SUPPLEMENTARY INFORMATION

Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics

Xingyu Lu[1, 2], Dali Han[1, 2], Boxuan Simen Zhao[1, 2], Chun-Xiao Song[1], Li-Sheng Zhang[1], Louis C. Doré[1], Chuan He[1*]

[1]Department of Chemistry and Institute for Biophysical Dynamics, Howard Hughes Medical Institute, The University of Chicago, 929 E. 57th Street, Chicago, Illinois 60637; [2]These authors contributed equally to this work

*To whom correspondence should be addressed. E-mail: chuanhe@uchicago.edu

Materials and methods

Generation of *Tdg*$^{-/-}$ mESC, Cell Culture and DNA Preparation. Generation of *Tdg*$^{-/-}$ mESC, cell culture and genomic DNA was prepared as previously described (Song et al., 2011). Briefly, the *Tdg* knockout ES cell lines were derived from inner cell mass of *Tdg*$^{fl/fl}$, *Tdg*$^{fl/-}$ and *Tdg*$^{-/-}$ embryos. To derive *Tdg*$^{-/-}$ ES cells, nuclear transfer was performed with the *Tdg*$^{-/-}$ iPS cells and enucleated MII oocytes. mESCs were cultured in feeder-free conditions. The genomic DNA was extracted by SDS/proteinase K digestion, phenol/chloroform extraction and ethanol precipitation.

Preparation of Synthetic DNA Oligonucleotides. Synthetic oligonucleotides containing 5hmC, 5fC and 5caC were prepared by using Applied Biosystems 392 DNA synthesizer with phosphoramidites from Glen Research. Subsequent purification was carried out using Sep-pak C18 cartridges (Waters) according to the manufacturer's protocol.

Dot-Blot for 5fC/5caC-specific Antibody Characterization. For dot blot assay, different amounts of 76mer DNA oligonucleotides (5'-CCTCACCATCTCAACCAATATTATATTATGTGTATATTXGATATTTTGTGTTAT AATATTGAGGGAGAAGTGGTGA -3', where X is either a C or modified C) were spotted on Amersham Hybond-N^{+} membrane (GE Healthcare) and fixed by Stratagene UV stratalinker 2400 using auto-crosslink option. The membrane was blocked in 5% nonfat milk incubated with 1:5,000 dilution of anti-5fC antiserum

(Active Motif) or 1:2,000 dilution of anti-5caC antiserum (gift from Dr. Yi Zhang) overnight at 4°C, respectively. After incubation with 1:2,000 dilution of HRP-conjugated anti-rabbit IgG secondary antibody, the membrane was supplied with SuperSignal West Pico Chemiluminescent Substrate (Thermo Scientific) and visualized by chemiluminescence exposure.

DIP-qPCR Assay for 5fC/5caC-Containing Amplicon Enrichment. For DIP-qPCR assays, the spiked-in control DNA was generated as previously described[27]. Briefly, C/5mC/5hmC/5fC/5caC-containing amplicon was PCR amplified by HotStarTaq DNA Polymerase (QIAGEN) with dNTP containing 10% modified dCTP and purified by gel electrophoresis. 50 pg of each spike-in DNA was added into 10 μg sonicated salmon sperm DNA (ssDNA) and sonicated to 200–600 bp. For each DIP assay, 1 μL (anti-5fC) or 0.5 μL (anti-5caC) of antiserum was used, and 1 μL of rabbit IgG was used as control. DNA and antibodies were incubated at 4°C for 4 h in a final volume of 500 μL 1×IP buffer (10 mM sodium phosphate [pH 7.0], 140 mM NaCl, 0.05% Triton X-100). 40 μL pre-washed protein G Dynabeads (Invitrogen) were added to the mix and incubated at 4°C overnight. These beads were washed three times with 1 mL of 1×IP buffer and three times with 1 mL of 1×IP buffer + 150 mM NaCl. After proteinase K digeston at 55°C for 2 h, the immunoprecipitated DNA was purified by phenol-chloroform extraction followed by ethanol precipitation. qPCR validation was performed in triplicates of 20 μL reactions each with 1× FastStart Essential DNA Green Master (Roche), 0.5 μM forward and reverse primers. Reactions were run on a

LightCycler96 (Roche) using the standard cycling conditions with primer sequences as follows: (dCTP) FW-CGGGAATGGCTTTGTGGTAA, RV-AATTCGCCTACACGCATCCT, (dmCTP) FW-AGTGGAGCAAGCGTGACAAGT, RV-CAGCGCGTAGGCTTCGA, (dhmCTP) FW-TGAATGCCGGGAATGGTTT, RV-TGGAGAGCACCACCACTGATT, (dfCTP) FW-CCGATTCCGCCTAGTTGGT, RV-TGCCTGCGATGGTTGGA, (dcaCTP) FW-CTGCGCCGCCACAAA, RV-CTGGAATTGGGCAGAAGAAAAC.

Selective 5fC/5caC Protection and Sodium Bisulfite Treatment. Hydroxylamine protection of 5fC on immunoprecipitated DNA was performed with 10 mM O-ethylhydroxylamine (Aldrich, 274992) in 100 mM MES buffer (pH = 5.0) at 37°C for 4 h. Protection of 5caC was performed with 20 mM NHS (Aldrich 130672), 2mM EDC (Pierce, 77149), in 75mM MES buffer (pH=5.0) at 37°C for 0.5 h. After buffer exchange to 100 mM sodium phosphate (pH=7.5) and 150 mM NaCl, the pre-activated 5caC was incubated with 10 mM (4-aminomethyl)benzylazide at 37°C for 1 h.

Protected DNA was purified by Zymo DNA Clean & Concentration kit and subject to the MethylCode bisulfite conversion kit (Invitrogen) following the manufacturer's instruction with the exception that the bisulfite thermal cycle program was run twice for 5fC.

HPLC and Mass-Spec Analysis of $EtONH_2$-Protected or $NaBH_4$-Treated 5fC

Model DNA. A 5-mer 5fC-containing DNA oligonucleotide (5'- AC(5fC)GT -3'; 0.2 nmol) was protected by $EtONH_2$ or treated by $NaBH_4$[1]. Reaction mixture was diluted 10-fold in 50 mM TEAA buffer (pH = 7.0) and injected into HPLC. The treated products were separated by C18 column (buffer A: 50 mM TEAA buffer, pH = 7.0; buffer B: 100% methanol; 0-15% buffer B in 40 mins). The substrate peak and product peaks were collected and combined for Maldi-Tof detection. For a longer model DNA, a 76-mer DNA oligonucleotide (0.2 nmol) was protected by $EtONH_2$ or treated by $NaBH_4$. The reaction mixtures were digested by nuclease P1 (Sigma) and alkaline phosphatase (Sigma) according to published protocols[2]. After a brief desalting with ammonium-equilibrated cation exchange resin (Poly-Prep Columns AG 50W-X8, Bio-Rad) and filtration, 10 μL (out of 40 μL) recovered solution was digested to single nucleosides. The nucleosides were separated by reverse phase ultra-performance liquid chromatography on a C18 column, with online mass spectrometry detection using Agilent 6410 QQQ triple-quadrupole LC mass spectrometer set to multiple reaction monitoring (MRM) in positive electrospray ionization mode. The nucleosides were quantified using the nucleoside to base ion mass transitions of 256 to 140 (5fC), and 228 to 112 (C).

Sanger Sequencing of Bisulfite-Treated Protected DIPed 5fC/5caC 76-mer Model DNA. 50 pg 76-mer 5fC/5caC DNA oligonucleotides were diluted 20-fold with the same 76-mer unmodified oligonucleotides. The mixed oligonucleotides were spiked into the 10 μg sonicated salmon sperm DNA (ssDNA). 5fC/5caC DIP, chemical

protection and bisulfite treatment were performed as described above. Products from the bisulfite treatment were PCR amplified by Zymo Taq Polymerase using the following primers: Forward: 5'-CCCTTTTATTATTTTAATTAATATTATATT-3'; Reverse:

5'-CTCCGACATTATCACTACCATCAACCACCCATCCTACCTGGACTACATTCT TATTCAGTATTCACCACTTCTCCCTCAAT-3'.

PCR conditions consisted of an initial denaturation step of 95°C for 10 min, followed by 40 cycles of 95°C for 30 s, 45°C for 30 s and 72°C for 1 min, and a final extension at 72°C for 7 min. PCR products were purified using Zymo DNA Clean & Concentration kit and subjected to Sanger sequencing.

For colony picking, PCR products were cloned into plasmid using TOPO TA Cloning (Invitrogen). After colony picking, the purified plasmids were subjected to sanger sequencing with the following primer: 5'-CTCCGACATTATCACT-3'.

DIP-CAB-seq Library Construction of 5fC- and 5caC-Containing Genomic DNA.

100 μg sonicated $Tdg^{fl/fl}$ or $Tdg^{-/-}$ mESC genomic DNA (average 400 bp) was subjected to four parallel 5fC/5caC DNA IP as described above. The immunoprecipitated DNA was combined and subjected to either traditional bisulfite treatment or chemical-assisted bisulfite treatment. 50 ng of bisulfite treated DNA was tagged with Illumina compatible adapter and amplified to an appropriate library concentration using EpiGnome Methyl-Seq kit (Epicentre-Illumina). Libraries were checked for quality and quantified using an Agilent 2100 Bioanalyzer DNA 1000

Chip.

Libraries were pooled and sequenced using rapid two lane flow-cells on an Illumina HiScan platform. Single-read 100-bp sequencing was completed with an Illumina TruSeq SBS kit v3-HS. Image analysis and base calling were performed with the standard Illumina pipeline.

Libraries were prepared and sequenced from two biological replicates per genotype from 5fC- and 5caC-enriched DNA. In parallel, genotype matched, nonenriched input genomic DNA libraries were generated and sequenced.

Data Processing and Analysis. Sequencing reads were mapped to the mouse genome (mm9) by bismark v0.8.3[3], using options -n 1 -l 40 -chunkmbs 512. To visualize sequencing signals in the genome browser, we generated bedGraph files for each dataset with HOMER[32]. The numbers of converted and unconverted cytosines were further extracted from each 5fC/5caC-enriched bisulfite-seq and DIP-CAB-seq datasets. CpGs with less than 10 reads were discarded. To detect modified CpGs that significantly enriched for 5fC/5caC, the difference between bisulfite-seq and DIP-CAB-seq datasets were tested by using a Fisher's exact test. CpGs with a significantly higher unconverted rate ($P<0.05$ and difference > 1%) in DIP-CAB-seq were taken as 5fC/5caC-modified cytosines. CpGs with the opposite pattern were considered false positives and used to estimate the false discover rate (FDR). To plot the distribution of H3K4me1 and H3K27ac (GSM881352 and GSM881349)[4] around modified cytosine sites, we used nucleosome sequencing signals (GSM1004653)[5] as

the background. To plot the distribution of Tet1 (GSM611192)[6], we used input signals (GSM723020) [6] as the background. To detect TF motifs around 5fC/5caC-modified cytosines, we performed *de novo* motif analysis with HOMER at +/- 100 bp region around 5fC and 5caC sites.

Definition of Enhancer Subgroups. Enhancers with evidence showing its interaction with distal genomic regions were considered as the most active enhancer subgroup (interacting enhancers). The genomic locations of these enhancers were obtained from published ChIA-PET dataset[5]. The active and poised enhancers were defined by using histone modification markers. Both H3K4me1 and H3K27ac enriched regions were obtained from previous publication[7]. Liftover[8] was used to convert the genomic location from mm8 to mm9. Regions overlapped with interacting enhancers were discarded. Regions enriched with H3K27ac were considered as active enhancers. Regions only enriched with H3K4me1 but not H3K27ac were considered as poised enhancers.

Chemical Synthesis. (4-aminomethyl)benzylazide was prepared according to previous literature[9].

Sequencing data accession codes. Sequence Read Archive: GSE56429. Identified 5fC and 5caC sites can be downloaded from this archive as well.

Figure S1

(A) General procedure for DIP-CAB-seq. The 5fC- and 5caC-containing DNA fragments are enriched through antibody-based DNA immunoprecipitation. The enriched DNA is divided into two halves and subject to CAB-seq or traditional bisulfite sequencing in parallel. Single-base information of 5fC/5caC is extracted through sequencing and comparison.

(B-E) Validation and characterization for anti-5fC and anti-5caC antibodies and DIP-CAB-seq strategy. **(B)** 5fC- and 5caC-containing 76-mer DNA oligonucleotides can be specifically recognized by the anti-5fC and anti-5caC antibodies, respectively. **(C)** DIP-qPCR assays demonstrated specific enrichment of 5fC- and 5caC-containing DNA fragments using these antibodies. Error bar is s.e.m., for n = 3 experiments. **(D)** Sanger sequencing of bisulfite-treated DNA (76-mer DNA containing either 5fC or 5caC at a specific position) showing high efficiency of 5fC and 5caC protection from deamination following immunoprecipitation and enrichment. **(E)** The 76-mer DNA with 5fC/5caC was diluted 20-fold by the same DNA without modification to obtain samples with 5% 5fC/5caC at the modified position. Under CAB-seq conditions we observed 5% signal difference between CAB-seq and regular bisulfite experiments. Detection limits for 5fC and 5caC of CAB-seq can be significantly improved (~50%) after immunoprecipitation/enrichment with corresponding antibodies.

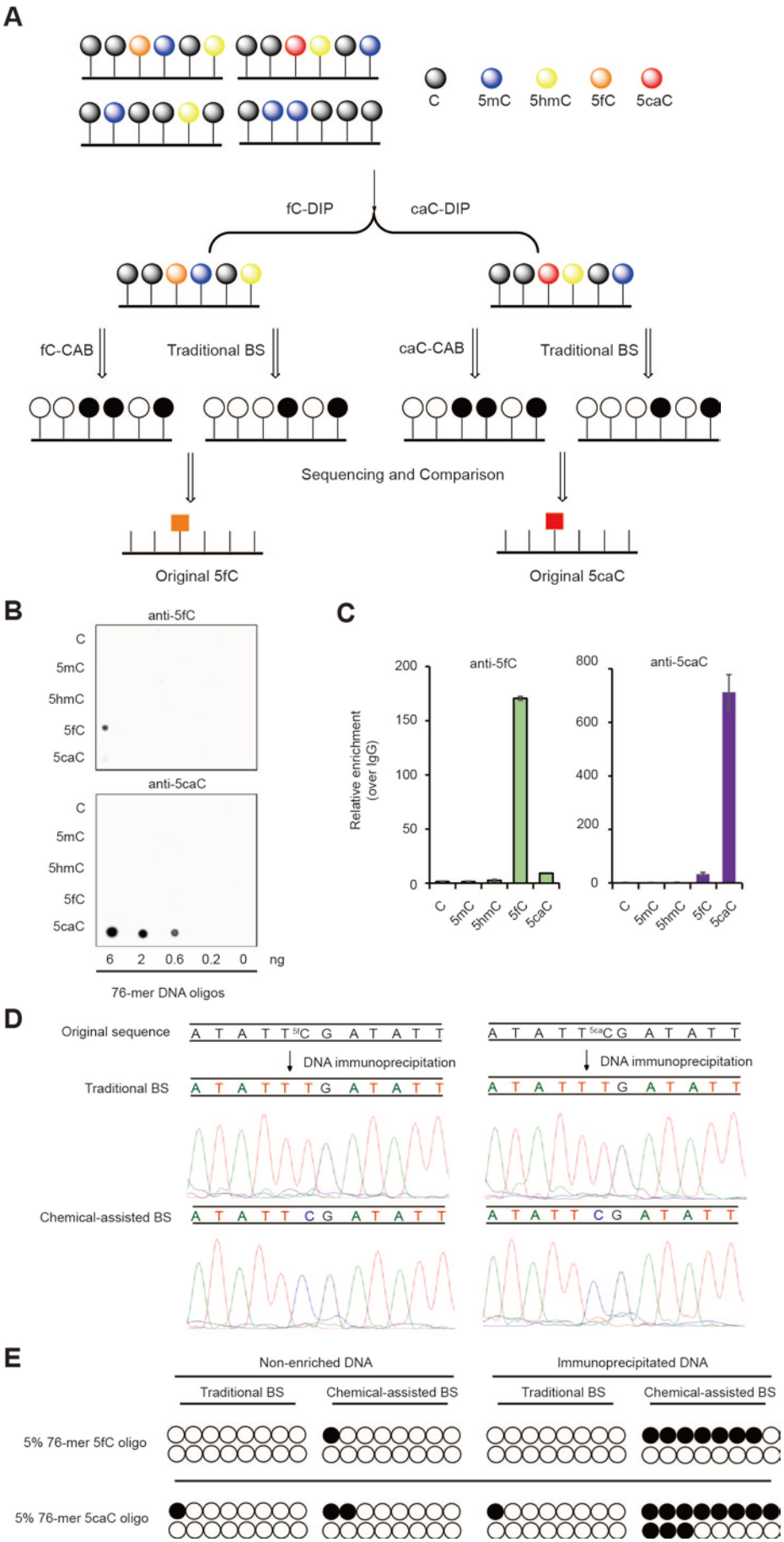
(F-H) Validation and characterization for 5fC protection. **(F)** 5-mer 5fC-containing DNA oligonucleotides and the modified oligonucletides can be distinguished though HPLC, showing high 5fC conversion under both $NaBH_4$ reduction and $EtONH_2$ protection conditions. **(G)** Maldi-Tof validated the reaction product of 5-mer 5fC-containing DNA oligonucleotides in both $NaBH_4$ reduction and $EtONH_2$ protection after HPLC separation. **(H)** HPLC-mass validated the reaction yield of $NaBH_4$ reduction and $EtONH_2$ protection on single nucleosides digested from 76-mer 5fC-containing DNA oligonucleotides.

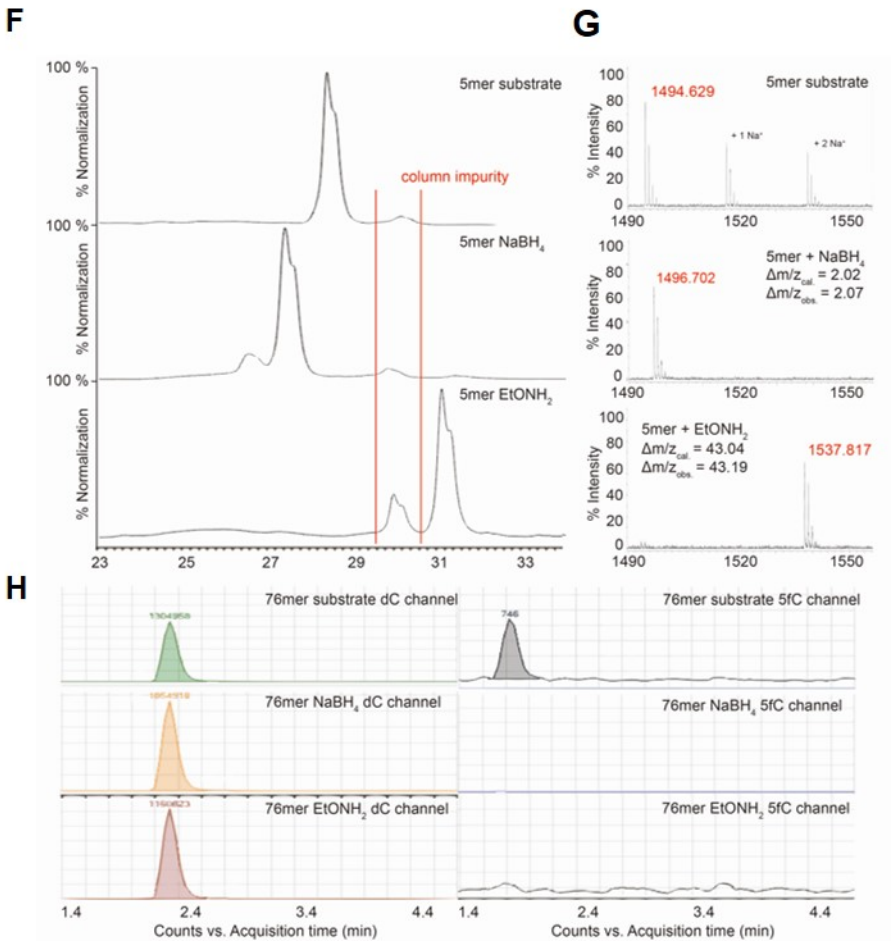
(I-L) Validations of 5fC and 5caC sites in $Tdg^{fl/fl}$ and $Tdg^{-/-}$ mouse ES cells. **(I)**

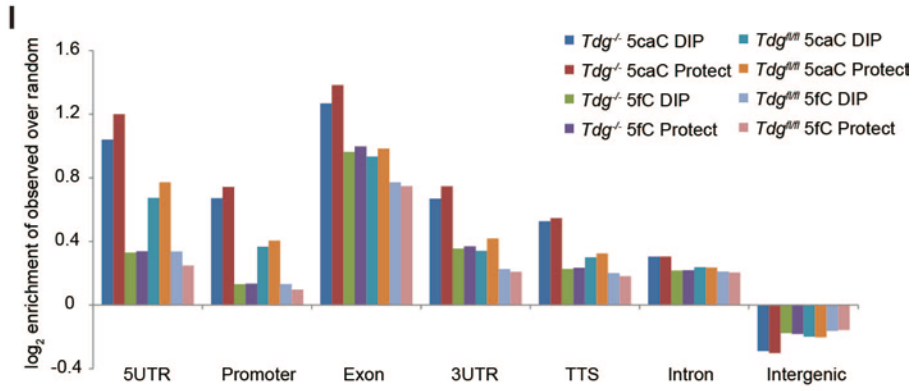
Enrichment of 5fC and 5caC signals in *Tdg*$^{fl/fl}$ and *Tdg*$^{-/-}$ mouse ES cells at various genomic regions. **(J)** Enrichment of 5fC and 5caC signals in *Tdg*$^{fl/fl}$ and *Tdg*$^{-/-}$ mouse ES cells at repeat elements. **(K)** Distribution of 5fC and 5caC signals in *Tdg*$^{fl/fl}$ and *Tdg*$^{-/-}$ mouse ES cells at a representative active enhancer loci. **(L)** The number (left panel) and FDR (right panel) of detected 5fC and 5caC single sites at various p value thresholds in each *Tdg*$^{fl/fl}$ and *Tdg*$^{-/-}$ dataset.

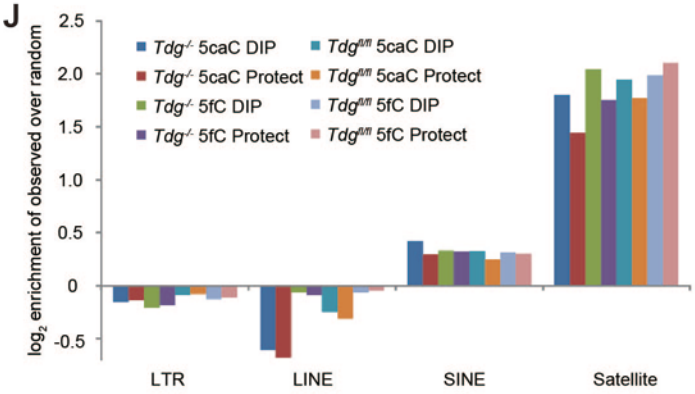
(M-O) The correlations of modified cytosine groups with enhancers and Tet1 signals. **(M)** The association percentages of different cytosine modification sites within interacting enhancers. **(N)** Pie chart showing the association percentages of enhancers with various groups of cytosine modifications sites. **(O)** Distribution of Tet1 signals at each modified cytosine groups.

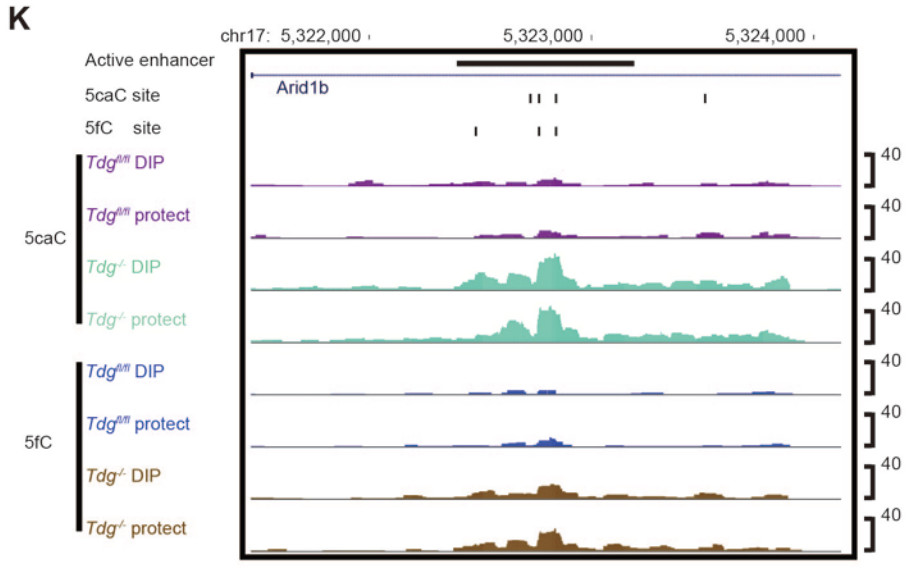
(P-Q) Sequence contexts around 5fC and 5caC sites. **(P)** The sequence context of +/- 150 bp regions around 5fC and 5caC sites, respectively. **(Q)** Density distribution of Esrrb, Klf4, Oct4, Sox2 motifs within +/- 500 bp regions around 5fC and 5caC sites, respectively.

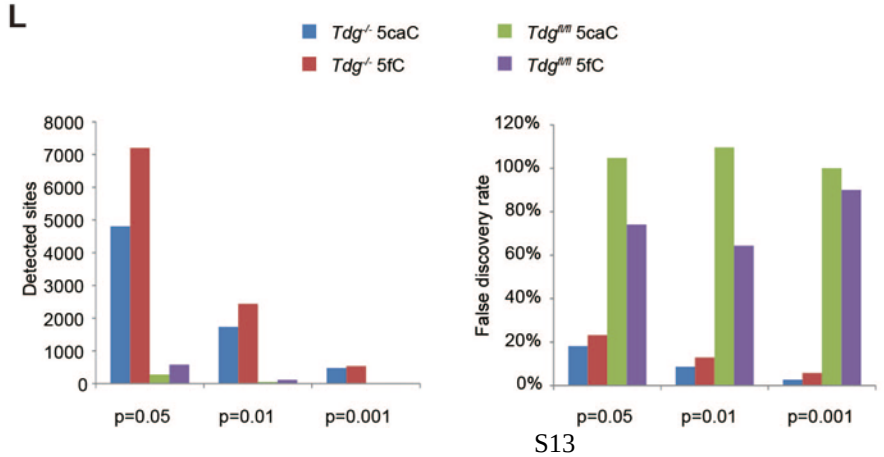

A
C
5mC
5hmC
5fC
5caC
fC-DIP
caC-DIP
fC-CAB
Traditional BS
caC-CAB
Traditional BS
Sequencing and Comparison
Original 5fC
Original 5caC
B
anti-5fC
anti-5caC
C
5mC
5hmC
5fC
5caC
6 2 0.6 0.2 0 ng
76-mer DNA oligos
C
anti-5fC
anti-5caC
Relative enrichment (over IgG)
200
150
100
50
0
800
600
400
200
0
D
Original sequence
A T A T T 5fC G A T A T T
A T A T T 5caC G A T A T T
DNA immunoprecipitation
DNA immunoprecipitation
Traditional BS
A T A T T T G A T A T T
A T A T T T G A T A T T
Chemical-assisted BS
A T A T T C G A T A T T
A T A T T C G A T A T T
E
Non-enriched DNA
Immunoprecipitated DNA
Traditional BS
Chemical-assisted BS
Traditional BS
Chemical-assisted BS
5% 76-mer 5fC oligo
5% 76-mer 5caC oligo

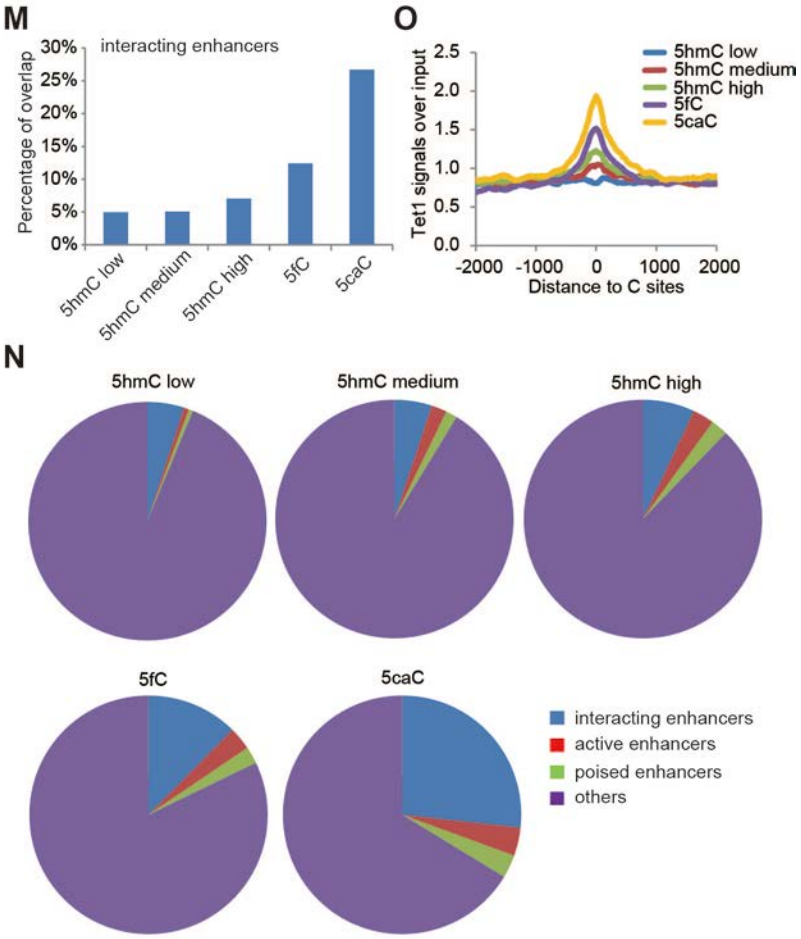

F
G
H
100 %
% Normalization
5mer substrate
column impurity
5mer NaBH4
5mer EtONH2
23
25
27
29
31
33
% Intensity
1494.629
+ 1 Na+
+ 2 Na+
1490
1520
1550
5mer + NaBH4
Δm/z cal. = 2.02
Δm/z obs. = 2.07
1496.702
5mer + EtONH2
Δm/z cal. = 43.04
Δm/z obs. = 43.19
1537.817
76mer substrate dC channel
76mer substrate 5fC channel
76mer NaBH4 dC channel
76mer NaBH4 5fC channel
76mer EtONH2 dC channel
76mer EtONH2 5fC channel
1.4
2.4
3.4
4.4
Counts vs. Acquisition time (min)

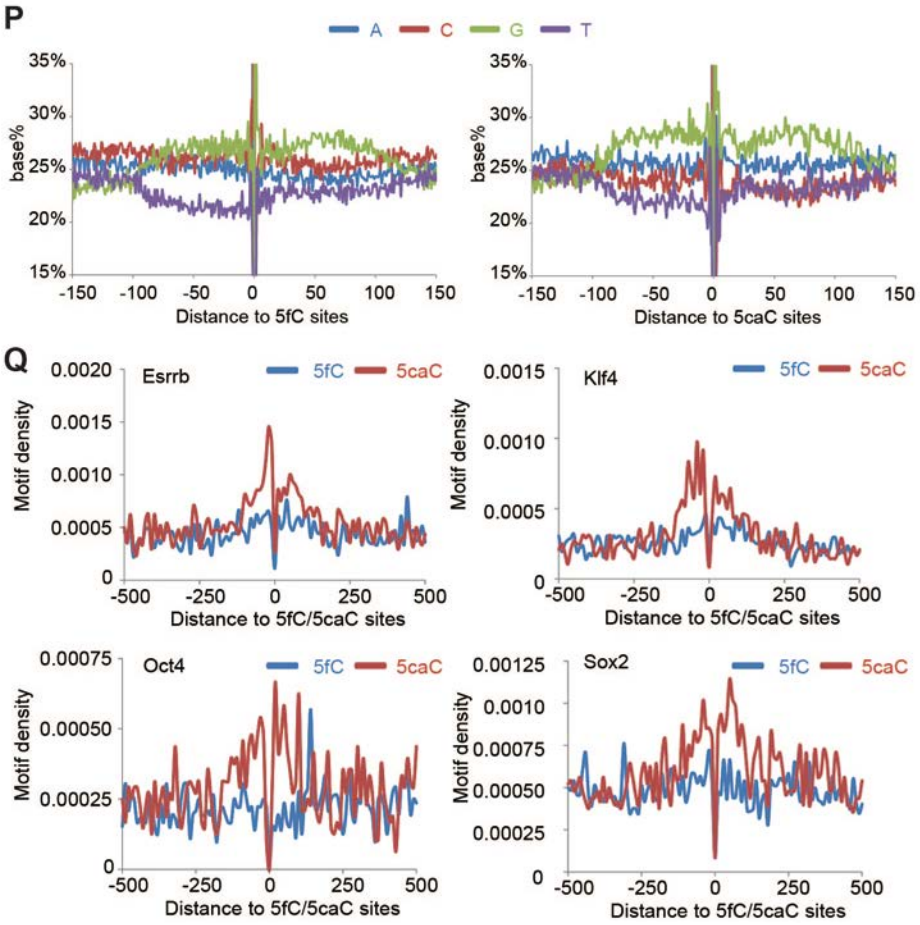

I
log2 enrichment of observed over random
1.6
1.2
0.8
0.4
0
-0.4
Tdg-/- 5caC DIP
Tdg-/- 5caC Protect
Tdg-/- 5fC DIP
Tdg-/- 5fC Protect
Tdgfl/fl 5caC DIP
Tdgfl/fl 5caC Protect
Tdgfl/fl 5fC DIP
Tdgfl/fl 5fC Protect
5UTR
Promoter
Exon
3UTR
TTS
Intron
Intergenic


J
log2 enrichment of observed over random
2.5
2.0
1.5
1.0
0.5
0
-0.5
Tdg-/- 5caC DIP
Tdg-/- 5caC Protect
Tdg-/- 5fC DIP
Tdg-/- 5fC Protect
Tdgfl/fl 5caC DIP
Tdgfl/fl 5caC Protect
Tdgfl/fl 5fC DIP
Tdgfl/fl 5fC Protect
LTR
LINE
SINE
Satellite


K
chr17: 5,322,000
5,323,000
5,324,000
Active enhancer
5caC site
5fC site
Arid1b
5caC
5fC
Tdgfl/fl DIP
Tdgfl/fl protect
Tdg-/- DIP
Tdg-/- protect
Tdgfl/fl DIP
Tdgfl/fl protect
Tdg-/- DIP
Tdg-/- protect
40


L
Tdg-/- 5caC
Tdg-/- 5fC
Tdgfl/fl 5caC
Tdgfl/fl 5fC
Detected sites
8000
7000
6000
5000
4000
3000
2000
1000
0
p=0.05
p=0.01
p=0.001
False discovery rate
120%
100%
80%
60%
40%
20%
0%
p=0.05
p=0.01
p=0.001


M
interacting enhancers
Percentage of overlap
30%
25%
20%
15%
10%
5%
0%
5hmC low
5hmC medium
5hmC high
5fC
5caC
O
Tet1 signals over input
2.5
2.0
1.5
1.0
0.5
0.0
-2000
-1000
0
1000
2000
Distance to C sites
5hmC low
5hmC medium
5hmC high
5fC
5caC
N
5hmC low
5hmC medium
5hmC high
5fC
5caC
interacting enhancers
active enhancers
poised enhancers
others


P
A
C
G
T
35%
30%
25%
20%
15%
base%
-150
-100
-50
0
50
100
150
Distance to 5fC sites
Distance to 5caC sites
Q
Esrrb
Klf4
Oct4
Sox2
5fC
5caC
Motif density
0.0020
0.0015
0.0010
0.0005
0
0.00075
0.00050
0.00025
0.00125
-500
-250
250
500
Distance to 5fC/5caC sites

Supplementary Table 1 Summary of sequencing reads and identified 5fC and 5caC sites. This table summarizes the total reads, mapped reads and the mapping efficiency for each sequencing libraries.

	total reads	mapped reads	mapping efficiency
5caC *Tdg*$^{-/-}$DIP	28098702	19740340	70.30%
5caC *Tdg*$^{-/-}$Protect	41363334	29710848	71.80%
5fC *Tdg*$^{-/-}$DIP	50455071	32003786	63.40%
5fC *Tdg*$^{-/-}$Protect	65768630	43461595	66.10%
5caC *Tdg*$^{fl/f-}$DIP	27143486	18212102	67.10%
5caC *Tdg*$^{fl/f-}$Protect	40073739	26543604	66.20%
5fC *Tdg*$^{fl/f-}$DIP	36753810	24554817	66.80%
5fC *Tdg*$^{fl/f-}$Protect	40815452	26585404	65.10%

Reference

1. Booth MJ, Marsico G, Bachman M, Beraldi D, Balasubramanian S. Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution. *Nat Chem* 2014; **6**: 435-40.
2. Song CX, Szulwach KE, Dai Q *et al.* Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. *Cell* 2013; **153**: 678-691.
3. Krueger F, Andrews SR Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 2011; **27**: 1571-1572.
4. Xiao S, Xie D, Cao X *et al.* Comparative epigenomic annotation of regulatory DNA. *Cell* 2012; **149**: 1381-1392.
5. Teif VB, Vainshtein Y, Caudron-Herger M *et al.* Genome-wide nucleosome positioning during embryonic stem cell development. *Nat Struct Mol Biol* 2012; **19**: 1185-1192.
6. Shen Y, Yue F, McCleary DF *et al.* A map of the cis-regulatory sequences in the mouse genome. *Nature* 2012; **488**: 116-120.
7. Creyghton MP, Cheng AW, Welstead GG *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci* 2010; **107**: 21931-21936.
8. Hinrichs AS, Karolchik D, Baertsch R *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* 2006; **34**: D590-598.
9. Lu X, Song CX, Szulwach K *et al.* Chemical modification-assisted bisulfite sequencing (CAB-Seq) for 5-carboxylcytosine detection in DNA. *J Am Chem Soc* 2013; **135**: 9315-9317.