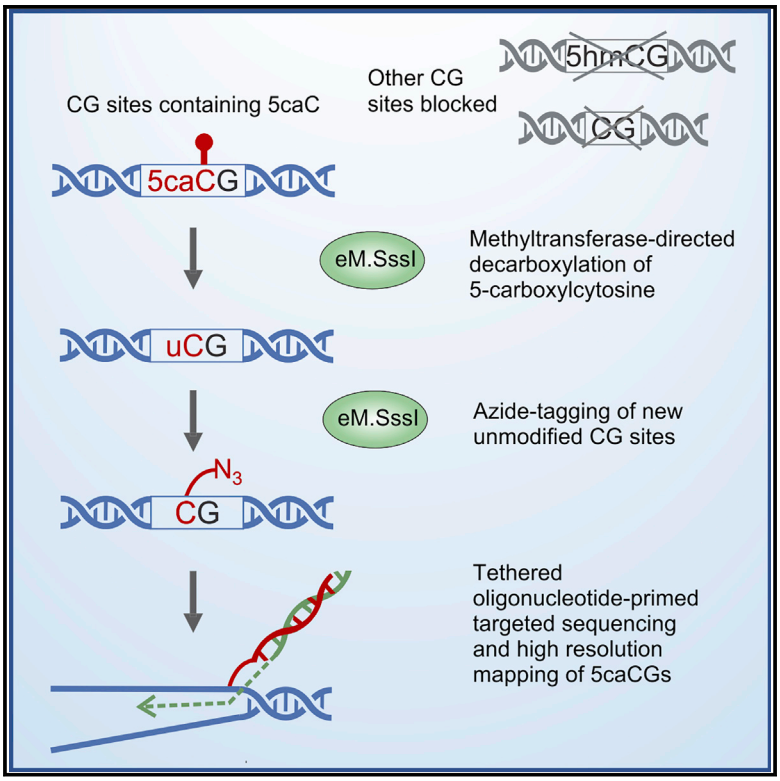
Cell Reports

Resource

A Bisulfite-free Approach for Base-Resolution Analysis of Genomic 5-Carboxylcytosine

Graphical Abstract



Highlights

- A bisulfite-free caCLEAR approach for profiling 5caCG sites genome-wide
- caCLEAR specifically targets only 5caCG sites and avoids whole-genome sequencing
- Two pluripotency states of mouse ESCs differ in the genomic distribution of 5caCGs
- 5caCGs tend to distribute in the antisense strand of highly expressed protein genes

Authors

Janina Ličytė, Povilas Gibas, Kotryna Skardžiūtė, Vaidotas Stankevičius, Audronė Rukšėnaitė, Edita Kriukienė

Correspondence

edita.kriukiene@bti.vu.lt

In Brief

Licyte et al. present a chemo-enzymatic approach for mapping 5-carboxylcytosine, which uses enzymatic removal of the carboxyl group from 5caCG sites and their detection via high-resolution sequencing. Profiling of 5caCGs in naive and primed pluripotent mouse ESCs demonstrates the state-specific distribution of 5caCGs and association with active gene expression.

Ličytė et al., 2020, Cell Reports *32*, 108155
September 15, 2020
https://doi.org/10.1016/j.celrep.2020.108155

CellPress

Cell Reports

CellPress
OPEN ACCESS

Resource

A Bisulfite-free Approach for Base-Resolution Analysis of Genomic 5-Carboxylcytosine

Janina Ličytė,[1,2] Povilas Gibas,[1,2] Kotryna Skardžiūtė,[1] Vaidotas Stankevičius,[1] Audronė Rukšėnaitė,[1] and Edita Kriukienė[1,3,*]

[1]Department of Biological DNA Modification, Institute of Biotechnology, Vilnius University, Vilnius 10257, Lithuania
[2]These authors contributed equally
[3]Lead Contact
*Correspondence: edita.kriukiene@bti.vu.lt
https://doi.org/10.1016/j.celrep.2020.108155

SUMMARY

Due to an extreme rarity of 5-carboxylcytosine (5caC) in the mammalian genome, investigation of its role brings a considerable challenge. Methods based on bisulfite sequencing have been proposed for genome-wide 5caC analysis. However, bisulfite-based sequencing of scarcely abundant 5caC demands significant experimental and computational resources, increasing sequencing cost. Here, we present a bisulfite-free approach, caCLEAR, for high-resolution mapping of 5caCGs. The method uses an atypical activity of the methyltransferase eM.SssI to remove a carboxyl group from 5caC, generating unmodified CGs, which are localized by uTOP-seq sequencing. Validation of caCLEAR on model DNA systems and mouse ESCs supports the suitability of caCLEAR for analysis of 5caCGs. The 5caCG profiles of naive and primed pluripotent ESCs reflect their distinct demethylation dynamics and demonstrate an association of 5caC with gene expression. Generally, we demonstrate that caCLEAR is a robust economical approach that could help provide deeper insights into biological roles of 5caC.

INTRODUCTION

5-methylcytosine (5mC) is the most conserved DNA modification from plants to animals. 5mC dynamics is regulated by the interplay between DNA methyltransferases, which transfer a methyl group from S-adenosylmethionine (SAM) onto the C5 position of the target cytosine residue and the TET family of dioxygenases which oxidize 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) via a stepwise manner, the latter two being replaced with unmodified cytosine by a thymine-DNA glycosylase (Tdg). It has been increasingly acknowledged that 5hmC plays a specific role in transcriptional regulation and DNA demethylation in different biological settings (Song and He, 2013; Wu and Zhang, 2014). Since the oxidized forms of 5mC (oxi-mCs) have, in part, distinct reader proteins (Spruijt et al., 2013; Iurlaro et al., 2013), they might play individual functions beyond that of the simple intermediates in the 5mC demethylation pathway. However, due to the low abundance of 5fC and 5caC (0.00018% 5caC of C; Ito et al., 2011), their functions are still underexplored. Profiling of 5caC in mouse embryonic stem cells (ESCs) demonstrated the preferential occurrence of 5caC (and 5fC) at active enhancers, pluripotency transcription factor (TF) binding sites, and low-methylated regions, including hypomethylated promoters of highly expressed genes. Importantly, 5fC and 5caC exhibit very limited overlap, suggesting their distinct roles in demethylation dynamics (Lu et al., 2015). Furthermore, involvement of the 5caC-dependent transcriptional regulation in the mechanisms of malignant transformation and its prognostic potential in cancer have been recently proposed (Eleftheriou et al., 2015; Zhou et al., 2018; Storebjerg et al., 2018).

The potential biological relevance of oxi-mCs explains why high efforts have been devoted to design technologies for measuring them globally or at single-base resolution. Several methods have been established for 5caC mapping, including low- and high-resolution genome profiling technologies (Shen et al., 2013; Wu et al., 2014). The high-resolution approaches rely on bisulfite conversion. Due to the non-selectivity of 5caC in bisulfite sequencing (both 5caC and cytosine are read as thymine), chemical or enzymatic pretreatment steps were used prior to conventional whole-genome bisulfite sequencing (WGBS) (Lu et al., 2015; Wu et al., 2014; Neri et al., 2015). These methods read 5caC and 5fC as a single signal or exploit the comparison between the 5caC- or 5fC-premodified and conventional WGBS and, thus, require double-sequencing efforts, which represent technical and economical constraints for genome-scale analysis. Furthermore, due to the low abundance of 5caC, a direct or derivatization-based WGBS of 5caC results in a vast majority of non-informative sequencing data.

Here, we present a bisulfite-free, single-base resolution method that enables targeted mapping of 5caC residues. The method makes use of a methyltransferase-promoted C-C bond cleavage reaction, leading to the decarboxylation of 5caC that yields unmodified cytosine (Liutkevičiūtė et al., 2014). By combining the decarboxylation activity of a CG-specific engineered bacterial MTase eM.SssI with targeted sequencing of the enzymatically introduced unmodified CG (uCG) sites (Staševskij et al., 2017),

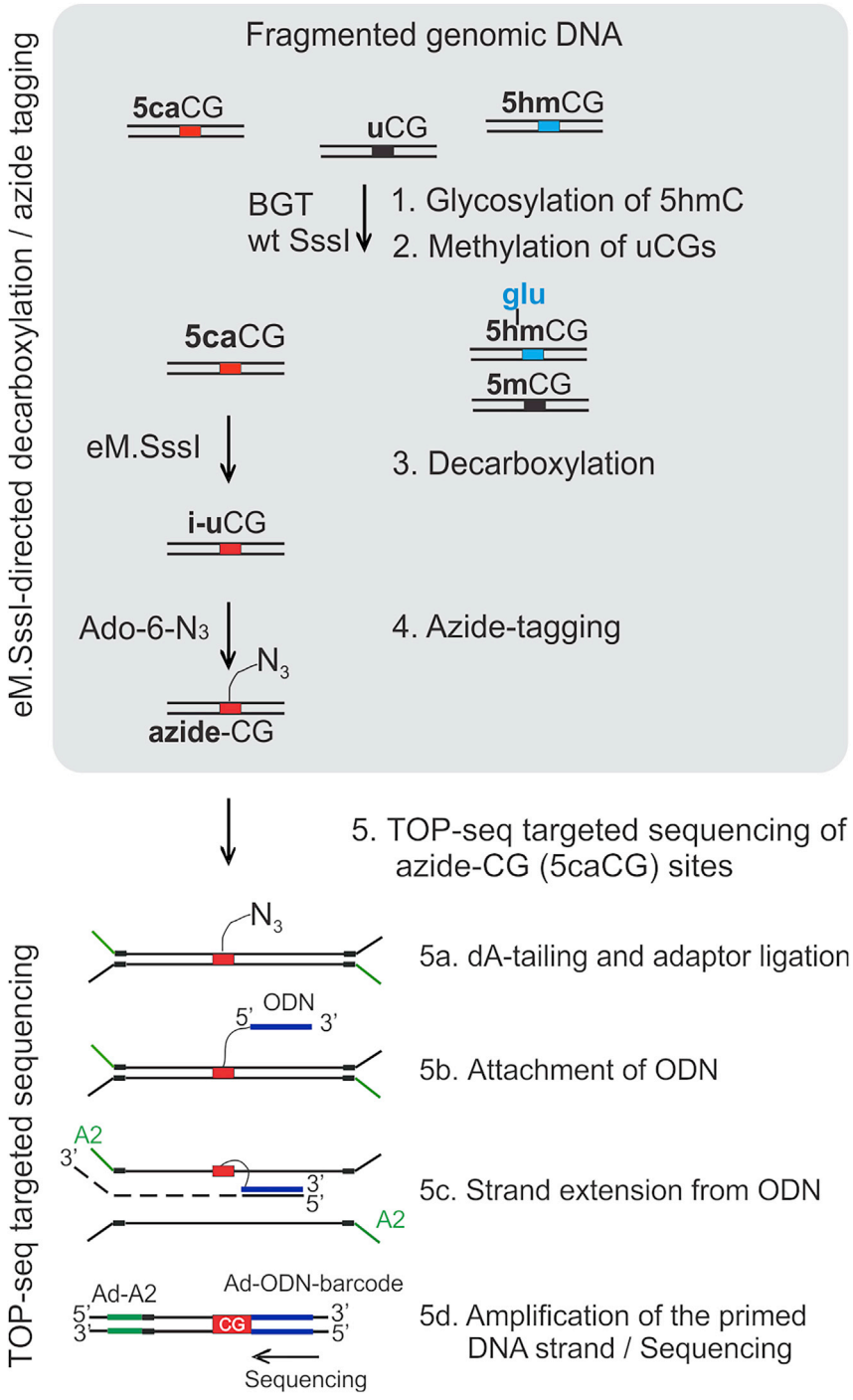


Figure 1. Outline of the caCLEAR Procedure

Step 1: glycosylation of 5hmC residues using T4 phage β-glucosyltransferase (BGT) and a cofactor UDP-glucose. Step 2: protection of genomic unmodified CG sites (uCGs) by methylation using the CG-specific MTase SssI and AdoMet. Step 3: decarboxylation of 5caC residues using the eM.SssI mutant enzyme introducing new unmodified CG sites, i-uCGs. Step 4: covalent tagging of i-uCGs with an azide group in the presence of eM.SssI and Ado-6-N_3. Step 5: profiling of the azide-containing CG sites by high-throughput targeted sequencing uTOP-seq. ODN, oligodeoxyribonucleotide; A2, one strand of a partially complementary adaptor.

we developed a technology for profiling genomic 5caCGs, termed caCLEAR ("5caC clearance"). We validated caCLEAR on model DNA systems and mouse ESCs. Given the broad interest in the 2i pluripotency state and the dramatic cellular state changes that are induced through the 2i inhibitors (PD0325901 and CHIR99021, which target mitogen-activated protein kinase [MEK] and glycogen synthase kinase-3 [GSK3], respectively), we decided to investigate 5caC profiles in the presence and absence of 2i, i.e., in the two pluripotency states of ESCs. Cultivation of cells under these distinct culture conditions presents an exciting model for studying general demethylation dynamics in naive and primed pluripotent states. caCLEAR revealed distinct properties of 5mC oxidation at various gene regulatory elements in the two closely related pluripotent ESCs and the involvement of 5caC in the state-specific gene expression regulation through modulating binding of TFs.

RESULTS

Development and Validation of caCLEAR

For the development of caCLEAR, we used two distinct enzymatic activities of eM.SssI MTase containing the expanded cofactor binding pocket (Kriukienė et al., 2013): the decarboxylation of 5caC yielding uCG sites and the transfer of an azide group from a synthetic cofactor Ado-6-N_3 onto uCG sites (Staševskij et al., 2017; Kriukienė et al., 2013). The conditional switch between these two activities of eM.SssI introduces azide labels into 5caCGs that can be localized by tethered oligonucleotide-primed targeted sequencing (uTOP-seq) (Figure 1). As eM.SssI is also capable of removing the 5-hydroxymethyl group from 5hmC, in step 1, genomic 5hmC residues are protected by glycosylation using phage T4 β-glucosyltransferase (BGT) (Song et al., 2011). To discriminate between genomic uCG sites and uCG sites introduced following decarboxylation (i-uCGs), genomic uCG sites are methylated with the wild-type (WT) M.SssI in the presence of AdoMet (step 2). To achieve complete modification of 5hmC and uCG sites in steps 1 and 2, we optimized these protocols in our model DNA systems and genomic DNA as measured by qPCR and a quantitative high-performance liquid chromatography coupled with tandem mass spectrometry (HPLC-MS/MS) (Figures

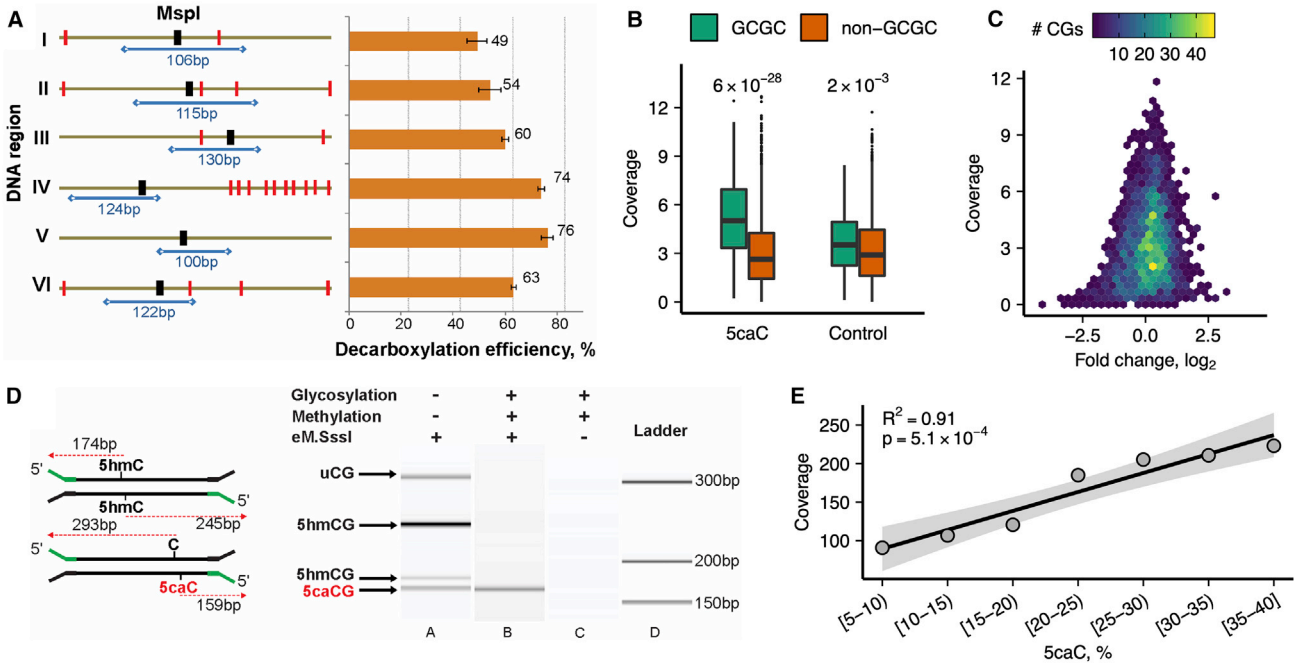


Figure 2. Validation of the caCLEAR Procedure in Model DNA Systems

(A) Region-specific decarboxylation of 5caC sites. Left: schematic representation of six genomic regions used for the assessment of decarboxylation by qPCR (target CG sites are indicated in black, other proximal CG sites are represented in red, and PCR products are indicated in blue). Right: efficiency of decarboxylation is indicated as a difference in the amount of a target site resistant to the R.MspI cleavage before and after decarboxylation.

(B) caCLEAR coverage (normalized by total read count) in relation to the extent of carboxylation at GCGC or CG sites (excluding GCGC), containing, on average, 60% and 30% of 5caC, respectively, of the model lambda phage DNA. Control samples indicate the maximum uTOP-seq coverage at CG sites (30% of uCGs, no 5caC). Numbers above the boxplots indicate p values calculated with two-sided paired t test.

(C) Fold change of normalized caCLEAR coverage at CG sites (excluding GCGC) between the caCLEAR and control uTOP-seq libraries at various coverage cutoffs. Color code denotes numbers of CG sites.

(D) caCLEAR analysis of two model DNA fragments (left) each containing differently modified cytosines in both strands of a single CG site. Analysis of the reaction products by Bioanalyzer (Agilent) (right) indicated a single correctly sized PCR amplicon in lane B out of 4 possible products (lane A). No PCR product was obtained if the decarboxylation step was omitted (lane C).

(E) Dependence between the caCLEAR coverage and 5caC levels determined by bisulfite sequencing in GCGC sites of model lambda DNA. Linear model was fitted using average coverage per 5caC percentage group (p value and adjusted R^2 are shown).

See also Figures S1, S2, and S3.

S1A and S1B). We determined that WT SssI MTase effectively methylated uCG sites and did not remove the carboxyl group from 5caC (Figure S1B). Additionally, we assessed the methylation efficiency genome-wide by WGBS using the lambda phage DNA samples. The analysis demonstrated nearly complete (~97.3%) methylation of all genomic uCG sites (Figure S1C), consistent with previous reports (Wu et al., 2014). Considering the known inappropriate bisulfite conversion of methylated cytosines to thymines (2.7% with some bisulfite kits [Holmes et al., 2014] and 3.5% as determined in our experiments), the achieved methylation level can be regarded as complete.

The evaluation of the eM.SssI-decarboxylation activity (step 3) on a model DNA fragment showed that the reaction was effective in a broad range of eM.SssI excess over its target sites (5×–20×). As the efficiency increased marginally above the 10× enzyme excess (Figure S1D), we utilized such conditions for all further protocols of 5caC-decarboxylation, if not specified otherwise. Next, we explored the eM.SssI-promoted decarboxylation on regional and whole-genome scales. Using the Tet1-oxidation of 5mC in human fibroblast IMR90 DNA, we prepared carboxylated model genomic DNA. Then, we evaluated the carboxylation level and eM.SssI-decarboxylation efficiency at the selected 5caCG sites utilizing the differential sensitivity of R.MspI restriction endonuclease to modification of its target site CCGG (R.MspI is immune to the presence of all DNA modifications in its target site, except of 5caC). Analysis of the resistance of the selected genomic sites to the MspI-cleavage by qPCR before and after treatment with eM.SssI revealed ~50%–80% decarboxylation efficiency across the regions (Figure 2A). Assuming that heavily carboxylated DNA was used in this experiment (60%–70% of CG sites are methylated in IMR90 DNA [Lister et al., 2009] which were oxidized by Tet1), we expect more efficient removal of the 5-carboxyl groups from 5caCGs, which rarely occurs in mammalian genomes (a percentage of 5caC/5fC distributes around 8%–10% in mouse ESCs, with some extremums reaching up to 80%; Neri et al., 2015).

We next explored eM.SssI-decarboxylation across 36 CG sites of a model DNA fragment prepared by PCR from a promoter region of human *c-fos* gene. By manipulation with WT

SssI and Tet1 enzymes, we carboxylated CG sites and analyzed the transformation 5caCG → i-uCG by Sanger bisulfite sequencing. To calculate the efficiency of decarboxylation, the differential readout of various modified cytosines in bisulfite sequencing was used: unmodified C and 5caC are converted to uracil and read as thymine, whereas 5mC resists the conversion and remains as C in sequencing (Yu et al., 2012). A good decarboxylation rate was detected, showing, on average, 70% and 60% across all CGs in the upper and the bottom strands, respectively (Figure S2A). Analysis of the bases that immediately flank CG sites revealed a minimal sequence preference for the eM.SssI-mediated decarboxylation; the rate remained 60%, on average, regardless of the proximal purine-pyrimidine composition (Figure S2B).

In caCLEAR, both decarboxylation (step 3) and azide-labeling reactions (step 4) can be successively performed in one tube without prior DNA purification. Addition of Ado-6-N_3 into the decarboxylation mixture resulted in the efficient covalent labeling of uCG sites in our model DNA systems and in c-*fos* fragment (Figures S1E and S2C).

Further, to evaluate the eM.SssI decarboxylation and azide labeling on a whole-genome scale, we carboxylated CG and GCGC sites of the lambda phage DNA to 30% and 60%, respectively, and applied steps 2–5 of caCLEAR (sequencing depth, 50× per each CG; 6,226 CGs in lambda genome). To test the completeness of decarboxylation at i-uCGs, we compared their coverage with that of the control lambda DNA uTOP-seq sample unmethylated to the same extent (30%) at CGs. Analysis revealed a relationship between the coverage of the sites and their carboxylation level (Figure 2B). Moreover, along the wide range of coverage cutoffs, no difference in coverage was observed between the caCLEAR and control uTOP-seq libraries for most of CG sites (due to the different carboxylation levels, GCGC sites were excluded from this analysis), indicating that 5-carboxyl groups were completely removed by eM.SssI (Figure 2C).

We then tested the full caCLEAR procedure on model DNA fragments. The first model system consisted of two DNA fragments, each containing a differentially modified CG site, and the second model system contained all DNA modifications at different CG sites of one DNA fragment (Figures 2D and S3A). With both model systems, caCLEAR generated a single PCR product originating from the 5caC-containing DNA strand (Figures 2D and S3A; lane B) out of four or six possible products, respectively (lane A).

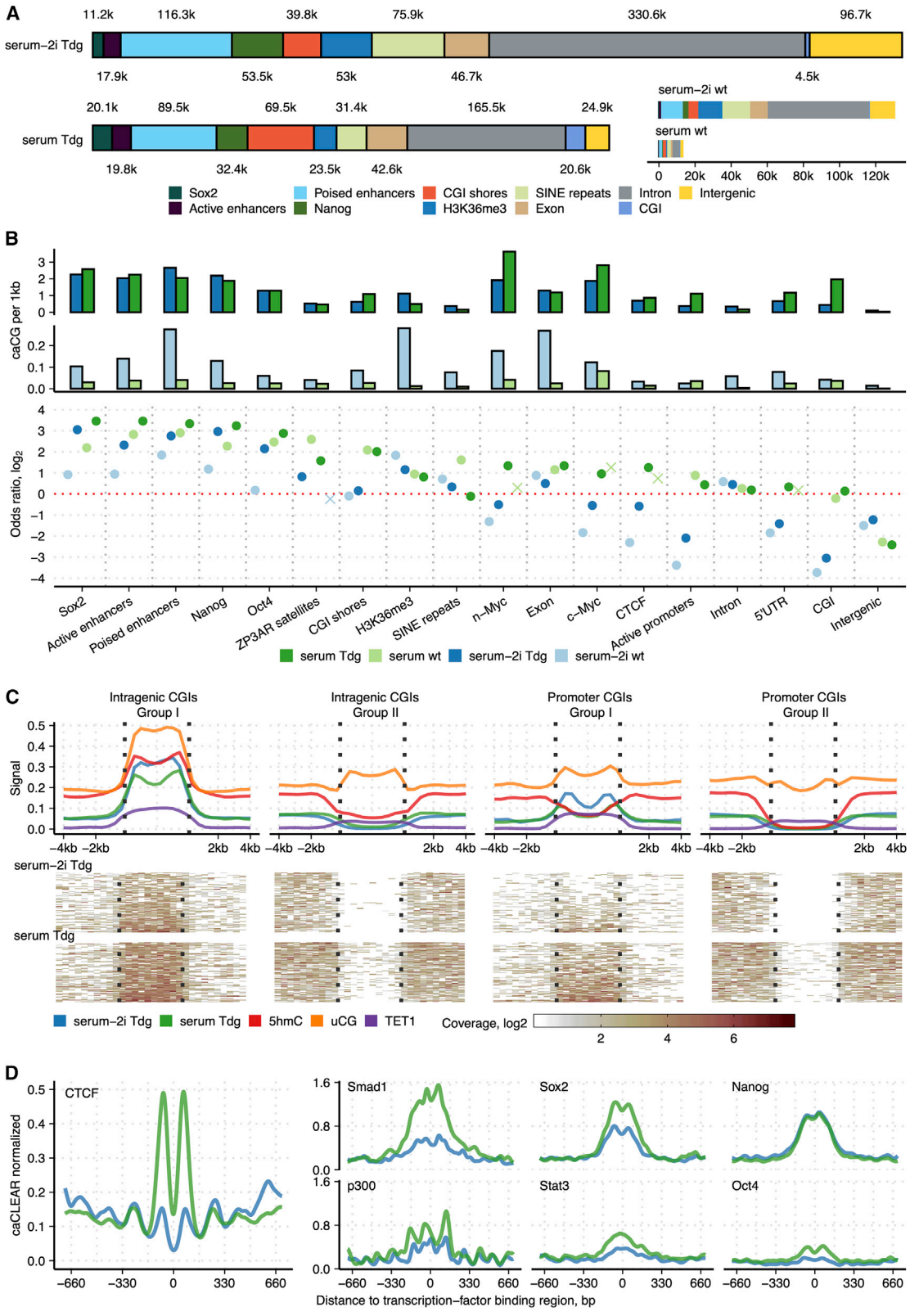
Further, we analyzed caCLEAR libraries of the lambda phage DNA pre-modified to 10% of 5hmC and 5caC at CCGG (328 genomic CGs) and GCGC sites (215 genomic CGs), respectively. In parallel, we carried out control caCLEAR experiments of the samples without 5caC. The analysis demonstrated a good correlation between technical replicates of the libraries (Pearson r = 0.96) (Figure S3B). Importantly, the majority of sequencing reads originated from the carboxylated GCGC sites, as expected, while other CG sites, including 5-hydroxymethylated CCGGs, generated only background read numbers in both target and control libraries (Figure S3B). For a final genome-wide validation, we introduced 5caC into lambda DNA using the Tet1-oxidation of pre-methylated GCGC sites to produce a DNA sample with, on average, 20% of 5caC in GCGC sites. We then measured the percentage of 5caC at each site of this sample by caMAB-seq bisulfite sequencing (the 5caC level varied in the range of 4%–40%) and also analyzed this model DNA by caCLEAR. Notably, assessment of the relationship between caCLEAR coverage and the extent of carboxylation (in 5% intervals of 5caC) revealed a linear relationship between bisulfite sequencing and caCLEAR (Figure 2E).

Altogether, these extensive experiments confirmed the validity of caCLEAR in detection of genomic 5caCG sites.

caCLEAR Analysis of Two Pluripotency States of Mouse ESCs

Having validated caCLEAR in the model DNA systems, we applied the method for analysis of 5caC in mouse ESCs. *Tdg*-depleted (*Tdg*−/−; Tdg) and Tet1/2/3 triple knockout (Tet TKO) ESCs were used as a positive and a negative control, respectively. Depending on culture conditions, ESCs can adopt two interconvertible states resembling two different developmental stages (Habibi et al., 2013; von Meyenn et al., 2016). ESCs grown in serum/leukemia inhibitory factor (LIF)-containing media are similar to cells from the early epiblast, while ESCs cultivated in a serum-free medium with two small-molecule inhibitors (2i ESCs) closely resemble cells from the inner cell mass and have been designated as the ground or naive state of pluripotency (Marks et al., 2012; Martello and Smith, 2014). As many published works (Yu et al., 2012; ENCODE Project Consortium, 2012; Song et al., 2013; Booth et al., 2012; Liu et al., 2019) used the mixture of serum/LIF/2i for the maintenance of a more homogeneous pluripotent population of ESCs, we cultivated the aforementioned three mouse cell lines in serum/LIF and serum/LIF/2i conditions (termed later as serum and serum-2i ESCs, respectively) and analyzed their developmental states through comparison of genome-wide 5caC patterns. We first investigated global amounts of 5mC and all oxi-mCs by HPLC-MS/MS assay (Figure S4A). Analysis revealed a globally hypomethylated state of serum-2i WT ESCs (~1% of 5mC), which is consistent with the reported data of 2i ESCs (Leitch et al., 2013). In contrast, serum WT ESCs showed ~5% levels of cytosine methylation: Leitch et al. (2013) and Stadler et al. (2011) reported ~4% of 5mC. While total amounts of 5hmC were similar in both cultivation conditions, serum-2i ESCs were enriched for other oxi-mC products. We found ~0.00008% 5caC of total guanine in serum WT ESCs, in line with previous reports (Ito et al., 2011; Wheldon et al., 2014), and ~0.00012% 5caC in serum-2i WT ESCs. As expected, analysis revealed considerably higher total 5caC amounts in Tdg ESCs (~0.00097% and ~0.0013% of total guanine in serum and serum-2i Tdg ESCs, respectively). Of note, 5caC and 5fC were undetectable in the Tet TKO cell line in both growth conditions, while 5hmC was found at trace amounts.

We then constructed genome-wide 5caC maps of all cell lines cultivated in both conditions. As global amounts of 5caC differ considerably among the cell lines, the final library sizes reflect the relative 5caC abundance (for sequencing statistics, see Table S1). We obtained approximately ~1.7 M processed single-end reads for each technical replicate of serum WT ESCs and ~15 M for serum Tdg ESCs. Accordingly, ~9 M and ~25 M reads were obtained for serum-2i WT and Tdg ESCs, respectively.



(legend on next page)

Importantly, technical replicates of Tdg ESC caCLEAR libraries showed high correlation (Pearson r = 0.85), while for WT ESC libraries, correlation was lower (range = 0.29–0.42; Figure S4B). caCLEAR generated very low amounts of reads (~0.4 M) in Tet TKO ESCs and identified ~196K CGs, which hardly overlapped between the replicates of the libraries (5,501 shared CGs, 2.8% of all identified CGs), suggesting that these CGs result from a random noise and, thus, are false-positive sites. To confidently call 5caCG sites in WT and Tdg ESCs, we used high-stringency filtering (see STAR Methods), which resulted in a total of 8,685 (mean coverage, 5×) and 374,488 (mean coverage, 20×) 5caCGs in serum WT and Tdg ESCs, respectively; and 110,753 (mean coverage, 8×) and 689,303 (mean coverage, 18×) 5caCGs in serum-2i WT and Tdg ESCs, respectively. Only 1,259 CGs passed our filtering criteria in Tet TKO cells.

We then compared caCLEAR with other 5caC profiling techniques. The 5caCG set of serum Tdg ESCs overlapped well with 5caC peaks detected in the same cell line by the antibody-based DNA immunoprecipitation (5caC-DIP) approach (Shen et al., 2013); 58% of 5caCGs overlapped with the 5caC-DIP-enriched regions (odds ratio [OR] = 18; $p < 2.2 \times 10^{-16}$ Fisher's exact test). In addition, 62% and 80% of 5caC-DIP peaks overlapped with at least one 5caCG in the serum Tdg and serum-2i Tdg datasets, respectively. Next, we compared the 5caCG sets identified in serum Tdg ESCs by caCLEAR and CAB-seq, a 5caC chemically assisted bisulfite sequencing method (4,806 5caCG sites were detected by CAB-seq) (Lu et al., 2013, 2015). We recovered 46% of CAB-seq data in our high-stringency dataset, and 78% of 5caCGs were recovered with less stringent criteria of 5caCG filtering (at least 1× coverage). We were unable to compare our 5caCG sets with the publicly available M.SssI methylase-assisted bisulfite sequencing (MAB-seq) data, as this bisulfite-based method aims at the simultaneous profiling of 5caC and 5fC modifications (Wu et al., 2014; Neri et al., 2015), whose genomic positions hardly overlap, as reported previously (Lu et al., 2015).

Although the distribution of the called 5caCG sites varied in gene regulatory elements and genomic features for different conditions and cell types (Figure 3A), a majority of 5caCG sites were enriched in poised and active enhancers (marked by H3K4me1 and H3K27ac/H3K4me1, respectively) and binding regions of various pluripotency-related TFs, such as Sox2, Nanog, Oct4 (Pou5Fl); some types of satellite repeats; and SINEs (Figure 3B). The two pluripotency states differed in 5caCG abundance at a wide range of important genomic features; in serum ESCs, 5caCGs were enriched at CG islands and shores, 5′ UTRs, active enhancer regions, active promoter regions marked by H3K4me3, and the insulator protein CTCF-binding sites, whereas serum-2i ESCs were 5caCG depleted or less enriched at these elements. In line with the more demethylated state of serum-2i ESCs, both WT and Tdg serum-2i cell lines were relatively more 5caCG enriched across intergenic regions, introns, and actively transcribed regions marked by H3K36me3, as compared to serum ESCs. WT and *Tdg*-depleted cell lines demonstrated similar genomic 5caCG enrichment for both cultivation conditions, except for TF binding sites, which were less 5caCG enriched in serum-2i WT ESCs. Despite the higher 5caCG abundance across the majority of genomic features in both WT and Tdg serum-2i ESCs (Figure S4C), 5caCGs are less specifically distributed in these cells, as demonstrated by the lower OR values for most of the explored genomic elements and features when compared to serum ESCs (Figure 3B).

The highest difference in 5caCG enrichment between serum-2i and serum ESCs was observed in CG islands and shores (Figure 3B). Analysis of the separate CGI classes (intragenic, promoter, and intergenic CGIs) in Tdg cell lines showed that 5caCGs mainly concentrated in gene-associated CGIs, while intergenic CGIs and their flanking regions were strongly depleted in 5caCGs (Figure S5A). We then ranked promoter and intragenic CGIs into two groups based on the abundance of 5caCGs inside CGIs in relation to their immediate vicinity (Figure 3C): group I demonstrated relative enrichment of 5caCGs compared to the flanking regions, whereas group II was mostly 5caCG poor at CGIs. Importantly, both groups of serum ESCs showed higher 5caCG density inside CGIs compared to serum-2i ESCs. This is further confirmed by the 5caCG profiles across CGIs in both cultivation conditions (Figure 3C). Next, we looked at the main methylation and hydroxymethylation levels of the same CGI groups acquired by uTOP-seq and hmTOP-seq (Gibas et al., 2020), respectively, in serum ESCs and the enrichment of the Tet1 chromatin immunoprecipitation sequencing (ChIP-seq) peaks. The uCG and 5hmCG distribution and Tet1 signal generally followed 5caCG patterns in the group I of intragenic CGIs, with the strongest signal concentrated at the boundaries of CGIs. Although slightly enriched in 5caCGs, promoter CGIs of the group I showed lower methylation levels and depletion in 5hmCGs in relation to the flanking regions. As the analysis demonstrated, group II of both intragenic and promoter CGIs represents the methylated and oxi-mC-poor but Tet1-bound CG islands located in relatively 5caCG-rich genomic areas (Figure 3C).

Localization of 5caC-Modified TF Binding Sites Alters between the Two Pluripotency States

We then investigated the distribution of 5caCGs at the binding sites of 12 TFs and regulators (Nanog, Oct4, Sox2, STAT3,

Figure 3. Distribution of 5caCGs in the Naive and Primed Pluripotency States of Mouse ESCs

(A) Distribution of the called 5caCG sites (in thousands) in genomic elements and features.

(B) Numbers of 5caCGs (per 1 kb) and odds ratio (OR; $\log_2$) from Fisher's exact test for enrichment of 5caCG sites across various genomic features in WT and Tdg ESC cell lines. Poised enhancers, regions with H3K4me1 histone marks; active enhancers, regions with H3K4me1 and H3K27ac marks; active promoters, 2-kb regions upstream of genes that overlap the H3K9ac histone mark. Non-significant estimates ($p \geq 0.05$) are marked with "X."

(C) Upper panel: profiles of 5caCG, 5hmCG, uCG (normalized to CG density), and Tet1 ChIP-seq data across intragenic and promoter CGI groups ranked by 5caCG density inside CGIs in relation to flanking regions. Lower panel: heatmap representation of 5caCG density in top 150 CGIs of each group.

(D) Distribution of 5caCGs around various TFRs (normalized for CG density). ChIP-seq data (Chen et al., 2008; Atlasi et al., 2019) were used for calculation of TFRs.

See also Figures S4, S5A, S5B, and S8 and Table S1.

Smad1, Zfx, c-Myc, n-Myc, E2f1, CTCF, p300, and Suz12) that play important roles in ESCs. We analyzed the presence of 5caCGs in the publicly available ChIP-seq regions (Chen et al., 2008; Atlasi et al., 2019; Galonska et al., 2015). The central peaks of TF binding regions (TFRs; average length of regions, 16–50 bp) (Chen et al., 2008; Atlasi et al., 2019) contained, on average, 1–2 5caCGs. The longer, ~600- to 1,000-bp binding sites of Nanog, Sox2, and Oct4 determined in both cultivation conditions (Galonska et al., 2015) were enriched in 5caCGs within 300 bp of the central areas and included 3–4 5caCGs on average. The caCLEAR profiles around the TFRs revealed 5caCG enrichment at the Sox2, Nanog, Oct4, Smad1, STAT3, CTCF, and p300 sites (Figure 3D). Interestingly, we found a periodically spaced distribution of 5caCGs around the CTCF-bound regions that peaked with 165-bp intervals characteristic of the length of the DNA linker between nucleosomes. A similar coincidence with the nucleosome array structure at CTCF sites was previously observed for the distribution of 5hmC and 5fC (Sun et al., 2015). Our results further confirmed the suggested influence of local chromatin structure to the activity of Tet proteins (Sun et al., 2013). The oscillating pattern of 5caCGs was also observed, though less evident, around the sites of the co-activator p300, which is known to interact with various TFs and functions as histone acetyltransferase.

In both analyzed cell lines, the 5caC-modified TFRs (5caC-TFRs) overlapped with H3K36me3 chromatin modification and poised enhancers (Figure 4A). In contrast to serum-2i ESCs, in serum ESCs, 5caCG-TFRs were enriched at active enhancers and H3K4me3/H3K9ac marks (characteristic of active promoter regions). Considering that TFs are recruited by enhancers, we calculated the distance distribution of 5caCG-TFRs according to the nearest annotated transcription start site (TSS). It is known that pluripotency factors can switch from the proximal to distal elements following the transition from serum to 2i conditions (Galonska et al., 2015; Tesar et al., 2007). We detected that 5caCG-TFRs of the majority of investigated TFs were positioned at longer distances to TSSs as compared to the 5caCG-depleted TFRs and that these distances were considerably longer in serum-2i ESCs (Figure 4B). Notably, in serum ESCs, the 5caCG-modified binding sites of the key pluripotency factors Nanog, Oct4, and Sox2 were located at the more proximal positions, as compared to the 5caC-depleted TFRs.

We then analyzed enrichment and fractional distribution of 5caCG-TFRs at promoters, genes, and intergenic areas using various intervals from TSSs (Figures 4C and S5B). In both states of ESCs, the 5caCG-TFRs of c-Myc, n-Myc, E2F1, Suz12, and Zfx tended to distribute in intergenic areas, whereas 5caCG-CTCF sites were exceptionally strongly enriched in exons (Figure 4C). As a considerable fraction of 5caCG-TFRs resided in introns and exons (Figure S5B), we evaluated expression levels of their host genes (Figure 4D). All investigated 5caCG-TFRs were positioned in moderately expressed gene groups, except for the c-Myc binding regions, which were enriched in highly expressed genes in serum-2i ESCs. A predominant function of c-Myc in ESCs is maintenance of pluripotency (Cartwright et al., 2005; Singh and Dalton, 2009). Recent data suggested that c-Myc partner Max acts as a sensor of 5caC, whose level can modulate the Myc-Max transcriptional network (Wang et al., 2017). Although Gene Ontology (GO) classification of c-Myc target genes upregulated in serum ESCs did not identify categories associated with developmental processes (Marks et al., 2012), our functional analysis showed that genes with 5caCG-c-Myc binding regions, indeed, have links to developmental processes in serum conditions (GO terms associated with embryo development, $p = 3 \times 10^{-5}$). Therefore, we hypothesize that 5caCG-modified sites attract c-Myc, which participates in shifting ESCs to the more primed state, preparing cells for developmental activation.

The aforementioned data demonstrated that 5caCG-CTCF regions were enriched in genes. CTCF regulates multiple genomic functions, promoting long-range interactions between enhancers and promoters and insulating areas of active transcription (Phillips and Corces, 2009; Ong and Corces, 2014). Many pluripotency gene-enhancer interactions in ESCs, which usually undergo transformations during differentiation, are anchored by CTCF (Beagan et al., 2017). Although 5caC modification levels at the neighboring areas of CTCF binding regions were more prominent in serum ESCs, both pluripotency states demonstrated similar 5caCG patterns, with the proximal areas more enriched in 5caCGs than the binding sites themselves (Figure 3D). Furthermore, 5caCG-CTCF loci were strongly enriched within active promoter marks in the primed state of ESCs (Figure 4A). To test biological associations of genes overlapping with 5caCG-CTCF binding sites in each state of ESCs, we performed GO functional annotation analysis. The serum-specific gene group revealed associations with developmental processes ($p = 5.5 \times 10^{-8}$) and regulation of signaling ($p = 6 \times 10^{-7}$), while the serum-2i set showed links to protein targeting to membrane categories ($p = 3.5 \times 10^{-5}$). This suggests that, through sensing the oxidation state of its binding sites, CTCF might modulate expression of genes, driving the switch between the naive ESCs and those that are primed for development. Of 5caCG-modified pluripotency TFRs, only Nanog site-containing genes demonstrated functional links with developmental processes (positive regulation of cellular process, $p = 3.5 \times 10^{-7}$; developmental process, $p = 4.7 \times 10^{-5}$).

5caCGs Distribute in the Antisense Strand of Actively Transcribed Genes

To discern whether 5caC characterizes the pluripotency-state-specific gene expression, we calculated differently carboxylated genes between the naive and primed ESCs. More 5caCG-enriched genes were identified in serum-2i ESCs, as expected (522 versus 83 genes; ANOVA q value < 0.05; absolute $\log_2$ FC > 4) (Figure 5A). However, GO functional analysis demonstrated no strong enrichment of this large cluster of genes, except for their weak association with nucleic acid metabolic processes. In contrast, the serum-specific genes revealed functional links with DNA-binding TF activity, multicellular organism development, and cell-fate specification (Figure 5A). These data are in line with the reported relationship of genes upregulated in 2i ESCs with metabolic processes, while upregulated genes in serum ESCs are linked to developmental processes (Marks et al., 2012), pointing to an involvement of active demethylation in the establishment of the pluripotency states.

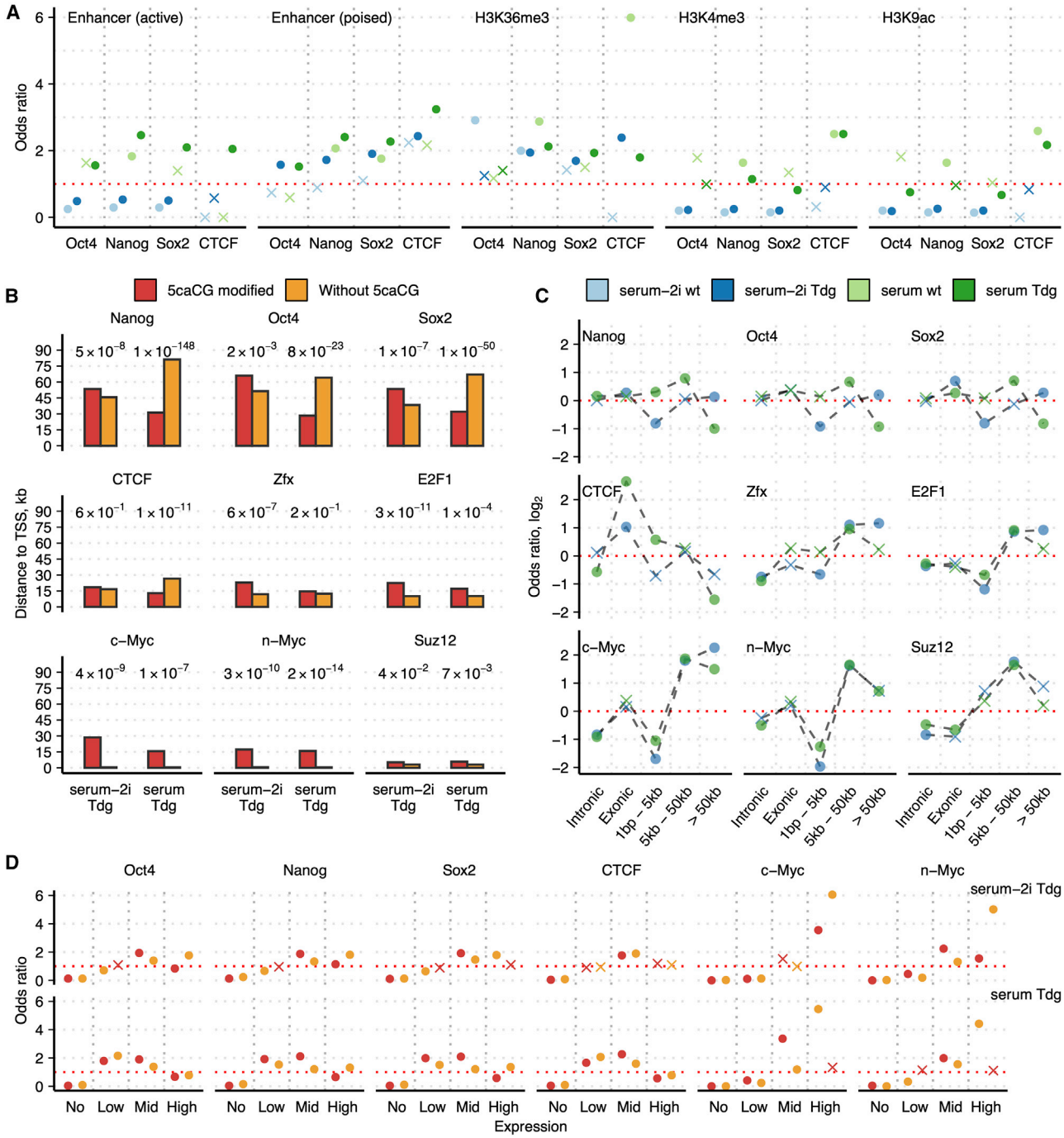


Figure 4. 5caCG-Modified TFRs Differently Distribute in Two Pluripotency States of ESCs and Are Associated with Active Expression

(A) OR from Fisher's exact test for enrichment of 5caCG-modified TFRs (5caCG-TFRs) in poised/active enhancers marked by H3K4me1 and H3K27ac/H3K4me1, respectively; active promoters marked by H3K4me3 and H3K9ac; and regions of active transcription marked by H3K36me3.

(B) Median distance of 5caCG-TFRs to transcription start sites (TSSs) in relation to 5caG-depleted TFRs. Values above the bar plots indicate p values calculated using the Mann-Whitney test.

(C) OR from Fisher's exact test for enrichment of 5caCG-TFRs in genic elements and gene proximal and distal sites.

(D) Enrichment distribution of 5caCG-TFRs and 5caCG-depleted TFRs across different gene expression groups. Non-significant estimates ($p \geq 0.05$) are marked with "X."

See also Figure S5B.

CellPress
OPEN ACCESS

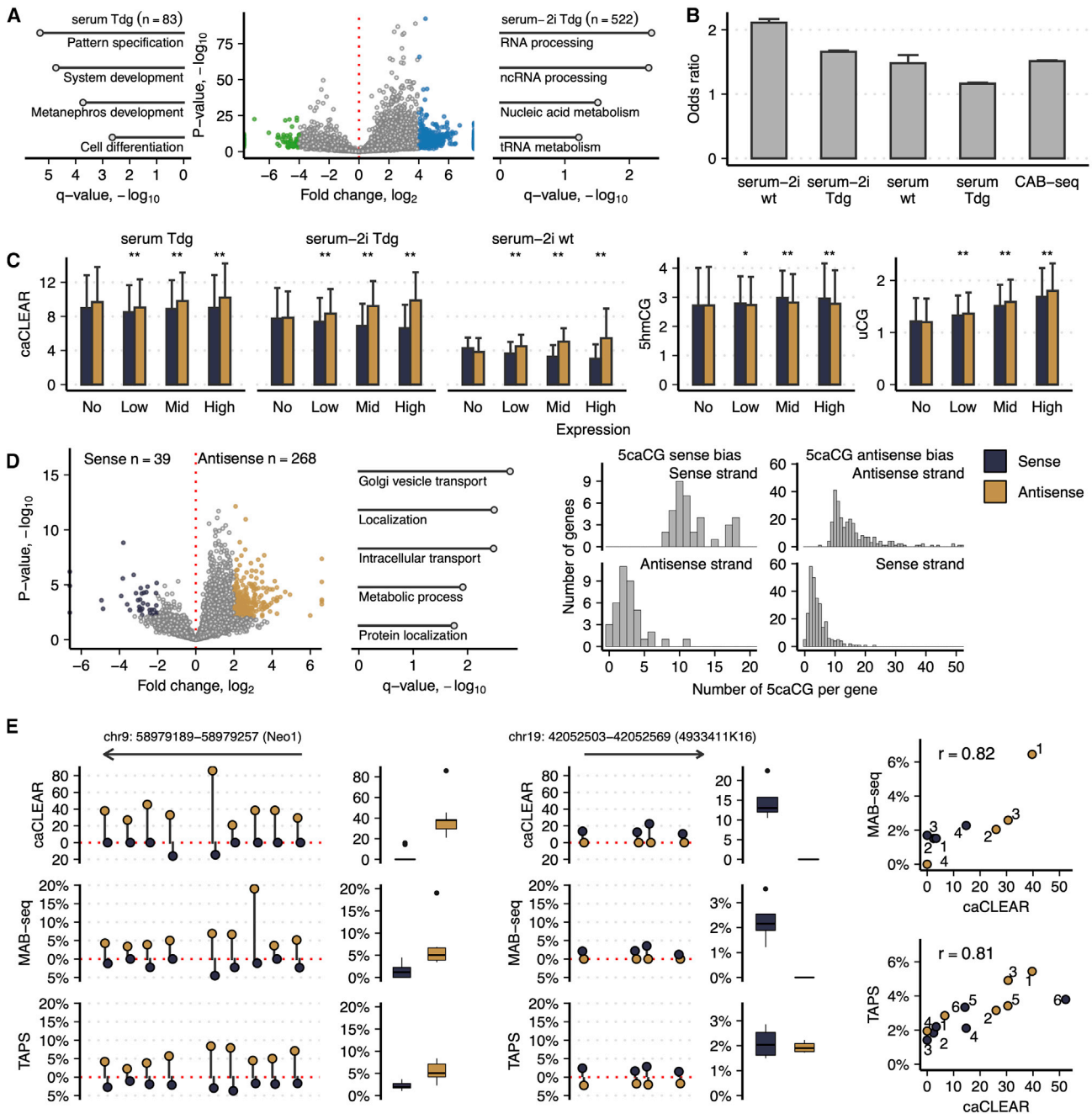


Figure 5. 5caCGs Distribute in the Antisense Strand of Actively Transcribed Genes

(A) Volcano plot indicates distribution of 5caCG-enriched genes in the two pluripotency states of ESCs. Functional annotation analysis is provided for the significant sets of genes, n = 83 and n = 522, shown by green and blue, in serum and serum-2i Tdg ESCs, respectively (ANOVA and F test q values < 0.05; absolute $\log_2$ FC > 4).

(B) Odds ratio for enrichment of 5caCGs at the antisense strand of genes for both WT and Tdg cell lines. The 5caCG bias toward the antisense strand of protein-coding genes was also detected for 2,803 5caCGs identified by the bisulfite-based CAB-seq method (Lu et al., 2015). For all reported estimates, $p < 2.2 \times 10^{-16}$.

(C) The strand bias of 5caCG, 5hmCG, and uCG sites across different gene expression groups. 5hmCG and uCG data are presented for serum WT ESCs. Two-sided paired t test p values are denoted by asterisks (*$p < 0.05$; **$p < 1 \times 10^{-7}$).

(D) Left: volcano plot demonstrates distribution of genes with the 5caCG bias in the sense and antisense strands in serum-2i ESCs. Functional annotation analysis is indicated for genes with the 5caCG bias in the antisense strand (Anova F-test q-values < 0.05; absolute $\log_2$ FC > 2). Right: histogram distribution of the genes with 5caCG bias indicates the difference in 5caCG amounts detected in both genic strands.

(legend continued on next page)

It is known that 5mC levels are symmetric across CG sites due to the maintenance mechanism of DNA methylation, but a major portion of 5hmC and 5fC distribute asymmetrically at CG sites (Wu et al., 2014; Yu et al., 2012; Booth et al., 2014). The MAB-seq approach has determined a strong CG modification asymmetry for the aggregate 5fC/5caC signal: only ~5% of the called CG sites (22,590 out of 454,400 called CG dyads) were symmetrically modified with 5fC/5caC. Since genomic positions of 5caC and 5fC barely overlap (Lu et al., 2015), we set out to investigate the strand-specific distribution of 5caCG sites in both cultivation conditions. First, using the data of the lambda phage experiment described in Figures 2E and S3B, we tested whether there is a potential strand bias for eM.SssI-mediated CG labeling and found no strand-specific difference in the caCLEAR signal across CG sites ($p = 0.18$, two-sided paired t test). To quantify the strand distribution of 5caCGs in ESCs, we first focused on the protein-coding genes with high abundance of 5caCGs. Strikingly, analysis of both WT and Tdg cell lines evidenced asymmetric distribution of 5caCG sites ($p < 2.2 \times 10^{-16}$, Fisher's exact test; Figure 5B) toward the antisense strand of transcribed genes (Figure 5C). Furthermore, the 5caCG strand bias was more prominent in highly expressed genes of serum-2i ESCs. In addition, we evaluated the average levels of unmethylated CGs and 5hmCGs across the same set of genes through the use of our uTOP-seq and 5hmC-specific tethered oligonucleotide–primed sequencing (hmTOP-seq) data of serum WT ESCs (Gibas et al., 2020). Similarly to the 5caCG data, more extensively expressed genes contained a higher abundance of uCG sites in their antisense strand. In contrast, 5hmCGs demonstrated a preference for the sense strand across the same gene sets (Figure 5C), in line with the reported data of mouse and human tissues (Booth et al., 2012; Wen et al., 2014; Gibas et al., 2020), indicating a distinct and highly regulated 5mC demethylation at both genic strands. The detected 5caC distribution bias was further confirmed by the analysis of 2,803 5caCGs identified by CAB-seq across protein-coding genes in serum Tdg ESCs (Figure 5B). Notably, we did not detect a significant strand-specific difference in 5caCG levels for any other genomic elements (CG islands, repeat elements, antisense RNAs, long non-coding RNAs, or processed pseudogenes), suggesting the presence of mechanisms maintaining the 5caC distribution bias exceptionally in protein-coding genes. Next, focusing on serum-2i Tdg ESCs data, we calculated genes with a significant difference in strand-specific 5caC levels: a volcano plot demonstrated a shift toward the antisense strand (Figure 5D). We identified 268 and 39 genes with a higher abundance of 5caCGs in the antisense and sense strands, respectively (absolute $\log_2$ fold change > 2; ANOVA q value < 0.05; Figure 5D). The genes enriched in 5caCGs across the antisense strand corresponded to highly expressed gene groups (Figure S5C) and were mainly associated with metabolic and intracellular transport categories in GO functional analysis (Figure 5D). No significant GO terms were associated with the sense strand set of genes.

To confirm the observed 5caCG distribution bias, we quantified 5caC levels at CGs across 4 selected loci using locus-specific MAB-seq for 5caC detection (caMAB-seq) (Song et al., 2013; Wu et al., 2016). To discriminate between 5caC and 5fC in bisulfite sequencing, 5fC was converted into 5hmC using sodium borohydride reduction (Song et al., 2013). The trace amounts of 5fC in HPLC-MS/MS analysis demonstrated efficient conversion of 5fC into 5hmC in genomic DNA (Figure S6A). Furthermore, genomic uCG sites were pre-methylated with WT SssI using the caCLEAR methylation protocol. Following this integrative approach, 5caC was red as T after bisulfite conversion, while all other cytosine modifications were resistant to bisulfite treatment and remained as C. In addition, for comparison with caCLEAR, we used one of the modifications of a bisulfite-free pyridine borane sequencing (TAPS; Liu et al., 2019) that allows the detection of 5caC. In this approach, following 5fC reduction to 5hmC, all cytosine states are recognized as C, except for 5caC, which is converted into dihydrouridine (Figure S6B) and is read as T. Consistent with caCLEAR, the strand-specific caMAB-seq and TAPS analysis confirmed the 5caCG asymmetry at the selected loci (Figures 5E and S7). All three methods demonstrated high consistency at this single-CG analysis; caCLEAR showed high correlation with both of the other methods (0.81–0.82; Figure 5E, right panel; correlation of caMAB-seq and TAPS was 0.79), confirming the sensitivity and robustness of the caCLEAR approach.

DISCUSSION

To understand mammalian active DNA demethylation, sensitive high-resolution methods are required for genome-wide mapping of the oxidized cytosine derivatives. Our ultimate goal was to develop a strategy for sensitive detection of all genomic 5caCG occurrences, avoiding bisulfite conversion, which causes many technical and analytical challenges. The caCLEAR technique identifies genomic 5caCG positions via mild enzymatic removal of the carboxyl groups from 5caC, yielding uCG sites that are detected by uTOP-seq profiling, while natural genomic uCG sites are excluded from analysis by WT SssI-directed methylation. As caCLEAR specifically targets only 5caCG sites, it requires less sequencing effort, compared to bisulfite-based 5caC mapping methods, and thus enables cost-efficient construction of 5caCG profiles.

The caCLEAR analysis of mouse ESCs suggested that 5caC may be more widespread than previously anticipated. The sensitivity of caCLEAR also allowed detection of 5caCGs in WT ESCs, thereby eliminating the requirement of *Tdg* depletion for the detection of carboxylated CG sites.

(E) Left: representation of caCLEAR coverage and the percentage of 5caC determined by caMAB-seq and pyridine borane sequencing in both strands of CGs across the loci in *Neo1* and *4933411K16* genes, which were identified as having the 5caCG bias in the antisense and sense strands, respectively. Genomic coordinates of the loci and the direction of the genes are indicated. Boxplots indicate the signal difference between the individual strands of the corresponding region. Right: pairwise correlation between caCLEAR and caMAB-seq or TAPS signal averaged across individual strands of all analyzed regions (shown in numbers 1–6) (correlation p values < 0.014).
See also Figures S5C, S6, and S7 and Table S2.

Genome-wide distribution of 5caCGs in the two pluripotent states of ESCs revealed their predominant accumulation at enhancers and other open genomic regions that are usually connected by a TF network, pointing to the dynamic turnover of 5mC at these loci. Interestingly, we detected the enrichment of 5caCGs across open chromatin loci of various mouse tissues in the primed but not naive ESCs (Figure S8), suggesting a similarity between the three-dimensional chromatin structures of the ESCs that were primed for development and differentiated mouse cells. We demonstrated that global 5caCG patterns are shaped by the pluripotency state and showed that serum-2i ESCs generally resemble the pure 2i state. Several studies have highlighted transcriptional and DNA modification changes following transition from one state to another (Marks et al., 2012; Ficz et al., 2013). We detected a switch from the gene proximal positions of 5caCG-TFRs in serum ESCs to the more distal elements in serum-2i ESCs for the majority of important TFs, except for Nanog, Oct4, and Sox2, which remained positioned closer to genes. Although 5caCGs are less abundant in serum ESCs in relation to serum-2i ESCs, 5caCGs specifically distribute across genes important for cell development processes in the primed ESCs. Moreover, 5caCG-TFRs accumulate in active enhancers and promoters of the primed ESCs, while such TFRs are restricted to poised enhancers in the naive state.

5caCG enrichment at the binding regions of various key TFs raised a hypothesis on the involvement of 5caC in TF-driven gene expression regulation. Strikingly, we detected a tendency for 5caCG-TFRs to reside in moderately expressed genes as compared to 5caCG-depleted TFRs. This might be associated with a potential impairment of transcriptional elongation due to the presence of 5caC (Kellinger et al., 2012). Interestingly, a fraction of the 5caCG-modified sites of CTCF, c-Myc, and Nanog tended to distribute across genes linked with cellular developmental processes exceptionally in the primed ESCs.

Generally, we found an association of 5caCGs and 5caCG-TFRs with active gene expression. Moreover, we detected a distribution of 5caCGs in the antisense strand of highly expressed genes that was more pronounced in serum-2i ESCs. As the antisense strand of the 5caCG-biased genes also contained higher levels of uCG sites, this points to the distinct Tet/Tdg demethylation processivity at both genic strands, which might be important for the maintenance of gene expression in ESCs. Whether this trend is inherent to differentiated mouse tissues or other organisms requires further investigation. The genes with a prominent strand-specific 5caCG bias in serum-2i ESCs are mainly linked to housekeeping functions, such as metabolic, signaling, and cell-cycle-control processes. On the whole, caCLEAR demonstrated that 5caC is an active mark of gene expression that, through recruiting TFs to Tet-oxidized open genomic targets, maintains or shifts the pluripotency states of ESCs. Therefore, 5caC may be a good indicator of dynamic reorganization of the DNA modification and chromatin landscape of a cell.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2020.108155.

ACKNOWLEDGMENTS

We are grateful to Prof. S. Klimašauskas and Prof. V. Masevičius for eM.SssI and Ado-6-N_3 cofactor and to Prof. G. Xu for Tet TKO and *Tdg*-depleted mouse ESC cell lines. We thank J. Gordevičius for consulting on bioinformatic analyses and Z. Staševskij for help with the uTOP-seq procedure. The work was supported by the European Social Fund according to the activity measure No. 09.3.3-LMT-K-712 "Improvement of researchers' qualification by implementing world-class R&D projects" under grant agreement with the Research Council of Lithuania (grant number 09.3.3-LMT-K-712-01-0041 to E.K.).

AUTHOR CONTRIBUTIONS

E.K. conceived the method and coordinated the experimental design of the caCLEAR technology, analytical procedures, and genomic 5caC analyses. J.L. established the caCLEAR protocol, performed method validation, prepared caCLEAR and caMAB-seq libraries, and analyzed HPLC-MS/MS data. P.G. established the data analysis pipelines and carried out all bioinformatic and statistical analyses. K.S. performed validation of caCLEAR on model DNA fragments using uTOP-seq and bisulfite sequencing. K.S. and J.L. prepared TAPS libraries. V.S. established cultivation protocols of mESCs. A.R. established HPLC-MS/MS methods, analyzed the data, and performed sequencing. E.K., J.L., and P.G. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: January 16, 2020
Revised: July 10, 2020
Accepted: August 25, 2020
Published: September 15, 2020

REFERENCES

Atlasi, Y., Megchelenbrink, W., Peng, T., Habibi, E., Joshi, O., Wang, S.-Y., Wang, C., Logie, C., Poser, I., Marks, H., and Stunnenberg, H.G. (2019). Epigenetic modulation of a hardwired 3D chromatin landscape in two naive states of pluripotency. Nat. Cell Biol. *21*, 568–578.

Beagan, J.A., Duong, M.T., Titus, K.R., Zhou, L., Cao, Z., Ma, J., Lachanski, C.V., Gillis, D.R., and Phillips-Cremins, J.E. (2017). YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. Genome Res. *27*, 1139–1152.

Booth, M.J., Branco, M.R., Ficz, G., Oxley, D., Krueger, F., Reik, W., and Balasubramanian, S. (2012). Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Science *336*, 934–937.

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

Cartwright, P., McLean, C., Sheppard, A., Rivett, D., Jones, K., and Dalton, S. (2005). LIF/STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism. Development *132*, 885–896.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106–1117.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46* (*D1*), D794–D801.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics *10*, 48.

Eleftheriou, M., Pascual, A.J., Wheldon, L.M., Perry, C., Abakir, A., Arora, A., Johnson, A.D., Auer, D.T., Ellis, I.O., Madhusudan, S., and Ruzov, A. (2015). 5-Carboxylcytosine levels are elevated in human breast cancers and gliomas. Clin. Epigenetics *7*, 88.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Ficz, G., Hore, T.A., Santos, F., Lee, H.J., Dean, W., Arand, J., Krueger, F., Oxley, D., Paul, Y.-L., Walter, J., et al. (2013). FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. Cell Stem Cell *13*, 351–359.

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47* (*D1*), D766–D773.

Galonska, C., Ziller, M.J., Karnik, R., and Meissner, A. (2015). Ground State Conditions Induce Rapid Reorganization of Core Pluripotency Factor Binding before Global Epigenetic Reprogramming. Cell Stem Cell *17*, 462–470.

Gibas, P., Narmontė, M., Staševskij, Z., Gordevičius, J., Klimašauskas, S., and Kriukienė, E. (2020). Precise genomic mapping of 5-hydroxymethylcytosine via covalent tether-directed sequencing. PLoS Biol. *18*, e3000684.

Habibi, E., Brinkman, A.B., Arand, J., Kroeze, L.I., Kerstens, H.H.D., Matarese, F., Lepikhov, K., Gut, M., Brun-Heath, I., Hubner, N.C., et al. (2013). Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. Cell Stem Cell *13*, 360–369.

Holmes, E.E., Jung, M., Meller, S., Leisse, A., Sailer, V., Zech, J., Mengdehl, M., Garbe, L.-A., Uhl, B., Kristiansen, G., and Dietrich, D. (2014). Performance evaluation of kits for bisulfite-conversion of DNA from tissues, cell lines, FFPE tissues, aspirates, lavages, effusions, plasma, serum, and urine. PLoS ONE *9*, e93933.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., and Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science *333*, 1300–1303.

Iurlaro, M., Ficz, G., Oxley, D., Raiber, E.-A., Bachman, M., Booth, M.J., Andrews, S., Balasubramanian, S., and Reik, W. (2013). A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation. Genome Biol. *14*, R119.

Kellinger, M.W., Song, C.-X., Chong, J., Lu, X.-Y., He, C., and Wang, D. (2012). 5-formylcytosine and 5-carboxylcytosine reduce the rate and substrate specificity of RNA polymerase II transcription. Nat. Struct. Mol. Biol. *19*, 831–833.

Kriukienė, E., Labrie, V., Khare, T., Urbanavičiūtė, G., Lapinaitė, A., Koncevičius, K., Li, D., Wang, T., Pai, S., Ptak, C., et al. (2013). DNA unmethylome profiling by covalent capture of CpG sites. Nat. Commun. *4*, 2190.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.

Leitch, H.G., McEwen, K.R., Turp, A., Encheva, V., Carroll, T., Grabole, N., Mansfield, W., Nashun, B., Knezovich, J.G., Smith, A., et al. (2013). Naive pluripotency is associated with global DNA hypomethylation. Nat. Struct. Mol. Biol. *20*, 311–316.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Liu, Y., Siejka-Zielińska, P., Velikova, G., Bi, Y., Yuan, F., Tomkova, M., Bai, C., Chen, L., Schuster-Böckler, B., and Song, C.-X. (2019). Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. Nat. Biotechnol. *37*, 424–429.

Liutkevičiūtė, Z., Kriukienė, E., Ličytė, J., Rudytė, M., Urbanavičiūtė, G., and Klimašauskas, S. (2014). Direct decarboxylation of 5-carboxylcytosine by DNA C5-methyltransferases. J. Am. Chem. Soc. *136*, 5884–5887.

Lu, X., Song, C.-X., Szulwach, K., Wang, Z., Weidenbacher, P., Jin, P., and He, C. (2013). Chemical modification-assisted bisulfite sequencing (CAB-Seq) for 5-carboxylcytosine detection in DNA. J. Am. Chem. Soc. *135*, 9315–9317.

Lu, X., Han, D., Zhao, B.S., Song, C.-X., Zhang, L.-S., Doré, L.C., and He, C. (2015). Base-resolution maps of 5-formylcytosine and 5-carboxylcytosine reveal genome-wide DNA demethylation dynamics. Cell Res. *25*, 386–389.

Lukinavičius, G., Lapinaitė, A., Urbanavičiūtė, G., Gerasimaitė, R., and Klimašauskas, S. (2012). Engineering the DNA cytosine-5 methyltransferase reaction for sequence-specific labeling of DNA. Nucleic Acids Res. *40*, 11594–11602.

Lukinavičius, G., Tomkuvienė, M., Masevičius, V., and Klimašauskas, S. (2013). Enhanced chemical stability of adomet analogues for improved methyltransferase-directed labeling of DNA. ACS Chem. Biol. *8*, 1134–1139.

Marks, H., Kalkan, T., Menafra, R., Denissov, S., Jones, K., Hofemeister, H., Nichols, J., Kranz, A., Stewart, A.F., Smith, A., and Stunnenberg, H.G. (2012). The transcriptional and epigenomic foundations of ground state pluripotency. Cell *149*, 590–604.

Martello, G., and Smith, A. (2014). The nature of embryonic stem cells. Annu. Rev. Cell Dev. Biol. *30*, 647–675.

Masevičius, V., Nainytė, M., and Klimašauskas, S. (2016). Synthesis of S-Adenosyl-L-Methionine Analogs with Extended Transferable Groups for Methyltransferase-Directed Labeling of DNA and RNA. Curr. Protoc. Nucleic Acid Chem. *64*, 1.36.1–1.36.13.

Neri, F., Incarnato, D., Krepelova, A., Rapelli, S., Anselmi, F., Parlato, C., Medana, C., Dal Bello, F., and Oliviero, S. (2015). Single-Base Resolution Analysis of 5-Formyl and 5-Carboxyl Cytosine Reveals Promoter DNA Methylation Dynamics. Cell Rep. *10*, 674–683.

Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. *15*, 234–246.

Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. Cell *137*, 1194–1211.

Shen, L., Wu, H., Diep, D., Yamaguchi, S., D'Alessio, A.C., Fung, H.-L., Zhang, K., and Zhang, Y. (2013). Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics. Cell *153*, 692–706.

Singh, A.M., and Dalton, S. (2009). The cell cycle and Myc intersect with mechanisms that regulate pluripotency and reprogramming. Cell Stem Cell *5*, 141–149.

Song, C.-X., and He, C. (2013). Potential functional roles of DNA demethylation intermediates. Trends Biochem. Sci. *38*, 480–484.

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

Song, C.-X., Szulwach, K.E., Dai, Q., Fu, Y., Mao, S.-Q., Lin, L., Street, C., Li, Y., Poidevin, M., Wu, H., et al. (2013). Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming. Cell *153*, 678–691.

Song, C.-X., Yin, S., Ma, L., Wheeler, A., Chen, Y., Zhang, Y., Liu, B., Xiong, J., Zhang, W., Hu, J., et al. (2017). 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. Cell Res. *27*, 1231–1242.

Spruijt, C.G., Gnerlich, F., Smits, A.H., Pfaffeneder, T., Jansen, P.W.T.C., Bauer, C., Münzel, M., Wagner, M., Müller, M., Khan, F., et al. (2013). Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives. Cell *152*, 1146–1159.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Staševskij, Z., Gibas, P., Gordevičius, J., Kriukienė, E., and Klimašauskas, S. (2017). Tethered Oligonucleotide-Primed Sequencing, TOP-Seq: A High-Resolution Economical Approach for DNA Epigenome Profiling. Mol. Cell *65*, 554–564.e6.

Storebjerg, T.M., Strand, S.H., Høyer, S., Lynnerup, A.-S., Borre, M., Ørntoft, T.F., and Sørensen, K.D. (2018). Dysregulation and prognostic potential of 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) levels in prostate cancer. Clin. Epigenetics *10*, 105.

Sun, M., Song, C.-X., Huang, H., Frankenberger, C.A., Sankarasharma, D., Gomes, S., Chen, P., Chen, J., Chada, K.K., He, C., and Rosner, M.R. (2013). HMGA2/TET1/HOXA9 signaling pathway regulates breast cancer growth and metastasis. Proc. Natl. Acad. Sci. USA *110*, 9920–9925.

Sun, Z., Dai, N., Borgaro, J.G., Quimby, A., Sun, D., Corrêa, I.R., Jr., Zheng, Y., Zhu, Z., and Guan, S. (2015). A sensitive approach to map genome-wide 5-hydroxymethylcytosine and 5-formylcytosine at single-base resolution. Mol. Cell *57*, 750–761.

Tesar, P.J., Chenoweth, J.G., Brook, F.A., Davies, T.J., Evans, E.P., Mack, D.L., Gardner, R.L., and McKay, R.D.G. (2007). New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature *448*, 196–199.

Vilkaitis, G., Dong, A., Weinhold, E., Cheng, X., and Klimasauskas, S. (2000). Functional roles of the conserved threonine 250 in the target recognition domain of HhaI DNA methyltransferase. J. Biol. Chem. *275*, 38722–38730.

von Meyenn, F., Iurlaro, M., Habibi, E., Liu, N.Q., Salehzadeh-Yazdi, A., Santos, F., Petrini, E., Milagre, I., Yu, M., Xie, Z., et al. (2016). Impairment of DNA Methylation Maintenance Is the Main Cause of Global Demethylation in Naive Embryonic Stem Cells. Mol. Cell *62*, 983.

Wang, D., Hashimoto, H., Zhang, X., Barwick, B.G., Lonial, S., Boise, L.H., Vertino, P.M., and Cheng, X. (2017). MAX is an epigenetic sensor of 5-carboxylcytosine and is altered in multiple myeloma. Nucleic Acids Res. *45*, 2396–2407.

Wen, L., Li, X., Yan, L., Tan, Y., Li, R., Zhao, Y., Wang, Y., Xie, J., Zhang, Y., Song, C., et al. (2014). Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. Genome Biol. *15*, R49.

Wheldon, L.M., Abakir, A., Ferjentsik, Z., Dudnakova, T., Strohbuecker, S., Christie, D., Dai, N., Guan, S., Foster, J.M., Corrêa, I.R., Jr., et al. (2014). Transient accumulation of 5-carboxylcytosine indicates involvement of active demethylation in lineage specification of neural stem cells. Cell Rep. *7*, 1353–1361.

Williams, K., Christensen, J., Pedersen, M.T., Johansen, J.V., Cloos, P.A., Rappsilber, J., and Helin, K. (2011). TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature *473*, 343–348.

Wu, H., and Zhang, Y. (2014). Reversing DNA methylation: mechanisms, genomics, and biological functions. Cell *156*, 45–68.

Wu, H., Wu, X., Shen, L., and Zhang, Y. (2014). Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing. Nat. Biotechnol. *32*, 1231–1240.

Wu, H., Wu, X., and Zhang, Y. (2016). Base-resolution profiling of active DNA demethylation using MAB-seq and caMAB-seq. Nat. Protoc. *11*, 1081–1100.

Yu, M., Hon, G.C., Szulwach, K.E., Song, C.-X., Zhang, L., Kim, A., Li, X., Dai, Q., Shen, Y., Park, B., et al. (2012). Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell *149*, 1368–1380.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhou, D., Alver, B.M., Li, S., Hlady, R.A., Thompson, J.J., Schroeder, M.A., Lee, J.-H., Qiu, J., Schwartz, P.H., Sarkaria, J.N., and Robertson, K.D. (2018). Distinctive epigenomes characterize glioma stem cells and their response to differentiation cues. Genome Biol. *19*, 43.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
IMR90, Human Caucasian fetal lung fibroblast, DNA	The European Collection of Cell Cultures (ECACC), UK https://www.phe-culturecollections.org.uk/Public Health England (UK)	Cat#85020204
Lambda phage DNA (dam–, dcm–)	Thermo Fisher Scientific	Cat#SD0021
Chemicals, Peptides, and Recombinant Proteins		
Proteinase K, recombinant, PCR grade	Thermo Fisher Scientific	Cat#EO0491
CpG Methyltransferase (M.SssI)	Thermo Fisher Scientific	Cat#EM0821
T4 beta-glucosyltransferase (BGT)	Thermo Fisher Scientific	Cat#EO0831
5mC Tet1 Oxidation kit	WiseGene	Cat#K003
NgTet1	Active Motif	Cat#81148
eM.SssI	Kriukienė et al., 2013	N/A
M.HhaI	Vilkaitis et al., 2000	N/A
M.HpaII	Lukinavičius et al., 2012	N/A
DBCO-S-S-PEG3-biotin	BroadPharm	Cat#BP-22453
Dynabeads MyOne Streptavidin C1	Thermo Fisher Scientific	Cat#65002
Maxima SYBR Green/ROX qPCR Master Mix (2X)	Thermo Fisher Scientific	Cat#K0221
Pfu DNA polymerase (recombinant)	Thermo Fisher Scientific	Cat#EP0502
Phusion U HS polymerase	Thermo Fisher Scientific	Cat#F555S
Nuclease P1 from Penicillium citrinum	Sigma-Aldrich	Cat#N8630
Ado-6-N_3 cofactor	Lukinavičius et al., 2013; Masevičius et al., 2016 Compound 3a	N/A
CuBr, 99.999%	Sigma-Aldrich	Cat#254185
TBTA	Sigma-Aldrich	Cat#678937
Pyridine borane	Sigma-Aldrich	Cat#179752
$NaBH_4$	Sigma-Aldrich	Cat#213462
FastAP Thermosensitive Alkaline Phosphatase	Thermo Fisher Scientific	Cat#EF0654
Platinum SuperFi PCR Master Mix	Thermo Fisher Scientific	Cat#12358010
DNA Clean & Concentrator kit	Zymo Research	Cat#D4014, Cat#D4034
Oligo Clean and Concentrator kit	Zymo Research	Cat#D4060
Genomic DNA Clean and Concentrator kit	Zymo Research	Cat#D4010
GeneJET Gel Extraction kit	Thermo Fisher Scientific	Cat#K0692
EZ DNA Methylation-Gold Kit	Zymo Research	Cat#D5005
EpiJET Bisulfite Conversion Kit	Thermo Fisher Scientific	Cat#K1461
MagJET NGS Cleanup and Size Selection Kit	Thermo Fisher Scientific	Cat#K2821
CloneJET PCR Cloning Kit	Thermo Fisher Scientific	Cat#K1231
GeneJET Plasmid Miniprep Kit	Thermo Fisher Scientific	Cat#K0503
Deposited Data		
IMR90 WGBS	Lister et al., 2009	GEO: GSM432687
Mouse genome sequence build mm10	UCSC database	https://genome.ucsc.edu
CpG islands	UCSC database	https://genome.ucsc.edu
Repeat Masker annotation	UCSC database	https://genome.ucsc.edu
ATAC-seq data	UCSC database	https://genome.ucsc.edu

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Mouse reference gene dataset	Frankish et al., 2019	https://www.gencodegenes.org/
Histone ChIP-seq data	ENCODE Project Consortium, 2012	https://www.encodeproject.org/
Serum-2i RNA data	Davis et al., 2018	https://www.encodeproject.org/
Serum RNA data	Gene Expression Omnibus	GEO: GSE23943
ChIP-seq of transcription factors	Chen et al., 2008	GEO: GSE11431
ChIP-seq of transcription factors	Atlasi et al., 2019	GEO: GSE92412
ChIP-seq of transcription factors	Galonska et al., 2015	GEO: GSE56312
CAB-seq data	Lu et al., 2015	GEO: GSE56429
5caC-DIP data	Shen et al., 2013	GEO: GSE42250
uCG data	Gibas et al., 2020	GEO: GSE140206
5hmCG data	Gibas et al., 2020	GEO: GSE140206
caCLEAR data	This work	GEO: GSE142319
Experimental Models: Cell Lines		
Mouse: ESC wt E14TG2a	Laboratory of prof. Guo-Liang Xu	N/A
Mouse: ESC Tdg-depleted	Laboratory of prof. Guo-Liang Xu	N/A
Mouse: ESC Tet1/2/3 triple knockout	Laboratory of prof. Guo-Liang Xu	N/A
Oligonucleotides		
TO (tethered oligonucleotide, with or without biotin): TXTTTTGTGTGGTTTGGAGACTGACTACCAGATGT AACA-biotin (X = C8-alkyne-dU)	Base-click	N/A
complementary priming strand: TGTTACATCTGGT AGTCAGTCTCCAAACCACACAA	Exiqon	N/A
Primers for locus-specific analysis of IMR90 DNA, see Region-specific analysis of eM.SssI-directed decarboxylation	Metabion	N/A
Primers for region-specific caMAB-seq bisulfite sequencing, see Table S2	Metabion	N/A
Primers for region-specific pyridine borane sequencing, see Table S2	Metabion	N/A
Ion Torrent barcoded adapters	Kapa Biosystems	Cat#KK8333
Software and Algorithms		
Bismark aligner and methylation caller	Krueger and Andrews, 2011	https://www.bioinformatics.babraham.ac.uk/projects/bismark/
GOrilla online tool	Eden et al., 2009	http://cbl-gorilla.cs.technion.ac.il

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Edita Kriukienė (edita.kriukiene@bti.vu.lt).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

Raw and processed caCLEAR data have been submitted to the NCBI Gene Expression Omnibus under the accession number GSE142319.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

ESCs cultivation

E14TG2a wt and mutant *Tdg*-depleted (*Tdg*−/−; Tdg) and Tet1/2/3 triple knockout (Tet TKO) mouse ESC lines were kindly provided by prof. Guo-Liang Xu. Cells were cultured either in serum media as in Wu et al. (2014) or in serum media supplemented with 2i as in Yu et al. (2012).

METHOD DETAILS

Preparation of model DNA fragments

188 bp and 200 bp model DNA fragments were produced from mouse gDNA by PCR amplification as in Liutkevičiūtė et al. (2014). For 230 bp model DNA fragment 4CG-dir 5'-GCCCATGTCGCTGTG and 4CG-rev 5'-AAGATGTGTCCCGGCT primers were used. 155 bp fragment was amplified from human DNA (BMX gene) with BMX-dir 5′-TGTGTTACTGTGTGGAAAAGACC and BMX-rev 5′-CCACTCCTTATAGTTTGGCTG primers. Model DNA fragments were gel-purified using GeneJET Gel Extraction kit (Thermo Scientific, TS).

To introduce 5hmC at GCGC sites, DNA was incubated with 5-fold molar excess of M.HhaI (Vilkaitis et al., 2000) over its targets in buffer (10 mM Tris-HCl (pH 7.4), 50 mM NaCl, 0.5 mM EDTA) supplemented with 13 μM formaldehyde at room temperature for 1 h. Then, DNA was treated with 0.2 mg/ml Proteinase K (ProtK, TS) and 0.1% of SDS for 1 h at 50°C, and then for 20 min at 65°C. DNA was purified with DNA clean and concentrator kit (Zymo research, ZR). 10 mM Tris-HCl (pH 8.5) was used for the elution of DNA in this and all further column-purification steps.

uCG sites in model DNA fragments were methylated using wild-type SssI (TS) according to the vendor's protocol. 5caC was introduced into both strands of DNA by Tet1-oxidation of 5mC as recommended by the vendor (5mC Tet1 Oxidation kit, WiseGene).

For the caCLEAR experiment with two model DNA fragments (Figure 2D), 5caC in 188 bp fragment was introduced through PCR, in which reverse primer was synthesized with 5caC in a CG site (Liutkevičiūtė et al., 2014). This DNA was used together with a 155 bp fragment containing 5hmC in GCGC site. For the caCLEAR experiment with a single model DNA fragment (Figure S3A), 218 bp DNA was prepared by PCR from mouse gDNA using Sp-PCR-dir 5'-GTGTTGGGGTGACTATTATG and 4CG-rev 5′-AAGAT GTGTCCXGGCT (X = 5caC) primers and then 5hmC was introduced into the GCGC site.

Assessment of M.SssI-directed decarboxylation and azide-labeling in a model DNA fragment

Model DNA fragments were mixed with sonicated lambda phage DNA with a ratio 1:150 ng to imitate low abundance of 5caC. Decarboxylation reaction was performed using 10-fold excess of eM.SssI (Kriukienė et al., 2013) over CG targets in buffer 10 mM Tris-HCl (pH 8), 50 mM NaCl, 0.5 mM EDTA at 22°C for 3 h, and 20 min at 65°C. Then, DNA was treated with ProtK and purified as in the section on Preparation of model DNA fragments. Residual 5caC level was determined as an amount of DNA resistant to R.MspI-cleavage using qPCR (see below).

For the high azide-labeling protocol, the decarboxylation reaction mixture was supplemented with Ado-6-N_3 cofactor (Lukinavičius et al., 2013; Masevičius et al., 2016, Compound 3a) to 0.2 mM and incubated at 30°C for 1 h, heated at 65°C for 20 min, treated with ProtK and purified. For medium labeling 0.3 μM Ado-6-N_3 was used, incubation was done at 30°C for 20 min. To evaluate the labeling efficiency, DBCO-S-S-PEG3-biotin (BroadPharm) was added to a 0.4 mM concentration in 10 mM Tris-HCl (pH 8.5), reactions were incubated at 37°C for 2 h, and DNA was purified with DNA clean and concentrator kit (ZR). Labeled DNA was enriched using Dynabeads MyOne Streptavidin C1 (TS) beads as in Kriukienė et al. (2013) and amounts of DNA in the bead and supernatant fractions were evaluated by qPCR.

DNA restriction and qPCR analysis

10-20 ng of DNA was incubated with 3-10 U of a relevant restriction enzyme (MspI, HpaII, Hin6I) (TS) in a recommended buffer at 37°C for 3 h and heat inactivation was performed. Samples were analyzed by qPCR with primers specific for analyzed DNA regions or fragments (described in Preparation of model DNA fragments and Region-specific analysis of eM.SssI-directed decarboxylation sections, and *Sanger bisulfite sequencing of the c-fos promoter DNA region*) and DNA amounts were calculated using a calibration curve of known DNA concentrations. For restriction analysis, an amount of intact DNA in relation to an uncleaved control sample was calculated. For enrichment analysis of 188 bp, 200 bp and 230 bp DNA fragments primers were used as in Kriukienė et al. (2013). qPCR was performed using Maxima SYBR Green qPCR Master Mix (TS) as recommended by the vendor.

Region-specific analysis of eM.SssI-directed decarboxylation

IMR90 gDNA was purchased from (ECACC, UK). DNA was sonicated on a Bioruptor UCD-200 (Diagenode) to yield ~500 bp fragments. 5mC oxidation to 5caC was performed using Tet1 (see Section on Preparation of model DNA fragments). For eM.SssI treatment, 20 ng of DNA was incubated in a decarboxylation buffer (see Assessment of M.SssI-directed decarboxylation and azide-labeling in a model DNA fragment) using 1.3 μM of eM.SssI and purified as described for model DNA fragments. Samples were analyzed by qPCR with primers specific for selected regions (see below). 5mC and 5caC levels were evaluated using R.HpaII and R.MspI, respectively, as an amount of uncleaved DNA in qPCR. Decarboxylation efficiency was calculated using R.MspI-cleavage

and evaluating intact DNA amounts by qPCR before and after decarboxylation. For restriction cleavage, 9 ng of genomic DNA was incubated with 10 U of a corresponding restriction endonuclease and qPCR was performed using Maxima SYBR Green qPCR Master Mix (TS) as recommended by the vendor. Tested regions: #1 Intergenic region, chr7: 117340531-117340636, (Inter1-dir 5′-tggtgtcccctaaagcataagt; Inter1-rev 5′-tcaagccacatttcccatcg); #2 *CDH13* gene, chr16: 82662108-82662222, (CDH1-dir 5′-cagaaatg cagtgatgggtga; CDH1-rev 5′-gggcagtccttaatcggga); #3 *AGTR1* gene, chr3: 148426487-148426616, (AGTR -dir 5′-tcttcttacatgggcc tatgtga; AGTR -rev 5′-ggactaggagaagctgaggg); #4 *TOP3A* gene, chr17: 18210312-18210435, (TOP3A-dir 5′-cagtacatgacaga gagccct; TOP3A-rev 5′-tgtgtcatgcagagggtcat); #5 *IGF1R* gene, chr15: 99324967-99325066, (IGFR-dir 5′-atgctccttatgtaccatgtgc; IGFR-rev 5′-tgctgtccaaattatccatcaag); #6 *CDH13* gene-2, chr16: 83191684-83191805, (CDH2-dir 5′-aagtgggttccctgtctcag; CDH2-rev 5′-gagggattgcctctgacctt).

Protection of 5hmC and unmethylated CG sites

5hmC was glycosylated using T4 beta-glucosyltransferase (BGT) (TS) at 37°C for 2 h according to vendor's recommendations, and then BGT was heat inactivated at 65°C for 20 min and treated with ProtK. Second round of BGT-glycosylation was done for 16 h at 37°C without ProtK-treatment.

Unmodified CG sites were methylated using 10-fold excess of WT SssI (TS) enzyme (1/10 volume of enzyme) over CG targets with 0.6 mM SAM in the vendor's buffer at 37°C for 4 h. After enzyme inactivation, samples were ProtK-treated and purified as in the section on Preparation of model DNA fragments. Identical second round of methylation and purification was performed, except that methylation was done at 37°C for 16 h.

Sanger bisulfite sequencing of the c-*fos* promoter DNA region

428 bp model DNA fragment was prepared from a promoter region of a human *c-fos* gene (inserted to pUC19 plasmid) using c-*fos*-gu-dir 5′-TTACACAGGATGTCCATATTAGG and c-*fos*-gu-rev 5′-CTGTGGAGCAGAGCTGGGTA primers and gel purified. 5mC was introduced into CG sites using methylation with wt SssI (SssI:CG ratio 0.3:1, 4h-incubation) and oxidized to 5caC, when required (see Preparation of model DNA fragments section). Then, decarboxylation and labeling was performed as in Assessment of M.SssI-directed decarboxylation and azide-labeling in a model DNA fragment. To separate 5caC from unmodified cytosine produced after decarboxylation in bisulfite (BS) analysis, DNA fragment was methylated after decarboxylation reaction and prior to BS.

BS conversion was performed with EZ DNA Methylation-Gold™ Kit (ZR) using 50 ng of fragment and alternative protocol 2. Upper and bottom strands were amplified separately using c-fgu-V-dir 5′-TTATATAGGATGTTTATATTAGGATATTTG, c-fgu-V-rev 5′-CAAAACTAAATAAAAACACRATCACTACT and c-fgu-A-dir 5′-TTACACAAAATATCCATATTAAAACATCT, c-fgu-A-rev 5′-GAGTT GGGTAGGAGTAYGGTTATTGTT primers, respectively. Phusion U HS and Taq polymerase (TS) were used for top and bottom strand, respectively. After gel purification, the fragment was cloned using CloneJET PCR Cloning Kit (TS). Plasmids from at least 10 clones were extracted using GeneJET Plasmid Miniprep Kit (TS) and analyzed by Sanger sequencing. 36 out of 37 CG sites were analyzed (excluding one overlapping with a primer) in each modification step. Methylation efficiency was calculated as % of C in each CG. 5caC level was calculated as % of T minus the percentage of incomplete methylation at each CG in the first methylation step. Decarboxylation efficiency was calculated as a difference in % of T between the initial 5caC and 5caC after decarboxylation.

Evaluation of wt SssI-methylation efficiency and 5caC level by BS of lambda phage DNA

Sheared DNA was end-repaired and methylated Ion Torrent adaptors were ligated. Then, the methylation protocol was applied (see Protection of 5hmC and unmethylated CG sites) and BS conversion performed using EpiJET Bisulfite Conversion Kit (TS). DNA was then amplified for 15 cycles using Phusion U HS polymerase (TS) and platform-specific primers. Concentrations of the libraries were evaluated using Agilent Bioanalyzer.

For evaluation of 5caC in lambda DNA, EZ DNA Methylation-Gold™ Kit (ZR) was used in the same procedure as above. The potential residual amount of 5fC left after Tet1-oxidation was reduced to 5hmC with $NaBH_4$ (see below). For assessment of inappropriate bisulfite conversion, 200 bp fragment (see Preparation of model DNA fragments) containing all methylated cytosines was prepared using PCR with 5-methylated dCTP and Pfu polymerase (TS) and mixed with lambda DNA samples before bisulfite conversion. Only sites outside primer binding regions were used when calculating the conversion.

Preparation of lambda DNA for caCLEAR analysis

DNA was sonicated to yield ~200-250 bp fragments with Covaris M220 instrument. In the first experiment (Figure 2B), 5mC was introduced into CG sites using 2.5-fold excess of wt SssI (TS) over its targets, reaction performed for 4 h. GCGC sites were methylated with a 10-fold excess of M.HhaI in its buffer (10 mM Tris-HCl (pH 7.4), 50 mM NaCl, 0.5 mM EDTA) with 300 μM AdoMet at 37°C for 2 h. Enzymes were heat-inactivated for 20 min at 65°C, treated with ProtK and purified (see Preparation of model DNA fragments section). Methylated DNA was mixed with unmodified lambda DNA to yield ~30% and ~60% of modified CG and GCGC sites, respectively. Tet1-oxidation of 5mC to 5caC was performed as described for model DNA fragments. For control samples, lambda DNA containing 30% uCG sites was prepared in parallel. For the experiment shown in Figure S3B, 5mC was introduced into GCGC sites through methylation with M.HhaI for 30 min using an enzyme to DNA ratio 0.125:1. 5mC was then oxidized with 10-fold excess of NgTet1 (Active Motif) in 50 mM Bis-tris (pH 6), 50 mM NaCl, 2 mM vitamin C, 1 mM alpha-ketoglutarate, 0.1 mM $FeSO_4$ (Roth) for 1.5 h at 34°C. Then, the reaction mixture was supplemented with 50 mM EDTA, treated with ProtK for 30 min at 50°C and purified.

The level of 5caC was evaluated using mass spectrometry. 5hmC was introduced into CCGG sites using 10-fold excess of M.HpaII (Lukinavičius et al., 2012) in M.HhaI buffer and incubated at 37°C for 3.5 h, and 20 min at 65°C. Then, DNA was ProtK-treated, and purified (see Preparation of model DNA fragments section). Hydroxymethylated and carboxylated DNA was methylated with wt SssI and mixed with 30% of unmethylated 5caC-containing DNA to yield a final sample containing 70% methylated CGs, 10% G5caCGC and 10% of C5hmCGG sites. Control lambda DNA without 5caC was also prepared. For the experiment shown in Figure 2E, 5mC was introduced using M.HhaI (ratio 0.125:1; M.HhaI:GCGC) and then, 5mC oxidation to 5caC was performed using NgTet1. To produce a DNA sample with 20% of 5caC at GCGC sites that also contains 70% methylated CG sites, carboxylated DNA was mixed with SssI-premethylated lambda DNA.

High-performance liquid chromatography – tandem mass spectrometry (HPLC-MS/MS)

mESC genomic DNA was purified by resuspending the cells in extraction buffer (200 mM Tris HCl pH 8.0, 250 mM NaCl, 25 mM EDTA, and 0.5%–2% SDS depending on lysis) and incubating with ProtK (TS) at 55°C for 4–6 h. DNA was purified by phenol-chloroform extraction.

After sonication to ~1000 bp fragments, DNA was purified using DNA clean and concentrator kit (ZR) and eluted with water. Then it was denatured for 10 min at 80°C and digested with Nuclease P1 (Sigma) using ~0.33 U for 1 μg DNA at 50°C for 4 h in P1 buffer (10 mM NaOAc pH 5.2, 1 mM ZnOAc), then dephosphorylated with FastAP phosphatase (TS) using ~1 U for 5 μg DNA at 37°C overnight. Reactions were stopped by heating at 75°C for 10 min and protein precipitate was spun down at 14000xg at 4°C for 30 min. Samples were analyzed on an integrated HPLC/ESI-MS/MS system (Agilent 1290 Infinity/ 6410B triple quadruple) equipped with a Supelco Discovery® HS C18 column (7.5 cm × 2.1 mm, 3 μm) using a linear gradient of solvents at a flow of 0.3 mL/min at 30°C.

For 5caC and 5fC ~22 μg of digested DNA was first fractionated in 4 runs with a linear gradient of solvents A (10 mM ammonium formate pH 5.5) and B (80% methanol) at a flow of 0.3 ml/min at 30°C as follows: 0 min, 0% B; 9–23 min, 0%–10% B; 23–27 min, 10%–100% B; 27–31 min, 100% B; 31,5–37 min, 0% B. Fractions corresponding 5caC and 5fC was collected at ~1,1-3 min and ~13,6-17,4 min respectively, vacuum-dried and reconstituted with water. For standard curves fractions of known amounts of 5caC and 5fC were collected from a mix with a 5 μg dam- dcm- lambda phage DNA.

For mass spectrometry analysis gradient of solvents A (0.0075% formic acid in water) and B (0.0075% formic acid in acetonitrile) was used. For 5caC gradient was: 0-6 min, 0% B; 6-10 min, 0%–100% B; 10-14 min, 100% B; 14,5-19,5 min, 0% B. Mass spectrometer was operating in the positive ion MRM mode and intensity of nucleoside-specific ion transition was recorded of m/z 272.1 → 156. Ionization capillary voltage 1800 V, drying gas temperature 150°C and flow rate 10 l/min, collision energy 15V. For 5fC gradient was: 0-24 min, 0% B; 24-25 min, 0%–100% B; 25-29 min, 100% B; 29,5-35 min, 0% B. Ion transition of 256.1 → 140 m/z was recorded, drying gas temperature of 300°C was used. For 5hmC analysis in ESCs 0.5-2 μg DNA was used, 50 ng for 5mC, for calibration known amounts of standards were mixed, 5hmC standards were mixed with corresponding amounts of lambda DNA. Gradient was: 0–6 min, 0% B; 6–18 min, 10% B; 20–24 min, 100% B; 25–33 min, 0% B, gas temperature was 300°C. Ion transitions were recorded: 5mC m/z 242.1 → 126.1; 5hmC m/z 258.1 → 142.1; G m/z 268.1 → 152.1, C m/z 228.1 → 112.1. Fragmentation was set at 80 V for 5fC, dhU and 100 V for other nucleosides. For assessment of 5fC reduction to 5hmC 2.5 μg of unfractionated serum-2i *Tdg*−/− DNA was used, gradient was: 0-24 min, 0% B; 24-27 min, 0%–100% B; 27-32 min, 100% B; 33-38 min, 0% B. For analysis of glycosylation of genomic 5hmC residues, 0.87 μg of HhaI-hydroxymethylated lambda DNA before and after glycosylation was used. For evaluation of TAPS the same gradient was used, transition of m/z 231.1 → 115.1 for dhU was recorded, drying gas temperature of 150°C was used.

Preparation of caCLEAR libraries

Genomic DNA was sheared with Covaris M220 to on average 200-250 bp fragments. ESCs DNA was then purified with DNA clean and concentrator kit (ZR).

Step 1. 5hmC sites were blocked by glycosylation with BGT (see section Protection of 5hmC and unmethylated CG sites).
Step 2. Genomic uCG sites were blocked by methylation using WT SssI (see section Protection of 5hmC and unmethylated CG sites). After the second methylation reaction, DNA was purified with Oligo Clean and Concentrator (ZR) kit.
Step 3. eM.SssI-decarboxylation was performed using 10-fold excess of eM.SssI (Kriukienė et al., 2013) over CG targets (see Assessment of M.SssI-directed decarboxylation and azide-labeling in a model DNA fragment).
Step 4. Decarboxylation reaction was supplemented with Ado-6-N_3 cofactor to 0.2 mM and incubated at 30°C for 1 h, then the enzyme was inactivated and DNA purified (see Assessment of M.SssI-directed decarboxylation and azide-labeling in a model DNA fragment).
Step 5. Tethered oligonucleotide-primed sequencing, uTOP-seq.
Step 5a. DNA ends were prepared and adaptors ligated as in Staševskij et al. (2017).
Step 5b. Purified DNA was supplemented with 20 μM biotinylated alkyne DNA oligonucleotide 5′-TXTTTTGTGTGGTTTGGAGACTGACTACCAGATGTAACA-biotin (X = C8-alkyne-dU, Base-click), $^1/_5$ volume of of CuBr (8 mM): TBTA (24 mM) mixture (Sigma) in DMSO (Roth) and DMSO was added to a final concentration of 50%. Then, the reaction mixture was incubated at 45°C for 20 min and subsequently diluted to ~1% DMSO before purification through GeneJET NGS Cleanup Kit (TS) columns. DNA with attached oligonucleotide was then enriched: 0.1 mg of Dynabeads MyOne Streptavidin C1 (TS) was incubated with DNA in 10 mM Tris-HCl (pH 8.5), 1 M NaCl and 0.2% of Tween 20 (Roth) at room temperature for 3 h on a roller. Beads were

washed as in (Song et al., 2017), using 150 μl buffer. To elute DNA, beads were incubated in water for 5 min at 95°C, 5 min at 4°C. For the experiments presented in Figures 2B and 2D and Figure S3, oligonucleotide without biotin was used and no enrichment was performed. ~50 ng of lambda and 827 ng of mESC DNA was used for enrichment.

Step 5c. All eluted DNA was supplemented with 0.5 μM complementary priming strand (5′-TGTTACATCTGGTAGTCAGTCTC CAAACCACACAA, with custom LNA modifications (Exiqon)) and 0.05 U/μl Pfu polymerase (TS) in Pfu $MgSO_4$ buffer with additional $MgSO_4$ to 1 mM, 0.2 mM dNTP (total volume 20 μl) and was incubated for 5 cycles at: 95°C 1 min, 65°C 10 min, 72°C 10 min.

Step 5d. All primed DNA in a 50 μl reaction containing Platinum SuperFi PCR Master Mix (TS) amplified as in Staševskij et al. (2017). The libraries were size-selected for 300 bp fragments (MagJet NGS Cleanup and Size-selection kit, TS), and were subjected to Ion Proton (TS) sequencing.

Locus-specific 5caC analysis by caMAB-seq

Genomic DNA was precleaned with Genomic DNA Clean and Concentrator kit (ZR) and methylated as in the Protection of 5hmC and unmethylated CG sites section, purified with the same kit and eluted in water. 5fC reduction performed according to Booth et al. (2014). 10 ng/μl DNA in water was reduced with 250 mM $NaBH_4$ (Sigma-Aldrich) and incubated for 1 h at room temperature, frequently vortexing and spinning down. Reaction quenched with 0.5 volume of 750 mM NaAc (pH 5). When no gas was released, buffer exchange to water using Microcon® Centrifugal Filters (MERCK) was performed. Two samples were prepared: serum-2i Tet TKO DNA as a negative control, serum-2i *Tdg*−/− showing only 5caC. 500 ng of DNA in each reaction was BS converted with EZ DNA Methylation-Gold™ Kit (ZR) using standard protocol and DNA was purified and eluted with 10 μl of M-Elution Buffer. After conversion, selected regions were amplified for 25 cycles with Phusion U HS polymerase (TS) using primers specific to each strand (Table S2) and DNA was purified with DNA clean and concentrator kit (ZR). Fragments were PAGE-purified using SYBR Gold staining (Invitrogen), crushed gel was incubated in an elution buffer (0.5 M CH_3COONH_4, 0.1 mM EDTA, 0.1% SDS) for 2 h at 37°C with shaking at 600 rpm. Samples were filtered through Ultrafree-MC Centrifugal Filter (Millipore) and DNA was purified using Ampure XP (Beckman Coulter) magnetic beads. DNA was end-repaired and Ion Torrent barcoded adapters (Kapa Biosystems) were ligated. After 4 cycles of amplification with Ion Torrent specific primers and Platinum SuperFi PCR Master Mix (TS), DNA was purified using GeneJET NGS Cleanup Kit (TS).

Locus-specific 5caC analysis by pyridine borane sequencing

For revealing 5caC, a modification of TAPS sequencing (Liu et al., 2019) that omits the Tet-treatment step was performed on serum-2i *Tdg*−/− DNA and serum-2i Tet TKO DNA as a negative control. mESC genomic DNA was precleaned with Genomic DNA Clean and Concentrator kit (ZR) and then 5fC reduction was performed as in caMAB-seq. DNA was then modified according to Liu et al. (2019): DNA (150 ng for 50 μl volume) was incubated with 1 M pyridine borane (Sigma-Aldrich) and 0.6 M sodium acetate (pH 4.3) at 37°C for 16 h with shaking at 850 rpm. After purification using Genomic DNA Clean and Concentrator kit (ZR), selected regions were amplified for 16 cycles with DreamTaq DNA Polymerase (TS) using primers specific for each region (Table S2). PCR products were PAGE-purified as for caMAB-seq. After the elution step, the solution was supplemented with 0.05 μg/μl glycogen (TS) and DNA precipitated using ethanol. DNA fragments were prepared for Ion Torrent sequencing as described for caMAB-seq.

QUANTIFICATION AND STATISTICAL ANALYSIS

Analysis of bacteriophage lambda caCLEAR and WGBS data

Processing of the lambda caCLEAR data was performed as described previously in (Gibas et al., 2020) except for the minimal length of the used reads (80 nt). Processed reads were mapped to a lambda genome and only reads starting exactly at CG sites were used for further analysis. To compare coverage between the GCGC and non-GCGC targets, coverage per CG was normalized by the total amount of reads in a specific sample. Lambda DNA WGBS data were analyzed using Bismark aligner and methylation caller with default settings, except for the non-directionality parameter (Krueger and Andrews, 2011).

Analysis of ESCs caCLEAR data

ESC caCLEAR data were processed as described for lambda caCLEAR with the following exceptions: reads were mapped to the mouse genome build mm10 with a minimal mapping quality of 30 and only reads starting within 4 bp around the CG sites were used. High-confidence 5caCG sites were selected in the following order: first, all CG sites identified in Tet TKO libraries (~0.39 M) were removed from the target libraries. Then, for Tdg samples, we used only those CGs that in both technical replicates had coverage equal or higher than the average coverage (5x and 6x coverage for serum-2i and serum samples, respectively); for wild-type samples, we selected those CGs that in both technical replicates had coverage equal or higher than the average coverage (4x and 2x for serum-2i and serum samples, respectively) and were also identified in a corresponding Tdg samples. Additionally, we removed 10% of identified CG sites (~1,100 in serum wt and ~12,400 in serum-2i wt) which had the largest coverage difference between the technical replicates.

caCLEAR enrichment in genomic elements and features was calculated by creating a contingency table for each CG site falling into a high-confidence caCLEAR CG set and overlapping a genomic element. Next, Fisher's exact test was performed to estimate the odds ratio (OR) and p value. Enrichment of 5caCG-modified TFRs was performed in a similar manner - TFR was tested for containing

5caCGs and overlapping a genomic element. To visualize caCLEAR signal around TFRs we smoothed CG coverage with a window size of 50 and normalized calculated estimates by a CG density value in a specific window. 5caCG-modified TFR distance analysis was done by selecting only intergenic TFRs and calculating their distance to a closest transcription start site of protein-coding genes. Distances between 5caCG-modified and 5caCG-depleted TFRs were compared using Mann-Whitney test.

To visualize DNA modification signals and Tet1 occupancy at and around CGIs we divided each CGI and it's -/+4kb flanking regions into 10 equally sized windows. Within each window average modification value (caCLEAR, uCG, 5hmCG) or Tet1 signal was calculated. Then all CGI regions were ranked by the difference between caCLEAR signal in the CGI and its flanking regions. Profiles along the CGI regions represent average modification signals normalized by the CG density (for caCLEAR, uCG and 5hmCG) or Tet1 occupation along the loci.

To identify pluripotency state-specific genes in Tdg ESCs we used Anova F-test for all genes with at least 10 5caCGs. Genes passing the test with FDR q-value < 0.05 and absolute fold-change higher than 4 were assigned to a specific condition. Condition specific genes were tested for gene ontology term enrichment using GOrilla online tool with default parameters (Eden et al., 2009). To identify genes with a strand-bias we selected all genes with at least 10 identified CGs and calculated average caCLEAR signal per strand. Next, we used Anova F-test to compare caCLEAR signal differences between the strands. All genes with FDR q-value < 0.05 and absolute fold-change higher than 2 were classified as having a strand-bias. To evaluate gene expression in genes with 5caCG-strand bias or genes with 5caCG-modified TFRs, all expressed genes were divided into three equally sized groups - low, medium (mid), high expression. Fisher's exact test was used to test whether a gene has a strand bias or 5caCG-modified TFRs and is within a specific expression group. ESC caMAB-seq and TAPS sequencing data were analyzed in the same manner as described above for lambda WGBS data.

Annotations

Genome sequence, CpG island (CGI), repeat and ENCODE ATAC-seq peak open chromatin regions were downloaded from the UCSC genome browser. 2 kb regions around a CGI were selected as CGI shores. Mouse reference gene dataset was downloaded from the GENCODE genes (Frankish et al., 2019). Histone ChIP-seq data and serum-2i RNA data were downloaded from the ENCODE data portal (ENCODE Project Consortium, 2012; Davis et al., 2018). Serum RNA data were downloaded from the Gene Expression Omnibus (GEO) - GSE23943. ChIP-seq of TFs, CAB-seq, 5caC-DIP data were obtained from the GEO (accession numbers for TFR - GSE11431 (Chen et al., 2008), GSE92412 (Atlasi et al., 2019), GSE56312 (Galonska et al., 2015), CAB-seq - GSE56429 (Lu et al., 2015), 5caC-DIP - GSE42250 (Shen et al., 2013)). Tet1 ChIP-seq data were downloaded from GEO (GSE24843) and peak-calling was performed with Macs2 (Zhang et al., 2008; Williams et al., 2011). uCG and 5hmCG data were taken from (Gibas et al., 2020).