data have accession numbers AFHZ00000000 (AAA001-B15), AFIB00000000 (AAA001-C10), AFHY00000000 (AAA007-O20), and AFIA00000000 (AAA240-J09). Raw sequences were deposited in the GenBank Short Read Archive under accession numbers SRA029592 and SRA035467 (AAA001-B15), SRA029604 and SRA035394 (AAA001-C10), SRA029593 and SRA035468 (AAA007-O20), and SRA029596 and SRA035470 (AAA240-J09).

Supporting Online Material
www.sciencemag.org/cgi/content/full/333/6047/1296/DC1
Materials and Methods
Figs. S1 to S19
Tables S1 to S15
References

1 February 2011; accepted 13 July 2011
10.1126/science.1203690

Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine

Shinsuke Ito,[1,2]* Li Shen,[1,2]* Qing Dai,[3] Susan C. Wu,[1,2] Leonard B. Collins,[4] James A. Swenberg,[2,4] Chuan He,[3] Yi Zhang[1,2]†

5-methylcytosine (5mC) in DNA plays an important role in gene expression, genomic imprinting, and suppression of transposable elements. 5mC can be converted to 5-hydroxymethylcytosine (5hmC) by the Tet (ten eleven translocation) proteins. Here, we show that, in addition to 5hmC, the Tet proteins can generate 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) from 5mC in an enzymatic activity–dependent manner. Furthermore, we reveal the presence of 5fC and 5caC in genomic DNA of mouse embryonic stem cells and mouse organs. The genomic content of 5hmC, 5fC, and 5caC can be increased or reduced through overexpression or depletion of Tet proteins. Thus, we identify two previously unknown cytosine derivatives in genomic DNA as the products of Tet proteins. Our study raises the possibility that DNA demethylation may occur through Tet-catalyzed oxidation followed by decarboxylation.

Although enzymes that catalyze DNA methylation process are well studied (*1*), how DNA demethylation is achieved is less known, especially in animals (*2*, *3*). A repair-based mechanism is used in DNA demethylation in plants, but whether a similar mechanism is also used in mammalian cells is unclear (*3*, *4*). Identification of hydroxymethylcytosine (5hmC) as the sixth base of the mammalian genome (*5*, *6*) and the capacity of Tet (ten eleven translocation) proteins to convert 5-methylcytosine (5mC) to 5hmC in an Fe(II) and alpha-ketoglutarate (α-KG)–dependent oxidation reaction (*6*, *7*) raised the possibility that a Tet-catalyzed reaction might be part of the DNA demethylation process.

A potential 5mC demethylation mechanism can be envisioned from similar chemistry for thymine-to-uracil conversion (*3*, *8*, *9*) (fig. S1A),

[1]Howard Hughes Medical Institute and Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599–7295, USA. [2]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599–7295, USA. [3]Department of Chemistry and Institute for Biophysical Dynamics, University of Chicago, Chicago, IL 60637, USA. [4]Department of Environmental Sciences and Engineering, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599–7295, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: yi_zhang@med.unc.edu

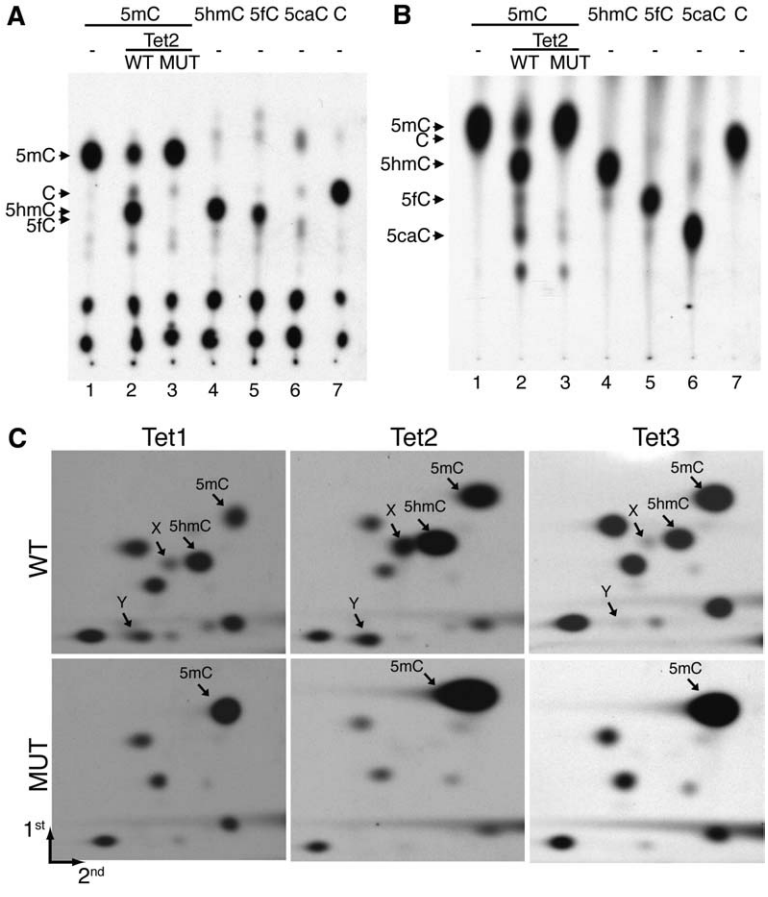

Fig. 1. Optimization of conditions for detection of cytosine and its 5-position modified forms by TLC. (**A**) Migration of labeled C and its 5-position modified forms by TLC under the first developing buffer. Lanes 1 to 3 serve as controls for the migration of 5mC and 5hmC generated from DNA oligos incubated with wild-type (WT) or catalytic mutant (MUT) Tet2. (**B**) The same samples used in (A) were separated by TLC under the second developing buffer. With the exception of 5mC and C, all of the other forms of C can be separated under this condition. (**C**) Autoradiographs of 2D-TLC analysis of samples derived from 5mC-containing TaqI 20-mer oligo DNA incubated with WT and catalytic-deficient mutant Tet1, Tet2, and Tet3.

with the Tet proteins oxidizing 5mC not only to 5hmC, but also to the aldehyde (5fC) and potentially the carboxylic acid (5caC) forms (fig. S1B). The failure to detect such reaction products may simply be due to the limitations of the previous assay used (*6*, *7*). To determine whether this might be the case, we synthesized 20-nucleotide oligomers (20-mers) with 5fC or 5caC in the internal C of an MspI site (*10*) and found that, although MspI is efficient in digesting the oligo DNAs with C, 5mC, or 5hmC in the internal C, it failed to digest the DNA containing 5fC or 5caC (fig. S2, A and B). Thus, if Tet proteins have the capacity to convert 5mC to 5fC or 5caC, these products would have evaded detection because of the inability of MspI to digest 5fC- or 5caC-containing DNA. To overcome this problem, we identified and demonstrated that TaqI

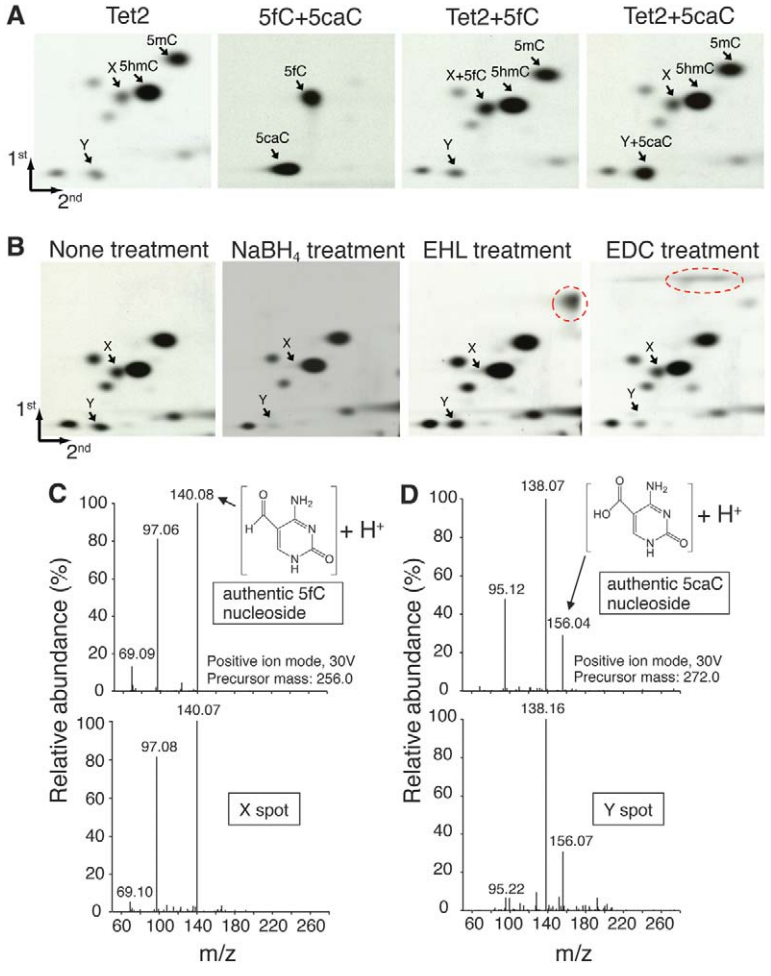

Fig. 2. Tet proteins are capable of converting 5mC to 5hmC, 5fC, and 5caC. (**A**) X and Y comigrate with 5fC and 5caC on 2D-TLC, respectively. Left image shows the migration pattern of the Tet2 reaction mixture on 2D-TLC. The locations of 5mC, 5hmC, X, and Y are indicated. Second image shows the locations of control 5fC and 5caC in a parallel 2D-TLC assay. Third and fourth images contain samples used in the first image plus radioactive 5fC and 5caC, respectively. (**B**) Confirmation of the identities of X and Y by chemical treatments. Left image shows the migration pattern of samples derived from incubation of Tet2 with the 5mC-containing TaqI 20-mer oligo DNA. Second image demonstrates treatment of the samples used in the left image with $NaBH_4$. Third and fourth images demonstrate that EHL and EDC react with the formyl group of 5fC and the carboxyl group of 5caC, respectively, to generate the new products indicated by the dotted circles. (**C**) Mass spectrometric analysis demonstrates that X has the same fingerprint as 5fC. (**D**) Mass spectrometric analysis demonstrates Y has the same fingerprint as 5caC.

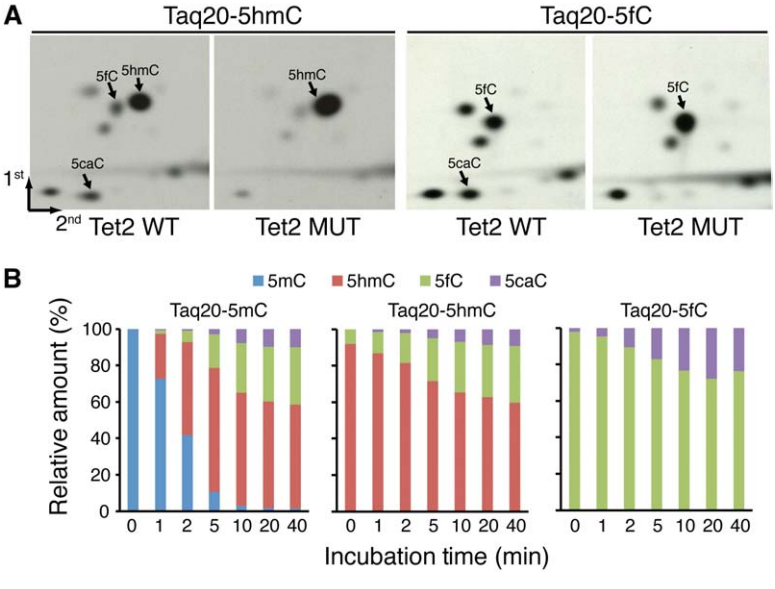

Fig. 3. Kinetic analysis of Tet2 using 5mC-, 5hmC-, and 5fC-containing oligo DNAs. (**A**) Autoradiographs of 2D-TLC analysis of samples derived from 5hmC- or 5fC-containing TaqI 20-mer DNA oligos incubated with WT or catalytic-deficient mutant Tet2. (**B**) Relative percentage of 5mC, 5hmC, 5fC, and 5caC at different time points after incubation of Tet2 proteins with 5mC-, 5hmC-, or 5fC-containing TaqI 20-mer DNA oligos.

is capable of digesting DNA modified with 5mC, 5hmC (*11*), 5fC, or 5caC (fig. S2, C and D).

In addition to restriction enzymes, thin-layer chromatography (TLC) conditions can also affect the detection of 5fC and 5caC. Under previous TLC conditions (*7*), 5hmC and 5fC have almost identical migration patterns (Fig. 1A, lanes 4 and 5), and 5caC failed to migrate (Fig. 1A, lane 6). Using a more acidic TLC buffer, all cytosine derivatives migrated (Fig. 1B, lanes 4 to 7). However, 5mC and C cannot be separated under this condition (Fig. 1B, lanes 1 and 7). Given that the TLC buffer used in Fig. 1A can separate C from 5mC, two-dimensional TLC (2D-TLC) using the two buffer conditions should allow for separation of cytosine and its derivatives.

By using TaqI digestion and 2D-TLC (fig. S3), we analyzed the enzymatic activity of the Tet proteins. Compared with the mutant control, incubation of the Tet1 protein with 5mC-containing substrate resulted in a decrease in the 5mC level concomitant with the appearance of a radioactive spot that correlates with 5hmC (Fig. 1C, left). Two additional radioactive spots, labeled "X" and "Y," whose appearance depend on Tet1 enzymatic activity were observed. Similarly, Tet2 and Tet3 also generated three enzymatic activity–dependent radioactive spots that were detected in Tet1-catalyzed reaction, although the signal that corresponds to the Y spot from the Tet3 reaction is extremely weak (Fig. 1C, middle and right).

If our hypothetical model for DNA demethylation is correct (fig. S1B), the X and Y spots are likely to be 5fC and 5caC. We compared the migration patterns of 5fC and 5caC with those of Tet2-treated 5mC-containing DNA substrates and found that the X and Y spots match 5fC and 5caC with respect to their migration (Fig. 2A, compare the first two images). We further confirmed this by mixing radioactive 5fC (third image) or 5caC (last image) with the samples used in the first image before performing 2D-TLC. To confirm the identities of the X and Y spots, we treated the Tet2-catalyzed reaction mixture with sodium borohydride ($NaBH_4$), which resulted in the disappearance of both X and Y spots concomitant with an increase in 5hmC (Fig. 2B, compare the first two images), indicating that both are oxidation products of 5hmC, consistent with the notion that they are 5fC and 5caC.

O-ethylhydroxylamine hydrochloride (EHL) and 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC) react with formyl and carboxyl groups to generate oximes and amides, respectively (*12*, *13*) (fig. S4). To determine the migration patterns of the reaction products, we performed reactions by using standard 5fC and 5caC and separated the products by 2D-TLC, establishing migration patterns for oxime (fig. S4A) and amide (fig. S4B). Similar EHL treatment of the Tet2 reaction mixture specifically converted the X spot to a new spot that comigrated with oxime (Fig. 2B, compare first and third images, and fig. S4A). In contrast, EDC treatment specifically converted the Y spot to a new signal that comigrated with amide (Fig. 2B, compare first and fourth images, and fig. S4B). To unequivocally define the identities of X and Y, we used mass spectrometry. Having established the mass spectrometry fingerprints of standard 5fC and 5caC (Fig. 2, C and D, top), we extracted the X and Y spots and subjected them to mass spectrometric analysis. The X spot shows the same major fragment ions as that of 5fC, whereas the Y spot shows the same major fragment ions as that of 5caC. Collectively, 2D-TLC comigration, chemical treatment, and mass spectrometry fingerprints demonstrate that Tet proteins not only can convert 5mC to 5hmC but also can further oxidize 5hmC to 5fC and 5caC.

To determine whether Tet proteins can use 5hmC- or 5fC-containing DNA as substrates, 20-mer DNA oligos with either 5hmC or 5fC in the TaqI site were incubated with Tet proteins. The 2D-TLC analysis demonstrated that incubation with wild-type Tet proteins, but not the catalytic mutants, resulted in a decrease in the level of 5hmC or 5fC concomitant with the appearance of 5fC and 5caC or 5caC alone (Fig. 3A and fig. S5), suggesting that Tet proteins can act on 5hmC- and 5fC-containing substrates. However, the 5caC signal generated by Tet3 is extremely weak.

We used a quantitative mass spectrometric assay to rule out the possibility that 5fC and 5caC are generated as a side reaction by Tet proteins. We generated a standard curve for each of the cytosine derivatives by mixing different amounts of each of 5mC, 5hmC, 5fC, and 5caC followed by liquid chromatography–mass spectrometry (LC-MS) (fig. S6). We then quantified the C derivatives at different time points after incubating Tet2 with 5mC-, 5hmC-, or 5fC-containing DNA substrates. Quantification of the relative amount of the substrate and the various products during the reaction process demonstrated that the reaction plateaued after 10 min of incubation regardless of whether 5mC-, 5hmC-, or 5fC-containing TaqI 20-mer DNA was used as a substrate (Fig. 3B). The reaction plateaued in 10 min because of the inactivation of the Tet2 enzyme during the incubation (fig. S7). During this period, Tet2 is able to convert more than 95% of the 5mC to 5hmC (~60%), 5fC (~30%), and 5caC (5%); but it can only convert about 40% or 25% when 5hmC- or 5fC-containing DNA was used as a substrate (Fig. 3B). From this data, we calculated the initial reaction rate of Tet2 for 5mC-, 5hmC-, and 5fC-containing substrates to be 429 nM/min, 87.4 nM/min, and 56.6 nM/min, respectively (fig. S8).

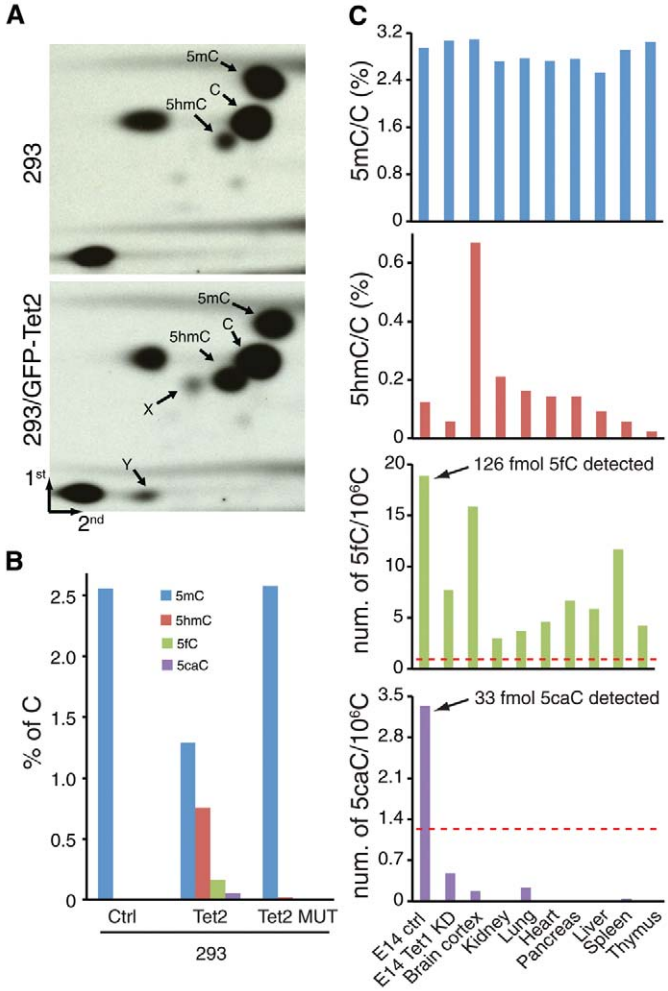

Fig. 4. 5fC and 5caC are present in genomic DNA, and their abundance is regulated by Tet proteins. (**A**) Genomic DNA prepared from either WT HEK293 cells or HEK293 cells expressing a GFP–tagged Tet2 were digested with TaqI, end-labeled with T4 polynucleotide kinase, digested with deoxyribonuclease I I and phosphodiesterase I, and analyzed by 2D-TLC. (**B**) Mass spectrometric quantification of genomic content of 5mC, 5hmC, 5fC, and 5caC relative to cytosine in HEK293 cells expressing GFP-tagged WT or a catalytic mutant Tet2. (**C**) Mass spectrometric quantification of genomic content of 5mC, 5hmC, 5fC, and 5caC relative to cytosine in mouse ES cells (E14 ctrl), Tet1 knockdown ES cells (E14 Tet1 KD), and mouse organs. The data represent the average of two independent experiments (table S1). The red dotted lines indicate the limits for accurate quantification, which are 0.8 5fC per 10^6 C and 1.2 5caC per 10^6 C in 20 μg of genomic DNA.

Although Tet2 has a preference for the 5mC-containing DNA substrate, its initial reaction rate for 5hmC- and 5fC-containing substrate is only 4.9-to 7.6-fold lower. The fact that there is an accumulation of 5fC and 5caC when 5mC is used as a substrate (Fig. 3B, left) strongly suggests that Tet-catalyzed iterative oxidation is likely a kinetically relevant pathway.

To determine whether Tet-catalyzed iterative oxidation of 5mC can take place in vivo, we transfected a mammalian expression construct containing the Tet2 catalytic domain fused to green fluorescent protein (GFP) into human embryonic kidney (HEK) 293 cells. After fluorescence-activated cell sorting, genomic DNA of GFP-positive cells was analyzed for the presence of 5hmC, 5fC, and 5caC by 2D-TLC (fig. S3). Compared with the untransfected control, cells expressing Tet2 not only have increased 5hmC levels but also contain two additional spots (Fig. 4A), which correspond to 5fC and 5caC, respectively. In addition, we quantified the genomic content of 5hmC, 5fC, and 5caC following the procedure depicted in fig. S9A (*14*). After establishing the retention times for each of the cytosine derivatives on high-performance liquid chromatography (HPLC) (fig. S9B, top), nucleosides derived from genomic DNA were subjected to the same HPLC conditions for fractionation. Fractions A and B (fig. S9B) that had the same retention times as that of 5caC and 5hmC or 5fC were collected. Mass spectrometry analysis demonstrates that both 5fC and 5caC are detected in the genomic DNA of cells overexpressing Tet2 (fig. S10A). By comparison to the standard curves (fig. S11A), overexpression of wild-type Tet2, but not a catalytic mutant, increased the genomic content of 5hmC, 5fC, and 5caC (Fig. 4B).

Next, we asked whether 5fC and 5caC are present in genomic DNA under physiological conditions. By using a similar approach as that used for the genomic DNA of Tet2-overexpressing HEK293, we show that not only 5hmC, but also 5fC and 5caC are present in the genomic DNA of mouse embryonic stem (ES) cells (fig. S10B). To quantify the genomic content of 5hmC, 5fC, and 5caC in mouse ES cells, we generated standard curves for each of the 5mC derivatives at low concentrations and determined the limit of detection for 5fC and 5caC to be 5 fmol and 10 fmol, respectively (fig. S11). We then quantified the genomic content of these cytosine derivatives in mouse ES cells to be about 1.3×10^3 5hmC, 20 5fC, and 3 5caC in every 10^6 C (Fig. 4C and table S1). Knockdown of Tet1 reduced the genomic content of 5hmC as well as 5fC and 5caC (Fig. 4C), indicating that Tet1 is at least partially responsible for the generation of these cytosine derivatives. The presence of 5fC is not limited to ES cells, because similar analysis also revealed their presence in genomic DNA of major mouse organs (Fig. 4C). However, 5caC can be detected with confidence only in ES cells (Fig. 4C and fig. S10B).

Here, we demonstrate that the Tet family of proteins have the capacity to convert 5mC not only to 5hmC, but also to 5fC and 5caC in vitro. In addition, we provide evidence for the presence of 5fC in the genomic DNA of mouse ES cells and organs and the presence of 5caC in mouse ES cells. We note that a similar study failed to detect their existence in genomic DNA of mouse organs (*15*) likely because of the differences in the detection limits between the two studies (pmol versus fmol). The Tet-catalyzed oxidation reaction is reminiscent of the thymine hydroxylase–catalyzed conversion of thymine to iso-orotate (*8*, *9*) (fig. S1), raising the possibility that 5mC demethylation could be potentially achieved through a process similar to the conversion of thymine to uracil, which is achieved by conversion of thymine to iso-orotate followed by decarboxylation by the iso-orotate decarboxylase (*8*, *9*). Although this hypothetical pathway for DNA demethylation is simple and appealing, the enzyme that is capable of decarboxylating 5caC-containing DNA has yet to be identified. Until such an enzyme is identified, we cannot rule out the possibility that the Tet family enzymes act together with other putative DNA demethylation pathways, such as the base-excision DNA repair pathway. Indeed, recent studies have provided some supporting evidence for such a possibility (*16*, *17*).

Note added in proof: A related paper appeared during revision of this paper: T. Pfaffeneder *et al.*, The discovery of 5-formylcytosine in embryonic stem cell DNA. *Angew. Chem. Int. Ed. Engl.* **50**, 7008 (2011); published online 30 June 2011.

References and Notes

1. M. G. Goll, T. H. Bestor, *Annu. Rev. Biochem.* **74**, 481 (2005).
2. S. K. Ooi, T. H. Bestor, *Cell* **133**, 1145 (2008).
3. S. C. Wu, Y. Zhang, *Nat. Rev. Mol. Cell Biol.* **11**, 607 (2010).
4. M. Gehring, W. Reik, S. Henikoff, *Trends Genet.* **25**, 82 (2009).
5. S. Kriaucionis, N. Heintz, *Science* **324**, 929 (2009); 10.1126/science.1169786.
6. M. Tahiliani *et al.*, *Science* **324**, 930 (2009); 10.1126/science.1170116.
7. S. Ito *et al.*, *Nature* **466**, 1129 (2010).
8. J. W. Neidigh, A. Darwanto, A. A. Williams, N. R. Wall, L. C. Sowers, *Chem. Res. Toxicol.* **22**, 885 (2009).
9. J. A. Smiley, M. Kundracik, D. A. Landfried, V. R. Barnes Sr., A. A. Axhemi, *Biochim. Biophys. Acta* **1723**, 256 (2005).
10. Q. Dai, C. He, *Org. Lett.* **13**, 3446 (2011).
11. L. H. Huang, C. M. Farnet, K. C. Ehrlich, M. Ehrlich, *Nucleic Acids Res.* **10**, 1579 (1982).
12. V. Y. Kukushkin, A. J. L. Pombeiro, *Coord. Chem. Rev.* **181**, 147 (1999).
13. A. Williams, S. V. Hill, I. T. Ibrahim, *Anal. Biochem.* **114**, 173 (1981).
14. G. Boysen *et al.*, *J. Chromatogr. B* **878**, 375 (2010).
15. D. Globisch *et al.*, *PLoS ONE* **5**, e15367 (2010).
16. S. Cortellino *et al.*, *Cell* **146**, 67 (2011).
17. J. U. Guo, Y. Su, C. Zhong, G. L. Ming, H. Song, *Cell* **145**, 423 (2011).

Acknowledgments: We thank Q. Zhang for suggestion of the $NaBH_4$ experiment and C. X. Song for help in oligo purification. This work was supported by NIH grants GM68804 (Y.Z.), GM071440 (C.H.), and P42ES5948 and P30ES10126 (J.A.S.). S.I. is a research fellow of the Japan Society for the Promotion of Science. Y.Z. is an Investigator of the Howard Hughes Medical Institute.

Supporting Online Material

www.sciencemag.org/cgi/content/full/science.1210597/DC1
Materials and Methods
Figs. S1 to S11
Table S1
References (*18*–*20*)

1 July 2011; accepted 11 July 2011
Published online 21 July 2011;
10.1126/science.1210597

Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA

Yu-Fei He,[1]* Bin-Zhong Li,[1]* Zheng Li,[1] Peng Liu,[1] Yang Wang,[1] Qingyu Tang,[2] Jianping Ding,[2] Yingying Jia,[2] Zhangcheng Chen,[2] Lin Li,[2] Yan Sun,[3] Xiuxue Li,[3] Qing Dai,[4] Chun-Xiao Song,[4] Kangling Zhang,[5] Chuan He,[4] Guo-Liang Xu[1]†

The prevalent DNA modification in higher organisms is the methylation of cytosine to 5-methylcytosine (5mC), which is partially converted to 5-hydroxymethylcytosine (5hmC) by the Tet (ten eleven translocation) family of dioxygenases. Despite their importance in epigenetic regulation, it is unclear how these cytosine modifications are reversed. Here, we demonstrate that 5mC and 5hmC in DNA are oxidized to 5-carboxylcytosine (5caC) by Tet dioxygenases in vitro and in cultured cells. 5caC is specifically recognized and excised by thymine-DNA glycosylase (TDG). Depletion of TDG in mouse embryonic stem cells leads to accumulation of 5caC to a readily detectable level. These data suggest that oxidation of 5mC by Tet proteins followed by TDG-mediated base excision of 5caC constitutes a pathway for active DNA demethylation.

Cytosine methylation is directly involved in the modulation of transcriptional activity and other genome functions (*1*), and DNA demethylation therefore plays important roles in transcriptional activation of silenced genes (*2*, *3*). Multiple mechanisms have been proposed to achieve DNA demethylation in mammals, which include direct removal of the exocyclic methyl group from the cytosine via C-C bond cleavage, enzymatic removal of the