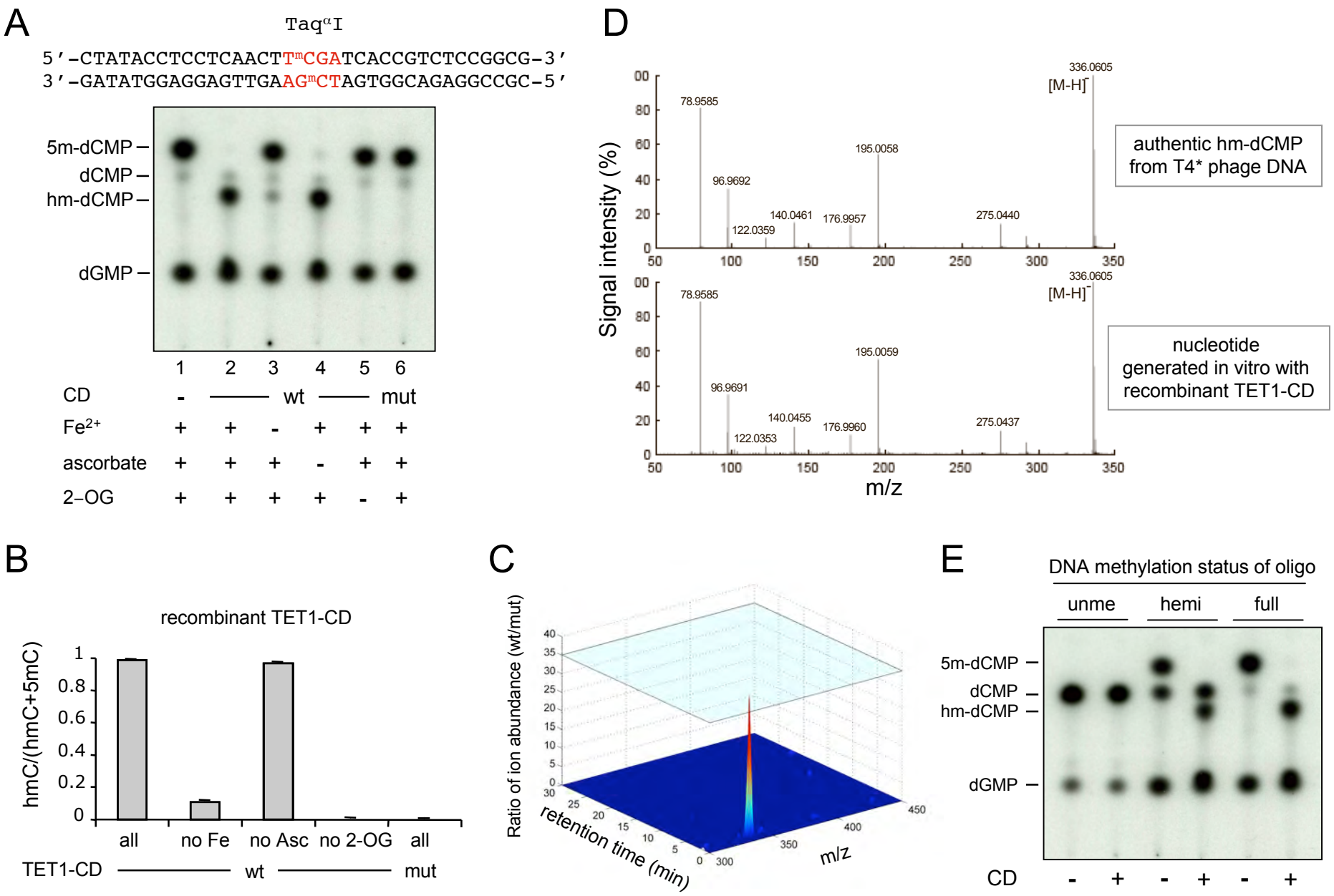
Tahiliani et al. Figure 2BC



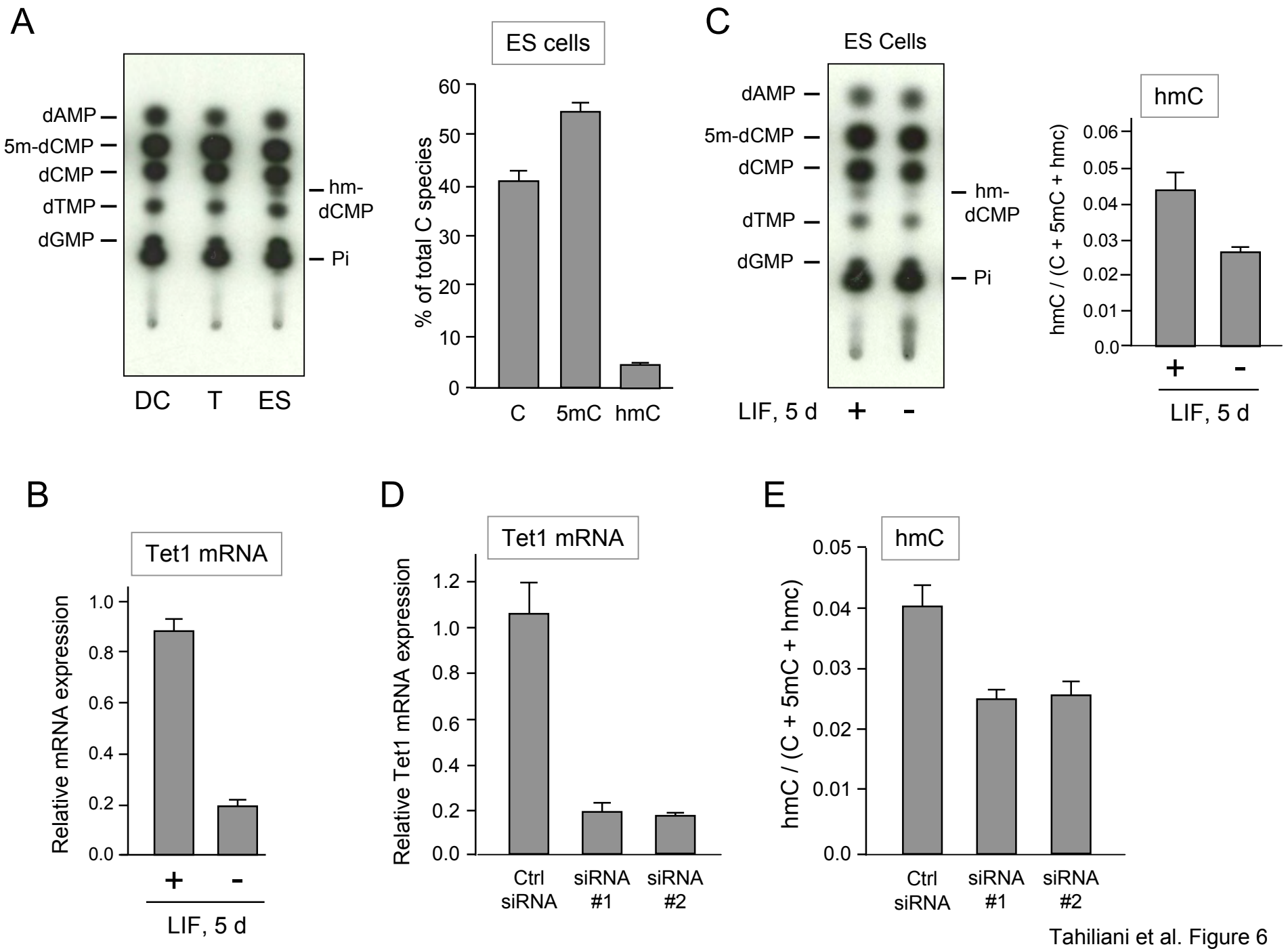
Tahiliani et al. Figure 3



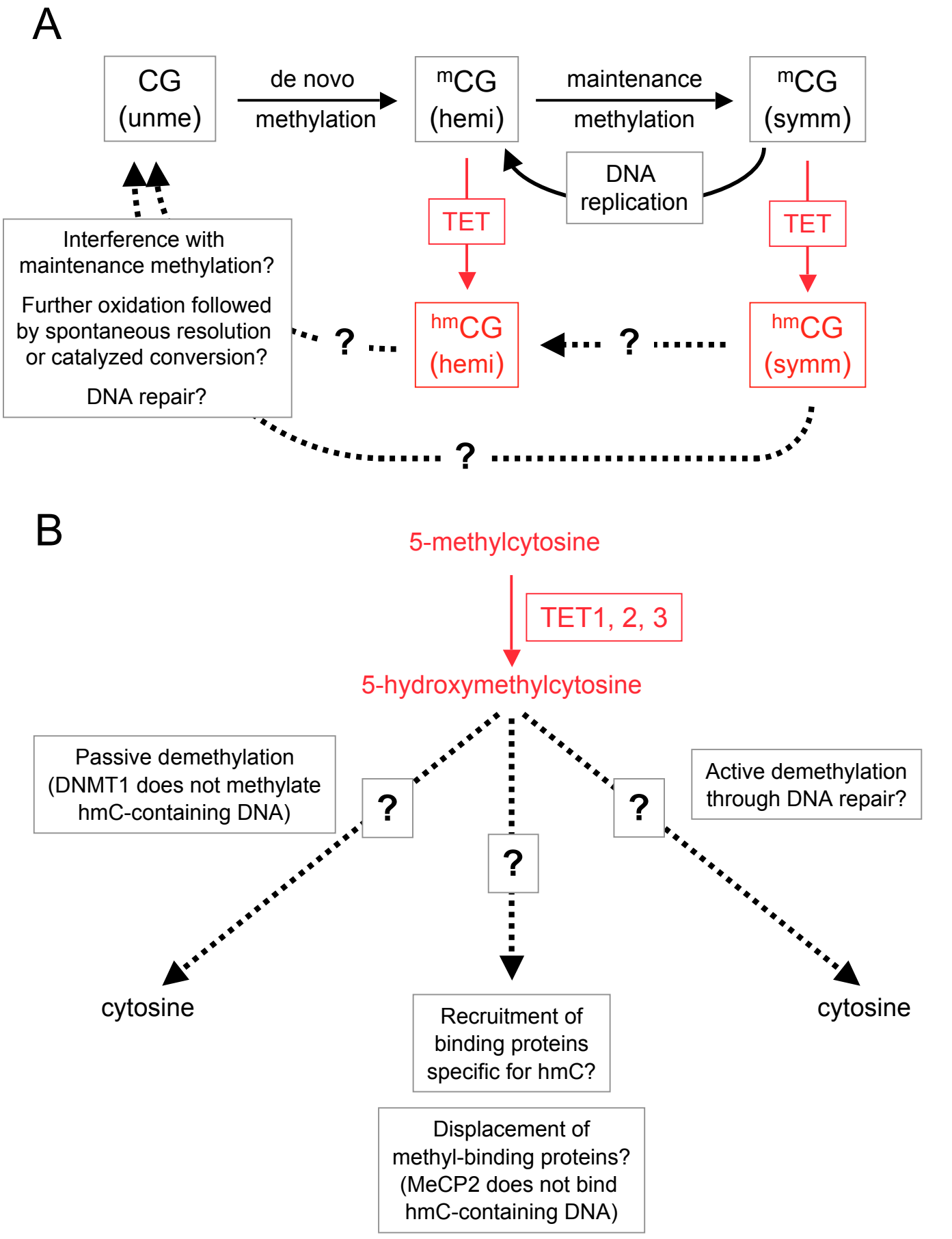
Tahiliani et al. Figure 4



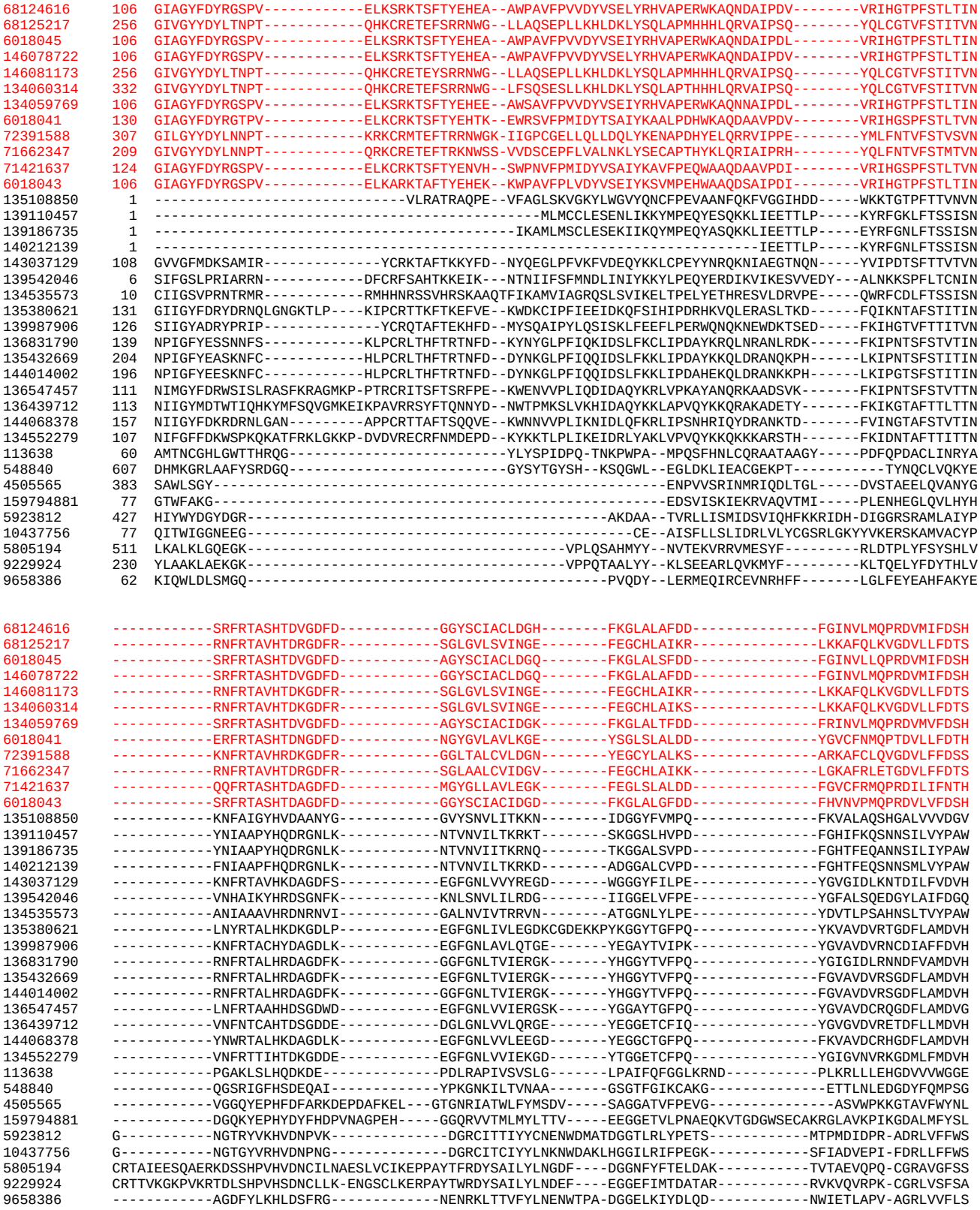
Tahiliani et al. Figure 5



Tahiliani et al. Figure 6



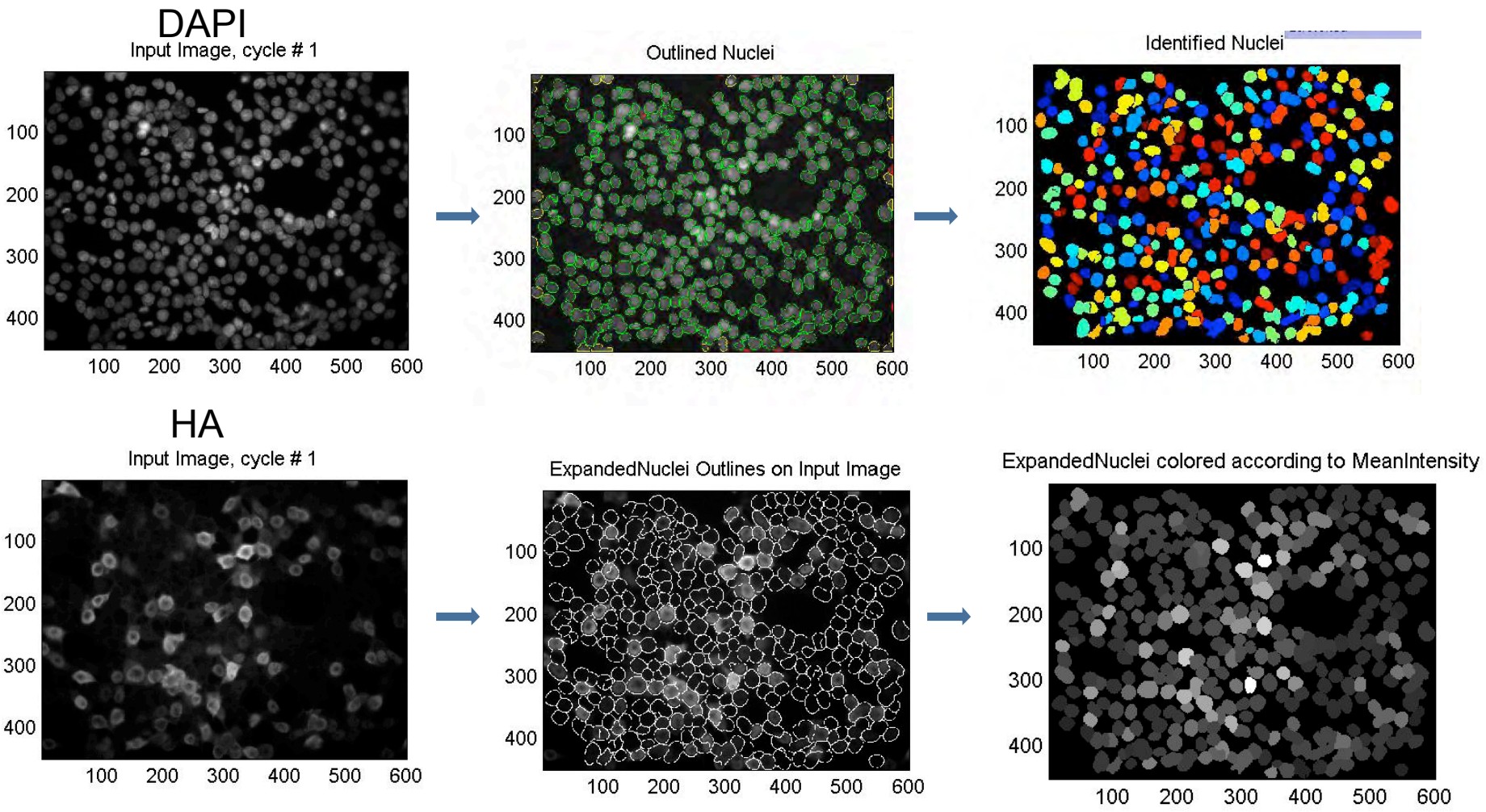
Tahiliani et al., Figure 7



Tahiliani et al., Suppl. Fig. S1

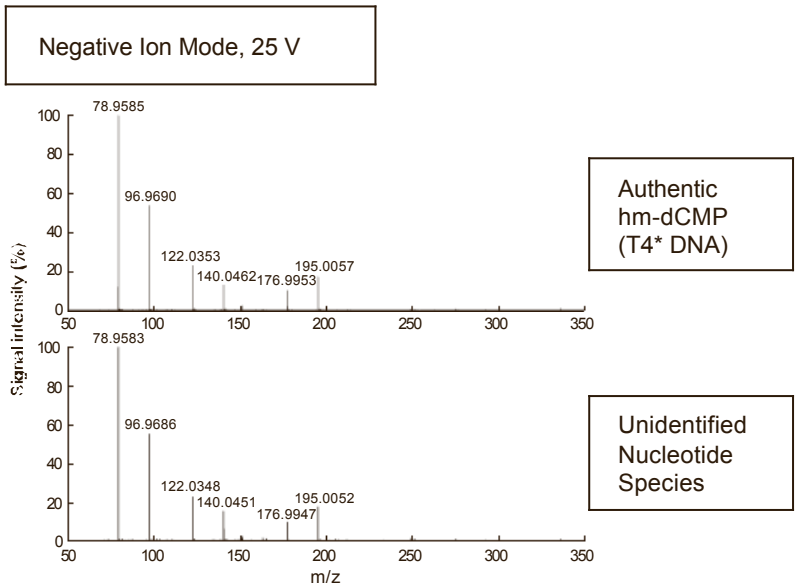
```
68124616     ---------HFHSNTEVELS----------------------FSGEDWKRLTCVFYYRA   264
68125217     ---------LEHGNTEVVN-----------------------PEIHWQRTSVVCYLRT   412
6018045      ---------HFHSNTEVELS----------------------FSGEDWKRLTCVFYYRA   264
146078722    ---------HFHSNTEVELS----------------------FSGEDWKRLTCVFYYRA   264
146081173    ---------LEHGNTEVVN-----------------------PEIHWQRTSVVCYLRT   412
134060314    ---------LEHGNTEVVH-----------------------PENHWQRTSIVCYLRT   488
134059769    ---------HFHSNTEVEVS----------------------CSEEDWKRLTCVFYYRT   264
6018041      ---------LFHSNTELEAK----------------------EANATWNRLSCVFYYRA   288
72391588     ---------LEHGNTEVHNR----------------------EGSWRRISIVCYLRC   464
71662347     ---------LEHGNTEVHNF----------------------DYCWKRVSVVCYLRN   366
71421637     ---------FFHSNTEPELNH----------------------PRDDWSRLTCVCYYRA   282
6018043      ---------YFHSNSELEISC----------------------PTEEWRRLTCVFYYRS   264
135108850    S--------IPHGVTPIIP-----------------------KAKNWERSSVVFYTLS   143
139110457    Y--------NIHGVTKIVR-----------------------ENEQSYRNSLIFYPLQ   129
139186735    F--------NIHGVTKIIK-----------------------EHEQGYRNSLIFYPLK   132
140212139    Y--------NIHGVTKIIK-----------------------HKEEGYRNSLIFYPLS   103
143037129    ---------KYHCNTGFTNFTD-------------------------------------   253
139542046    T--------EIHGVMPIYQ-----------------------TKENPYRASIVYYSLE   167
134535573    R--------NYHGVTPIEP-----------------------THPGGYRNSLIWYALD   172
135380621    ---------EFHCNTELTG-----------------------DNYSRLSLVSYLRK   303
139987906    ---------ELHGNTQTISKK---------------------PYERISIICYYRK   283
136831790    ---------QWHSNTPIIETDEDKLFNNTLNNDYKDNPNIGTEGIYTKYTRLSFVCYLRE   323
135432669    ---------QWHSNTDIYETEEDKIYNDTIDYAFNDNPEVGTVGLDKKYTRLTFVCYLRE   388
144014002    ---------QWHSNTDIYETEEDKIFNNTIDYAFNDNPEVGTVGLDKKYTRLTFVCYLRE   380
136547457    ---------RLHGNCPMIP-----------------------GDDTSQRISLVCYLRK   281
136439712    ---------QLHANTKLKL-----------------------IGKDSIRLSIVSYLRT   283
144068378    ---------EWHCNTKIKP-----------------------ITKDYSRLSLVAYLRE   318
134552279    ---------QPHGNLEMKK-----------------------KHPDVERLSVVCYLRK   276
113638       ------SRLFYHGIQPLKAGF---------------------HPLTIDCRYNLTFRQAG   213
548840       F-----QETHKHNVVA-------------------------------VTPRLSFTFRSTV   743
4505565      FASGEGDYSTRHAACPVL-----------------------------VGNKWVSNKWLHE   519
159794881    KPDGSNDPASLHGSCPTLK-----------------------------GDKWSATKWIHV   226
5923812      D------RRNPHEVMPVF-----------------------------RHRFAITIWYMD   566
10437756     D------RRNPHEVQPSY-----------------------------ATRYAMTVWYFD   214
5805194      G------TENPHGVKAVT-----------------------------RGQRCAIALWFTL   670
9229924      G------KECLHGVKPVT-----------------------------KGRRCAMALWFTM   388
9658386      -------ERFPHEVLEAH-----------------------------ADRVSIAGWFRT   193
```

CellProfiler™ cell image analysis: Quantification of HA and 5mC signals in individual cells

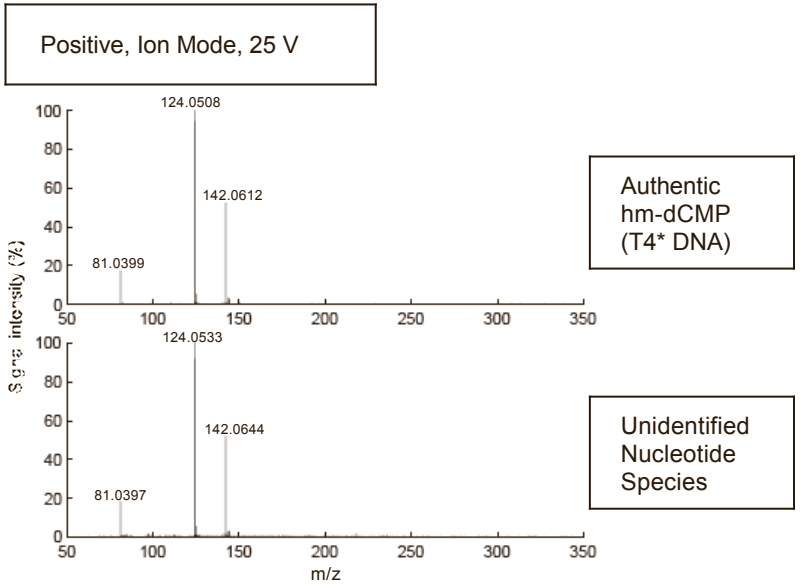


Tahiliani et al. Suppl. Figure 2

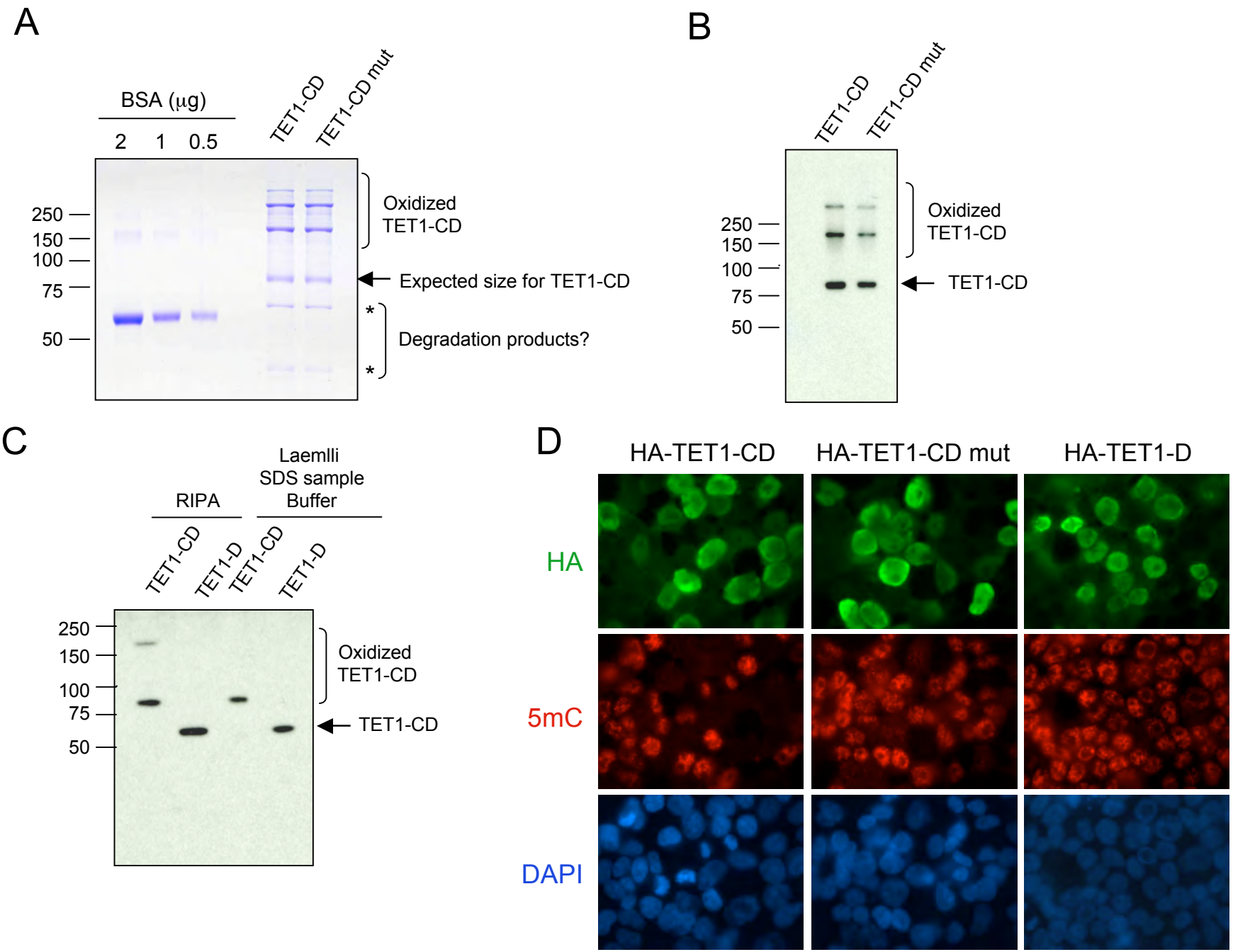
A



B



Tahiliani et al. Suppl Figure 3



Tahiliani et al., Suppl. Fig. 4

A

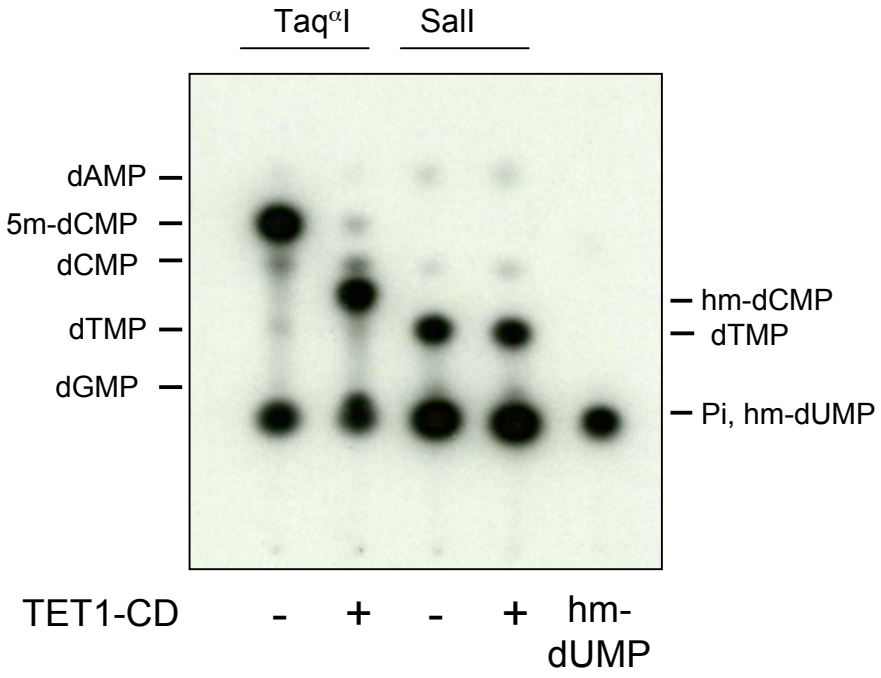
SalI

5'-TGCTACCTCCTCAACGTCGACCACCGTCTCCTGCA-3'
3'-ACGATGGAGGAGTTGCAGCTGGTGGCAGAGGACGT-5'

Taq$^{\alpha}$I

5′-CTATACCTCCTCAACTT^{m}CGATCACCGTCTCCGGCG-3′
3′-GATATGGAGGAGTTGAAG^{m}CTAGTGGCAGAGGCCGC-5′

B



Tahiliani et al., Suppl. Fig. 5

Supporting online material

Supplemental Figure Legends

Supplemental Figure S1. **Sequences used to generate the position-specific score matrix (PSSM).** The sequences provided in the alignment represent those used to create the position specific score matrix for searches with the PSI-BLAST program. Each sequence is identified by the gi number allowing its recovery from the GenBank database and the numbers flanking the alignment provide the limits of the aligned region in the sequence. The JBP proteins from kinetoplastids are marked in red. This alignment was used in the manner described in the article for the sequence search strategy leading to the detection of the TET proteins and their homologs. The searches were performed using the checkpoint start option –B, with the –h set to 0.01 and –F was set to F and the all JBP sequences were cycled through the –i options for individual searches.

The TET/JBP family is defined by all 2OG-Fe(II) oxygenases that are closer to the kinetoplastid JBP proteins and the metazoan than any other family of dioxygenases. They are characterized by a synapomorphic extended α-helix just N-terminal to the first core strand. The TET/JBP family can further be divided based on sequence features and similarity score based clustering in several distinct subfamilies with highly distinctive phyletic patterns. These include: 1) the TET subfamily found in metazoans; 2) the JBP subfamily found in kinetoplastids and uncharacterized marine microbes; 3) the fungal-algal subfamily found in basidiomycete fungi and chlorophyte algae; 4) gp2 subfamily found in actinophages Cooper and Nigel of Mycobacteria and *Frankia* prophage and several uncultured viruses from marine samples. The fungal-algal subfamily is greatly expanded in the basidiomycete fungal genomes and is present in 35-60 paralogous versions per genome. The ParB protein, which is associated with members of the bacteriophage gp2 subfamily in conserved operons, belongs to a superfamily of proteins implicated in DNA-binding and chromosome segregation, suggesting that gp2 might interact with the ParB protein in these functions. In other phages the ParB protein shows fusions with DNA methylases, and these enzymes have been implicated in replication or chromosome partitioning in enterobacteriophages such as P1 (*1*). This suggests that the actinophage gp2 might further modify methylated bases or reverse their methylation to regulate these processes in viruses possessing them. As these viral

versions are the smallest members of the TET/JBP family and comprise more or less of the minimal 2OG-Fe(II) oxygenases catalytic domain, they could potentially be the ancestral versions, which spawned the eukaryotic versions.

Suppl. Fig. S2. **Image analysis using CellProfiler**. Nuclei were outlined based on DAPI fluorescence using the IdentifyPrimAutomatic module and denoted as Outlined Nuclei. Objects within a pre-set diameter range of 10-35 pixel units are outlined in green and included in analysis. Objects outside the range (including cell clusters) as well as those touching the borders are outlined in red and excluded from analysis. To account for HA staining at nuclear boundaries, an IdentifySecondary module was added to expand the nuclear outlines by 2 pixels, denoted as ExpandedNuclei Outlines. The MeasureObjectIntensity module was then used to apply the ExpandedNuclei Outlines on the corresponding HA image, and the Outlined Nuclei on the corresponding 5mC image (not shown), to measure pixel intensities of HA and 5mC staining respectively. Pixel numbers are indicated at the axes of images, which are taken with at a magnification of 20x.

Supplemental Figure S3. Mass spectrometry fragmentation analyis (MS/MS) of authentic hm-dCMP from T4* phage (*top*) and the unidentified nucleotide species present in genomic DNA from HEK293 cells overexpressing TET1-CD (*bottom*). **(A)** MS/MS analysis was performed in negative ion mode with a collision energy of 25 V. **(B)** MS/MS analysis was performed in positive ion mode with a collision energy of 25 V. Observed masses are shown (anticipated mass accuracy was within 0.003 Da).

Supplemental Figure S4. Purification of recombinant Flag-HA-TET1-CD from Sf9 cells and characterization of TET1-CD fragment. (A) Coomassie-stained SDS-PAGE gel of wild-type and mutant Flag-HA-TET1-CD purified from Sf9 cells to near homogeneity by affinity chromatography with anti-Flag antibody conjugated beads. Known amounts of BSA were loaded on the same gel for comparison. Wild-type and mutant TET1-CD (79 KDa, indicated by the arrow) both demonstrate a strong tendency to oxidize and form disulfide-linked dimers (158 KDa) and higher-order multimers, which are resistant to DTT. Asterisks denote degradation products that are not detected by immunoblotting (see **(B)**), probably due to loss of the N-terminal Flag-HA epitope tag.

(B) The bands of higher apparent molecular weight were identified as TET1 oxidation products by immunoblotting with an anti-HA antibody. **(C)** Immunoblot with anti-Flag antibody showing that the multimeric forms TET1-CD increase with increased processing time of lysates and concomitant exposure to oxidizing conditions. A twenty minute lysis in 10 mM DTT-containing RIPA buffer (non-denaturing) results in the detectable presence of a TET1-CD dimer. This dimer is not detected when cell pellets from Flag-HA-TET1-CD overexpressing cells are lysed directly in Laemlli SDS sample buffer (denaturing). The tendency of TET1-CD to oxidize appears to be due at least in part to disulfide bond formation by the Cys-Rich region (C) that is N-terminal to the DSBH (D) region. Removal of the N-terminal Cys-Rich region resulted in expression of a protein (Flag-HA-TET1-D) that runs at its expected molecular weight (57.6 KDa), even when proteins in the lysates are exposed to oxidizing conditions. Together the data shown in **(A-C)** suggest that intermolecular disulfide bonds ~~that~~ are formed during oxidation, and involve a physiological homodimerization interface that is not accessible to reducing agents under native conditions where protein-protein interactions are maintained. **(D)** Overexpression of Flag-HA-TET1-CD in HEK293 cells, but not Flag-HA-TET1-D, results in decreased staining by an anti-5mC antibody.

Supplemental Figure S5. Recombinant Flag-HA-TET1-CD purified from Sf9 cells does not convert thymine (another 5-methylpyrimidine) to hmU *in vitro*. **(A)** 2 μg of double-stranded DNA oligonucleotides containing a fully-methylated Taq$^{\alpha}$I (T^^{m}CGA) or a SalI (G^TCGAC) site were incubated with 3 μg of purified Flag-HA-TET1-CD or mutant Flag-HA-TET1-CD in a buffer containing 1 mM 2OG, 2 mM ascorbic acid, 75 μM Fe(II) for 5 hours at 37 C (1:10 enzyme to substrate ratio). Recovered oligonucleotides were digested with Taq$^{\alpha}$I or SalI, end-labeled, hydrolyzed to dNMP's and resolved using TLC. TET1-CD is able to hydroxylate 5mC, but is not able to act on thymine in the context of a SalI site in a double-stranded oligonucleotide substrate in vitro. A double-stranded DNA oligonucleotide terminating in hmU was end-labelled and hydrolyzed to generate a standard for hm-dUMP migration (lane 5). SalI has been demonstrated to cleave G(hmU)CGAC equivalently to GTCGAC (2).

Methods

Computational and bioinformatic analyses. The non-redundant (NR) database of protein sequences (National Center for Biotechnology Information, NIH, Bethesda) was searched using the PSI-BLAST programs (*3*). Profile searches using the PSI-BLAST program were conducted either with a single sequence or a sequence with a PSSM used as the query, with a profile inclusion expectation (E) value threshold of 0.01, and were iterated until convergence (*3*). For all compositionally biased queries the correction using composition-based statistics was used in the PSI-BLAST searches (*4*). Multiple alignments were constructed using the Kalign program (*5*), followed by manual correction based on the PSI-BLAST results. The multiple alignment was used to create a HMM using the Hmmbuild program of the HMMER package (*6*). It was then optimized with Hmmcaliberate and the resulting profile was used to search a database of completely sequenced genomes using the Hmmsearch program of the HMMER package (*6*). Profile-profile searches were performed using the HHpred program (*7, 8*). The JPRED program (*9*) and the COILS program were used to predict secondary structure. Globular domains were predicted using the SEG program with the following parameters: window size 40, trigger complexity=3.4; extension complexity=3.75 (*10*).

The Swiss-PDB viewer (*11*) and Pymol programs were used to carry out manipulations of PDB files. Reconstruction of exon-intron boundaries was done using the NCBI Splign program with the tblastn searches against chromosomes as a guide. Gene neighborhoods were determined using a custom script that uses completely sequenced genomes or whole genome shot gun sequences to derive a table of gene neighbors centered on a query gene. Then the BLASTCLUST program is used to cluster the products in the neighborhood and establish conserved co-occurring genes. These conserved gene neighborhood are then sorted as per a ranking scheme based on occurrence in at least one other phylogenetically distinct lineage (“phylum” in NCBI Taxonomy database), complete conservation in a particular lineage (“phylum”) and physical closeness on the chromosome indicating sharing of regulatory -10 and -35 elements. Phylogenetic trees were constructed with the MEGA4 package.

TET1 expression plasmids TET1 ORF was amplified from SY5Y cDNA and the human clone and inserted into XhoI and NotI sites of an Flag-HA tagged pOZ-N. Mutant TET1 (H1671D, Y1673A) was generated using the QuikChange Mutagenesis Kit

(Stratagene). The sequences of all clones were confirmed by conventional DNA sequencing. Wild-type and mutant Flag-HA-TET1-CD was amplified and cloned into Acc651 and XbaI sites of pEF1 (Invitrogen). The IRES-CD25 of pOZ-N was amplified and cloned into the the Acc65I and BstB1 sites of Flag-HA-TET1-CD-pEF1. Wild-type and mutant Flag-HA-TET1-CD was amplified from Flag-HA-TET1-CD-pOZ and inserted into SalI and NotI sites of pFastBac (Invitrogen).

Immunocytochemistry Cells were plated on sterile coverglass in 24-well plates at 1.5 x 10^5 cells/well and grown overnight before transient transfection with pEF1-TET1 expression constructs or empty vector (mock) using *Trans*IT™-293 transfection reagent (Mirus, Madison, WI) according to manufacturer's instructions. At 42-44 hr post-transfection, cells were fixed for 15 min in 4% paraformaldehyde in PBS and permeabilized with 0.2% Triton X-100 in PBS for 15 min at room temperature. For detection of 5-methylcytosine, cells were treated with 2N HCl at room temperature for 30 min and subsequently neutralized for 10 min with 100 mM Tris-HCl buffer, pH 8. After extensive washes in PBS, cells were blocked for 1 hour at room temperature in 1%BSA, 0.05% Tween 20 in PBS. Rabbit anti-HA polyclonal antibody (diluted at 1:400; Santa Cruz Biotechnology, Santa Cruz, CA) and mouse anti-5 methylcytosine clone 162 33 D3 antibody (diluted at 1:2500-3000; Calbiochem, San Diego, CA) were added in blocking buffer for 2-3 hours at room temperature and detected concurrently by secondary antibodies coupled with Cy2 or Cy3 respectively. DNA was stained with 250 ng/ml of4',6-diamidino-2-phenylindole (DAPI) and mounted in SlowFade® Gold antifade reagent (Molecular Probes, Eugene, OR). Images were recorded digitally on a Zeiss Axiovert 200 inverted microscope equipped with a CCD cameraby using OpenLab imaging software (Improvision, Coventry, UK). The quantification of nuclear size was performed on high magnification (63x objective) images using Volocity software (Improvision); the HA-stained images were used to distinguish between transfected (HA positive) and non-transfected (HA-negative) cells and nuclear profiles were drawn on the DAPI-stained images for nuclear area measurements.

CellProfiler™ cell image analysis. Three fields, containing 200-400 cells each, were imaged from each well of transfected cells using an 20X objective. Greyscale images (tiffs) were uploaded on CellProfiler as three individual files for every field captured under the three excitation wavelengths respectively for DAPI, GFP (detection of HA) and

Cy3 (detection of 5mC). Nuclear outlines were profiled based on the DAPI staining, with a secondary module included to expand the nuclear outlines by another 2 pixels (denoted as “expanded nuclei”) to account for HA staining at nuclear boundaries. Clustered cells and cells at the edge of fields were excluded. Staining intensities of HA and 5mC within individual cells profiled were measured as mean pixel intensities of GFP and Cy3 signals, respectively, within the “expanded nuclei” and original nuclei profiles, respectively. The raw data were exported on Excel spreadsheets and plotted as dot plots of 5mC mean pixel intensities against HA mean pixel intensities for each transfection sample.

Transfection and Sorting of HEK293T cells expressing hCD25 HEK293T cells were transfected with TET1-CD-IRES-CD25-pEF1 vectors using TransIT transfection reagent (MIrius). After 48 hours, 30 x 10^6 cells were stained with anti-hCD25-PE antibody (1:200) (Becton Dickinson) in 1X FACS Buffer (1XPBS, 2% FBS, 1 mM EDTA, 0. 1% NaAzide) for 30 min at 4 C. Cells were washed twice with 1X FACS Buffer. Cells were then stained with anti-PE microbeads (Miltenyi) for 25 min, 4 C and then washed two times with 1X FACS Buffer and then once with 1X MACS buffer (1X PBS, 0.5% BSA, 0.09% Na Azide and 2 mM EDTA). Cells were resuspended in 2 ml of ice-cold 1X MACS Buffer and CD25-positive cells were sorted using AutoMACS cells sorter (Possel. Input, flow-through and collected samples were analyzed by FACS to confirm enrichment for CD25-positive cells in collected sample.

Analysis of 5mC levels using thin-layer chromatography Nuclei were prepared from CD25-positive cells by resuspension in 1 ml NPB (240 mM sucrose, 7.5 mM Tris, pH 7.5, 3.75 mM $MgCl_2$, 0,75% Triton-X-100, with 100 μg RNAseA/ml (Qiagen)) and placing on ice for 20 minutes. Cells were spun at 1300 g for 15 min, 4 C and then washed once in NPB. Nuclei were lysed in 650 μl of 1X LB ((10 mM Tris, pH 8.0, 300 mM NaAcetate, pH 5.2, 0.5% SDS, 5 mM EDTA, 100 μg RNAseA/ml and 300 μg/ml Proteinase K (Roche)) and incubated overnight at 55 C. An extra 300 μg/ml Proteinase K was added in the morning and the samples were left at 55 C for 5 hours. Samples were extracted with equal volumes of phenol, phenol: chloroform: isoamyl alcohol (25:24:1) and chloroform: isoamyl alcohol (24:1) and then precipitated with 2 volumes of ethanol. Genomic DNA was washed twice with 1 ml of 70% EtOH, dried and resuspended in 10 mM Tris, 0.1 mM EDTA, pH 8.0 and allowed to resuspend overnight at 32 C.

2 μg of genomic DNA was digested with 100 units of MspI, HpaII or Taq^{a}1 and 100 μg of RNaseA (Qiagen) overnight. An extra 100 units of restriction enzyme was added in the morning incubations were continued for 6 hours. 10 units of calf intestinal phosphatase (CIP) (NEB) was added and incubated for 1 hour at 37 C. DNA was purified using Qiaquick Nucleotide Removal Kit (Qiagen) as per the manufacturer's instructions. 400 ng of eluted DNA fragments were end-labeled with T4 Polynucleotide Kinase (T4 PNK) (NEB) and 10 μCi of [γ^{32-}P]-ATP for 1 hour at 37 C. Labeled fragments were precipitated by the addition of 30 μg of linear polyacrylamide, 1/10 volume of 3 M Sodium Acetate, pH 5.2 and 2.5 volumes of ethanol at left at -80 C for 1 hour. Samples were spun at 14,000 rpm, for 20 minutes at 4 C and washed twice with 70% EtOH at 25 C. Pellets were resuspended in 30 mM Tris, pH 8.9,15 mM $MgCl_2$, 2 mM CaCl, with 10 μg of DNaseI (Worthington) and 10 μg SVPD (Worthington) and incubated for 3 hours at 37 C. 3 μl was spotted on cellulose TLC plates (20 cm x 20 cm, Merck) and developed in isobutyric acid: H_20: NH_3 (66:20:1). Plates were analyzed by phosphorimager scanning using Phosphorimager Storm 860 scanner software.

Preparation of unglucosylated T4 phage DNA for preparation of hm-dCMP standard T4 phage stock was titred by spotting 10 μl of serial 10X dilutions on an LB plate on which 100 μl of an overnight culture of *E.coli* CR63 in 3 ml of T4 top agar was poured and allowed to solidify. The plate was incubated overnight at 37 C. 10 ml of *E.coli* CR63 OD_{600} of 0.5 was infected with a single plaque of T4 phage and incubated with shaking at 37^0C until the culture cleared (about 2.5 hours). The culture was incubated on ice for 10 minutes and then lysed was completed by the addition of several drops of chloroform and gentle mixing. The lysate was titred as described above.

E.coli ER1656 was grown in LB to OD_{600} of 0.5 and then infected with 0.2 phage per bacterium and incubated at 37^0C with shaking until the culture cleared (about 8 hours).. The culture was chilled on ice for 10 minutes and then lysis was completed by the addition of 1 ml of chloroform. DNase I was added to 1 mg/ml and the culture was incubated for 2 hours at 4^0C. The lysate was centrifuged at 12,000g for 10 min at 4 C to pellet debris. The supernatant was collected and phage were pelleted by centrifugation at 23,500g for 1.5 hours at 4 C. The phage pellet was left covered in TE overnight to resuspend. Phage DNA was extracted using an equal volume of phenol, phenol: chloroform: isoamyl alcohol (25:24:1) and chloroform: isoamyl alcohol (24:1). The extracted phage DNA was dialyzed into TE overnight with 2 changes of buffer.

Mass spectrometry experiments. Genomic DNA from HEK293 cells transfected with TET1 wild-type or mutant CD or T4 phage grown in E.coli 13656 were hydrolyzed to dNMP's with SVPD and DNaseI and resolved using TLC. Spots corresponding to paricular dNMP's were scraped, extracted with water, lyophilized, and re-suspended in water for liquid chromatography/mass spectrometry (LC/MS) analysis using an Acquity UPLC/Q-TOF Premier electrospray LC/ESI-MS system (Waters Corp., Milford, MA). Liquid chromatography (LC) was performed with a Waters HSS C18 column (1.0mm i.d. x 50mm, 1.8-um particles) using a linear gradient of 0% to 50% methanol in 0.1% aqueous ammonium formate, pH 6.0. The flow rate was 0.05 mL per min and the eluant was directly injected into the mass spectrometer. Mass spectra were recorded in continuum mode and converted to centroid mode to generate accurate mass spectra. Data was analyzed with Masslynx 4.1 software (Waters).

Recombinant Protein Expression and Purification Bacmid DNA was generated using DH10Bac™ *E. coli* E.coli (Invitrogen) as directed by the manufacturer. Transposition into the correct site was confirmed using PCR. Baculovirus was amplified for three generations using suspension adapted Sf9 cells. Sf9 cells were then infected with baculovirus for 4 days. The resulting cell pellet was kept on ice for 30 minutes in 40 mM Tris, pH 7.4, 300 mM NaCl, 0.2% NP40, 0.4% Triton, 5 mM DTT, 1X protease inhibitors without EDTA (Roche) and then at 12,000 rpm (SLA-TC600), 30 min, 4 C. The supernantant was then incubated with anti-Flag antibody-conjugated beads (Invitrogen) for 5 hours at 4 C. The beads were washed 4 times in 40 mM Tris, pH 7.4, 300 mM NaCl, 0.2% NP40, 8% Glycerol, 1X PI, 5 mM DTT and then eluted in 195 mM Tris, pH 7.4, 110 mM NaCL, 0.14% NP40, 5.8% Glycerol, 0.37X PI, 3.7mM DTT, 365 μg/ml Flag peptide. The homogeneity of the eluted protein was determined using SDS-PAGE followed by Coomassie blue staining and immunoblotting using an anti-Flag antibody (Sigma).

Preparation of double-stranded oligonucleotide substrates

Synthetic oligonucleotides were purchased from IDT. All oligonucleotides were 35 nucleotides in length with the modifications shown below.

F: 5'-CTATACCTCCTCAACTTCGATCACCGTCTCCGGCG-3'

F^{Me}: 5'-CTATACCTCCTCAACTT(mC)GATCACCGTCTCCGGCG-3'

R: 5'-Biotin-CGCCGGAGACGGTGATCGAAGTTGAGGAGGTATAG-3'

R^{Me}: 5'-Biotin-CGCCGGAGACGGTGAT(mC)GAAGTTGAGGAGGTAT AG-3'

Oligonucleotides were annealed to the appropriate complementary oligonucleotide in 100 mM KAc, 30 mM HEPES, pH 7.5. The mixture was boiled for 5 minutes then slowly cooled to room temperature overnight. Double-stranded oligonucleotides were purified by polyacrylamide gel electrophoresis.

In vitro Enzymatic Assays 7.5 μl of recombinant protein (about 3 μg) was incubated with 2 μg of oligonucleotide substrates in 50 mM HEPES, pH 8, 50 mM NaCl, 2 mM Ascorbic Acid, 1mM 2OG, 75 μM FAS (Fe^{2+}), and 1 mM DTT for 3 hours at 37 C. Oligonucleotide substrates were purified using Qiaquick Nucleotide Removal Kit (Qiagen) and then digested with Taq^a1 overnight, treated with CIP for 1 hour and purified once more with Qiaquick Nucleotide Removal Kit (Qiagen). Purified DNA oligonucleotides were end-labeled with T4 Polynucleotide Kinase (NEB) and 10 μCi of γ^{32P}-ATP for 1 hour at 37 C. Labeled fragments were precipitated by the addition of 30 μg of linear polyacrylamide, 1/10 volume of 3 M Sodium Acetate, pH 5.2 and 2.5 volumes of ethanol followed by incubation at -80 C for 1 hour. Samples were spun at 14,000 rpm, for 20 minutes at 4 C in a refrigerated microcentrifuge. Unincorporated radionucleotide was removed by washing two times with 70% EtOH and spinning at room temperature for 10 minutes. Pellets were resuspended in 10 μl 30 mM Tris, 15 mM $MgCl_2$, 2 mM CaCl, pH 8.9 with 10 μg of DNaseI (Worthington) and 10 μg SVPD (Worthington) and incubated for 3 hours at 37 C. 3 μl was spotted on cellulose TLC plates (Merck) and developed in isobutyric acid: water: ammonia (66:20:1). Plates were analyzed by phosphorimager scanning using Phosphorimager Storm 860 scanner software.

ES cell culture V6.5 mouse ES cells were maintained on mitomycin c-inactivated primary mouse embryonic fibroblasts in ES medium containing DMEM (Invitrogen, Carlsbad, CA), 15% ES FBS (Omega Scientific, Tarzana, CA) , 0.1 mM each of nonessential amino acids (Invitrogen), 2 mM L-glutamine (Invitrogen), 0.1 mM β-mercaptoethanol (Invitrogen), 50 units/ml penicillin/streptomycin (Invitrogen) and 1000 U/ml ESGRO® (LIF; Chemicon). For all experiments described, cells were trypsinized and plated for 30 min on standard tissue

culture dishes to remove feeder cells before floating ES cells were collected and re-plated on gelatin-coated dishes or wells. For LIF withdrawal assays, cells were plated at a density of 2-3 x 10^5 cells per 10-cm dish and LIF was removed the day after (day 0). RNA interference (RNAi) experiments were performed as previously described using Dharmacon siGENOME siRNA duplexes (Thermo Fisher Scientific Inc, Boulder, CO) against mouse Tet1 (Cat. # D-062861-01/02). The Dharmacon siGENOME non-targeting siRNA#2 (Cat. # D-001210-02) was used as a negative control. Mouse ES cells were seeded in gelatin-coated 12-well at a density of 1 x 10^5 cells per well and transfected the day after (day 0) with 50 nM siRNA using Lipofectamine RNAiMAX reagent (Invitrogen) according to the manufacturer's instructions. Retransfections were performed on pre-adherant cells at day 2 at a split of 1:4 and finally at day 4 at a split of 1:2 in 6-well plates. Cells were harvested at Day 5 for RNA and thin-layer chromatography analyses.

RNA Isolation, cDNA synthesis and Quantitative Real-Time PCR Total RNA was isolated with an RNeasy kit (Qiagen, Chatsworth, CA) with on-column DNase treatment. cDNA was synthesized with 0.5 mg total RNA using SuperScript III reverse transcriptase (Invitrogen). Quantitative PCR was performed using FastStart Universal SYBR Green Master mix (Roche, Mannheim, Germany) on a StepOnePlus real-time PCR system (Applied Biosystems, Foster City, CA) according to the manufacturer's instructions. The levels of gene expression were normalized to Gapdh. Primer sequences are: Tet1 forward 5'-GAGCCTGTTCCTCGATGTGG-3', Tet1 reverse 5'-CAAACCCACCTGAGGC TGTT-3'; Oct3/4 forward 5'-TCTTTCCACCAGGCCCCCGGCTC-3', Oct3/4 reverse 5'-TGCGGGCGGACATGGGGAGATCC-3', Gapdh forward 5'-GTGTTCCTACCCCCAATG TGT-3', Gapdh reverse 5'-ATTGTCATACCAGGAAATGAGCTT-3'.

References

1. M. B. Lobocka *et al.*, *J Bacteriol* **186**, 7032 (Nov, 2004).
2. M. Hori *et al.*, *Nucleic Acids Res* **31**, 1191 (Feb 15, 2003).
3. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389 (Sep 1, 1997).

4. A. A. Schaffer *et al.*, *Nucleic Acids Res* **29**, 2994 (Jul 15, 2001).

5. T. Lassmann, E. L. Sonnhammer, *Nucleic Acids Res* **34**, W596 (Jul 1, 2006).

6. S. R. Eddy, *Bioinformatics* **14**, 755 (1998).

7. J. Soding, *Bioinformatics* **21**, 951 (Apr 1, 2005).

8. J. Soding, A. Biegert, A. N. Lupas, *Nucleic Acids Res* **33**, W244 (Jul 1, 2005).

9. J. A. Cuff, G. J. Barton, *Proteins* **40**, 502 (Aug 15, 2000).

10. J. C. Wootton, *Comput Chem* **18**, 269 (Sep, 1994).

11. N. Guex, M. C. Peitsch, *Electrophoresis* **18**, 2714 (Dec, 1997).