PLOS ONE

RESEARCH ARTICLE

The sequence preference of DNA methylation variation in mammalians

Ling Zhang[1,2], Chan Gu[2,3], Lijiang Yang[1,2], Fuchou Tang[1], Yi Qin Gao[1,2]*

1 Biodynamic Optical Imaging Center (BIOPIC), School of Life Sciences, Peking University, Beijing, China, **2** Institute of Theoretical and Computational Chemistry, College of Chemistry and Molecular Engineering, Peking University, Beijing, China, **3** Prenatal Diagnosis Center, Department of Obstetrics and Gynecology, Ministry of Education Key Laboratory of Obstetric, Gynecologic and Pediatric Diseases and Birth Defects, West China Second University Hospital, Sichuan University, Chengdu, Sichuan, China

* gaoyq@pku.edu.cn

Citation: Zhang L, Gu C, Yang L, Tang F, Gao YQ (2017) The sequence preference of DNA methylation variation in mammalians. PLoS ONE 12(10): e0186559. https://doi.org/10.1371/journal.pone.0186559

Editor: Dajun Deng, Beijing Cancer Hospital, CHINA

Received: June 2, 2017

Accepted: October 3, 2017

Published: October 18, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the National Natural Science Foundation of China under 21573006, 21233002. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Methylation of cytosine at the 5 position of the pyrimidine ring is the most prevalent and significant epigenetic modifications in mammalian DNA. The CpG methylation level shows a bimodal distribution but the bimodality can be overestimated due to the heterogeneity of per-base depth. Here, we developed an algorithm to eliminate the effect of per-base depth inhomogeneity on the bimodality and obtained a random CpG methylation distribution. By quantifying the deviation of the observed methylation distribution and the random one using the information formula, we find that in tetranucleotides 5'-N_5CGN_3-3' (N_5, N_3 = A, C, G or T), $GCGN_3$ and $CCGN_3$ show less apparent deviation than $ACGN_3$ and $TCGN_3$, indicating that $GCGN_3$ and $CCGN_3$ are less variant in their level of methylation. The methylation variation of N_5CGN_3 are conserved among different cells, tissues and species, implying common features in the mechanisms of methylation and demethylation, presumably mediated by DNMTs and TETs in mammalians, respectively. Sequence dependence of DNA methylation variation also relates to gene regulatory and promotes the reexamination of the role of DNA sequence in fundamental biological processes.

Introduction

Eukaryotic chromosomes carry genomic information of individual growth and development, which is stored not only in DNA sequence, but also in epigenetic information such as DNA methylation and histone modifications [1, 2]. Methylation of cytosine at the 5 position of the pyrimidine ring (5mC) is the most prevalent and significant epigenetic modification in mammalian DNA [3], which is generally associated with cellular processes such as cell differentiation, X chromosome inactivation, transposon silencing, genomic imprinting and tumourigenesis [4, 5]. About 1.5 percent of the cytosines in mammalian genomes are methylated [6]. The vast majority of 5mC exists at CpG dinucleotides, and more than 60 percent of CpG dinucleotides are methylated in mammalian genome [7]. One interesting phenomenon is that the distribution of methylcytosines in the genome of many species and cell types is not random [8], and the CpG methylation levels exhibit a bimodal distribution indicating a sequence/site

specificity. The measured CpG methylation levels distribution peaking at 0 and 1 [9, 10], which implies that during sequencing a large population of the CpG sites remain either unmethylated or fully methylated. Most of the CpG dinucleotides are hypermethylated, whereas those in CpG islands keep hypomethylated in adult cells [11]. One of the possible functions of this bimodal pattern is to keep the factor-mediated basal transcription profile of the preimplantation embryo [12].

The development of the next generation sequencing (NGS) technology has made the methylome of many species available. Lister *et al*. performed the first genome-wide single-base resolution sequencing of *Arabidopsis* methylome in 2008 [13]. followed by the methylome of h1 human embryonic stem cells (ESCs) and human induced pluripotent stem cells (iPSCs) [14]. In 2013 the same group implemented the sequencing of DNA methylation in the frontal cortex of human and mice covering their life span [15]. Contemporaneously, many other researchers also obtained methylomes of other tissues and species. For example, Guo *et al*. carried out the sequencing of human early embryos and human primordial germ cells (PGCs) [16], and another important progress is that the epigenomes of 18 tissue types from 4 individuals of high coverage were obtained by Schultz *et al*. [17]. 111 reference human epigenomes were obtained by the NIH Roadmap Epigenomics Consortium in 2015 [18]. The abundant methylome data from previous work enable us to exploit the bioinformatics approach for a better understanding of DNA methylation.

Despite the modern experimental sequencing technologies have achieved significant success, researchers are still confronted with great challenges. For example, the non-uniform sequencing depth of each CpG can lead to the overestimation of the bimodality of CpG methylation. We use sequencing depth to denote the coverage of a CpG dinucleotide in sequencing. In this paper we eliminated the bias of the bimodality of CpG methylation caused by the heterogeneous sequencing depths to the bimodality of CpG methylation and analyzed the bimodal distributions of CpG methylation for a large variety of mammalian cell and tissue types. Our goal in this study is to investigate whether the flanking bases have an effect on the CpG methylation and whether such an effect, if exists, could shed light on the mechanism of methylation and demethylation. Interestingly, these analyses did reveal a sequence dependent feature that is common for all samples analyzed. In tetranucleotides 5'-N_5CGN_3-3' (N_5, N_3 = A, C, G or T), the $GCGN_3$ and $CCGN_3$ show a lower tendency of methylation variation, compared to $ACGN_3$ and $TCGN_3$. Molecular dynamics simulations were then used to understand the origin of such differences. The intrinsic DNA structure parameters of CpG/5mCpG sites are found to be significantly affected by the flanking bases N_5 and N_3. The structural differences between different sequences provide a possible explanation of the variation of methylation properties and further may clarify the gene regulatory mechanisms.

Materials and methods

Data resources

Methylomes of mouse and human brain cells are from Lister *et al*. [15]. Human embryonic stem cells (ESCs) and human induced pluripotent stem cells (iPSCs) are taken from Lister *et al*. [14], and methylomes of human normal somatic cells are from Schultz *et al*. [17]. Methylomes of human primordial germ cells (PGCs) and the neighboring somatic cells (gonadal somatic cells, SOMAs) are taken from Guo *et al*. [16]. The details of each sample are listed in S1–S6 Tables. We categorized the data into six groups: mouse brain cells, human brain cells, human ESCs and iPSCs, human normal somatic cells, human PGCs and human SOMAs.

Analysis of methylation level

Following the literature, we use β_i to represent the measured methylation level of the ith CpG site.

$$\beta_i = \frac{M_i}{T_i} \tag{1}$$

where T_i is the sequencing depth of the ith CpG site and M_i is its measured methylation frequency. Accordingly, one can obtain the observed probability distribution of β through a simple count of the appearing frequency, which is normally represented by a discretized function $r_{obs}(\beta)$,

$$r_{obs}(\beta) = \frac{\sum_i \delta(\beta - \beta_i)}{N} \tag{2}$$

Where N is the total number of all CpG sites and $\delta(\cdot)$ is the delta function.

If all the sequencing depth T is large enough, one can get the accurate methylation level β of each CpG site as well as the according β distribution (denoted as $r_{acc}(\beta)$). However, due to the limited and non-uniform sequencing depth in the experiments, one gets only a measured methylation level of each CpG site (β_i) (Eq 1) and its distribution $r_{obs}(\beta)$ (Eq 2).

In Eq 2, both the denominator and nominator are integers, therefore $r_{obs}(\beta)$ can only come out among a group of fraction numbers, that means the continuous $r_{acc}(\beta)$ is "discretized" as $r_{obs}(\beta)$. Due to the non-uniform sequencing depth, $r_{obs}(0)$ and $r_{obs}(1)$ can be overestimated compared to the $r_{acc}(0)$ and $r_{acc}(1)$, respectively. For a specific sequencing depth T, β can only come out as $0, \frac{1}{T}, \frac{2}{T}, \ldots, 1$, but not any other values. For example, a sequencing depth of 2 yields possible methylation levels of 0, ½, and 1, while a depth of 4 yields 0, ¼, ½, , and 1. If one directly counts the observed values, there would be a bias towards more favored values of 0, ½, and 1 compared to ¼ and . On the other hand, every sequencing depth can generate a β values of 0 and 1, thus β values of 0 and 1 appear more frequently than other values in a simple counting strategy, which will result in an artifact in the observed bimodal distribution, especially when a bunch of low sequencing depth CpG sites are included in the analyses.

To eliminate the bias caused by the non-uniform sequencing depth, we computationally generate an artificial sample in which the sequencing depth of each CpG site is identical to the experimental value but the methylation frequency of each CpG site obeys a binomial distribution, which means that the methylation level of each CpG site is random. The methylation distribution of the randomized sample is denoted as $r_{ran}(\beta)$. The difference between the direct observation ($r_{obs}(\beta)$) and the random counterpart ($r_{ran}(\beta)$) is free of sequencing depth bias and it characterizes how much the observed methylation level distribution deviates from the random distribution.

In the randomized sample, for the ith CpG site with a sequencing depth T_i, the binomial distribution [19] of the methylation frequency $n(n = 0,1 \ldots T_i)$ can be formulated as,

$$P(T_i, n) = \mathrm{C}_{T_i}^{n} p^n (1-p)^{T_i - n} \tag{3}$$

where p is defined as the average of the observed methylation level of all CpG sites (Eq 4).

$$p = \frac{\sum_{i=1}^{N} \beta_i}{\mathrm{N}} \tag{4}$$

The probability distribution of β in the random case for the methylome can then be

calculated by Eq 5

$$r_{ran}(\beta) = \frac{\sum_i \sum_n P(T_i, n)\delta(n/T_i - \beta)}{N} \quad (5)$$

To quantify the deviation of the observed methylation distribution ($r_{obs}(\beta)$) and that in the random case ($r_{ran}(\beta)$), one can use the direct difference of them ($d(\beta) = r_{obs}(\beta) - r_{ran}(\beta)$) or another choice is the relative entropy (also known as Kullback-Leibler divergence) [20]. In the latter approach, larger relative entropy indicates a stronger deviation between the two functions. The deviation of the observed distribution from the random distribution is expressed in Eq 6.

$$KL(\beta) = KL(r_{obs}(\beta)||r_{ran}(\beta)) = \sum_\beta r_{obs}(\beta)\log_2 \frac{r_{obs}(\beta)}{r_{ran}(\beta)} \quad (6)$$

Specifically, for $KL(0)$ and $KL(1)$, we have

$$KL(0) = KL(r_{obs}(0)||r_{ran}(0)) = r_{obs}(0)\log_2 \frac{r_{obs}(0)}{r_{ran}(0)} \quad (7)$$

$$KL(1) = KL(r_{obs}(1)||r_{ran}(1)) = r_{obs}(1)\log_2 \frac{r_{obs}(1)}{r_{ran}(1)} \quad (8)$$

The biological implication of the deviation between the observed methylation level distribution and the random distribution at 0 ($d(0)$ or $KL(0)$) and 1 ($d(1)$ or $KL(1)$) are that they reflect the methylation and demethylation variation (or conservation conversely), respectively. As in the normal bisulfite sequencing, the mixed cell populations were sequenced and the measured methylation level β_i for the ith CpG site approximately indicates the ratio of the reads with the ith CpG site methylated to the total reads with ith CpG site detected, the greater difference between the observed methylation level distribution and the random distribution in the 0 ($d(0)$ or $KL(0)$) means the CpG is prone to keep its unmethylated state.

For tetranucleotide N_5CGN_3 (N_5, N_3 = A, C, G or T), there are 16 possible different sequences with the two DNA chains considered separately. We calculate all the 16 $KL(\beta)$ to investigate the influence of the flanking sequences to the CpG sites. The p in Eq 3 is thus the average methylation level of the corresponding tetranucleotide (ACGA, CCGT *et al.*).

Results

The bimodal distribution of CpG methylation level and methylation variation

As reported earlier, the distribution of the mammalian CpG methylation level is bimodal, with peaks seen at $\beta = 0$ and $\beta = 1$. We calculated the CpG methylation level distribution for all six categories of samples (66 samples in total, Materials and methods). The observed methylation level distribution does show a typical bimodal pattern (Fig 1A, red). For comparison, we also show in this figure the randomized methylation level following the binomial distribution with the experimental sequence depth taken into account (Materials and methods, Eq 5, Fig 1A, green). The randomized methylation level distribution also exposits the bimodal feature, which shows that even if a simple binomial distribution assumption with uneven sequencing depth can result in bimodality, although appears to a less extent than the experimental observation. The more pronounced peaks at $\beta = 0$ and $\beta = 1$ in the experimental data (Fig 1A, red)

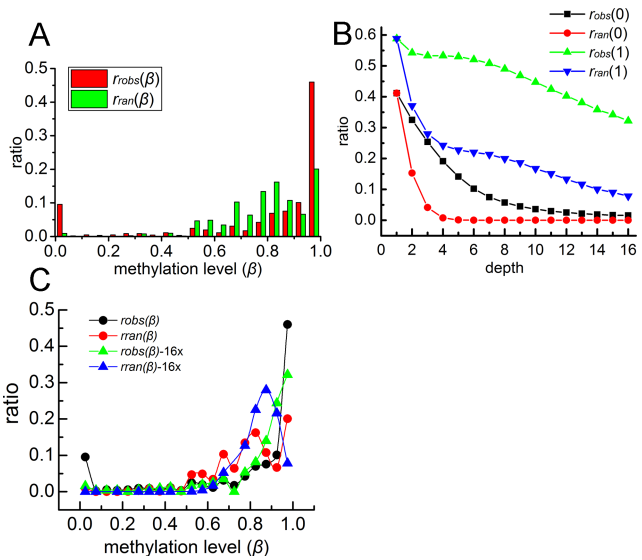

Fig 1. The bimodal distribution of CpG methylation in human brain samples. (A) The bimodal bias caused by the inhomogeneity of sequencing depth in human brain cells (chromosome 1 of the 12yr sample). **(B)** The observed ratio and the random ratio of 0 and 1 end of the methylation level distribution in different sequencing depths. **(C)** The observed and random methylation level distribution of all sequencing depth (black circle and red circle respectively) and the observed and random methylation level distribution under the sequencing depth of 16x is in green triangle and blue triangle respectively.

https://doi.org/10.1371/journal.pone.0186559.g001

indicates that the methylation level is indeed biased toward 0 or 1, even after the removal of the bias caused by the sequencing depth.

To illuminate the effect of sequencing depth on the observed bimodality of CpG methylation level distribution, the $r_{obs}(0)$, $r_{ran}(0)$, $r_{obs}(1)$ and $r_{ran}(1)$ of different sequencing depth is shown in Fig 1B, which clearly shows that with the increase of sequencing depth, the bimodality of methylation level become less obvious. We also show in the Supporting information the observed and random distribution of CpG methylation in each sequencing depth (S1 Fig). It can be seen that the non-uniform sequencing depth does lead to an overestimated distribution at various β values, especially at $\beta = 0$ and $\beta = 1$, even though a binomial distribution is assumed. Therefore, it is important to eliminate the disadvantage of uneven sequencing depth. For comparison, we show that the methylation level distribution is not biased toward for $\beta = 0$ or $\beta = 1$ when a uniform sequencing depth is used for each CpG site, as seen in Fig 1C.

The methylation variation of tetranucleotides in human brain cells

The average methylation levels of N_5CGN_3 (N_5, N_3 = A, C, G or T) in chromosome 1 of brain samples are shown in S2 Fig. For N_5CGC (N_5 = A, C, G or T), the average methylation levels of ACGC and TCGC are higher than that of GCGC and CCGC. This ranking order remains for N_5CGG (N_5 = A, C, G or T) while N_5CGA and N_5CGT (N_5 = A, C, G or T) have different average methylation level ranking orders. Therefore, the average methylation level of N_5CGN_3 (N_5, N_3 = A, C, G or T) does not show a consensus dependence on the flanking bases.

We then performed a detailed analysis of $KL(0)$ and $KL(1)$ (Materials and methods, Eqs 7 and 8) to investigate how the methylation/demethylation pattern of a particular CpG site is conserved and whether a simple flanking sequence dependence can be identified. We characterized the difference between observed and randomized methylation level distribution functions for the 16 tetranucleotides in chromosome 1 of human brain samples using the relative

entropy approach. The calculated results (Figs 2 and 3) exhibit a clear trend that both $KL(0)$ and $KL(1)$ of $GCGN_3$ and $CCGN_3$ is greater than that of $TCGN_3$ and $ACGN_3$, independent of the N_3 base, indicating that the CpG methylation of $GCGN_3$ and $CCGN_3$ is more conserved than that of $ACGN_3$ or $TCGN_3$. The results also make clear that the 5' but not 3' base of the CpG affects the methylation variation of the CpG in a consensus way (Fig 4). In the following text, we focus on the N_5 base and examine the methylation in the N_5CG trinucleotides. For comparison, we also calculated the direct deviation of $r_{obs}(\beta)$ and $r_{ran}(\beta)$, which is denoted as $d(\beta)$ ($\beta = 0, 1$) (S3 Fig). The $d(0)$ and $d(1)$ show a similar trend as $KL(0)$ and $KL(1)$.

In addition, we calculated the $KL(0)$ and $KL(1)$ of trinucleotide N_5CG of other 23 chromosomes in human brain samples. The results (S4 and S5 Figs) shows that the pattern is conserved among all of the 22 autosomes and the 2 allosomes, indicating that the pattern is not chromosome specific.

The methylation and demethylation variation of trinucleotides are conserved among different tissues and species

As trinucleotides in human brain samples show a conserved pattern that TCG and ACG are more variable in methylation level than GCG and CCG for all chromosomes, we wonder if this pattern is conserved in other tissues and species. Firstly, in 8 human somatic cells (including lung, gastric, spleen *et al.*), the $KL(0)$ and $KL(1)$ of GCG and CCG is also greater than that of TCG and ACG, indicating lower methylation variation of GCG and CCG. Next, in human ESCs and iPSCs the trend of $KL(0)$ and $KL(1)$ that GCG, CCG > ACG, TCG is also conserved. Next we investigate whether the methylation variation is conserved during early embryonic development and analyzed samples from human PGCs and SOMAs. Human PGCs are hypomethylated due to the large scale of demethylation during the embryogenesis. 12 different stages or gender of human PGCs and 8 SOMAs were analyzed. Although the methylation levels of these human samples vary dramatically, we found that all $KL(0)$ and $KL(1)$ values follow the same order of GCG, CCG > TCG, ACG. Finally, to examine whether the simple trend found in these analyses persists across different mammalian species, we also analyzed methylome data of the mouse brain cells. The $KL(0)$ and $KL(1)$ of all analyzed samples are shown in Figs 5 and 6. The ranking order, GCG, CCG > TCG, ACG, for $KL(0)$ and $KL(1)$ are strictly followed by these cells.

In summary, the sequence dependence of CpG methylation variability is conserved across different tissues, developmental stages, ages, genders and species, which indicates that the variations in methylation and demethylation of GCG and CCG are lower than that of ACG and TCG.

The methylation variation in partially methylated domains and the effect of sequencing depth on methylation variation

Partially methylated domains (PMDs) are the hypomethylated regions in specific cells such as IMR90 cell lines, human placenta and certain cancer cells and gene expression in PMDs are repressed [6, 21, 22]. We calculated the $KL(0)$ and $KL(1)$ of PMDs in the IMR90 cell lines (Fig 7) (PMD regions are from Lister *et al.* [6].). For $KL(1)$, the methylation variation of GCG and CCG is smaller than that of ACG and TCG, consistent with the results obtained for the whole chromosome. However, the $KL(0)$ of N_5CGA and N_5CGT shows an opposite trend. These orders of methylation variability order are conserved among all 23 chromosomes of IMR90 cell lines (S6 and S7 Figs). We speculate that the higher methylation variability ($KL(0)$) of $GCGN_3$ and $CCGN_3$ (N_3 = A,T) in PMDs may reflect that these tetranucleotides are selectively unmethylated (S8 Fig).

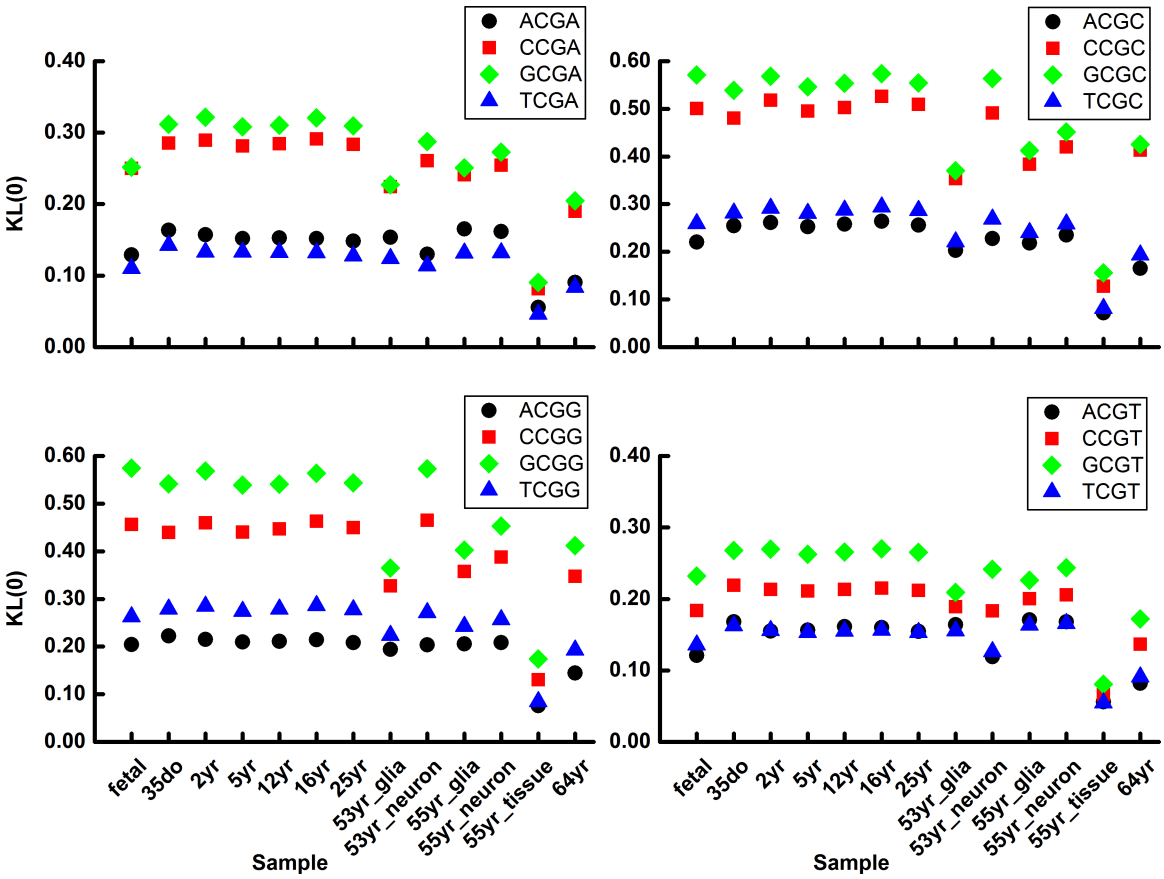

Fig 2. Methylation variation of N_5CGA (upper left), N_5CGC (upper right), N_5CGG (lower left) and N_5CGT (lower right) of chromosome 1 in human brain cells. The *KL*(0) of $ACGN_3$, $CCGN_3$, $GCGN_3$ and $TCGN_3$ are represented as black circle, red square, green diamond and blue triangle, respectively. N_5, N_3 = A, C, G or T.

https://doi.org/10.1371/journal.pone.0186559.g002

To further investigate the effect of sequencing depth to the methylation and demethylation variation, we calculated the methylation and demethylation variation of different sequencing depth. In Fig 8, we show the *KL*(0) and *KL*(1) of the sequencing depth 4x, 6x and 8x. The *KL*(0) and *KL*(1) increases with the larger sequencing depth, but the methylation and demethylation trend GCG, CCG > TCG, ACG is also conserved.

Discussion

The current study shows that the variability of DNA methylation level of a particular CpG site is noticeably affected by its flanking bases. In particular, since the methylation and demethylation processes are mediated by homologous DNMTs [23, 24] and TETs [25, 26] in both human and mouse, the identical ranking order of variability may indicate similar molecular mechanisms used by the two different species. It is therefore interesting to look into how methylation and demethylation are affected by the sequence-dependent intrinsic DNA structural properties. We therefore performed MD simulations to examine how flanking bases affect the local structural properties of the CpG step (S1 Text).

We first note that a number of crystal structures of DNMTs and TETs have been reported and helped illuminate the molecular mechanisms of methylation and demethylation in mammalians [23–25, 27–29]. Both DNMTs and TETs make use of the base flipping mechanism in

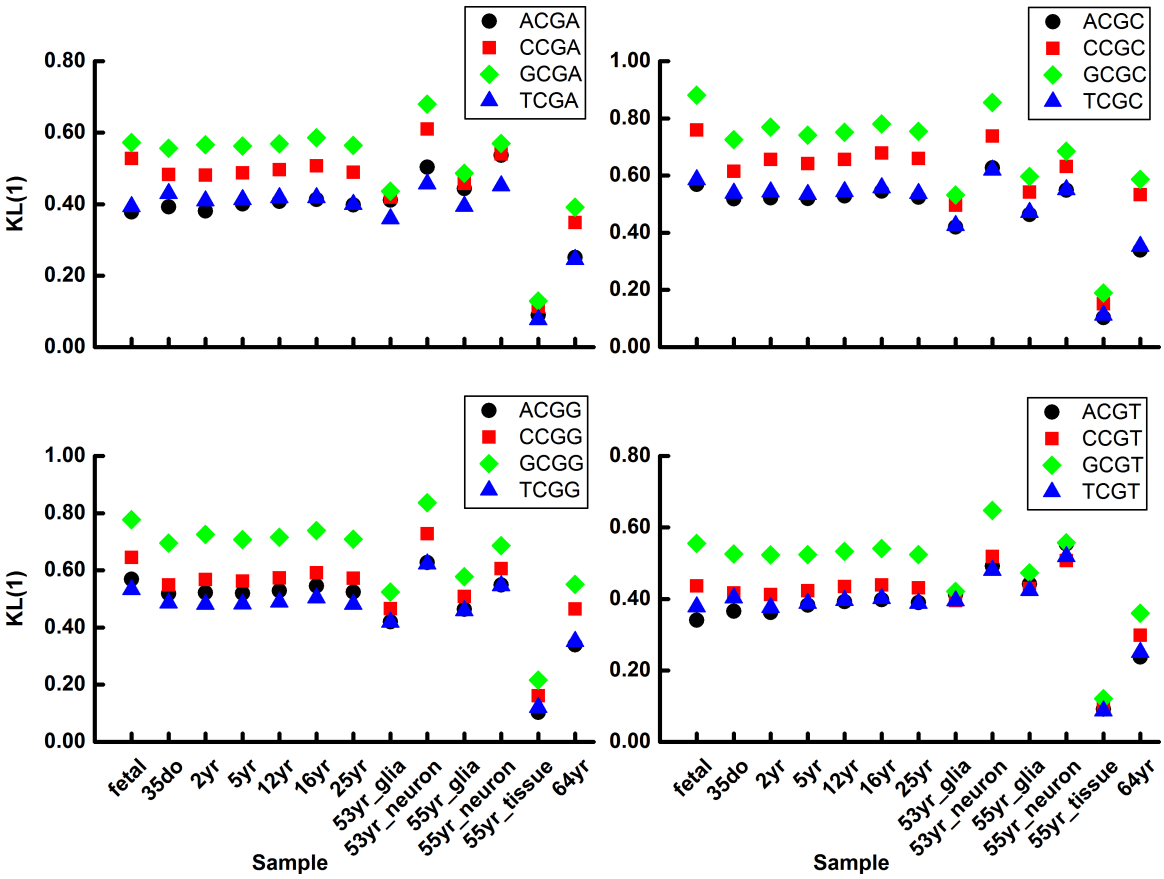

Fig 3. Demethylation variation of N_5CGA (upper left), N_5CGC (upper right), N_5CGG (lower left) and (D) N_5CGT (lower right) in human brain cells.

https://doi.org/10.1371/journal.pone.0186559.g003

order to modify the target cytosine/methylcytiosine. Several steps involve in the base flipping mechanism: first, the target sites are recognized by the recognition domain of protein; the target cytosine or methylcytosine flips out via their interactions with the catalytic domain of the protein; and finally the target is modified in the active pocket[30–32]. Since the substrates of modification are cytosine for methylation and methylcytosine for demethylation, the chemical reactions of methylation (or demethylation) in the catalytic pocket are the same among N_5CGN_3 (or N_5mCGN_3), leaving the base flipping process a possible cause of the flanking base dependence. Since the target cytosine (or methylcytosine) mainly interacts with the DNMTs (or TETs) from the DNA minor groove[23, 29], it is expected that an accessible minor groove would facilitate the base flipping. In the B form DNA structure, the atoms of cytosine/methylcytosine can be classified as the "minor groove atoms" or "major groove atoms" based on the groove they face. We compared the probability distributions of Solvent Accessible Surface Area (SASA) of cytosine in N_5CGN_3 (S9 Fig) and methylcytosine in N_5mCGN_3 (S10 Fig) of the different DNA sequences. The SASA describes the surface area of a molecule which is accessible to the solvent. The SASA in this paper were calculated using the Naccess program (http://www.bioinf.man.ac.uk/naccess/). As seen in S6 and S7 Figs, the average SASA of minor groove atoms in $GXGN_3$ and $CXGN_3$ (X = C or mC) are in general smaller than those in $AXGN_3$ and $TXGN_3$ (X = C or mC) for both cytosine or the methylcytosine, indicating that the target cytosine/methylcytosine in the former two types of tetranucleotides are less

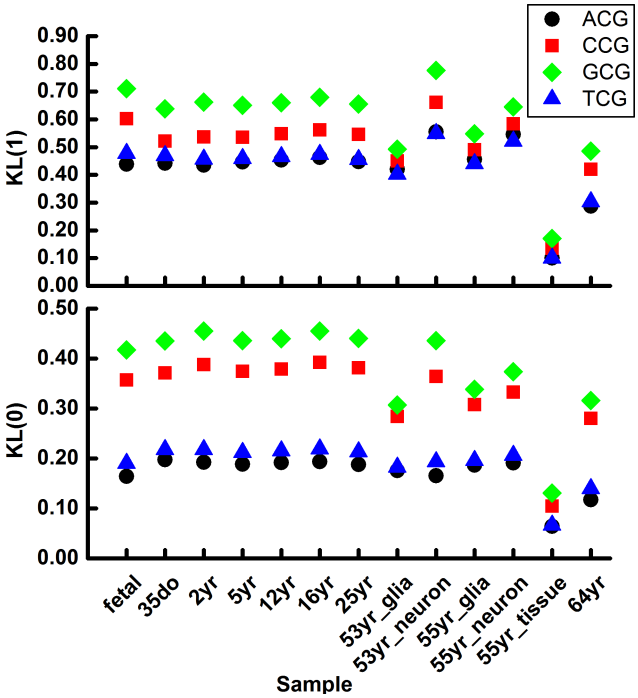

Fig 4. Methylation variation (above) and demethylation variation (below) of trinucleotide N_5CG of chromosome 1 in human brain samples.

https://doi.org/10.1371/journal.pone.0186559.g004

accessible to the proteins, and thus a less favored DNA/protein interaction, than that in the latter two. The SASA distributions of O2 atom of C/5mC are shown in S9 and S10 Figs,

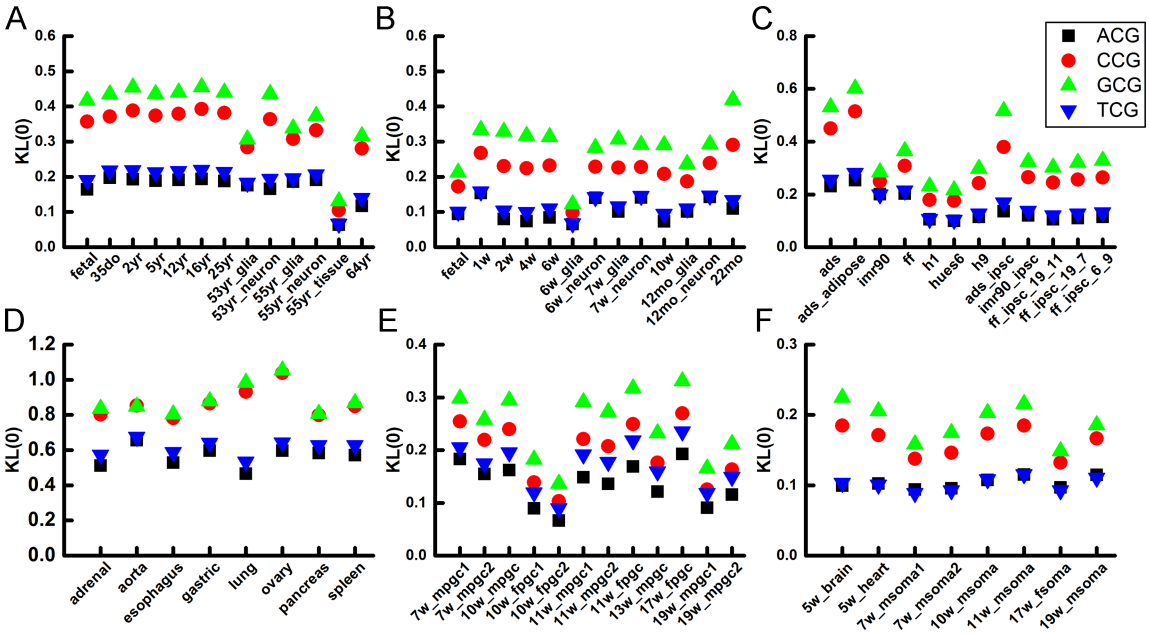

Fig 5. Methylation variation of N_5CG in human and mouse cells. (A) Human brain samples. **(B)** Mouse brain cells. **(C)** Human ESCs and iPSCs. **(D)** Human normal somatic cells. **(E)** Human PGCs. **(F)** Human gonadal somatic cells (SOMAs).

https://doi.org/10.1371/journal.pone.0186559.g005

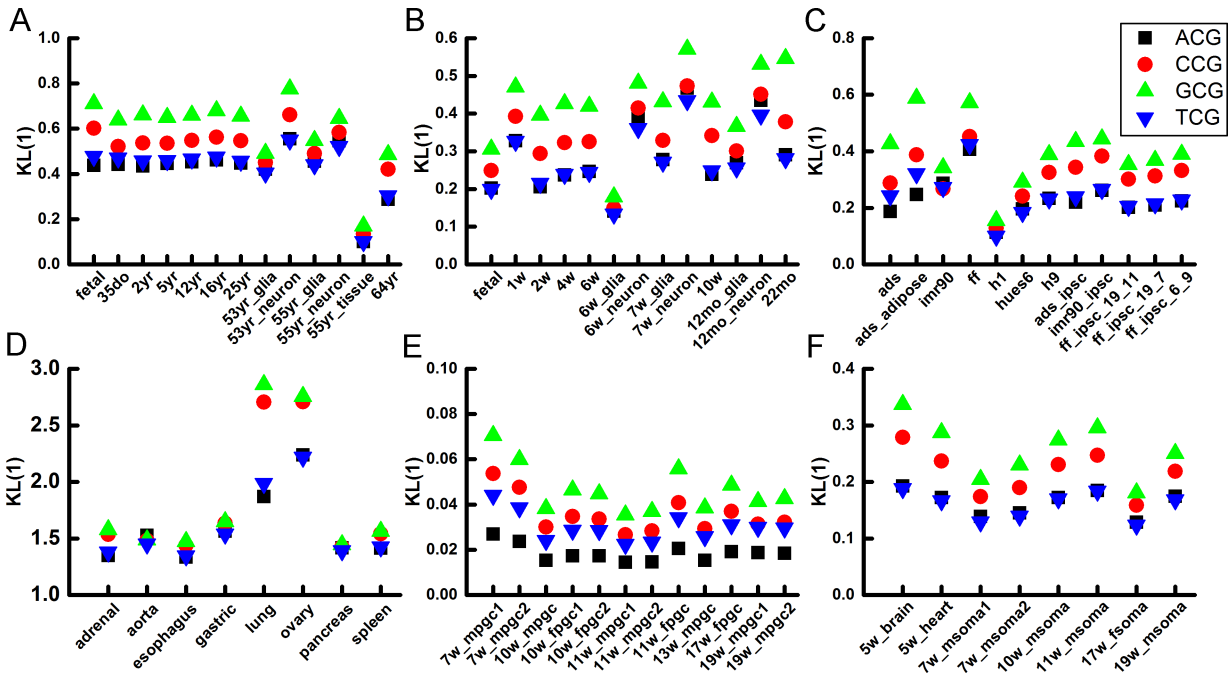

Fig 6. Demethylation variation of N_5CG in human and mouse cells. (A) Human brain samples. **(B)** Mouse brain cells. **(C)** Human ESCs and iPSCs. **(D)**Human normal somatic cells. **(E)** Human PGCs. **(F)** Human gonadal somatic cells (SOMAs).

https://doi.org/10.1371/journal.pone.0186559.g006

respectively. It can be clearly seen from these figures that $AXGN_3$ and $TXGN_3$ are characterized by smaller O2 SASA values than $GXGN_3$ and $CXGN_3$. These results are consistent with $AXGN_3$ and $TXGN_3$ being more prone to form pre-flipping states than $GXGN_3$ and $CXGN_3$.

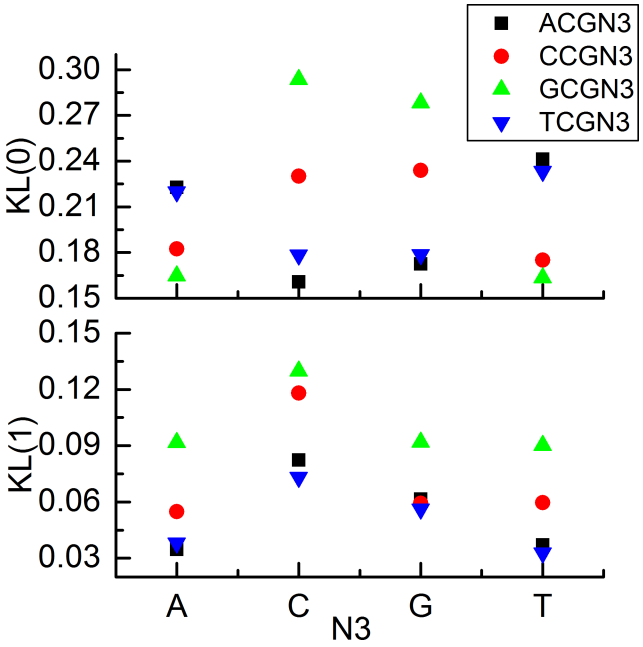

Fig 7. Methylation variation (above) and demethylation variation (below) in PMDs (chr1 of IMR90 cell lines).

https://doi.org/10.1371/journal.pone.0186559.g007

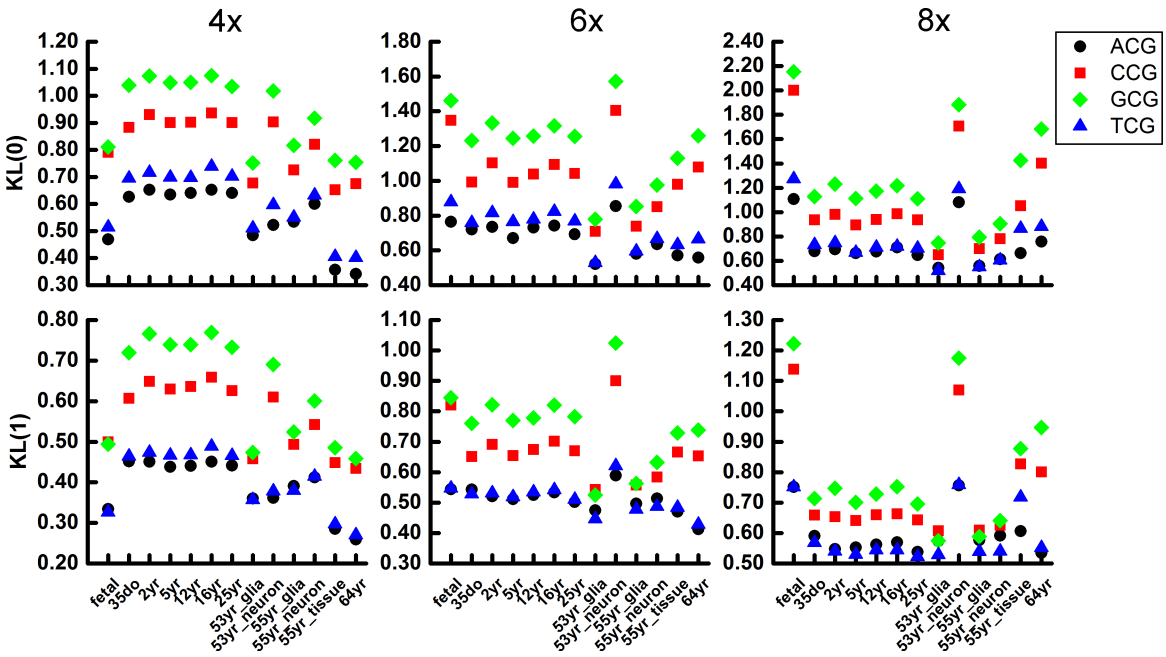

Fig 8. Methylation variation (above) and demethylation variation (below) in human brain cells of difference sequencing depth.

https://doi.org/10.1371/journal.pone.0186559.g008

Generally, the most stable hydrogen bond is embedded inside the base pair, and the least stable one is exposed to the major groove. By comparing the stabilities of hydrogen bonds between O2 of C/5mC-N2 of G, which directly contacts with the active-site loop, we found that the formation probability of hydrogen bonds between O2 of C/5mC-N2 of G for $GXGN_3$ and $CXGN_3$ are typically lower than that of $AXGN_3$ and $TXGN_3$, suggesting that the hydrogen bonds O2 of C/5mC-N2 of G in $AXGN_3$ and $TXGN_3$ are normally less stable than that in $GXGN_3$ and $CXGN_3$. Such a result is again in accordance with the higher variation of $AXGN_3$ and $TXGN_3$ than $GXGN_3$ and $CXGN_3$.

DNMT3A/3B is anchored to chromatins through the co-factor DNMT3L, and TETs interact with chromatins by the CXXC domain at the N terminal. DNA wrapped on histones adopts bending conformations, implying that bending may be beneficial to base flipping, which is also suggested by the bending DNA conformations in the ngTET1-DNA and hTET2-DNA crystal structures [25, 26] as well as the two DNA segments from the HhaI-DNA complex superimposed onto the DNMT3A-DNMT3L tetramer [23, 33]. The bending magnitude is usually adopted to describe the degree of DNA bending, which is calculated with roll and tilt angles [34]. We calculated the distributions of bending magnitudes around the target CpG/5mCpG sites in N_5XGN_3 (S11 and S12 Figs), with roll and tilt angles obtained by 3DNA program [35] and found that in general $AXGN_3$ and $TXGN_3$ have higher average bending magnitudes than $GXGN_3$ and $CXGN_3$.

These results all indicate that compared to $GXGN_3$ and $CXGN_3$, $AXGN_3$ and $TXGN_3$ adopt conformations that are likely more susceptible to interact with the environment, consistent with the latter two having higher methylation and demethylation variability. In summary, the structural properties of different sequences are consistent with their differences on the bimodality of methylation levels on the genome scale, providing a possible explanation to our sequencing data analysis.

Besides the relation to the structural properties, the methylation variation among different tetranucleotides also provides possible functional implications. Gene regulation is one of the most fundamental issues in understanding biological process such as cell differentiation, tumorigenesis and embryonic development, and recent studies provided important hints on the relationship between DNA methylation and gene expression [18]. The most prevalent paradigm of the regulation of gene expression levels through DNA methylation is that the hypermethylation of gene promoters or CpG islands relates to the repression of gene expression. But recent research reveals that methylation of distal regulatory sites (such as enhancers) also affects the gene expression, especially in transformed cells [36, 37]. Our result indicates that with CCG and GCG the methylation level is better maintained, which is expected to be important for the maintenance of biological properties including gene expression level. Accordingly, the promotors, the methylation of which affects significantly gene expression, is richer in CCG/GCG than ACG/TCG. For example, in promoters of chromosome 1 in the 12yr brain sample, the numbers of CCG and GCG trinucleotides are 63242 and 60158, respectively, whereas those of ACG and TCG are 32034 and 32294, respectively. Meanwhile, the average methylation level of CCG (0.263) and GCG (0.258) are also significantly lower than those of ACG (0.37) and TCG (0.359), suggesting their differences in regulating gene expression through methylation. As the methylation state of promoters in a specific cell type is relatively stable, the higher abundance and lower methylation level of CCG and GCG in promoter regions agree with its relatively stable methylation state.

Nevertheless, the maintenance of DNA methylation pattern relies on the dynamic balance between methylation and demethylation and other processes, and gene expression is a complex event involving hierarchical regulatory mechanisms. For example, the sequence of genes and their regulatory elements such as promoters and enhancers, the epigenetic modifications including DNA methylation and histone modifications, and the three-dimentional chromatin structure can all regulate the temporal and spatial specific expression of genes. We believe that DNA sequence is the infrastructure of gene regulatory and the different methylation variation of tetranucleotides N_5CGN_3 illuminated in this work may provide further useful information in the relation among DNA sequence and structure, DNA methylation and gene expression regulation.

Conclusions

We show in this study that the variability of the CpG methylation level is significantly affected by the bases flanking the CpG base step. We analyzed the CpG methylation level distribution and especially the observed bimodality. For tetranucleotides N_5CGN_3, the methylation variation of $GCGN_3$ and $CCGN_3$ are less pronounced than that of $ACGN_3$ and $TCGN_3$, suggesting $GCGN_3$ and $CCGN_3$ tend to be more conserved in cytosine methylation and demethylation. This flanking base dependence of CpG methylation variability is conserved among different cells, tissues and species, strongly suggesting a common mechanism of methylation and demethylation, which are mediated by DNMTs and TETs in mammalian, respectively. In summary, a quantitative description of the bimodal methylation level distribution and its sequence dependence were provided by the analyses of a large amount of methylomes, which provides implications to connect the sequence dependent methylation conservation and the DNA local structure.

Supporting information

S1 Text. Simulation details of molecular dynamics.
(PDF)

S1 Fig. The observed and random distribution of CpG methylation in different sequencing depth.
(PDF)

S2 Fig. Average methylation level of N_5CGA, N_5CGC, N_5CGG and N_5CGT in human brain cells.
(PDF)

S3 Fig. Comparison between the $d(0)$, $KL(0)$ and $d(1)$, $KL(1)$ in the chromosome 1 of human brain samples.
(PDF)

S4 Fig. The methylation variation of all chromosomes in human brain samples.
(PDF)

S5 Fig. The demethylation variation of trinucleotides in all chromosomes in human brain samples.
(PDF)

S6 Fig. The methylation variation of tetranucleotide in PMDs of all chromosomes in IMR90 cell line.
(PDF)

S7 Fig. The demethylation variation of tetranucleotide in PMDs of all chromosomes in IMR90 cell line.
(PDF)

S8 Fig. The average methylation level of the tetranucleotides of chromosome 1 in IMR90 cell lines.
(PDF)

S9 Fig. The distributions of SASA of O2 for (A) N_5CGA, (B) N_5CGC, (C) N_5CGG and (D) N_5CGT.
(PDF)

S10 Fig. The distributions of SASA of O2 for (A) N_5mCGA, (B) N_5mCGC, (C) N_5mGG and (D) N_5mCGT.
(PDF)

S11 Fig. The distribution of bending magnitudes of CpG sites relate to (A) N_5CGA, (B) N_5CGC, (C) N_5CGG and (D) N_5CGT.
(PDF)

S12 Fig. The distribution of bending magnitudes of 5mCpG sites relate to (A) N_5mCGA, (B) N_5mCGC, (C) $N_5$5mCGG and (D) N_5mCGT.
(PDF)

S1 Table. The detailed information of human brain samples.
(PDF)

S2 Table. The detailed information of mouse brain samples.
(PDF)

S3 Table. The detailed information of human ESCs and iPSCs samples.
(PDF)

S4 Table. The detailed information of human somatic cell samples.
(PDF)

S5 Table. The detailed information of human PGC samples.
(PDF)

S6 Table. The detailed information of human gonadal somatic (SOMA) samples.
(PDF)

Author Contributions

Conceptualization: Yi Qin Gao.

Data curation: Ling Zhang, Chan Gu.

Formal analysis: Ling Zhang, Chan Gu.

Funding acquisition: Yi Qin Gao.

Methodology: Yi Qin Gao.

Project administration: Yi Qin Gao.

Resources: Lijiang Yang, Fuchou Tang.

Software: Ling Zhang, Lijiang Yang.

Supervision: Yi Qin Gao.

Writing – original draft: Ling Zhang, Chan Gu.

Writing – review & editing: Ling Zhang, Yi Qin Gao.

References

1. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003; 33:245–54. https://doi.org/10.1038/ng1089 PMID: 12610534
2. Jenuwein T, Allis CD. Translating the histone code. Science. 2001; 293(5532):1074–80. https://doi.org/10.1126/science.1063127 PMID: 11498575
3. Lister R, Ecker JR. Finding the fifth base: genome-wide sequencing of cytosine methylation. Genome Res. 2009; 19(6):959–66. https://doi.org/10.1101/gr.083451.108 PMID: 19273618
4. Bestor TH, Bourc'his D. Transposon silencing and imprint establishment in mammalian germ cells. Cold Spring Harb Symp Quant Biol. 2004; 69:381–7. https://doi.org/10.1101/sqb.2004.69.381 PMID: 16117671
5. Klose RJ, Bird AP. Genomic DNA methylation: the mark and its mediators. Trends Biochem Sci. 2006; 31(2):89–97. https://doi.org/10.1016/j.tibs.2005.12.008 PMID: 16403636
6. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009; 462(7271):315–22. https://doi.org/10.1038/nature08514 PMID: 19829295
7. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 2002; 16(1):6–21. https://doi.org/10.1101/gad.947102 PMID: 11782440
8. Ehrlich M, Gamasosa MA, Huang LH, Midgett RM, Kuo KC, McCune RA, et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic Acids Res. 1982; 10 (8):2709–21. https://doi.org/10.1093/nar/10.8.2709 PMID: 7079182
9. Wang Y, Wang X, Lee TH, Mansoor S, Paterson AH. Gene body methylation shows distinct patterns associated with different gene origins and duplication modes and has a heterogeneous relationship with gene expression in Oryza sativa (rice). New Phytol. 2013; 198(1):274–83. https://doi.org/10.1111/nph.12137 PMID: 23356482
10. Falckenhayn C, Boerjan B, Raddatz G, Frohme M, Schoofs L, Lyko F. Characterization of genome methylation patterns in the desert locust Schistocerca gregaria. J Exp Biol. 2013; 216(Pt 8):1423–9. https://doi.org/10.1242/jeb.080754 PMID: 23264491

11. Straussman R, Nejman D, Roberts D, Steinfeld I, Blum B, Benvenisty N, et al. Developmental programming of CpG island methylation profiles in the human genome. Nat Struct Mol Biol. 2009; 16(5):564–71. https://doi.org/10.1038/nsmb.1594 PMID: 19377480

12. Cedar H, Bergman Y. Programming of DNA methylation patterns. Annu Rev Biochem. 2012; 81:97–117. https://doi.org/10.1146/annurev-biochem-052610-091920 PMID: 22404632

13. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008; 133(3):523–36. https://doi.org/10.1016/j.cell.2008.03.029 PMID: 18423832

14. Lister R, Pelizzola M, Kida YS, Hawkins RD, Nery JR, Hon G, et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature. 2011; 471(7336):68–73. https://doi.org/10.1038/nature09798 PMID: 21289626

15. Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, et al. Global epigenomic reconfiguration during mammalian brain development. Science. 2013; 341(6146):1237905. https://doi.org/10.1126/science.1237905 PMID: 23828890

16. Guo F, Yan L, Guo H, Li L, Hu B, Zhao Y, et al. The transcriptome and DNA methylome landscapes of human primordial germ cells. Cell. 2015; 161(6):1437–52. https://doi.org/10.1016/j.cell.2015.05.015 PMID: 26046443

17. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015; 523(7559):212–6. https://doi.org/10.1038/nature14465 PMID: 26030523

18. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518(7539):317–30. https://doi.org/10.1038/nature14248 PMID: 25693563

19. Ziller MJ, Gu H, Muller F, Donaghey J, Tsai LT, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013; 500(7463):477–81. https://doi.org/10.1038/nature12433 PMID: 23925113

20. Kullback S. Information theory and statistics: Courier Corporation; 1968.

21. Schroeder DI, Blair JD, Lott P, Yu HO, Hong D, Crary F, et al. The human placenta methylome. Proc Natl Acad Sci U S A. 2013; 110(15):6037–42. https://doi.org/10.1073/pnas.1215145110 PMID: 23530188

22. Berman BP, Weisenberger DJ, Aman JF, Hinoue T, Ramjan Z, Liu Y, et al. Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains. Nat Genet. 2012; 44(1):40–6. https://doi.org/10.1038/ng.969 PMID: 22120008

23. Jia D, Jurkowska RZ, Zhang X, Jeltsch A, Cheng X. Structure of DNMT3A bound to DNMT3L suggests a model for de novo DNA methylation. Nature. 2007; 449(7159):248–51. https://doi.org/10.1038/nature06146 PMID: 17713477

24. Ooi SK, Qiu C, Bernstein E, Li K, Jia D, Yang Z, et al. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature. 2007; 448(7154):714–7. https://doi.org/10.1038/nature05987 PMID: 17687327

25. Hu L, Li Z, Cheng J, Rao Q, Gong W, Liu M, et al. Crystal structure of TET2-DNA complex: Insight into TET-mediated 5mC oxidation. Cell. 2013; 155(7):1545–55. https://doi.org/10.1016/j.cell.2013.11.020 PMID: 24315485

26. Hu L, Lu J, Cheng J, Rao Q, Li Z, Hou H, et al. Structural insight into substrate preference for TET-mediated oxidation. Nature. 2015; 527(7576):118–22. https://doi.org/10.1038/nature15713 http://www.nature.com/nature/journal/v527/n7576/abs/nature15713.html#supplementary-information. PMID: 26524525

27. Song JK, Rechkoblit O, Bestor TH, Patel DJ. Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. Science. 2011; 331(6020):1036–40. https://doi.org/10.1126/science.1195380 PMID: 21163962

28. Song JK, Teplova M, Ishibe-Murakami S, Patel DJ. Structure-based mechanistic insights into DNMT1-mediated maintenance DNA methylation. Science. 2012; 335(6069):709–12. https://doi.org/10.1126/science.1214453 PMID: 22323818

29. Hashimoto H, Pais JE, Zhang X, Saleh L, Fu ZQ, Dai N, et al. Structure of a Naegleria TET-like dioxygenase in complex with 5-methylcytosine DNA. Nature. 2014; 506(7488):391–5. https://doi.org/10.1038/nature12905 PMID: 24390346

30. Klimasauskas S, Kumar S, Roberts RJ, Cheng XD. HhaI methyltransferase flips its target base out of the DNA helix. Cell. 1994; 76(2):357–69. https://doi.org/10.1016/0092-8674(94)90342-5 PMID: 8293469

31. Matje DM, Coughlin DF, Connolly BA, Dahlquist FW, Reich NO. Determinants of precatalytic conformational transitions in the DNA cytosine methyltransferase M.HhaI. Biochemistry. 2011; 50(9):1465–73. https://doi.org/10.1021/bi101446g PMID: 21229971

32. Jin L, Ye F, Zhao D, Chen S, Zhu K, Zheng M, et al. Metadynamics simulation study on the conformational transformation of HhaI methyltransferase: An induced-fit base-flipping hypothesis. Biomed Res Int. 2014; 2014:1–13. https://doi.org/10.1155/2014/304563 PMID: 25045662

33. Cheng XD, Blumenthal RM. Mammalian DNA methyltransferases: A structural perspective. Structure. 2008; 16(3):341–50. https://doi.org/10.1016/j.str.2008.01.004 PMID: 18334209

34. Young MA, Ravishanker G, Beveridge DL. A 5-nanosecond molecular dynamics trajectory for B-DNA: analysis of structure, motions, and solvation. Biophys J. 1997; 73(5):2313–36. https://doi.org/10.1016/S0006-3495(97)78263-8 PMID: 9370428

35. Lu XJ, Olson WK. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat Protoc. 2008; 3(7):1213–27. https://doi.org/10.1038/nprot.2008.104 PMID: 18600227

36. Sur I, Taipale J. The role of enhancers in cancer. Nat Rev Cancer. 2016; 16(8):483–93. https://doi.org/10.1038/nrc.2016.62 PMID: 27364481

37. Kulis M, Heath S, Bibikova M, Queiros AC, Navarro A, Clot G, et al. Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. Nat Genet. 2012; 44(11):1236–42. http://www.nature.com/ng/journal/v44/n11/abs/ng.2443.html#supplementary-information. https://doi.org/10.1038/ng.2443 PMID: 23064414