Biochemistry

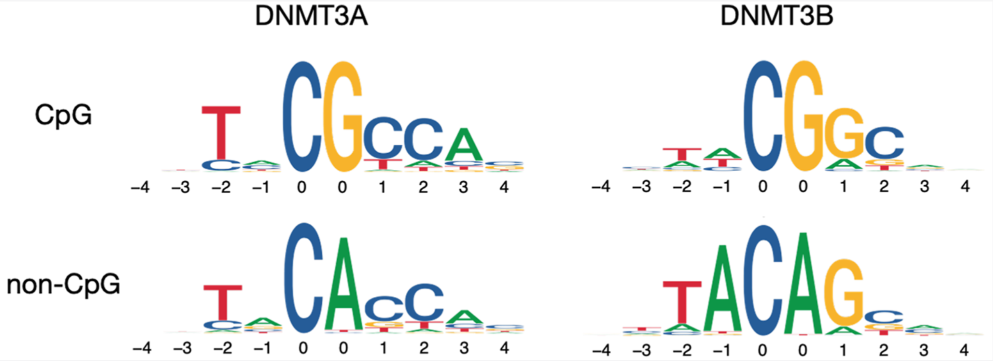
Genome-wide DNA Methylation Signatures Are Determined by DNMT3A/B Sequence Preferences

Shi-Qing Mao,[||] Sergio Martínez Cuesta,[||] David Tannahill, and Shankar Balasubramanian*

Cite This: *Biochemistry* 2020, 59, 2541−2550

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information



ABSTRACT: Cytosine methylation is an important epigenetic mark, but how the distinctive patterns of DNA methylation arise remains elusive. For the first time, we systematically investigated how these patterns can be imparted by the inherent enzymatic preferences of mammalian *de novo* DNA methyltransferases *in vitro* and the extent to which this applies in cells. In a biochemical experiment, we subjected a wide variety of DNA sequences to methylation by DNMT3A or DNMT3B and then applied deep bisulfite sequencing to quantitatively determine the sequence preferences for methylation. The data show that DNMT3A prefers CpG and non-CpG sites followed by a 3′-pyrimidine, whereas DNMT3B favors a 3′-purine. Overall, we show that DNMT3A has a sequence preference for a TNC[G/A]CC context, while DNMT3B prefers TAC[G/A]GC. We extended our finding using publicly available data from mouse Dnmt1/3a/3b triple-knockout cells in which reintroduction of either DNMT3A or DNMT3B expression results in the acquisition of the same enzyme specific signature sequences observed *in vitro*. Furthermore, loss of DNMT3A or DNMT3B in human embryonic stem cells leads to a loss of methylation at the corresponding enzyme specific signatures. Therefore, the global DNA methylation landscape of the mammalian genome can be fundamentally determined by the inherent sequence preference of *de novo* methyltransferases.

In mammals, most DNA methylation occurs at C-5 of cytosine bases. Cytosine methylation is a well-established epigenetic mark and is involved in the regulation of key biological processes, including tissue specific gene expression patterns, X-chromosome inactivation, transposon silencing, and genomic imprinting.[1−3] There are three main DNA methyltransferase enzymes, DNMT3A, DNMT3B, and DNMT1. DNMT3A and DNMT3B are *de novo* methylases that operate on both unmethylated and hemimethylated DNA.[4,5] In contrast, DNMT1 is a maintenance methylase that preserves methylation patterns during replication due to an inherent requirement for hemimethylated DNA.[6−8]

Cytosine methylation is often described in terms of CpG and non-CpG contexts (i.e., CH, where H = A, T, or C), with the latter extended to include CHG and CHH categories based on symmetry.[9] Overall, this leads to palindromic (CpG), partially palindromic (CHG), and nonpalindromic (CHH) methylation sites. Most CpG sites gain methylation on both DNA strands in early mammalian embryonic development[1,10] and then remain highly methylated throughout development. Cytosines in CpG islands (genomic regions with a high frequency of CpG sites) associated with promoters are dynamically regulated and closely linked with gene expression.[11−14] Different tissues have distinct profiles of non-CpG methylation, and the highest levels are found in pluripotent stem cells and in the central nervous system.[11,15−18] In human embryonic stem cells (hESCs), ∼25% of total cytosine methylation occurs in a non-CpG context with 71% at CHG and 29% at CHH sites, while in human neurons, 53% of total methylated cytosines are at non-CpGs, of which >80% are at

Received: April 24, 2020
Revised: June 16, 2020
Published: June 16, 2020

https://dx.doi.org/10.1021/acs.biochem.0c00339

CHH sites.[18] As the maintenance methylation enzyme DNMT1 has no reported activity at non-CpG sites,[15] this raises basic questions about how distinctive DNA methylation landscapes are established and maintained.

Several factors shape the DNA methylome, including nucleosome positioning[19,20] and histone modifications.[21,22] The methylation-deficient DNMT family member DNMT3L has been reported to stimulate DNMT3A/B activity by enhancing the stability of enzyme complex recruitment to DNA or by an increased level of cofactor *S*-adenosyl-L-methionine binding.[23−25] Genome engineering experiments of inserted artificial sequences in mouse stem cells have begun to uncover the contribution to methylation of the underlying genomic sequence, namely, CpG density and GC content.[26,27] Furthermore, transcription factor (TF) binding[27−29] and G-quadruplex DNA secondary structures[30] are implicated in protecting TF-bound regions and certain CpG islands from methylation, respectively. While such factors are critical for regulating DNA methyltransferases and influencing the distribution of DNA methylation, a key unanswered question that remains is how differential methylation patterns are imparted to different genomic sequences in the first place.

The preferential methylation of unmethylated CpG by DNMT3A/B and of hemimethylated CpG sites by DNMT1 has been studied biochemically and structurally.[7,31] Early work attempted to determine the flanking sequence preferences of DNMT3A/B at CpGs using four synthetic oligos with no assessment of non-CpG methylation.[32] Notably, no studies have considered a large and unbiased pool of competing substrates as a fair test of methylation preferences. It has also not been established whether CpG methylation is installed in a sequence specific manner by DNMT3A/B in the mammalian genome under physiological conditions. Recent DNA methylation maps show non-CpG methylation in nearly all human tissues,[16,33] but the question of whether DNMT3A or DNMT3B establishes these non-CpG methylation signatures is still elusive.

Herein, we describe a novel assay and systematic analyses that quantitatively interrogate DNMT3A and -3B enzyme specificity on a large and diverse set of cytosine contexts using unmethylated *Escherichia coli* genomic DNA as the substrate coupled with high-depth bisulfite sequencing analysis. We find that each enzyme shows distinct target sequence signatures that are unchanged upon boosted methylation activity by the inactive cofactor DNMT3L. We find that these signatures are naturally observed within the mouse and human DNA methylomes, demonstrating that the intrinsic substrate preferences of DNMT3A/B are critical for determining the distribution of DNA methylation in mammalian genomes.

■ MATERIALS AND METHODS

***In Vitro* Methylation Assay.** Full-length recombinant human DNMT3A (Abcam, ab170408), DNMT3B (Abcam, ab170410), and DNMT3L (active motif, catalog no. 31414) were purchased from commercial providers; 100 ng of unmethylated *E. coli* genomic DNA (D5016, Zymo Research) was incubated at 37 °C with 500 ng of DNMT3A, DNMT3B, or DNMT3L and 160 μM *S*-adenosylmethionine (SAM, catalog no. B9003S, NEB) in reaction buffer (50 mM Tris-HCl, 1 mM EDTA, 1 mM dithiothreitol, 5% glycerol, and 100 μg/mL bovine serum albumin) for 30, 120, and 240 min. For DNMT3L stimulation experiments, 200 ng of DNMT3A or DNMT3B and 200 ng of DNMT3L were incubated with 100 ng of *E. coli* DNA for 120 min. For comparison, 1 unit of bacterial CpG methyltransferase *M.Sss*I (New England Biolabs), which has high methylation activity *in vitro*, was also incubated with DNA for 10, 30, and 240 min. After incubation, the reaction was terminated by the mixture being heated at 65 °C for 20 min. DNA was then purified using a DNA Clean & Concentrator Kit (D4030, Zymo Research) and processed for high-throughput bisulfite sequencing.

Bisulfite Sequencing. Bisulfite libraries were prepared using a Pico Methyl-Seq Library Prep Kit (D5456, Zymo Research) by following the manufacturer's protocol. Briefly, DNA was treated with bisulfite conversion reagent at 98 °C for 8 min and then at 54 °C for 60 min. Converted DNA was purified and amplified using random priming. Amplified DNA was purified, adapted, and indexed. Libraries were pooled and sequenced on an Illumina NextSeq-500 platform using High-Output Kit ver. 2.5 (75 cycles) in single-end mode. The nonconversion rate was estimated to be 0.5% using *E. coli* DNA incubated with the inactive DNMT3L.

Calculating Sequence Context Occurrences Genome-wide. The observed numbers of unique sequence contexts (flanking cytosine, CG, or CA dinucleotides) present in the forward and reverse strands of the λ, *E. coli*, and human reference genomes were obtained using bedtools ver. 2.27.0[34] and custom Python scripts. The observed number of occurrences for a given n k-mer was compared to the total number (t) of all possible sequence contexts, e.g., NCGN ($n = 2$; $t = 16$), NNCGNN ($n = 4$; $t = 256$), and NNNCGNNN ($n = 6$; $t = 4096$), and is represented in Figure S1.

Processing and Analysis of *E. coli* Bisulfite Sequencing Data. The quality of raw sequencing reads was evaluated using FastQC ver. 0.11.3 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Low-quality base calls were filtered, and Illumina TruSeq adapters were trimmed from the read's 3′ end using cutadapt ver. 1.123.[35] No reads smaller than 10 bp were kept (after adapter and base quality trimming). Following the read quality assessment, the first six bases of every read were also trimmed.

Bisulfite-converted reads were aligned to the *E. coli* K-12 MG1655 ASM584v2 reference genome (Ensembl Genomes release 41) using bismark ver. 0.19.0[36] with options *non_directional*−*unmapped*, and duplicated alignments were removed using deduplicate_bismark. Methylation calls were obtained using bismark_methylation_extractor with the option −*CX_context*. The sequence context for each cytosine in the *E. coli* genome was obtained using bedtools slop and bedtools getfasta. Only cytosines with at least 10 aligned sequencing reads were considered for further analysis.

To visualize methylation levels, boxplots and sequence logos were generated in different sequence contexts using the libraries data.table v1.10.4, ggplot2 v2.2.1[37] and ggseqlogo v0.1[38] in the R programming language (https://www.r-project.org/).

Processing and Analysis of Mouse and Human Bisulfite Sequencing Data Sets. Public whole genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS) data sets used in this study are listed in Table S1. Raw WGBS data sets from GEO were processed like the *E. coli* libraries, whereas RRBS data sets were quality trimmed and further clipped by three bases from the 5′ end using Trim Galore ver. 0.6.4_dev (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). The GENCODE reference genomes[39] used were human release

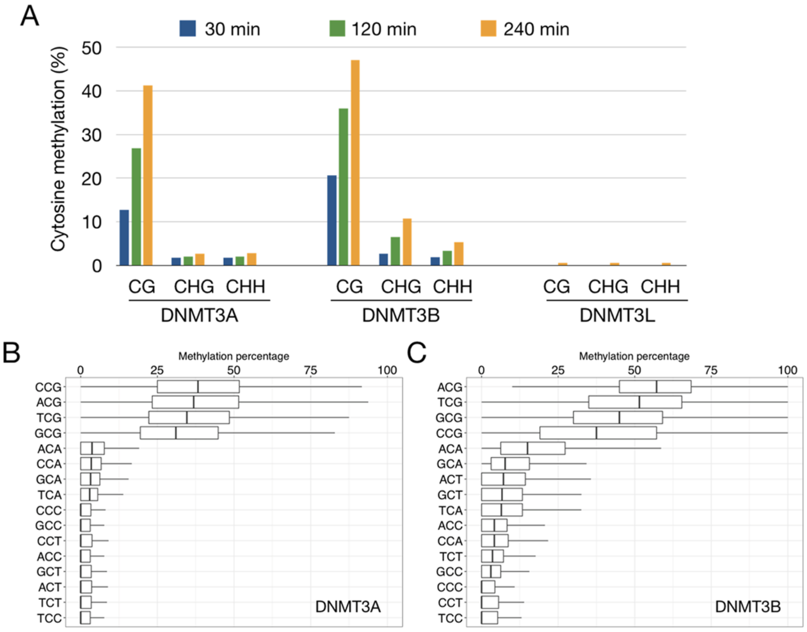


Figure 1. *In vitro* methylation assay using recombinant full-length human DNMT3A and -3B. (A) Average level of methylation introduced by DNMT3A, DNMT3B, or DNMT3L in CpG and non-CpG sites. Box plots of methylation levels at NCN sequences after incubation with (B) DNMT3A or (C) DNMT3B for 240 min, ranked by median methylation level. Sequences are written in 5′ to 3′ order.

28 (GRCh38.p12) and mouse release M18 (GRCm38.p6). For WGBS data sets, cleaned-up fastq files from human and mouse were aligned against GRCh38.p12 and GRCm38.p6, respectively, and subsequent reads were processed on a chromosome-by-chromosome basis. Unless otherwise stated, methylation on chromosome 1 of both mouse and human data sets was reported. For RRBS data sets, non-deduplicated aligned reads were processed for all chromosomes simultaneously, and methylation counting and visualization were performed as in the *E. coli* libraries.[40]

For details about the bioinformatics data analysis, see https://github.com/sblab-bioinformatics/dnmt3a-dnmt3b.

■ RESULTS

DNMT3A and -3B Enzyme Sequence Preferences Revealed by a High-Throughput Biochemical Methylation Assay. For a comprehensive study of methyltransferase enzyme sequence preferences, we aimed to biochemically capture a wide range of substrates that display sufficient sequence diversity and coverage to provide a fair and systematic collection of possible sequence targets. The 4.6 million bp *E. coli* genome is 51% G/C rich and contains 346670 CpG sites, which represents 96.6% of all possible NNNNCGNNNN (N = A, T, C, or G) sequences (63295 of 4^8 = 65536 total combinations), 96.6% of all NNNNC-ANNNN, and 99% of all NNNNCNNNN (Figure S1). Thus, unlike previous studies using a limited range of CpG substrates (275 CpG sites altogether),[32] the *E. coli* genome has sufficient sequence context diversity to serve as an essentially unbiased substrate to investigate the sequence preferences of different methylases.

We then developed a biochemical assay to evaluate the methylation activity of recombinant full-length human DNMT3A or DNMT3B using unmethylated *E. coli* genomic DNA as the substrate, followed by methylation assessment through whole genome bisulfite sequencing and subsequent computational analysis. We sought assay conditions (10−60% total methylation) that avoided saturated methylation that would mask any differential activity. Either DNMT3A or DNMT3B was incubated with *E. coli* DNA for different time ranges (30, 120, and 240 min) to provide a range of methylation levels for subsequent analysis. After bisulfite sequencing, average methylation at CpG and non-CpG contexts was calculated. The level of methylation at CpG sites increased with incubation time and ranged from 11% to 20% at 30 min and from 40% to 46% at 240 min for DNMT3A and DNMT3B, respectively (Figure 1A). After 240 min, the level of non-CpG methylation was 2.7% at CHGs and 2.8% at CHHs for DNMT3A and 10.7% at CHGs and 5.3% at CHHs for DNMT3B. These results show that while DNMT3A and DNMT3B show a broadly similar level (41% and 47%, respectively) of CpG methylation after incubation for 240 min, DNMT3B has a relatively greater methylation activity for non-CpG sites (∼1.9−4-fold) compared to DNMT3A. Excluding biases introduced by bisulfite conversion, we also showed that the nonconversion level was 0.5% for both CpG and non-CpG contexts after incubating inactive DNMT3L with *E. coli* genomic DNA for 240 min (Figure 1A). Furthermore, our results also show that DNMT3B but not DNMT3A has a >2-fold methylation activity for CHG over CHH sequences (Figure 1A), which implies an inherent sequence-dependent preference of DNMT3B.

To investigate the influence of sequence context on cytosine methylation by DNMT3A and DNMT3B, we ranked the median cytosine methylation levels for trinucleotide sequences with cytosine as the middle base [i.e., NCN (Figure 1B,C)]. Both DNMT3A and DNMT3B showed a strong preference for CpG dinucleotides, resulting in more than 30% and 37% methylation, respectively. The next most methylated sequence context was for CpA dinucleotides, which was less than 4% and 15% methylation on all non-CpG sites after incubation for 240 min as in DNMT3A and DNMT3B, respectively (Figure 1B,C). DNMT3B also showed more variable methylation than DNMT3B on CpG or CpA. The preference for CpG sites is

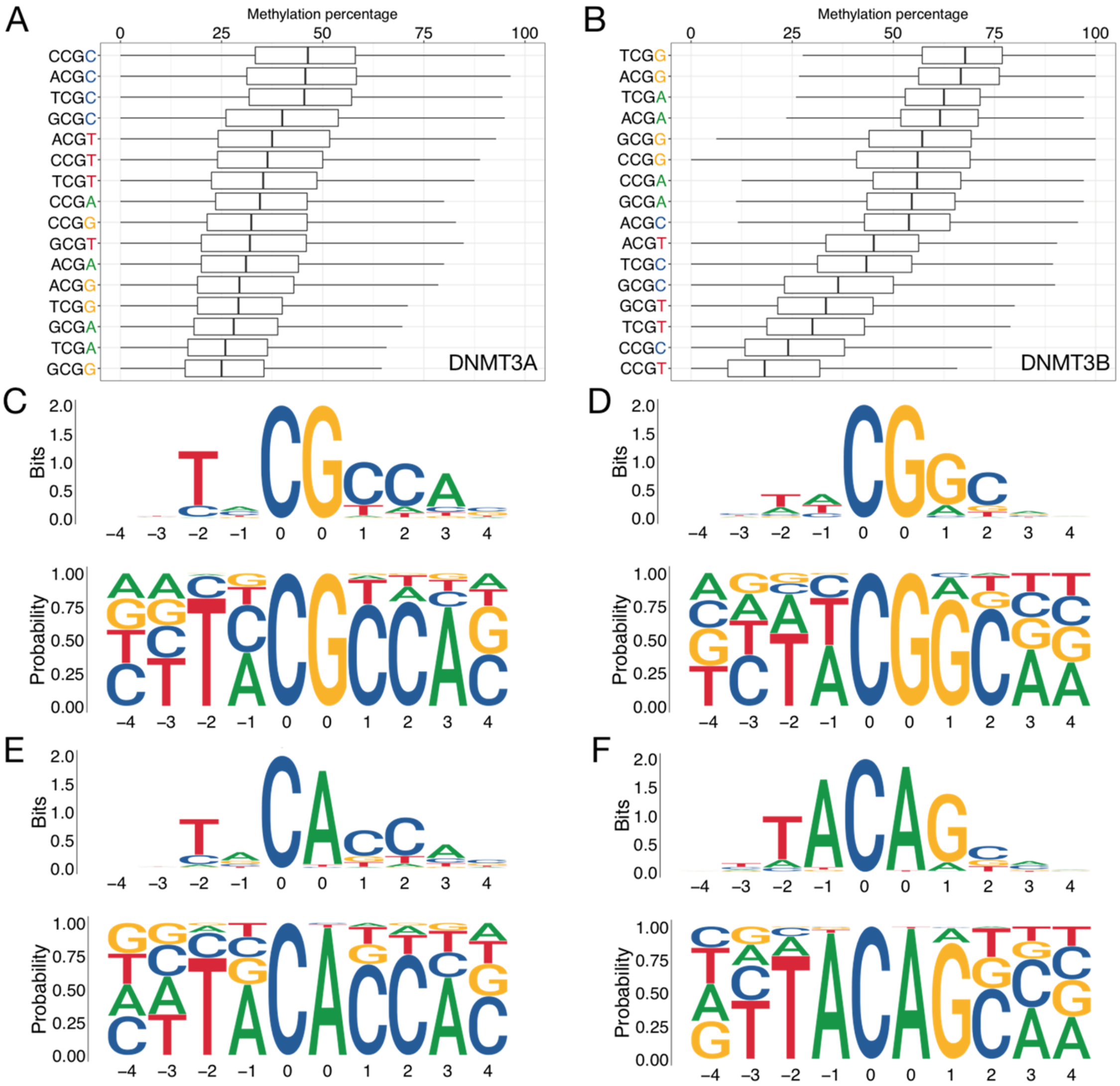


Figure 2. DNMT3A and -3B have different preferences for flanking sequences for CpG and non-CpG sites. Box plot of methylation levels in NCGN context after incubation with (A) DNMT3A or (B) DNMT3B for 240 min, ranked by median methylation level. Sequence logo of the 1000 most methylated 10-mer CG sequences after incubation with (C) DNMT3A or (D) DNMT3B for 30 min. Sequence logo of the 1000 most methylated 10-mer non-CpG sequences after incubation with (E) DNMT3A or (F) DNMT3B for 120 min.

independent of incubation time (Figure S2A,B). An important control using *M.Sss*I showed high methylation activity, and all trinucleotide sequence contexts are equally available for methylation without any preference (Figure S2C), which also rules out any biases caused by sample processing and data analysis. Altogether, the results reveal the importance of bases flanking the substrate cytosine and the already known preference for CpG over non-CpG sites.

Distinct DNMT3A and DNMT3B Sequence Preferences Are Directed by Flanking Sequences for both CpG and Non-CpG Contexts. To further investigate the sequence preferences of DNMT3A and -3B, we then explored the influence of both the 5′ and 3′ flanking bases for CpG sites by ranking the median methylation level at all known NCGN sequences. Notably, DNMT3A generally favors a pyrimidine (C or T) as the 3′ adjacent base with NCGC and NCGT sequences gaining the most (44%) and second most (38%) methylations, whereas conversely, DNMT3B prefers a 3′ purine base (G or A), with NCGG and NCGA sequences being most methylated (62% and 58%, respectively). We also observed that DNMT3B showed a preference for sequences with a T or A at the 5′ position in NCGG or NCGA contexts, respectively, whereas DNMT3A favors sites with C/A at the 5′ position (Figure 2A,B). DNMT3B also showed a greater spread in median methylation levels, from 17% to 67%, across different sequence contexts, while DNMT3A was more restricted ranging from 25% to 40% (Figure 2A,B). When longer flanking sequences (NNCGNN) were considered, a clear pattern of sequence preferences and differences between the two enzymes emerged (Figure S4). For example, DNMT3A prefers TACGCC sequences (N = 3206; median level of methylation of 66.7%) and disfavors AGCGGG sequences (N = 2585; 12.9%), whereas DNMT3B prefers GTCGGC sequences (N = 2641; 73.9%) and disfavors GCCGTG sequences (N = 2570; 8.3%) (Figure S4A,B). The differences in methylation range and sequence preference were independent of incubation time. There was no observed flanking sequence preference for the *M.Sss*I control methylase (Figure S3 and Figure 1C).

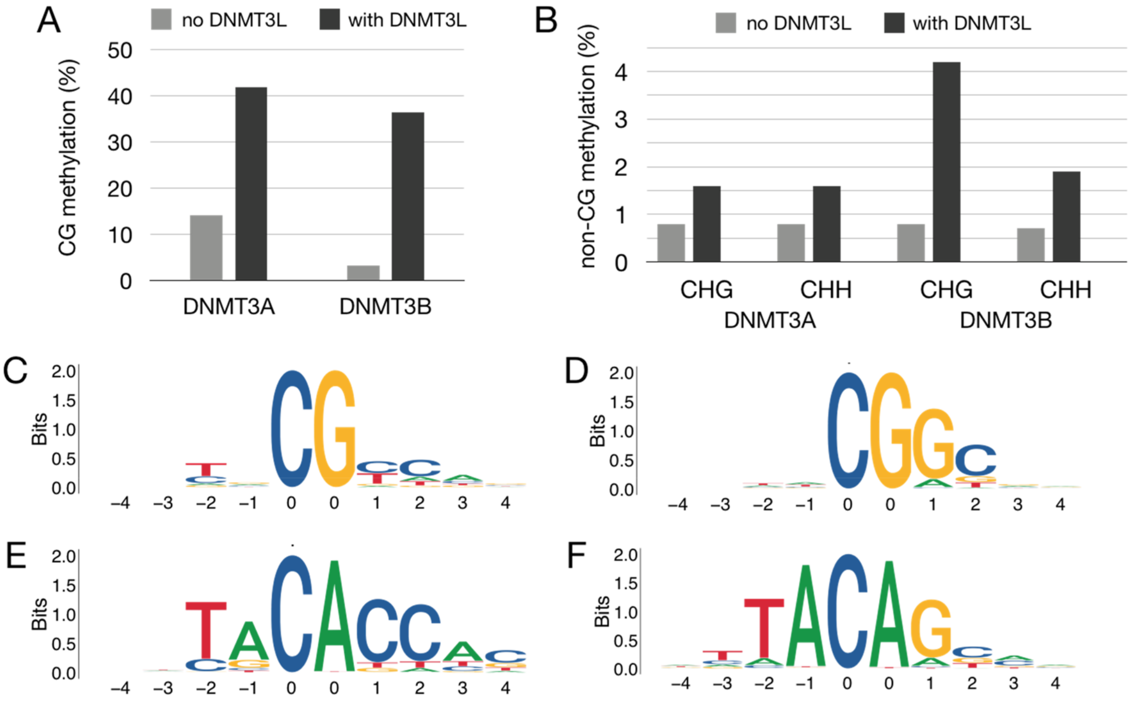


Figure 3. DNMT3L stimulates DNMT3A and DNMT3B activity without altering their sequence preference. Average methylation by DNMT3A and DNMT3B in (A) CpG and (B) non-CpG sites. Sequence logo of the 1000 most methylated 10-mer CG sequences after incubation with (C) DNMT3A/3L or (D) DNMT3B/3L for 120 min. Sequence logo of the 1000 most methylated 10-mer non-CpG sequences after incubation with (E) DNMT3A/3L or (F) DNMT3B/3L for 120 min.

To determine whether additional flanking bases have an influence on preference, we extended our analyses to include four bases 5′ and 3′ of the CpG (i.e., NNNNCGNNNN) by calculating the consensus sequence logo of the top 1000 most methylated sequences after incubation for 30 min (Figure 2C,D). Following from the strong enzymatic preference at the adjacent 3′ position for CpG substrates as highlighted before, DNMT3A also showed a strong preference for a T at the −2 position 5′ with NN**T**NCGNNNN representing 75% in all methylated sequences (Figure 2C). In contrast, DNMT3B had a preference for T or A in both the −1 and −2 positions 5′ with NN[**T**/**A**]NCGNNNN or NNN[**T**/**A**]CGNNNN sequences representing >75% (Figure 2D). Both DNMT3A and DNMT3B showed similar preferences for C at the +2 position 3′, and DNMT3A also showed a preference for A at the +3 position 3′ (Figure 2C,D). Longer incubation times ultimately led to full methylation at a wide range of sequence contexts (Figure S4), obscuring intrinsic sequence preferences (Figure S5A,B).

For non-CpG dinucleotides, DNMT3A and DNMT3B showed higher activity at CpA than at CpC or CpT sites, with NNNN**CA**NNNN representing 97% of all methylated sequences (Figure 2E,F). The sequence preference of DNMT3A/B at non-CpG sites is similar to that at CpG sites, which was also independent of the incubation time before reaching the saturation level (Figure S5C,D). Furthermore, DNMT3A and -3B each showed a similar preference for flanking sequences at the less methylated CT or CC dinucleotide sites compared to that of CA or CG sequences (Figure S6). Overall, these *in vitro* methylation analyses unveil distinctive methylation signatures for human *de novo* methyltransferases in both CpG and non-CpG contexts, which reveals intrinsic enzymatic substrate specificities.

To examine the possible asymmetry of sequence preferences within a duplex context, we identified the 10-mer CpG sites that were both >60% methylated at the C (forward strand) and G (reverse strand) position. Then, 1748 heavily methylated duplex sites were found after incubation with DNMT3A, and 18062 sites for DNMT3B. Sequence logo analysis reveals a core [A/G]CG[T/C] signature for DNMT3A and a [C/T]CG[G/A] signature for DNMT3B (Figure S7). We found no evidence to support asymmetry in sequence preference. These signatures were self-complementary, in concordance with the flanking sequence signature of DNMT3A/B.

DNMT3L Stimulates DNMT3A/B Activity without Altering Sequence Preference. DNMT3L is highly related to DNA methyltransferases, and though it does not have any methyltransferase activity per se, it is a key factor that stimulates *de novo* methylation.[23−25] Early work on DNMT3L suggested that it can modulate DNMT3A/B activity without changing the sequence preferences of DNMT3A/B.[32] However, this study focused on only a limited number of CpG sites and used near-saturation levels of methylation; thus, an unbiased and accurate assessment of the effects of DNMT3L remains open.

To further investigate how DNMT3L may affect DNMT3A/B sequence preferences, we added full-length human recombinant DNMT3L to the methylation reaction together with DNMT3A or DNMT3B. To avoid methylation saturation due to increased overall methylation levels, 200 ng instead of 500 ng of DNMT3A or DNMT3B was used, which resulted in 14% of CG methylation for DNMT3A and 3.2% for DNMT3B (Figure 3A). DNMT3L increased DNMT3A methylation activity by 3-fold and DNMT3B methylation activity by 11-fold in a CpG context (Figure 3A), which is consistent with previous reports.[23−25] Methylation at non-CpG sites was also enhanced (Figure 3B). Sequence logo analysis shows an

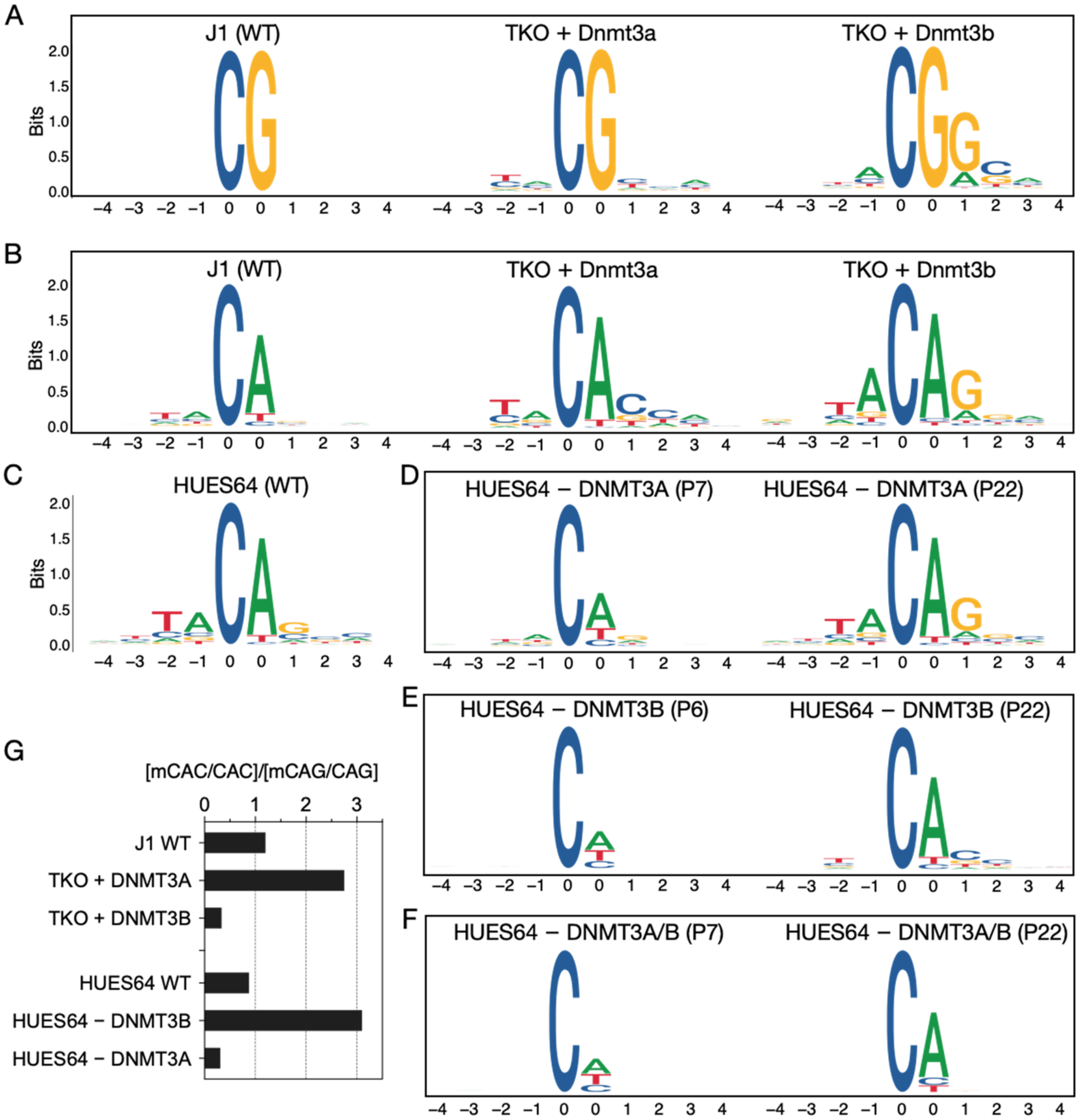


Figure 4. Methylation signatures in the mouse and human genome depend on the presence of DNMT3A or DNMT3B. Sequence logos of the most methylated sequence contexts in chromosome 1 of mouse stem cells. (A) Top methylated 10-mer CG sequences: 100% methylation for WT (left; N = 325622), >60% methylation for TKO with Dnmt3a (middle; N = 628), and >40% for TKO with Dnmt3b (right; N = 399). (B) Top methylated 10-mer CH sequences: >40% methylation for WT (left; N = 16392), >30% for TKO with Dnmt3a (middle; N = 533), and >20% for TKO with Dnmt3b (right; N = 177). Sequence logos of the most methylated (>20% methylation) CH sequence contexts in chromosome 1 of hESCs. (C) WT hESCs (N = 163050). (D) DNMT3A-KO hESCs at early passage 7 (left; N = 28101) and late passage 22 (middle; N = 59323). (E) DNMT3B-KO hESCs at passage 6 (left; N = 24327) and passage 22 (middle; N = 20603). (F) DNMT3A/B double-knockout hESCs at passage 7 (left; N = 12736) and passage 22 (right; N = 6859). (G) Ratio between CAC and CAG methylation in mouse Dnmt-TKO cells with Dnmt3a/Dnmt3b reintroduced and hESCs with either DNMT3A or DNMT3B knocked out.

unaltered sequence preference for DNMT3A/3B after addition of DNMT3L in both CpG and non-CpG contexts (Figure 3C−F; see also Figure 2C−F and Figures S5 and S8). This suggests that the stimulatory effect of DNMT3L does not alter the flanking sequence preference for DNMT3A/B, which is consistent with the absence of any direct interaction between DNMT3L and DNA within a DNMT3A−DNMT3L tetramer complex.[41,42]

Methylation Signatures of DNMT3A and DNMT3B in Mammalian Cells. To further expand our *in vitro* findings that revealed DNMT3A/B sequence preferences, we asked if the observed patterns hold true in cellular and physiological conditions. Subsequently, we explored the extent to which endogenous mammalian DNA methylomes are explained by the distinct specificities of DNMT3A and DNMT3B.

Mammalian DNMT3 protein sequences are highly conserved with 96% of amino acids (875 of 912) being identical between mouse and human DNMT3A, including 100% identical C-terminal residues and catalytic domains (508−912). Human and mouse DNMT3B protein sequences are 88% identical (717 of 817). Due to this level of conservation, we anticipate that human and mouse DNMT3 enzymes will show equivalent sequence preferences, and therefore, we used

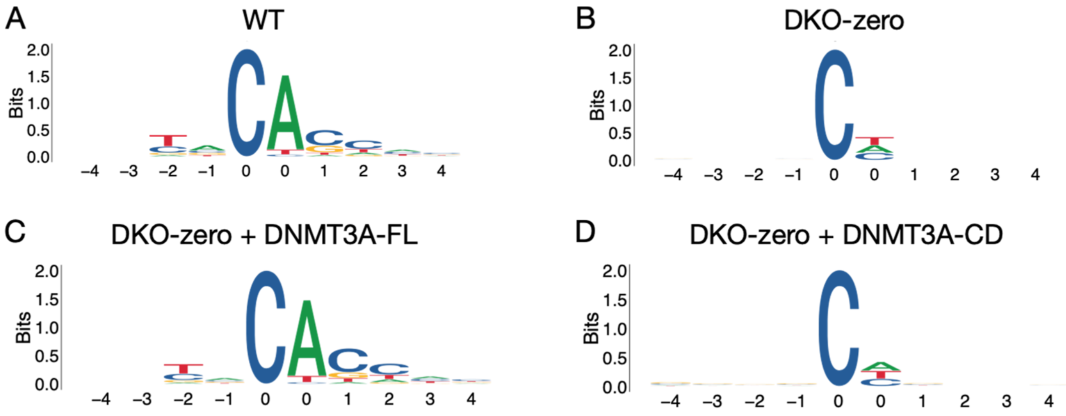


Figure 5. N-Terminal domain accounting for the sequence preferences of DNMT3A. Sequence logos of the most methylated 10-mer non-CpG sequences ($N = 5000$) in (A) WT cells (WT), (B) DKO-zero cells, (C) DKO-zero cells expressing full-length DNMT3A (DNMT3A-FL), or (D) DKO-zero cells expressing the DNMT3A catalytic domain (DNMT3A-CD).

human or mouse methylation data sets interchangeably in the following analyses.

We examined the patterns of highly methylated sequences in wild-type (WT) J1 mouse embryonic stem cells (mESCs) compared to mESCs in which Dnmt1, -3a, and -3b have been genetically deleted (Dnmt triple-knockout or TKO cells)[43] with either Dnmt3a or Dnmt3b subsequently reintroduced ectopically.[44] In WT stem cells, no obvious pattern was observed in the top methylated 10-mer sequences (i.e., CpG ± 4 bases), which most likely reflects the close-to saturation levels of CG methylation (81.7%). In contrast, in TKO cells with reintroduced Dnmt3a or Dnmt3b, there is an average methylation level of 7.1% and 2.8%, respectively, in CpG contexts, which are nonsaturating and therefore allow the further analysis of sequence preferences (see Materials and Methods). The preferred sequence contexts were readily apparent at the methylated sites in TKO cells expressing either Dnmt3a or Dnmt3b (Figure 4A,B). These signatures correspond closely to the patterns identified in the *in vitro* assay (Figure 2C−F); namely, CpG and non-CpG sites followed by a 3′ pyrimidine gained more methylation when Dnmt3a was expressed and 3′ purine when Dnmt3b was expressed (Figure 4A,B).

To explore if the same preferences persist in human cells, we then profiled methylation signatures in HUES64 human ESCs (hESCs) with either wild-type or DNMT knockout genotypes.[45] WT hESCs displayed a mixture of DNMT3A-type and DNMT3B-type methylation signature (Figure 4C), which was not observed in mouse WT cells. We attributed this to the higher level of expression of DNMT3A/B in human HUES64 cells compared to mouse cells (Figure S9A). Moreover, we observed that the DNMT3B-type signature emerges when DNMT3A is depleted, with later cell culture passages leading to more prominent effect (Figure 4D). Similarly, removal of DNMT3B leads to the loss of the DNMT3B signature in early passages with the subsequent appearance of the DNMT3A signature, which suggests the slow dilution of DNMT3B-type methylation and accumulation of DNMT3A type over a period of 15 passages (Figure 4E). Finally, DNMT3A and DNMT3B double knockout leads to a substantial loss of CA methylation (from 1.8% to 0.2%) and loss of DNMT3 signatures (Figure 4F).

The clear difference in sequence preferences between DNMT3A and DNMT3B is at the 3′ base directly adjacent to the substrate dinucleotides. To further infer whether the methylation levels in CAC and CAG contexts are a good representation of the DNMT3A and DNMT3B methylation signatures, we calculated the average methylation at trinucleotides CAN (N = A, T, C, or G) in mouse and human stem cells and found that CAC gained more methylation compared to other trinucleotides when Dnmt3a was introduced. On the contrary, more methylation at CAG was observed when Dnmt3b was reintroduced, which is consistent with the preferences discovered before (Figure S9B). In both human and mouse WT ESCs, the ratio between CAC and CAG methylation is close to 1, suggesting a balancing act between DNMT3A and -3B (Figure 4G). Additionally, introduction of DNMT3A into mouse TKO cells (or removal of DNMT3B in human WT cells) led to ∼2−3-fold more CAC methylation; however, introduction of DNMT3B into mouse TKO cells (or removal of DNMT3A in human WT cells) led to more CAG methylation. In line with the inherent sequence preferences flanking CpG sites revealed *in vitro*, we also noted that TKO cells gain more CGC/CGT methylation after reintroduction of Dnmt3a and more CGG/CGA methylation after reintroduction of Dnmt3b (Figure S9C,D). No significant change was observed in WT or DNMT3A- and DNMT3B-depleted hESCs in sequence contexts adjacent to CpG dinucleotides, which may be due to saturation levels (Figure S9C,D).

The DNMT3A N-Terminal Domain Imparts Sequence Preferences. The distinct patterns of flanking sequence preferences for DNMT3A or DNMT3B at both CpG and non-CpG sites suggest that there are intrinsic enzyme structural features determining their specificity. To determine whether the N-terminal or the catalytic domain is a determinant for the sequence preferences of DNMT3A, we analyzed publicly available RRBS data sets generated in the Dnmt3a/b double knockout and Dnmt1 knocked down mESCs (DKO-zero) expressing either full-length (FL) or the catalytic domain (CD) of Dnmt3a.[46] The expression of either FL- or CD-Dnmt3a reinstated CpG methylation levels similar to that of WT cells.[46] The most methylated non-CpG sites in WT cells revealed a TN**CA**[C/G]C methylation signature combining the DNMT3A and DNMT3B's methylation signatures observed *in vitro* (Figure 5A, and also Figures 2E and 4B). The knockout of Dnmt3a/b and knockdown of Dnmt1 abrogated methylation in DKO-zero cells, which resulted in no methylation signature (Figure 5B). The reintroduction of full-length

DNMT3A (but not the DNMT3A catalytic domain) restored the characteristic DNMT3A methylation signature observed in WT cells (Figure 5C,D). Overall, this suggests that the N-terminal domain is a determinant for the sequence preference of DNMT3A.

DISCUSSION

A key challenge is to build an understanding of how the *de novo* methyltransferases DNMT3A and DNMT3B cooperate to establish the mammalian DNA methylome in early embryonic development. Evidence suggests that the underlying primary genomic sequence could be involved in the dynamic and recurring deposition of cytosine methylation in regulatory regions by sequence specific recruitment of transcription factors.[29,47] However, how sequence context affects the activity of *de novo* DNA methyltransferases is elusive. By quantitatively examining the *in vitro* methylation activity of full-length human recombinant DNA methyltransferases on a diverse set of sequence contexts present in a small bacterial genome, we have uncovered the inherent enzymatic preferences for sequences flanking the substrate dinucleotides. DNMT3A favors a TNC[G/A]CC signature, while DNMT3B prefers TAC[G/A]GC. Our observations are corroborated by our findings of similar Dnmt3a or Dnmt3b methylation signatures in mouse Dnmt-TKO cells that express either Dnmt3a or Dnmt3b ectopically. Furthermore, depletion of DNMT3A in human HUES64 cells enhances a DNMT3B-type methylation pattern, especially in a CA context, while removal of DNMT3B leads to the appearance of a DNMT3A-type signature. Taken together, we propose that the intrinsic sequence preferences of DMNT3A/B should be taken into consideration when studying the establishment of tissue specific methylation patterns.

From our analysis of mouse TKO stem cells and human DNMT3 knockout cells, it is evident that DNMT3A and DNMT3B impose methylation patterns in cells that resemble those seen *in vitro* from the corresponding purified recombinant enzymes in the absence of additional factors. This suggests that while the interaction with DNMT3L,[48−50] histone modifications,[21,22,44] or transcription factors[29] could modulate or guide the methylation capacity of DNMT3s at certain regions, the inherent enzyme sequence preferences shape a substantial part of the underlying methylation patterns globally.

While human DNMT3A and DNMT3B share ∼45% conservation across the whole protein, ∼80% of amino acids are conserved in the catalytic domain. This points to regulatory features outside the catalytic domain having evolved to provide each protein selectivity to methylate distinct genomic loci in different tissues and developmental stages. Epigenetic enzymes such as DNMTs and TETs are being deployed in a range of epigenetic engineering and biotechnological setups with potential clinical utility, and our examination of the intrinsic sequence preference of these enzymes could help guide the selection of DNMT3s for optimal activity.

CONCLUSIONS

In summary, we provide a comprehensive and robust quantitative analysis of the intrinsic sequence preferences for the enzymatic activities of *de novo* DNA methyltransferases on CpG and non-CpG target sites *in vitro* and in mammalian stem cells. The accurate determination of sequence preferences of *de novo* methyltransferases provides a new understanding of the origin of specific DNA methylation patterns in different cell lineages and regulatory regions.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.biochem.0c00339.

Figures S1−S9 and Table S1 (PDF)

Accession Codes

The data are available in ArrayExpress with accession code E-MTAB-8406.

AUTHOR INFORMATION

Corresponding Author

Shankar Balasubramanian − *Cancer Research UK Cambridge Institute, Li Ka Shing Centre, Cambridge CB2 0RE, U.K.; Department of Chemistry and School of Clinical Medicine, University of Cambridge, Cambridge CB2 1EW, U.K.*; orcid.org/0000-0002-0281-5815; Email: sb10031@cam.ac.uk

Authors

Shi-Qing Mao − *Cancer Research UK Cambridge Institute, Li Ka Shing Centre, Cambridge CB2 0RE, U.K.*

Sergio Martínez Cuesta − *Cancer Research UK Cambridge Institute, Li Ka Shing Centre, Cambridge CB2 0RE, U.K.; Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, U.K.*

David Tannahill − *Cancer Research UK Cambridge Institute, Li Ka Shing Centre, Cambridge CB2 0RE, U.K.*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.biochem.0c00339

Author Contributions

‖S.-Q.M. and S.M.C. contributed equally to this work.

Funding

This work is supported by Cancer Research UK (C14303/A17197) and a Wellcome Trust Investigator Award (209441/z/17/z).

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Bird, A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev. 16*, 6−21.

(2) Jones, P. A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet. 13*, 484−92.

(3) Schübeler, D. (2015) Function and information content of DNA methylation. *Nature 517*, 321−326.

(4) Okano, M., Xie, S., and Li, E. (1998) Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat. Genet. 19*, 219−220.

(5) Okano, M., Bell, D. W., Haber, D. A., and Li, E. (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell 99*, 247−257.

(6) Gowher, H., and Jeltsch, A. (2001) Enzymatic properties of recombinant Dnmt3a DNA methyltransferase from mouse: the enzyme modifies DNA in a non-processive manner and also methylates non-CpA sites. *J. Mol. Biol. 309*, 1201−1208.

(7) Song, J., Rechkoblit, O., Bestor, T. H., and Patel, D. J. (2011) Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. *Science 331*, 1036−1040.

(8) Song, J., Teplova, M., Ishibe-Murakami, S., and Patel, D. J. (2012) Structure-based mechanistic insights into DNMT1-mediated maintenance DNA methylation. *Science 335*, 709−712.

(9) Law, J. A., and Jacobsen, S. E. (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat. Rev. Genet. 11*, 204−220.

(10) Reik, W., Dean, W., and Walter, J. (2001) Epigenetic reprogramming in mammalian development. *Science 293*, 1089−93.

(11) 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., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., and Ecker, J. R. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature 462*, 315−322.

(12) Smith, Z. D., Chan, M. M., Mikkelsen, T. S., Gu, H., Gnirke, A., Regev, A., and Meissner, A. (2012) A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature 484*, 339−344.

(13) Ren, W., Gao, L., and Song, J. (2018) Structural Basis of DNMT1 and DNMT3A-Mediated DNA Methylation. *Genes 9*, 620.

(14) Ziller, M. J., Gu, H., Müller, F., Donaghey, J., Tsai, L. T. Y., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., Gnirke, A., and Meissner, A. (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature 500*, 477−481.

(15) Ramsahoye, B. H., Biniszkiewicz, D., Lyko, F., Clark, V., Bird, A. P., and Jaenisch, R. (2000) Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc. Natl. Acad. Sci. U. S. A. 97*, 5237−5242.

(16) Guo, J. U., Su, Y., Shin, J. H., Shin, J., Li, H., Xie, B., Zhong, C., Hu, S., Le, T., Fan, G., Zhu, H., Chang, Q., Gao, Y., Ming, G. L., and Song, H. (2014) Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. *Nat. Neurosci. 17*, 215−222.

(17) He, Y., and Ecker, J. R. (2015) Non-CG Methylation in the Human Genome. *Annu. Rev. Genomics Hum. Genet. 16*, 55−77.

(18) Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., Lucero, J., Huang, Y., Dwork, A. J., Schultz, M. D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J. C., Rao, A., Esteller, M., He, C., Haghighi, F. G., Sejnowski, T. J., Behrens, M. M., and Ecker, J. R. (2013) Global epigenomic reconfiguration during mammalian brain development. *Science 341*, 1237905.

(19) Chodavarapu, R. K., Feng, S., Bernatavichute, Y. V., Chen, P.-Y., Stroud, H., Yu, Y., Hetzel, J. A., Kuo, F., Kim, J., Cokus, S. J., Casero, D., Bernal, M., Huijser, P., Clark, A. T., Krämer, U., Merchant, S. S., Zhang, X., Jacobsen, S. E., and Pellegrini, M. (2010) Relationship between nucleosome positioning and DNA methylation. *Nature 466*, 388−392.

(20) Kelly, T. K., Liu, Y., Lay, F. D., Liang, G., Berman, B. P., and Jones, P. A. (2012) Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res. 22*, 2497−2506.

(21) Cedar, H., and Bergman, Y. (2009) Linking DNA methylation and histone modification: Patterns and paradigms. *Nat. Rev. Genet. 10*, 295−304.

(22) Du, J., Johnson, L. M., Jacobsen, S. E., and Patel, D. J. (2015) DNA methylation pathways and their crosstalk with histone methylation. *Nat. Rev. Mol. Cell Biol. 16*, 519−532.

(23) Chedin, F., Lieber, M. R., and Hsieh, C.-L. (2002) The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a. *Proc. Natl. Acad. Sci. U. S. A. 99*, 16916−16921.

(24) Suetake, I., Shinozaki, F., Miyagawa, J., Takeshima, H., and Tajima, S. (2004) DNMT3L Stimulates the DNA Methylation Activity of Dnmt3a and Dnmt3b through a Direct Interaction. *J. Biol. Chem. 279*, 27816−27823.

(25) Gowher, H., Liebert, K., Hermann, A., Xu, G., and Jeltsch, A. (2005) Mechanism of stimulation of catalytic activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L. *J. Biol. Chem. 280*, 13341−13348.

(26) Wachter, E., Quante, T., Merusi, C., Arczewska, A., Stewart, F., Webb, S., and Bird, A. (2014) Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. *eLife 3*, No. e03397.

(27) Krebs, A. R., Dessus-Babus, S., Burger, L., and Schübeler, D. (2014) High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. *eLife 3*, No. e04094.

(28) Macleod, D., Charlton, J., Mullins, J., and Bird, A. P. (1994) Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes Dev. 8*, 2282−2292.

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

(30) Mao, S.-Q., Ghanbarian, A. T., Spiegel, J., Martínez Cuesta, S., Beraldi, D., Di Antonio, M., Marsico, G., Hänsel-Hertsch, R., Tannahill, D., and Balasubramanian, S. (2018) DNA G-quadruplex structures mold the DNA methylome. *Nat. Struct. Mol. Biol. 25*, 951−957.

(31) Yokochi, T., and Robertson, K. D. (2002) Preferential methylation of unmethylated DNA by mammalian de novo DNA methyltransferase Dnmt3a. *J. Biol. Chem. 277*, 11735−11745.

(32) Wienholz, B. L., Kareta, M. S., Moarefi, A. H., Gordon, C. A., Ginno, P. A., and Chédin, F. (2010) DNMT3L modulates significant and distinct flanking sequence preference for DNA methylation by DNMT3A and DNMT3B in vivo. *PLoS Genet. 6*, No. e1001106.

(33) Schultz, M. D., He, Y., Whitaker, J. W., Hariharan, M., Mukamel, E. A., Leung, D., Rajagopal, N., Nery, J. R., Urich, M. A., Chen, H., Lin, S., Lin, Y., Jung, I., Schmitt, A. D., Selvaraj, S., Ren, B., Sejnowski, T. J., Wang, W., and Ecker, J. R. (2015) Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature 523*, 212−216.

(34) Quinlan, A. R., and Hall, I. M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics 26*, 841−842.

(35) Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal 17*, 10.

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

(37) Ginestet, C. (2011) ggplot2: Elegant Graphics for Data Analysis. *J. R. Stat. Soc. Ser. A (Statistics Soc. 174*, 245−246.

(38) Wagih, O. (2017) ggseqlogo: A versatile R package for drawing sequence logos. *Bioinformatics 33*, 3645−3647.

(39) Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., Gonzalez, J. M., Grego, T., Hardy, M., Hourlier, T., Hunt, T., Izuogu, O. G., Lagarde, J., Martin, F. J., Martínez, L., Mohanan, S., Muir, P., Navarro, F. C. P., Parker, A., Pei, B., Pozo, F., Ruffier, M., Schmitt, B. M., Stapleton, E., Suner, M. M., Sycheva, I., Uszczynska-Ratajczak, B., Xu, J., Yates, A., Zerbino, D., Zhang, Y., Aken, B., Choudhary, J. S., Gerstein, M., Guigó, R., Hubbard, T. J. P., Kellis, M., Paten, B., Reymond, A., Tress, M. L., and Flicek, P. (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res. 47*, D766−D773.

(40) Guenatri, M., Duffié, R., Iranzo, J., Fauque, P., and Bourc'his, D. (2013) Plasticity in Dnmt3L-dependent and -independent modes of de novo methylation in the developing mouse embryo. *Development 140*, 562−572.

(41) Jia, D., Jurkowska, R. Z., Zhang, X., Jeltsch, A., and Cheng, X. (2007) Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature 449*, 248−251.

(42) Zhang, Z.-M., Lu, R., Wang, P., Yu, Y., Chen, D., Gao, L., Liu, S., Ji, D., Rothbart, S. B., Wang, Y., Wang, G. G., and Song, J. (2018) Structural basis for DNMT3A-mediated de novo DNA methylation. *Nature 554*, 387−391.

(43) Tsumura, A., Hayakawa, T., Kumaki, Y., Takebayashi, S., Sakaue, M., Matsuoka, C., Shimotohno, K., Ishikawa, F., Li, E., Ueda, H. R., Nakayama, J., and Okano, M. (2006) Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. *Genes Cells 11*, 805−814.

(44) Baubec, T., Colombo, D. F., Wirbelauer, C., Schmidt, J., Burger, L., Krebs, A. R., Akalin, A., and Schübeler, D. (2015) Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature 520*, 243−247.

(45) Liao, J., Karnik, R., Gu, H., Ziller, M. J., Clement, K., Tsankov, A. M., Akopian, V., Gifford, C. a, Donaghey, J., Galonska, C., Pop, R., Reyon, D., Tsai, S. Q., Mallard, W., Joung, J. K., Rinn, J. L., Gnirke, A., and Meissner, A. (2015) Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nat. Genet. 47*, 469−478.

(46) Galonska, C., Charlton, J., Mattei, A. L., Donaghey, J., Clement, K., Gu, H., Mohammad, A. W., Stamenova, E. K., Cacchiarelli, D., Klages, S., Timmermann, B., Cantz, T., Schöler, H. R., Gnirke, A., Ziller, M. J., and Meissner, A. (2018) Genome-wide tracking of dCas9-methyltransferase footprints. *Nat. Commun. 9*, 597.

(47) Quante, T., and Bird, A. (2016) Do short, frequent DNA sequence motifs mould the epigenome? *Nat. Rev. Mol. Cell Biol. 17*, 257−262.

(48) Neri, F., Krepelova, A., Incarnato, D., Maldotti, M., Parlato, C., Galvagni, F., Matarese, F., Stunnenberg, H. G., and Oliviero, S. (2013) Dnmt3L Antagonizes DNA Methylation at Bivalent Promoters and Favors DNA Methylation at Gene Bodies in ESCs. *Cell 155*, 121−134.

(49) Suetake, I., Morimoto, Y., Fuchikami, T., Abe, K., and Tajima, S. (2006) Stimulation effect of Dnmt3L on the DNA methylation activity of Dnmt3a2. *J. Biochem. 140*, 553−559.

(50) Ooi, S. K. T., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.-P., Allis, C. D., Cheng, X., and Bestor, T. H. (2007) DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature 448*, 714−717.