Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale

Bo Xia[1,2,8], Dali Han[3,4,8], Xingyu Lu[3,4,8], Zhaozhu Sun[1,2], Ankun Zhou[1,2], Qiangzong Yin[5], Hu Zeng[1,2], Menghao Liu[1,2], Xiang Jiang[1,2], Wei Xie[5], Chuan He[3,4,6,7] & Chengqi Yi[1,2,6,7]

Active DNA demethylation in mammals involves oxidation of 5-methylcytosine (5mC) to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC). However, genome-wide detection of 5fC at single-base resolution remains challenging. Here we present fC-CET, a bisulfite-free method for whole-genome analysis of 5fC based on selective chemical labeling of 5fC and subsequent C-to-T transition during PCR. Base-resolution 5fC maps showed limited overlap with 5hmC, with 5fC-marked regions more active than 5hmC-marked ones.

Discovery of the ten-eleven translocation (TET)-dependent generation and removal of oxidized derivatives of 5mC, namely, 5hmC, 5fC and 5caC, revealed a new paradigm of active DNA demethylation in mammalian genomes[1–3]. Besides acting as demethylation intermediates, these oxidized variants of 5mC may also have functional roles[4]. Emerging evidence suggests that 5hmC is a stable epigenetic modification implicated in many biological processes and various diseases[4,5]. 5fC and 5caC accumulate at distal regulatory elements[6–8] and can be removed through base-excision repair by mammalian thymine DNA glycosylase (TDG)[3,9].

5fC is found in many cell types and all major organs[2,10], but it is present at a level of 0.02% to 0.002% of cytosines, approximately 10-fold to 100-fold lower than that of 5hmC[2,10]. Therefore, highly sensitive and selective methods are required for genome-wide detection of this derivative. We and others have developed chemical-, enzyme- or antibody-based methods for enrichment of 5fC in genomic DNA[6–8], but such affinity-based approaches have only limited resolution. More recent base-resolution methods all rely on harsh bisulfite treatment for effective deamination of 5fC[8,11–13], which can result in further DNA degradation[14]. Whole-genome mapping of 5fC using bisulfite-based methods requires an unusually high sequencing depth and thus is cost-prohibitive[15,16]. So far, 5fC has been profiled only in partial genomes of wild-type mouse embryonic stem cells (mESCs)[11,13].

Here we present a bisulfite-free method that detects whole-genome 5fC signals in mESCs at single-base resolution. We screened for chemicals that could form intramolecular cyclization with 5fC, similar to the Friedländer synthesis that uses 2-aminobenzaldehyde and ketones to form quinoline derivatives (**Supplementary Fig. 1**). We successfully identified several chemicals that reacted readily with 5fC (**Supplementary Figs. 2** and **3** and **Supplementary Note 1**) and formed the intended cyclization products involving the exocyclic amino group of 5fC; such products are read as C during PCR amplification and were not informative for this study (**Supplementary Fig. 4a**). One adduct between 5fC and 1,3-indandione (5fC-I) is read as a T instead of a C during PCR (**Supplementary Fig. 4b–e**), because the original 4′ amino group of 5fC is no longer a competent proton donor in 5fC-I and thus may fail to form a canonical base pair with dG (**Supplementary Fig. 5** and **Supplementary Note 2**). Although the mechanism of the C-to-T transition awaits future investigation, we proposed that such a transition could be used as a direct readout of 5fC.

To enrich 5fC-containing genomic DNA, we synthesized an azido derivative of 1,3-indandione (AI) (**Supplementary Note 3**). AI completely converted 5fC to the 5fC-AI adduct under very mild conditions, without causing detectable DNA degradation, and hence allowing high DNA recovery (**Fig. 1a** and **Supplementary Fig. 6**). The reaction was also highly selective for 5fC among all modified cytosines (**Supplementary Fig. 7**). We then coupled a cleavable biotin to the AI-labeled 5fC via click chemistry (**Fig. 1a** and **Supplementary Fig. 8**). We also screened different polymerases to minimize PCR bias and washed away DNA strands that did not contain 5fC, aiming to maximize the C-to-T signals in the sequencing reads (**Supplementary Fig. 9**). Such cyclization-enabled C-to-T transition of 5fC (fC-CET) was used from the sequencing reads to obtain genome-wide maps of 5fC at single-base resolution (**Fig. 1b**).

We used several spike-in DNA sequences (**Supplementary Table 1**) to confirm the specificity and sensitivity of fC-CET by quantitative PCR (qPCR). The results proved that AI showed no cross-reactivity to C, 5mC, 5hmC or 5caC (**Fig. 1c**). Moreover, our chemical-assisted pulldown demonstrated efficient

[1]State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing, China. [2]Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China. [3]Department of Chemistry and Institute for Biophysical Dynamics, The University of Chicago, Chicago, Illinois, USA. [4]Howard Hughes Medical Institute, The University of Chicago, Chicago, Illinois, USA. [5]Tsinghua-Peking Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing, China. [6]Synthetic and Functional Biomolecules Center, College of Chemistry and Molecular Engineering, Peking University, Beijing, China. [7]Department of Chemical Biology, College of Chemistry and Molecular Engineering, Peking University, Beijing, China. [8]These authors contributed equally to this work. Correspondence should be addressed to C.Y. (chengqi.yi@pku.edu.cn) or C.H. (chuanhe@uchicago.edu).

RECEIVED 22 FEBRUARY; ACCEPTED 20 JULY; PUBLISHED ONLINE 7 SEPTEMBER 2015; CORRECTED ONLINE 21 SEPTEMBER 2015 (DETAILS ONLINE); DOI:10.1038/NMETH.3569

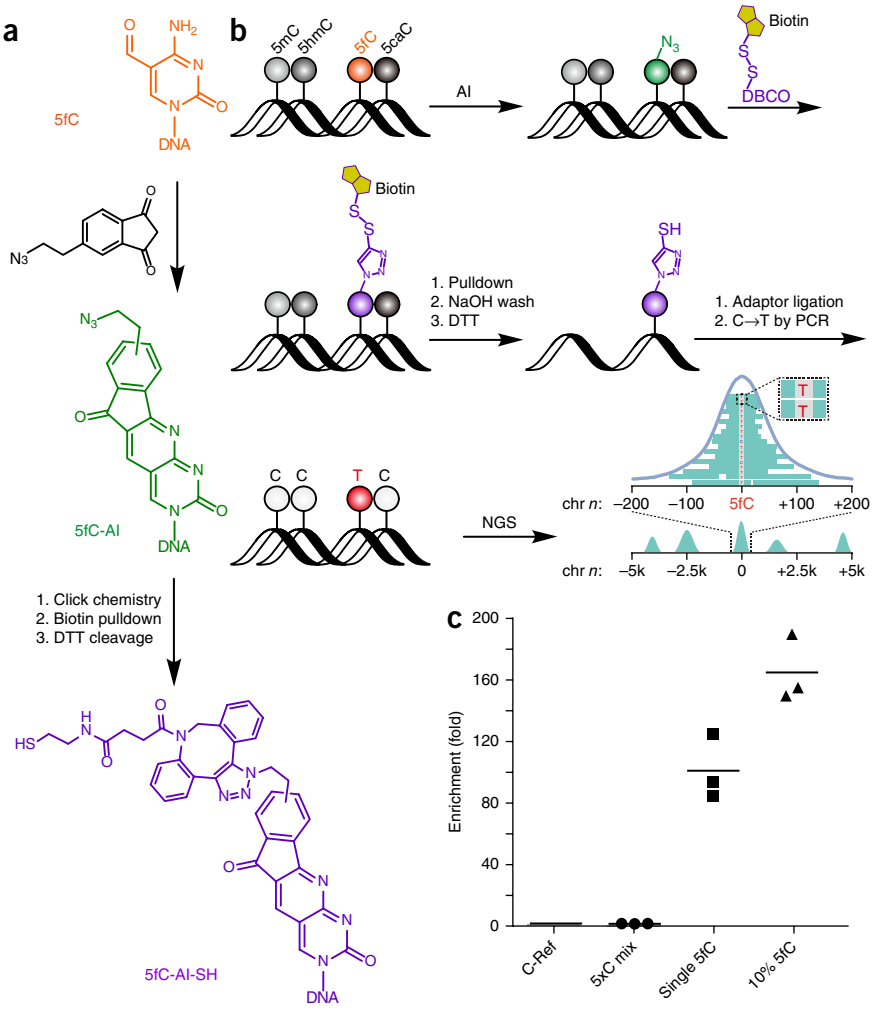

Figure 1 | Cyclization labeling of 5fC and fC-CET. (**a**) AI-mediated cyclization labeling of 5fC and subsequent conjugation of biotin via click chemistry. (**b**) Schematic diagram of fC-CET. Genomic DNA is sequentially labeled with AI, conjugated to biotin for pulldown enrichment and ligated to adaptors for high-throughput sequencing. C-to-T transitions are specifically searched for to define 5fC sites in the whole genome. NGS, next-generation sequencing. (**c**) Enrichment of spike-in sequences by qPCR. Values represent fold enrichment over the input (n = 3), normalized to the C-Ref sequence (reference with only regular C's). 5xC mix, PCR-amplified DNA with 70% dCTP, 15% dmCTP, 10% dhmCTP and 5% dcaCTP; single 5fC, synthetic DNA with a single 5fC site; 10% 5fC, PCR-amplified DNA with 10% dfCTP (within dCTP).

enrichment for 5fC-containing DNAs (**Fig. 1c**); even for sequences with only a single 5fC, fC-CET enriched the sequence by ~100-fold, with little of the density bias commonly associated with antibody-based strategies.

We then applied fC-CET to wild-type ($Tdg^{fl/fl}$) and *Tdg*-null ($Tdg^{-/-}$) mESCs and readily identified 29,501 and 77,750 5fC-enriched regions, respectively (**Fig. 2a** and **Supplementary Figs. 10** and **11a**,**b**). The majority of the 5fC-enriched regions in $Tdg^{fl/fl}$ mESCs fell within those in $Tdg^{-/-}$ mESCs (**Supplementary Fig. 11c**), confirming extensive TDG-dependent active DNA demethylation. 5fC-enriched regions detected by fC-CET were also in good agreement with those detected by fC-Seal[8] (**Supplementary Fig. 11d**,**e**). Additionally, we compared 5fC-enriched regions with 5hmC-enriched regions[16] and found that ~71.8% of 5fC-enriched regions overlapped with 5hmC-enriched regions (**Fig. 2b**), consistent with our previous observations[8].

We next sought to create a base-resolution map of 5fC in the whole genome of mESCs. Requiring positive hits in both replicates, we identified 32,685 and 139,027 high-confidence 5fC sites in $Tdg^{fl/fl}$ and $Tdg^{-/-}$ mESCs, respectively (**Supplementary Fig. 12a**). Previous base-resolution maps of 5fC were obtained either by reduced-representation bisulfite sequencing from wild-type mESCs or by whole-genome bisulfite sequencing from $Tdg^{-/-}$ mESCs[14,17]. In comparison, fC-CET readily identified a comprehensive view of 5fC in the whole genome of mESCs. In the *Nanog* gene (**Fig. 2a**),

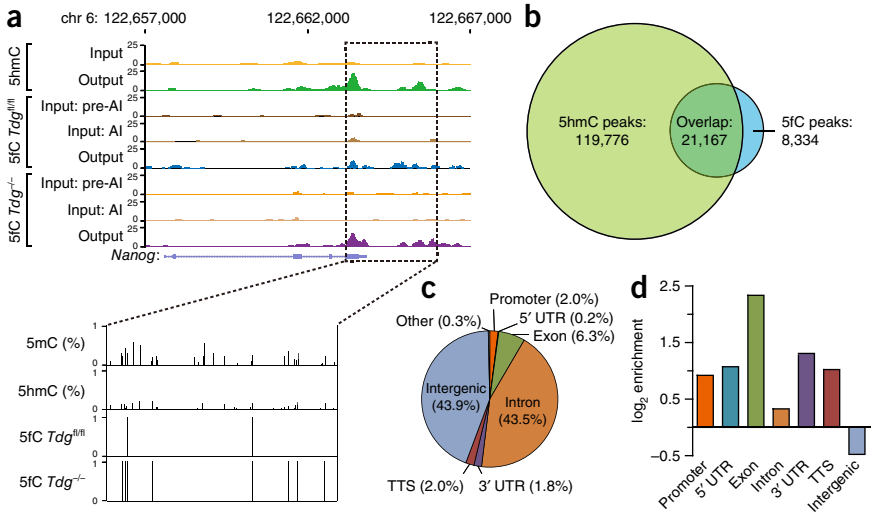

Figure 2 | fC-CET reveals base-resolution 5fC maps in the whole genome. (**a**) Genome browser view of representative 5fC-enriched regions in the *Nanog* gene. The data shown here represent results from two biological replicates. Results from hmC-Seal in the same region are also plotted for comparison. (**b**) Venn diagram showing that fC-CET–detected 5fC-marked regions largely overlap with hmC-Seal–detected 5hmC regions. (**c**,**d**) Overall distribution of 5fC sites in genomic elements of wild-type mESCs (**c**) and their relative enrichment (**d**). TTS, transcription termination site.

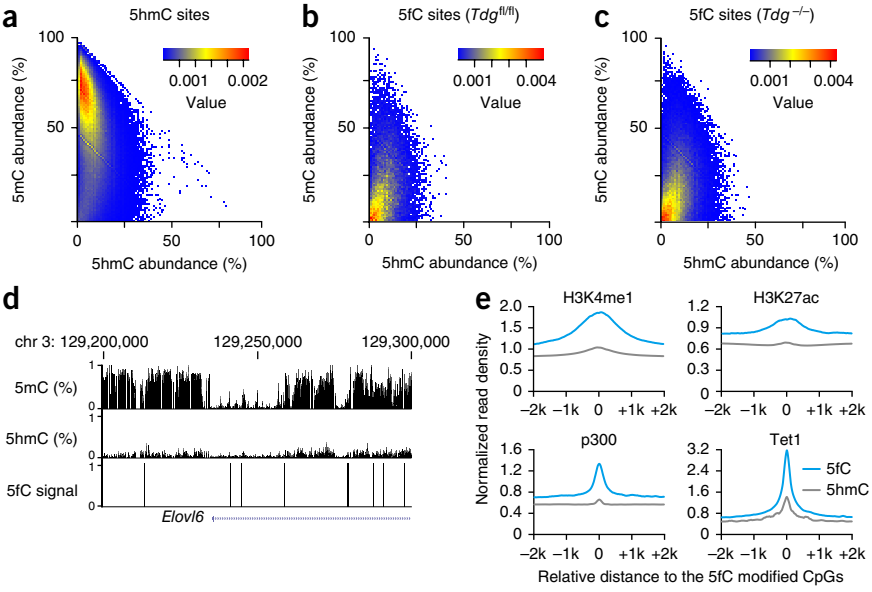

Figure 3 | 5fC represents a more active marker than 5hmC. (**a**–**c**) Heat maps of the abundance of 5hmC (horizontal axis) and 5mC (longitudinal axis) for the TAB-Seq–detected 5hmC sites (**a**) and fC-CET–detected 5fC sites in wild-type (**b**) and *Tdg*$^{-/-}$ (**c**) mESCs. (**d**) A representative view comparing 5fC-marked sites with 5mC and 5hmC abundance. 5mC and 5hmC data are shown as the mean of two biological replicates; 5fC data represent results from two biological replicates. (**e**) Normalized read densities of 5fC (blue, fC-CET) and 5hmC (gray, hmC-Seal) at H3K4me1-, H3K27ac-, p300- and Tet1-enriched regions in wild-type mESCs.

fC-CET detected 5fC sites in *Tdg*$^{fl/fl}$ mESCs and in *Tdg*$^{-/-}$ mESCs, all of which were also previously identified as 5mC sites[18]. Moreover, 5fC-enriched regions could contain one or multiple 5fC sites (**Fig. 2a** and **Supplementary Fig. 10b**), demonstrating the sensitivity of fC-CET in detecting 5fC in both loosely and densely modified regions. In *Tdg*$^{fl/fl}$ mESCs, a large fraction of 5fC sites were located in intragenic regions, with particular enrichment in exons (**Fig. 2c**,**d**). A similar pattern of 5fC distribution was observed in the *Tdg*$^{-/-}$ mESCs (**Supplementary Fig. 12b**,**c**). Furthermore, we selected nine 5fC sites for locus-specific validation and validated five and nine 5fC sites by formyl chemically assisted bisulfite sequencing (fCAB-Seq) and fC-CET, respectively (**Supplementary Fig. 13** and **Supplementary Table 2**).

With base-resolution maps of 5fC and TET-assisted bisulfite sequencing (TAB-Seq)-detected 5hmC[18], we next sought to investigate their spatial relationship. Although 5fC- and 5hmC-enriched regions largely overlapped (**Fig. 2b**), 5fC and 5hmC sites shared limited overlap on the single-base level: only ~22.2% of 5fC sites (7,249 out of 32,685) were previously identified as 5hmC. These observations suggest that 5fC and 5hmC sites may have different steady-state features and that the degree of TET-mediated oxidation reactions may be subject to further regulation. Given that 5fC and 5hmC may be recognized by different reader proteins[19,20], the limited overlap between 5fC and 5hmC sites on the single-base level further hints at their different biological roles.

To characterize the relationship of 5fC with 5mC and 5hmC, we calculated the abundance of 5hmC and 5mC at the TAB-Seq–detected 5hmC sites (**Fig. 3a**) and fC-CET–detected 5fC sites (**Fig. 3b**,**c**). On the single-base level, 5hmC sites showed high levels of 5mC (**Fig. 3a**). Previous profiling results showed a decrease of 5mC abundance in 5fC-marked regions[7,8]. We found that on the single-base level, the abundance of 5mC in 5fC sites was indeed very low (**Fig. 3a**,**b**,**d**) (mean value of 18.67% for *Tdg*$^{fl/fl}$ mESCs, compared with 54.56% for 5hmC-marked sites). Observations were similar for the *Tdg*$^{-/-}$ mESCs (mean value of 18.78%) (**Fig. 3c**). The markedly lower abundance of 5mC in the 5fC-occupied sites suggests that 5fC-marked genomic regions may be more active than 5hmC-marked regions.

We calculated the chromatin-immunoprecipitation sequencing (ChIP-Seq) signals of active histone-modification markers H3K4me1 and H3K27ac at the corresponding genomic regions. Enhancer marker H3K4me1 exhibited enrichment for both 5hmC and 5fC, whereas active enhancer marker H3K27ac was enriched for 5fC but exhibited only weak signals for 5hmC[8]. In fact, signals for both H3K4me1 and H3K27ac were much higher for 5fC than for 5hmC (**Fig. 3e** and **Supplementary Fig. 14**). Moreover, compared with the 5hmC regions, 5fC-marked regions were more enriched for the transcriptional coactivator p300, Tet1, and DNase I hypersensitive regions, although CTCF-bound regions were similarly enriched (**Fig. 3e** and **Supplementary Fig. 15**). Taken together, our results show that 5fC marks distinct regulatory elements and represents a more active marker than 5hmC.

fC-CET uses selective chemical labeling to achieve bisulfite-free, base-resolution sequencing of 5fC at the genome scale. Because it demonstrates no noticeable DNA degradation, fC-CET also has potential for analyses of precious DNA, including clinical samples, in addition to the applications explored in this study. Furthermore, if selective conversion of 5mC or 5hmC to 5fC is combined with fC-CET, this bisulfite-free method could have wider applications in epigenome sequencing.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Sequencing data have been deposited into the Gene Expression Omnibus (GEO) under accession number GSE66144.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

The authors thank R. Meng, S.T. Huang, J.Y. Liu, J.Y. Li, X.T. Shu, X.Y. Li and C.X. Zhu for technical assistance; X.X. Zhang and H.S. Guo (Peking University, Beijing, China) for providing genomic DNA at the beginning of the project; C.F. Xia for synthetic suggestions; and O. Stovicek for editing the manuscript. This work was supported by the National Basic Research Foundation of China (grant 2014CB964900 to C.Y.), the National Natural Science Foundation of China (grants 31270838 and 21472009 to C.Y.), and the US National Institutes of Health (grant R01 HG006827 to C.H.). C.H. is supported by the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

B.X. and C.Y. conceived the original idea and designed the experiments with the help of X.L. and C.H.; B.X. performed the experiments with the help of X.L., H.Z., M.L. and X.J.; D.H. performed bioinformatics analysis; Z.S. and A.Z. synthesized the chemicals; Q.Y. and W.X. helped with the library preparation; C.H. and C.Y. supervised the project; and B.X. and C.Y. wrote the manuscript with contributions from D.H., X.L. and C.H.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Tahiliani, M. *et al. Science* **324**, 930–935 (2009).
2. Ito, S. *et al. Science* **333**, 1300–1303 (2011).
3. He, Y.F. *et al. Science* **333**, 1303–1307 (2011).
4. Song, C.X. & He, C. *Trends Biochem. Sci.* **38**, 480–484 (2013).
5. Bachman, M. *et al. Nat. Chem.* **6**, 1049–1055 (2014).
6. Raiber, E.A. *et al. Genome Biol.* **13**, R69 (2012).
7. Shen, L. *et al. Cell* **153**, 692–706 (2013).
8. Song, C.X. *et al. Cell* **153**, 678–691 (2013).
9. Maiti, A. & Drohat, A.C. *J. Biol. Chem.* **286**, 35334–35338 (2011).
10. Pfaffeneder, T. *et al. Angew. Chem. Int. Ed. Engl.* **50**, 7008–7012 (2011).
11. Booth, M.J., Marsico, G., Bachman, M., Beraldi, D. & Balasubramanian, S. *Nat. Chem.* **6**, 435–440 (2014).
12. Lu, X. *et al. Cell Res.* **25**, 386–389 (2015).
13. Wu, H., Wu, X., Shen, L. & Zhang, Y. *Nat. Biotechnol.* **32**, 1231–1240 (2014).
14. Booth, M.J. *et al. Science* **336**, 934–937 (2012).
15. Rivera, C.M. & Ren, B. *Cell* **155**, 39–55 (2013).
16. Neri, F. *et al. Cell Rep.* **10**, 674–683 (2015).
17. Song, C.X. *et al. Nat. Biotechnol.* **29**, 68–72 (2011).
18. Yu, M. *et al. Cell* **149**, 1368–1380 (2012).
19. Iurlaro, M. *et al. Genome Biol.* **14**, R119 (2013).
20. Spruijt, C.G. *et al. Cell* **152**, 1146–1159 (2013).

ONLINE METHODS

Oligonucleotide synthesis and model DNA preparation. Oligonucleotides containing 5fC, 5mC, 5hmC or 5caC were synthesized using the ABI Expedite 8909 nucleic acid synthesizer. The modified nucleotides were site-specifically incorporated at the desired positions (**Supplementary Table 1**) with commercially available phosphoramidites (Glen Research). Subsequent deprotection and purification were carried out with Glen-Pak cartridges (Glen Research) according to the manufacturer's instructions. Purified oligonucleotides were characterized by matrix-assisted laser desorption/ionization–time of flight (MALDI-TOF) (<40-mer). Regular oligonucleotides (and PCR primers) were purchased from Sangon Biotech.

Long duplex DNAs (**Supplementary Tables 1** and **2**) were prepared through ligation of short duplexes (20–40 bp) with sticky overhangs[21]. In brief, the ligation-site oligonucleotides were phosphorylated with T4 polynucleotide kinase (New England BioLabs) and then annealed with the corresponding complementary strands. Annealed duplexes with sticky overhangs were mixed and ligated with T7 DNA ligase (New England BioLabs) at 16 °C for 4 h and then purified by native PAGE (10%).

10% 5fC dsDNA and 5xC-mix dsDNA used in the qPCR assay were prepared through PCR amplification as previously described[18]. All modified dCTPs were purchased from Trilink.

Cell lines and genomic DNA. To generate the *Tdg*$^{-/-}$ mESCs, we transferred the nucleus of a *Tdg*$^{-/-}$ induced pluripotent stem cell into an enucleated oocyte to produce a *Tdg*$^{-/-}$ embryo. The *Tdg*$^{-/-}$ mESC was then derived from the inner cell mass of the *Tdg*$^{-/-}$ embryo. Wild-type (*Tdg*$^{fl/fl}$) mESCs were prepared in parallel. The genomic DNA was prepared by SDS–proteinase K digestion followed by phenol-chloroform extraction and ethanol precipitation.

5fC cyclization labeling and click chemistry. Typically, 5fC cyclization-labeling chemicals can be divided into two groups (**Supplementary Fig. 1**). For 1,3-indandione (J&K) and AI (self-synthesized), the reaction was performed in a suspension of 1,3-indandione or AI in 100 mM MES buffer (pH 6.0). For diethyl malonate (J&K), methyl-ethyl acetoacetate (J&K) or ethyl 6-azido-3-oxohexanoate (self-synthesized), the reaction was performed in 100 mM NaOH methanol solution with 100 mmol of the corresponding chemical. We used 4 μg oligonucleotide or model DNA per 100-μL reaction and incubated the reaction mixture at 37 °C for 24 h in an Eppendorf tube in a thermomixer (Eppendorf, 850 r.p.m.). After the reaction, we used ethanol precipitation to purify the short DNAs with the help of glycogen (Invitrogen), whereas genomic DNA samples were purified with the QIAquick PCR purification kit (Qiagen). For click chemistry we added DBCO-S-S-PEG_3-biotin (Click Chemistry Tools, A112-10) to a final concentration of 400 mM and incubated the mixture at 37 °C for 2 h. Purification steps were performed with the QIAquick PCR purification kit.

We applied the precipitated DNA to Micro Bio-Spin P-6 gel columns (Bio-Rad) to remove any additional chemicals. Products were characterized with MALDI-TOF, and the 1,3-indandione and AI reaction products were enzymatically digested to nucleosides and further analyzed with HPLC[17].

Sanger sequencing and TOPO cloning tests on model DNA. Chemically labeled model DNA was prepared as described above. Bisulfite treatment was performed with an EpiTect Fast bisulfite conversion kit (Qiagen) according to the manufacturer's instructions. PCR amplification was performed under common reaction conditions (Model-F and Model-Seq-R)[18], except for the bisulfite-treated products, which were amplified (Model-BS-F and Model-Seq-R) with Hotstar Taq polymerase (Qiagen). PCR products were purified with the QIAquick PCR purification kit (Qiagen) and Sanger-sequenced with unified sequencing primer, or they were used directly in TOPO cloning tests with the pEASY-T5 Zero cloning kit (TransGen) according to the manufacturer's instructions. Oligonucleotides and primers are presented in **Supplementary Table 1**.

Single-nucleotide primer extension assay. Templates and primer (**Supplementary Table 1**) were adapted from ref. 22. Primer was labeled with [γ^{32}P]-ATP according to the standard protocol. For each reaction, 100 nM of ^{32}P-primer, 130 nM of template, 800 μM of one of the dNTPs and 0.5 U of *Bsu* DNA polymerase large fragment (New England BioLabs) were used in 10 μL of 1× NEB buffer 2 (50 mM NaCl, 10 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM DTT, pH 7.9). Reaction mixtures were incubated at 37 °C for different times and quenched by the addition of 22.5 μL of stop solution (20 mM EDTA, 80% (vol/vol) formamide, 0.25% (wt/vol) xylene cyanol and 0.25% (wt/vol) bromophenol blue). The quenched reaction mixtures were analyzed by 12% denaturing PAGE containing 8 M urea. Gels were visualized by phosphoimaging on a Typhoon FLA 7000 biomolecular imager (GE Healthcare).

AI-mediated enrichment of 5fC-containing DNA. AI-mediated labeling and purification are described above. Typically, 2 μg of model DNA or fragmented genomic DNA (~100–400 bp, with dsDNA Fragmentase (New England BioLabs)) was used per reaction. Dynabeads MyOne Streptavidin C1 (Invitrogen) was used to pull down the biotin-labeled DNA with minor modifications to the suggested immobilizing procedure for nucleic acids. Specifically, 1× binding and washing (B&W) buffer (pH 7.5) was added with 0.1% Tween-20, and the canonical washing step was repeated five times and followed by washing with 50 μL of 1× SSC buffer (pH 7.0). Then the beads were resuspended and incubated at room temperature for 10 min in freshly prepared 0.15 M NaOH. Beads with biotinylated DNA strands were then sequentially washed once with 50 μL of 0.1 M NaOH, once with 50 μL of 1× B&W buffer and once with 50 μL of 1× TE buffer (pH 7.5). Beads were then resuspended and incubated at 37 °C for 2 h in freshly prepared 50 mM DTT to release the 5fC-containing strands. Then the supernatant containing the desired DNA was purified with Micro Bio-Spin P-6 gel columns to remove DTT.

Dot blot assay. Chemical labeling of 76-mer model DNA containing single 5fC or *Tdg*$^{-/-}$ mESC genomic DNA was performed as described above. For the dot blot assay, different amounts of model DNA or genomic DNA were denatured in advance with NaOH solution (final concentration of 0.15 M), spotted on Amersham Hybond-N+ membrane (GE Healthcare) and air-dried for 5 min. The membrane was UV-crosslinked and then blocked with 5%

nonfat milk in 1× Tris-buffered saline with Tween-20 (TBST) at room temperature for 2 h. The membrane was then incubated overnight at 4 °C with antiserum to 5fC (Active Motif, 61223, 1:2,500 dilution) and washed three times with 1× TBST. After incubation with secondary horseradish peroxidase–conjugated anti-rabbit IgG (CW Biotech, CW0103, 1:10,00 dilution) at room temperature for 1 h and three washes with 1× TBST, the membrane was supplied with 1 mL of SuperSignal West chemiluminescent substrate (Thermo Scientific) and then visualized by chemiluminescence exposure.

FspI restriction enzyme digestion. The FspI restriction site (TG5fC/GCA)-containing 70-mer oligonucleotide was prepared and chemically labeled as described above. Labeling product was PCR amplified (Model-F and Model-R; **Supplementary Table 1**) to introduce the C-to-T transition in the cutting site. PCR products were purified with the QIAquick PCR purification kit and subjected to FspI restriction enzyme (New England BioLabs) digestion according to the manufacturer's instructions. The digested products were analyzed with 4% agarose gel.

Pulldown specificity test with quantitative PCR. The model DNAs and primers for qPCR (**Supplementary Table 2**) were prepared as described above. We added 2 pg of each spike-in DNA per 1 μg of fragmented genomic DNA background. The qPCR test was run in triplicate with SYBR Premix Ex TaqTM II (Takara) according to the manufacturer's instructions. Reactions were run on the ABI viiA7 instrument. Three biological replicates were repeated to validate the results.

Library preparation and next-generation sequencing of fC-CET–enriched DNA samples. The fC-CET–enriched genomic DNAs were used directly for library preparation with the TELP (tailing, extension, ligation and PCR) protocol[23]. A minor modification was the use of MightyAmp DNA polymerase (Takara) for one round of on-bead primer extension before PCR amplification. The adaptor-ligated samples were then PCR amplified using NEB Next 2× PCR Master Mix (New England BioLabs) and indexed primers (New England BioLabs). Libraries were checked using the Agilent 2100 bioanalyzer before being loaded onto the Illumina HiSeq 2500 platform. A single-end (100 bp or longer) sequencing mode was suggested for maximal data collection.

Two biological replicates of each type of mESC were prepared and sequenced, which means that in parallel, two nonenriched input DNAs (input: pre-AI), two AI labeled samples (input: AI) and two pulldown output samples were sequenced simultaneously according to the same procedure.

Validation of locus-specific 5fC sites. Nine 5fC sites (**Supplementary Table 3**) from the *Tdg*$^{-/-}$ mESC genome, containing both previously identified sites (by fCAB-Seq or methylase-assisted bisulfite sequencing (MAB-Seq))[8,13] and sites newly identified by fC-CET, were chosen for validation. For fC-CET, different amounts of starting DNA material (1 μg or 100 ng) were tested. For fCAB-Seq, 1 μg of input DNA was protected with 10 mM *O*-ethylhydroxylamine (Aldrich, 274992) in 100 μL of 100 mM MES buffer (pH 5.0) at 37 °C for 4 h and then purified by ethanol precipitation. For MAB-Seq, 1 μg of input DNA was treated with M.SssI (New England BioLabs) in 50 μL for four rounds (each round consisted of a 2-h initial treatment (1.5 μL of M.SssI and 1 μL of *S*-adenosyl-L-methionine (SAM)) and subsequent 4-h treatment after the addition of 0.5 μL of M.SssI and 1 μL of SAM). Subsequent DNA purification was performed with phenol-chloroform-isoamyl alcohol (25:24:1) extraction followed by ethanol precipitation. Bisulfite treatment of normal, *O*-ethylhydroxylamine–protected or M.SssI-treated DNAs was performed using the EpiTect Fast DNA bisulfite kit (Qiagen) according to the manufacturer's instructions, except that two thermo cycles were run. Bisulfite-treated or fC-CET–treated DNAs (genomic loci or model DNA) were PCR amplified with bisulfite primers or normal primers, respectively. The PCR-amplified samples were purified and then subjected to Illumina library preparation using the NEBNext Ultra DNA Library Prep kit (New England BioLabs). The libraries were then pooled together and sequenced on the Illumina MiSeq platform with single-end reads of 150 bp.

Data processing and analysis. Raw reads were first trimmed for the poly-C sequence at the 3′ end. Reads shorter than 60 bp were discarded. Processed reads were then mapped to the mouse genome (mm9) by Bismark v0.8.3 (ref. 24) using options -n 1 -l 40 -chunkmbs 512. The 5fC-enriched regions in each output sample were detected using the model-based analysis of ChIP-Seq (MACS) peak-calling algorithm[25], with the corresponding input AI sample serving as the input control. The numbers of converted (N_T) and unconverted (N_C) cytosines were further extracted from each output data set. CpGs with fewer than ten reads or two converted reads were discarded. We used the binomial distribution with parameter N as (N_T + N_C) and r as the normal cytosine conversion rate (evaluated by spike-in sequence; r = 1.87% and 1.41% for wild-type replicates, and r = 2.01% and 1.42% for *Tpg*-null replicates) to calculate the probability of observing N_T or more C-to-T conversions by chance. To identify modified CpGs that were significantly enriched for 5fC, we considered CpGs with Holm-Bonferroni method–adjusted $P < 0.05$ in both replicate samples that were located within 5fC-enriched regions as genuine 5fC sites. Genome annotation analysis and read visualization were done with Homer software[26]. To plot the distribution of ChIP-Seq signals around 5fC sites, we discarded the sites located at repeat regions.

External data. H3K4me1 (GSM881352) and H3K27ac (GSM881349) ChIP-Seq data sets were obtained from ref. 27. The Tet1 (GSM611192) data set was obtained from ref. 28. The P300 (GSM1019072) data set was obtained from ref. 8. MethylC-Seq and TAB-Seq data were obtained from ref. 29.

21. Wang, D. *et al. Biochemistry* **42**, 6747–6753 (2003).
22. Obeid, S. *et al. EMBO J.* **29**, 1738–1747 (2010).
23. Peng, X. *et al. Nucleic Acids Res.* **43**, e35 (2015).
24. Krueger, F. & Andrews, S.R. *Bioinformatics* **27**, 1571–1572 (2011).
25. Zhang, Y. *et al. Genome Biol.* **9**, R137 (2008).
26. Heinz, S. *et al. Mol. Cell* **38**, 576–589 (2010).
27. Xiao, S. *et al. Cell* **149**, 1381–1392 (2012).
28. Williams, K. *et al. Nature* **473**, 343–348 (2011).
29. Hon, G.C. *et al. Mol. Cell* **56**, 286–297 (2014).

doi:10.1038/nmeth.3569

Corrigendum: Bisulfite-free, base-resolution analysis of 5-formylcytosine at the genome scale

Bo Xia, Dali Han, Xingyu Lu, Zhaozhu Sun, Ankun Zhou, Qiangzong Yin, Hu Zeng, Menghao Liu, Xiang Jiang, Wei Xie, Chuan He & Chengqi Yi

Nat. Methods; **doi:10.1038/nmeth.3569; corrected online 21 September 2015**

In the version of this article initially published online, Chuan He is incorrectly affiliated with Tsinghua-Peking Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing, China, and is missing an affiliation with the Department of Chemical Biology, College of Chemistry and Molecular Engineering, Peking University, Beijing, China. This error has been corrected for the print, PDF and HTML versions of this article.