博士論文

Functional analysis of non-CpG methylation in mammalian cells via integrative approach

（哺乳動物細胞の非 CpG メチル化に関する統合的アプローチによる機能解析）

Jong-Hun Lee

李鍾勳

Laboratory of functional analysis *in Silico*

Department of Computational biology and medical science,

Graduate school of Frontier Science,

The University of Tokyo

Contents

Abstract

DNA methylation, an addition of methyl group on fifth carbon at cytosine, is one of the most important epigenetic modifications. In mammalian cells, the methylation preferentially occurs at CpG dinucleotides, regulating cell development and maintenance. Recently, however, methylated non-CpGs (mCpHs; H means A, C, and T) are spotlighted as a key epigenetic mark regardless of the small amount of it. Those are especially abundant in pluripotent stem cells (PSCs) and non-dividing cells (i.e. neuron), and associated to cell type specific progresses such as embryonic stem cell differentiation and brain maturation.

The research of mCpHs, however, is limited by lack of the detection techniques, since existing experiments mostly focuses on CpG sites. Although the whole genome bisulfite sequencing (WGBS), the advanced sequencing technique combined with bisulfite treatment, could capture genome-wide methylation pattern with single nucleotide resolution, the low accuracy hinders it from being widely used. Many computer programs have been developed to overcome the shortcomings. However, those merely succeeded in improving either quantity or quality of WGBS data.

The other limitation is the correlation between methylated CpGs (mCpGs) and mCpHs. Since those are mediated by common enzymes, DNA methyltransferase 3a and 3b (DNMT3a and DNMT3b, respectively), the genome wide distribution of those are closely correlated, resulting in difficulty on finding the role of mCpHs that independent to mCpGs. Based on the correlation and higher affinity of DNMTs on CpG sites, some researchers insist that the mCpHs are merely a by-product from hyper-activity of DNMTs that originally targets CpGs. However, substantial evidences support that the mCpHs have independent role on cell functions. In this way, the role of mCpH over biological processes is controversial because of their spatial correlation with mCpGs.

In this study, we attempted to uncover the independent role of mCpHs by resolving the two difficulties. First, we developed an integrative approach for methylation detection. The integrative method effectively increased both accuracy and amount of methylome. In addition, it facilitated combining of public WGBS data by reducing experimental bias (Section 1). Second, through comprehensive analysis on the high-quality methylome, we found that the CpHs ±100 base pair (bp) from CpGs are methylated in highly correlated way to CpGs. Remarkably, the CpH methylation pattern at those CpHs is highly distinct between ESCs and neurons because of the differential activity of DNMT3a and DNMT3b (Section 2). Lastly, by extracting mCpG-independent mCpHs, we found that the CpH methylation pattern across sample ages is closely related to brain specific progresses such as “mental retardation” and “zinc finger protein activity”. Collectively, our results uncovered cell type specific formation and function of mCpHs in mammalian cells by integrating and analyzing public WGBS data. The study shed light on the methylation research by improving quality and quantity of methylome and suggesting cell type specific methylation mechanism via functional analysis *in silico*.

General background

DNA methylation, an addition of methyl groups on 5th carbon at cytosine, is one of the most important epigenetic marks. It regulates gene expression by hindering (or accelerating) interaction between DNA strand and transcription factors (TFs). Unlike genome strand, mostly common among cells in an organism, the DNA methylation pattern over genome is highly distinguishable among cell types. The distinct methylation pattern results in cell type specific processes such as morphogenesis, by regulating the amount of transcriptomes. In addition, irregular methylation patterns lead to generation of abnormal cells such as cancer cells. Thus, the DNA methylation is regarded as an identification of cells governing cell development and maintenance.

In mammalian, the DNA methylation mainly occurs at CpG (Cytosine followed by Guanine) dinucleotides. The average methylation level at CpGs is up to 85% [1], whereas that in non-CpG sites is mostly under 1% [2]. Thus, for decades, researchers have focused on methylated CpGs (mCpGs) and found that those regulate cell differentiation, maintenance, and retardation [3-5]. Recently, however, the methylated non-CpG sites (mCpHs; H means adenine, cytosine, and thymine) are being emerged as key epigenetic regulator of cell type specific functions. The mCpHs are highly abundant in pluripotent stem cells (PSCs) and non-dividing cells (i.e. neuron) [2], and regulates cell type specific functions such as synaptogenesis [6], or embryonic stem cell differentiation [7]. In this way, the mCpHs, as well as mCpGs, are important epigenetic mark on cell processes in mammalian.

The methylation is induced by DNA methyltransferases (DNMTs) in mammalian. Those methyltransferases approach to cytosines on DNA strand and attach methyl-groups donated by S-adenosyl methyionine (SAM) [8]. The DNMTs are largely classified into two types by their functions; maintenance methyltransferase (i.e. DNMT1) and *de novo* methyltransferase (i.e. DNMT3).

The DNMT1 attaches to hemi-methylated strand (meaning only one strand is methylated) and transfers the methyl-groups to the cytosine at the other strand. Thus, the DNMT1 contributes on maintaining methylation pattern across cell division by methylating the newly synthesized DNA strands. The DNA strands interacting with DNMT1 should be symmetric, so that only CpG sites (the opposite side is also CpG) are possible to be methylated by DNMT1. On the other hand, the DNMT3 is mainly responsible for *de novo* methylation. The known members of the DNMT3 family are DNMT3a, DNMT3b, and DNMT3l. The DNMT3a and DNMT3b (DNMT3a/b) directly interact with DNA strands, whereas DNMT3l is known as an allosteric enzyme that co-operate with the DNMT3a/b. Those are able to interact with CpHs, even though the affinity is ten times higher to CpGs than CpHs. Thus, it is known that CpGs are methylated by DNMT1 and DNMT3a/b, whereas CpHs are methylated by DNMT3a/b.

Since both CpGs and CpHs are *de novo*-methylated by DNMT3a/b, the genome-wide distribution of mCpGs and mCpHs is highly analogous to each other [9, 10]. Since the average methylation level is much higher at CpGs, some researchers insist that mCpHs are merely non-targeted by-products of hyper-activity of DNMT3a/b that originally targets CpGs. The fact that mCpH-existing cells show

generally higher expression level of DNMT3a/b also supports the opinion [11].

However, the mCpHs show functional independence to mCpGs in several researches. In brain, for example, the mCpHs are gradually increased as aged, in a same pattern with the progress of synaptogenesis [6]; whereas mCpGs are remained as stable. In addition, the methyl-CpG binding protein 2 (MeCP2), mutation of which causes Rett syndrome, binds to not only mCpGs but also mCpHs. Considering that postnatal onset of Rett syndrome coincides with the emergence of mCpH in brain tissues, there is a possibility that MeCP2-related neuro-diseases are governed not by mCpGs, but by mCpHs [12]. Also, there are mega-base mCpH deserts in induced pluripotent stem cells (iPSCs), in which genes are less transcribed compared to those in ESCs [13]. The CpG methylation level in those regions is not distinguishable between iPSCs and ESCs, implying that the failure of epigenetic reprogramming at CpHs leads to genetic aberration in iPSCs. Altogether, even though the mCpHs are spatially correlated to mCpGs, there is functional independence over mCpGs, especially on regulating cell type specific phenomena.

One of the underlying mechanisms that mCpHs regulate cell type specific progresses is differential distribution over genome among cell types. Remarkably, the distributions of mCpHs in PSCs and neuron, in which mCpHs are mostly abundant, are hugely distinguishable. For example, in PSCs, The mCpHs prefer to be at intragenic regions, whereas those at intergenic regions in neurons [6, 7]. Especially, the mCpHs tend to be abundant at actively expressed gene-bodies in PSCs, whereas those are at poised gene-bodies in neurons, resulting in positive and negative correlation between mCpH-abundance and gene expression level in PSCs and neurons, respectively. In general, the methylation in genic regions tends to repress gene expression, so that the positive correlation between mCpH abundance and gene expression in PSCs has been mysterious among researchers. Meanwhile, the DNA motif that abundant nearby mCpHs is also distinct between PSCs and neurons; that is “CAG” in PSCs, whereas “CAC” in neurons [5, 10]. It implies that the methylation could be induced by at least two mechanisms that function in cell type specific way [2]. Since the abundant DNA motif in other somatic cells is also “CAC”, the motif CAG is considered as PSC-specific mCpH mark. Also, those appears in mixed way in oocyte [14]. In this way, the mCpH shows cell type specific formational and functional mechanisms, and specifics of it are waiting to be uncovered.

However, the research of CpH methylation is largely limited by its detection experiments. Since most of the experiments for detecting methylation is concentrating on CpG sites, large portion of mCpHs are not able to be detected by existing experiments. For example, the bisulfite treated microarrays, such as Infinium Human Methylome 450K, covers only 3000 CpH sites, whereas it detect 450,000 CpG sites [15]. In addition, the Chip-seq based methylation detection methods such as Methylated DNA immunoprecipitation (MeDIP) contains bias toward CpG sites because of the preference of anti-bodies on hyper-methylated cytosines [16]. The reduced representation bisulfite sequencing (RRBS) [17], utilizing enzyme bindings to specific sites to detect information-rich regions, is also not allowed for detecting mCpHs, since the CpHs are widely distributed over whole genome, unlike mCpGs that tend to exist as cluster called CpG Island (CGI),

Thus, for now, whole genome bisulfite sequencing (WGBS) is considered as the only way to detect mCpHs. It detects genome-wide CpH methylation status with single base resolution by utilizing next generation sequencing (NGS) combined with bisulfite treatment. The bisulfite treatment converts un-methylated cytosine into uracil which eventually becomes thymine by Polymerase Chain Reaction (PCR) [18]. Then, the DNA fragments are read by NGS technique and aligned into reference genome. However, problem happens on the aligning step. Since the thymines on fragments have to be aligned into both cytosine and thymine loci, aligning complexity increases, resulting in low accuracy of the methylation detection.

A number of aligning tools have been developed to improve the accuracy. Largely, there are two types of bisulfite-read mappers, wild-card type and three-letter type [19]. The wild-card type mappers employ a new latter, Y, for read aligning. It converts all the cytosines (Cs) in reference genome to Y, and aligns both Cs and thymines (Ts) in sequenced read to the Y. Since both Cs and Ts are aligned to a single latter Y, the mapping accuracy is not greatly improved, but relatively large number of sequenced reads is aligned to reference genome. The representative of wild-card type mapper is BSMAP [20]. It has been used for many epigenetic researches, including the research about mCpHs. However, the low mapping accuracy remains as the weakness of it. On the other hand, the three-letter type mappers convert all Cs to Ts and uses only three letters (A, G and T) for read aligning. Since it uses only three letters, the sequenced reads are frequently aligned to multiple loci. By applying strict policy on multi-aligned reads, the mappers could achieve great mapping accuracy, however, lose large portion of reads (low mapping rate). The representative of three-letter type bisulfite read mapper is Bismark [21]. It has been used most widely in epigenetic research field, but small-size bisulfite read data is not affordable for being analyzed by it because of the low mapping rate. BS-seeker2 was developed to recover the low mapping rate, flaw of Bismark, by applying local alignment on read aligning [22]. However, there was trade-off of the mapping rate against mapping accuracy, resulting in lower mapping accuracy compared to that by Bismark. In this way, even though large number of bisulfite read mappers has been developed (or evolved), they barely succeeded in improving either amount or accuracy of methylation detection.

In this study, we attempted to find cell type specific mCpH distribution by improving both quality and quantity of methylome from WGBS data. To improve the methylation detection, we developed an integrative method that combines the outputs from three most widely used bisulfite read mappers, Bismark, BSMAP, and BS-seeker2. By scoring read depth against artifacts, we succeeded in improving both amount and accuracy of methylation detection. In addition, the integrative method facilitated combining of WGBS samples generated from various experiments by reducing experimental bias caused different read conditions among experiments. In consideration of the difficulties on generating WGBS samples such as costly and time-consuming steps, our integrative approach contributes to methylation research by facilitating re-use of public WGBS data. The specifics about the integrative approach are described in Section 1, with some results of bisulfite read mapper analysis.

Applying the integrative approach, we analyzed the cell type specific mCpH patterns in ESC and neuron. We confirmed that the mCpH are spatially correlated with mCpGs, especially when those are within ±100bp distance. Remarkably, we found that the methylation at CpHpH and CpHpG contexts are preferentially methylated by DNMT3a and DNMT3b, respectively, resulting in distinct characteristics of mCpHs in embryonic stem cells (ESCs) and neuron. In addition, we proved that the positive correlation between mCpH and gene expression level in ESCs is caused by the active interaction between DNMT3b and histone mark, H3k36me3, abundant at highly expressed gene bodies. The specifics of the analysis are written in Section 2.

Lastly, we found that the mCpHs independently formed to mCpGs are highly functional on brain maturation. The genes related to “mental retardation” and “zinc finger activity” are clustered by mCpG-independent mCpH pattern across sample ages. The specifics of the analysis are written in Section 3.

Altogether, the results contribute to DNA methylation research by developing method for accurate methylation detection and uncovering the roles of mCpHs that independent to mCpGs. This research shed light on the mechanism of epigenetic regulation on cell type specific processes.

Section1: An integrative approach for efficient analysis of whole genome bisulfite sequencing data

Abstract

Whole genome bisulfite sequencing (WGBS) is a high-throughput technique for profiling genome-wide DNA methylation at single nucleotide resolution. Especially, it is specified for detecting methylation at non-CpG context (CpH; H means A, C, and T) by scanning whole genome with little bias to CpGs. However, the WGBS is limited by low accuracy that caused by bisulfite-induced damage on DNA fragments. Although many computer programs have been developed for accurate detection, most of the programs have barely succeeded in improving either quantity or quality of the methylation results.

To improve both, we developed an integration method that combines methylomes from most widely used bisulfite-read mappers, Bismark, BSMAP, and BS-seeker2. A comprehensive analysis of the three mappers revealed that the mapping results of the mappers were mutually complementary under diverse read conditions. Therefore, we sought to integrate the characteristics of the mappers by scoring them to gain robustness against artifacts. As a result, the integration significantly increased detection accuracy compared with the individual mappers. In addition, the amount of detected cytosine was higher than that by Bismark. Furthermore, the integration successfully reduced the experimental bias of detection accuracy that induced by read heterogeneity. We applied the integration to real WGBS samples and succeeded in classifying the samples according to the originated tissues by both CpG and CpH methylation patterns.

In this study, we improved both quality and quantity of methylation results from WGBS data by integrating the methylomes of the three bisulfite-read mappers. Also, we facilitated the re-analyzing of public WGBS data by reducing the effects of read heterogeneity on methylation detection. This study contributes to DNA methylation researches by improving efficiency of methylation detection from WGBS data and facilitating the comprehensive analysis of public WGBS data.

Introduction

DNA methylation, defined as an addition of methyl group on 5-carbon in cytosine, is an epigenetic mark on cell processes. The DNA methylation pattern can serve to identify cells and guides cell development and tissue maintenance [4]. For decades, researchers have focused on methylation at CpG sites (mCpG) and found that those are regulating cell-specific functions, aging and deceases [3, 23-25]. Recently, however, methylation at CpH (mCpH; where H can be A, C, or T) sites is being confirmed to be a key regulator of cell-type specific functions. Those are known to be involved with brain development and embryonic stem cell (ESC) differentiation [6, 7, 11, 12]. Therefore, profiling both mCpG and mCpH in a genome scale is crucial for understanding various biological processes.

To analyze the methylation modifications, high-throughput methods coupled with microarray and next-generation sequencing have been widely used. Bisulfite microarray is a specially designed genotyping microarray combined with bisulfite treatment. Although this method is a useful strategy for targeted DNA methylation analyses, it is not suitable for genome-scale studies due to low genome coverage; only 0.8% of CpGs and 0.02% of CpHs have been covered in the newest version [15]. Another experimental method, reduced representation bisulfite sequencing (RRBS) [17], utilizes enzyme bindings to “CCGG” sites to detect information-rich regions. However, the enzyme binding leads experimental bias and limits the detection of mCpH [19]. Alternatively, as the widely-accepted gold standard method, whole genome bisulfite sequencing (WGBS) can detect both mCpG and mCpH at single nucleotide resolution in a genome scale [19].

For efficiently detecting the methylated sites with WGBS data, many computer programs have been developed. In particular, Bismark [21], BSMAP [20], and BS-seeker2 [22] are the most widely used bisulfite-read mappers that employ distinct strategies. BSMAP is a wild-card type mapper that converts all cytosine bases (Cs) of a reference genome to a letter Y and then aligns sequenced Cs and thymine bases (Ts) to the Ys [19] by using SOAP [26]. Bismark and BS-seeker2 are three-letter type mappers that convert all Cs to Ts in both sequenced reads and a reference genome. However, different with Bismark, employing Bowtie2 [27] with global alignment mode, BS-seeker2 employs Bowtie2 with local alignment mode to increase mapping rate (percentage of reads being aligned). It has been reported that wild-card type mapper (i.e. BSMAP) tend to show better mapping rate but struggle with mapping accuracy (percentage of reads mapped at correct positions) [19], whereas the three-letter type mappers (i.e. Bismark and BS-seeker2) show exactly opposite tendency [19]. In this way, even though many bisulfite-read mappers have been developed for better methylation detection, those have succeeded in improving either quantity or quality of it, not both of it.

In this study, we attempted to improve both quantity and quality of methylation detection from WGBS data by developing a novel integrative method. First, we investigated the performances of the three mappers on virtual WGBS dataset that has been simulated under various conditions. Through gathering detailed information, we confirmed that the mappers exhibit (dis)similar behaviors depending on the properties of simulated reads, which is consistent with results from previous

studies [19, 22]. Since the results showed that the behaviors of the three mappers were complementary to each other, we sought to integrate the characteristics of the mappers by scoring them to gain robustness against artifacts (e.g. sequencing errors and aligning errors). As a result, our integrative approach improved both quality (i.e. the accuracy of the methylation detection at each C) and quantity (i.e. the number of detected Cs) of the methylation data with less dependency on the read properties (Figure 1.1). We also applied our approach to public WGBS datasets of 13 tissues, and successfully grouped them according to their originated tissues by the patterns of mCpG and mCpH. Altogether, this study contributes to DNA methylation research by efficiently analyzing the WGBS data and facilitating comprehensive analyses of methylation patterns under the public WGBS data. In addition, this study gives a clue to algorithmic improvement of bisulfite-read mappers.

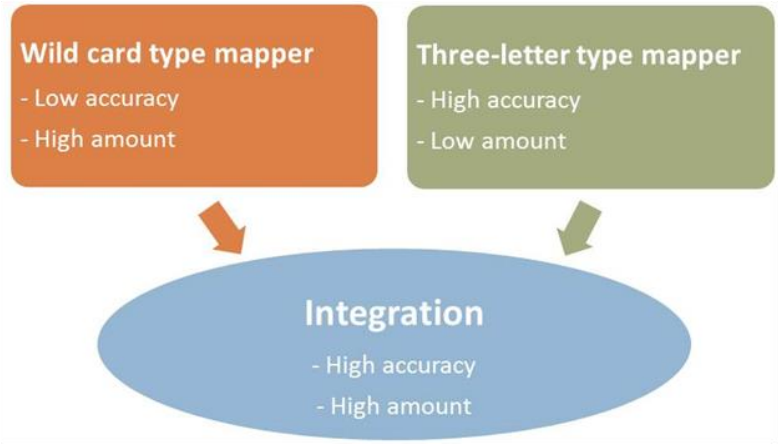


Figure1.1 Concept map for integrative approach.

Results

1. Overview of integrative approach

We integrated the methylation results from three bisulfite-read mappers: Bismark, BSMAP, and BS-seeker2. Bismark and BSMAP are the most widely used three-letter type and wild-card type mappers, respectively [28-30], and BS-seeker2 is the newest three-letter type mapper, which has shown a higher mapping rate than Bismark [22].

The evaluation and integration of Bismark, BSMAP, and BS-seeker2 were conducted as described below (Figure 1.2-a). First, bisulfite-read sets were generated by Sherman [31], with randomly designated methylation levels for every block of 500 base pairs (bps) in human chromosome 19. Then we mapped the reads by Bismark, BSMAP, and BS-seeker2 and evaluated the performances of the three mappers with respect to mapping rate and accuracy; the mapping rate is the portion of mapped read number over total read number, whereas the mapping accuracy is the portion of correctly mapped read number over mapped read number. Lastly, we integrated the methylation results from the mappers with three strategies and evaluated the performances in terms of detection accuracy (d-accuracy) and amount of detected Cs (d-amount). The d-accuracy was determined by the similarity between generated and detected methylation levels at each block (Figure 1.2-b).

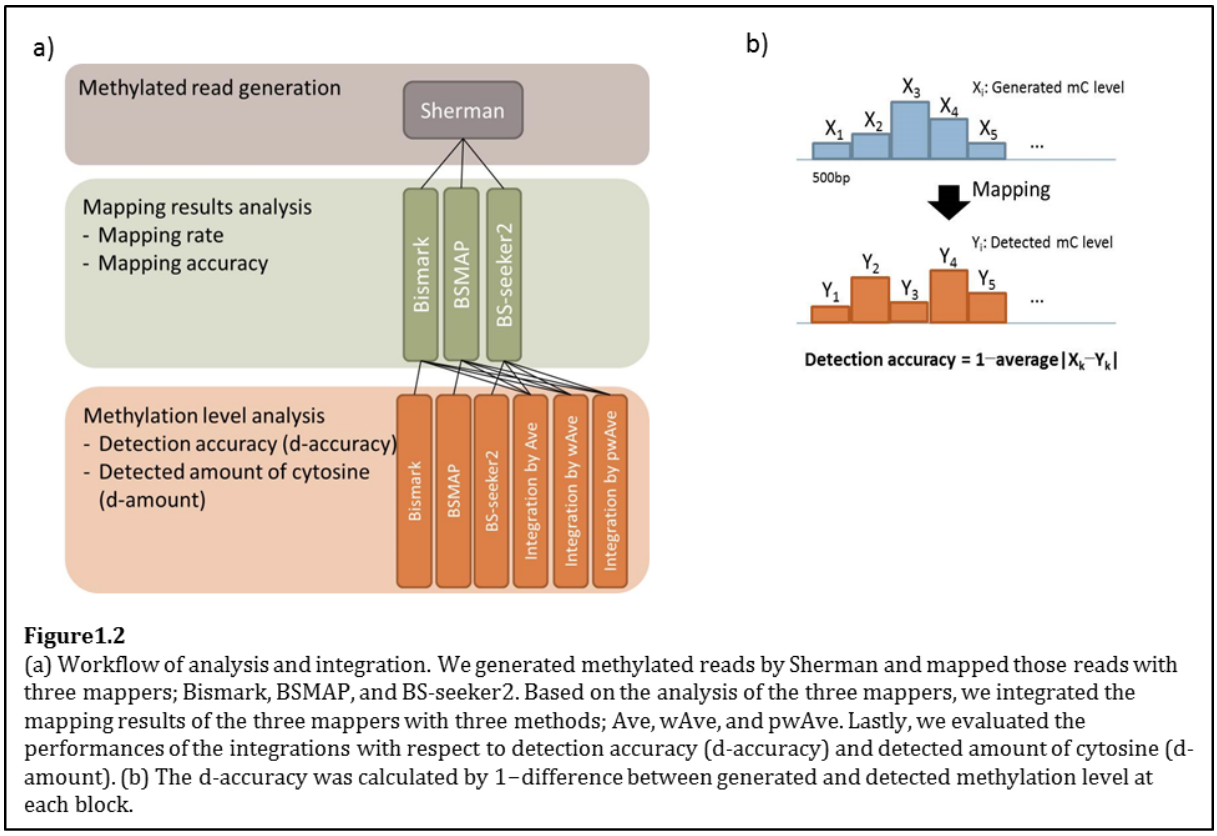


Figure1.2
(a) Workflow of analysis and integration. We generated methylated reads by Sherman and mapped those reads with three mappers; Bismark, BSMAP, and BS-seeker2. Based on the analysis of the three mappers, we integrated the mapping results of the three mappers with three methods; Ave, wAve, and pwAve. Lastly, we evaluated the performances of the integrations with respect to detection accuracy (d-accuracy) and detected amount of cytosine (d-amount). (b) The d-accuracy was calculated by 1−difference between generated and detected methylation level at each block.

2. Read-dependent performances of the three mappers

To investigate the performances of the three mappers under diverse read conditions, we analyzed the mapping results of the three mappers in the context of varied read quality, read length, and methylation levels.

For all three mappers, the mapping rate and mapping accuracy fluctuated with changes in read quality (Figure 1.3-a, b). When reads contained little errors (<4%), BSMAP showed a higher mapping rate and lower mapping accuracy compared with others, consistent with previous studies [19, 22]. As the read error rate increased (6-8%), however, the mapping rate of the BSMAP decreased dramatically, to lower than that of Bismark, implying that the mapping rate of wild card-type mappers could be lower than that of three letter type mappers when the read quality is extremely bed. Interestingly, for BS-seeker2, both mapping rate and mapping accuracy did not decreased substantially, implying that BS-seeker2 shows robustness upon the fluctuation of read quality.

The read length also affected the performances of the three mappers (Supplementary figure 1.1). We compared mapping results of 50bp-long reads (short reads) with those of 100bp-long reads (long reads). When read error rate was low (2%), both mapping rate and mapping accuracy were higher within long reads, which is coincident with previous results [28]. When read error rate was high (8%), however, mapping rate of Bismark and BSMAP were higher within short reads than within long reads implying that the read length and quality is jointly affecting the mapping rate. Remarkably, both mapping rate and mapping accuracy of Bismark was higher in short reads when read quality is bed (8%), implying that Bismark performs great for short read mapping.

In addition, we found that the performances of the three mappers varied greatly within repeat regions. In particular, the reads generated from short interspersed nuclear elements (SINEs) tended to be unmapped by Bismark and BS-seeker2 (Figure 1.3-c) and incorrectly mapped by BSMAP (Figure 1.2-d), which clearly showed the difference in performances between wild-card type and three-letter type mappers.

Lastly, we found that hypo-methylated reads tended to be incorrectly mapped by BSMAP and BS-seeker2 (Figure 1.3-e, f). This tendency was not found in the mapping results of Bismark. This may be explained in part by that the increased number of Ts, induced by the bisulfite conversion of unmethylated Cs, hindered the correct mapping of BSMAP and BS-seeker2. To confirm that, we measured the percentage of Ts in reads that correctly and incorrectly mapped by the three mappers. For BSMAP and BS-seeker2, the incorrectly mapped reads contained higher amount of Ts than the correctly mapped ones (Figure 1.3-g).

In summary, Bismark, BSMAP, and BS-seeker2 performed differently in different read conditions. Bismark mapped reads with great accuracy and was not affected by the density of Ts in reads. However, Bismark tended to lose both mapping rate and accuracy when read error rate was higher in longer reads. BSMAP generally mapped a large number of reads to incorrect positions. Additionally, the mapping accuracy of BSMAP was affected by the density of Ts in reads. Both the mapping rate and mapping accuracy of BS-seeker2 were only slightly affected by the read error rate, whereas the

mapping accuracy was affected by the density of Ts in reads (Table 1.1).

	Mapping rate	Mapping accuracy	Read error dependency	Bisulfite-conversion dependency
Bismark	Low	High	Yes	No
BSMAP	High	Low	Yes	Yes
BS-seeker2	Middle	Middle	No	Yes

Table1.1: Summary of the performances of the tree mappers in varied read conditions

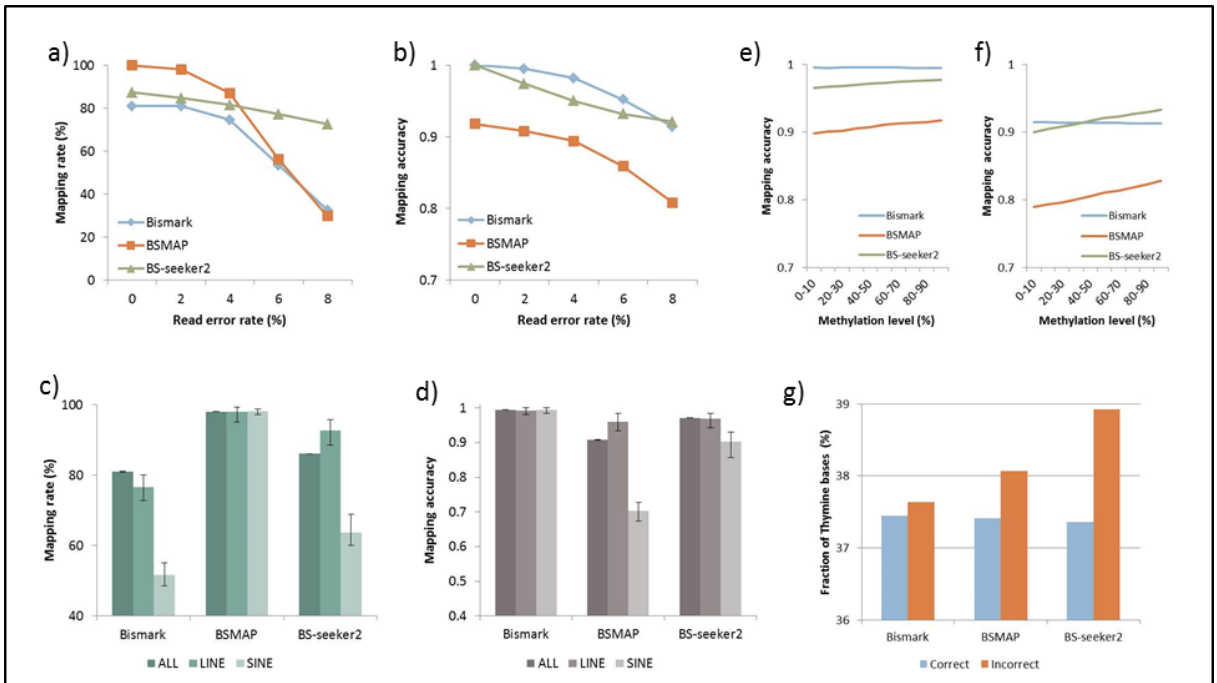


Figure1.3 Mapping results of the three mappers with short reads (length=50bps)
(a) Mapping rate of the three mappers across read error rates. The mapping rate was calculated as mapped read number over total read number. (b) Mapping accuracy of the three mappers across read error rates. The mapping accuracy was calculated as the rate of correctly mapped read number over mapped read number. Mapping rate (c) and mapping accuracy (d) of the three mappers within all of hg19 chr19, LINEs, and SINEs. In this case, the read error rate was 2%. Mapping accuracy of the three mappers with reads in which error rate equals to 2% (e) and 8% (f). X-axis means methylation level in reads set at the bisulfite-read generation step. (g) Percentage of Ts in each read that were (in)correctly mapped by the three mappers.

3. Integrative approach improved both the accuracy and amount of methylation detection

Based on the different performances of Bismark, BSMAP, and BS-seeker2 in varying read conditions, we integrated the mapping results of the three mappers using three strategies: Ave - average of the methylation levels from the three mappers, wAve - weighting by read depths, and pwAve – weighting by probabilistic method (See Method).

First, we examined the overlap of correctly mapped reads by the mappers. We found that 88.6% of high-quality 100-bps reads (2% read error rate) were correctly mapped by all the mappers, but this dramatically decreased to 6.7% in the case of low-quality reads (8% read error rate, Supplementary figure 1.2). It implies the reads that successfully mapped to the same position by three mappers could have high quality and high chance for being mapped to the right position. Indeed, as the number of covering mappers (i.e. n_i in Methods) increased, wAve improved the d-accuracy (Figure 1.4-a). However, taking reads that only mapped by all three mappers made the d-amount dramatically decrease, even becoming far lower than the average of the three mappers. Taking account of this tradeoff, we choose $n_i \geq 2$ that yields constantly higher d-amount than Bismark, and higher d-amount than BS-seeker2 or BSMAP in some cases (Figure 1.4-b).

As shown in Table 1.2, among the three integration methods, wAve marked the highest d-accuracy in most read conditions, whereas pwAve showed the best d-accuracy in limited cases that short reads contain few errors (≤4%). The wAve remarkably improved d-accuracy compared with the individual mappers (Figure 1.4-c, right). The Wilcoxon single-rank test over 500 bps blocks revealed significantly low P value (≤5.0E-2) in most of read conditions (Supplementary table 1). Especially, the wAve increased d-accuracy within SINEs in which the mappers showed low mapping rate or accuracy (Figure 1.4-c, left). Taken together, the wAve successfully improved the methylation detection compared with using individual mappers.

It is noteworthy that the d-accuracy of wAve exhibited the reduced dependency on read conditions (i.e. read length and quality, Figure 1.4-d). We checked the distribution of d-accuracies of the three mappers and wAve, gathered from mapping results of 100 read sets (see Method).The wAve shows relatively less variance of d-accuracies among varied read conditions. Especially, the wAve decreased the difference of d-accuracy between high and low read error cases (Figure 1.4-c). This implies that the wAve successfully reduces the effects of heterogeneous read conditions on methylation detection, facilitating comprehensive analyses of methylation patterns among public WGBS samples from various experiments.

	Long reads (100bps)					Short read (50bps)				
Read error (%)	**0**	**2**	**4**	**6**	**8**	**0**	**2**	**4**	**6**	**8**
Ave	96.04	95.67	95.12	94.22	93.19	95.78	94.88	93.80	92.33	90.90
wAve	**96.05**	**95.72**	**95.24**	**94.74**	**94.43**	95.76	94.85	93.81	**92.55**	**91.50**
pwAve	96.05	95.69	95.14	94.32	93.43	**95.80**	**94.91**	**93.82**	92.38	91.02

Table 1.2: d-accuracy of the integration methods

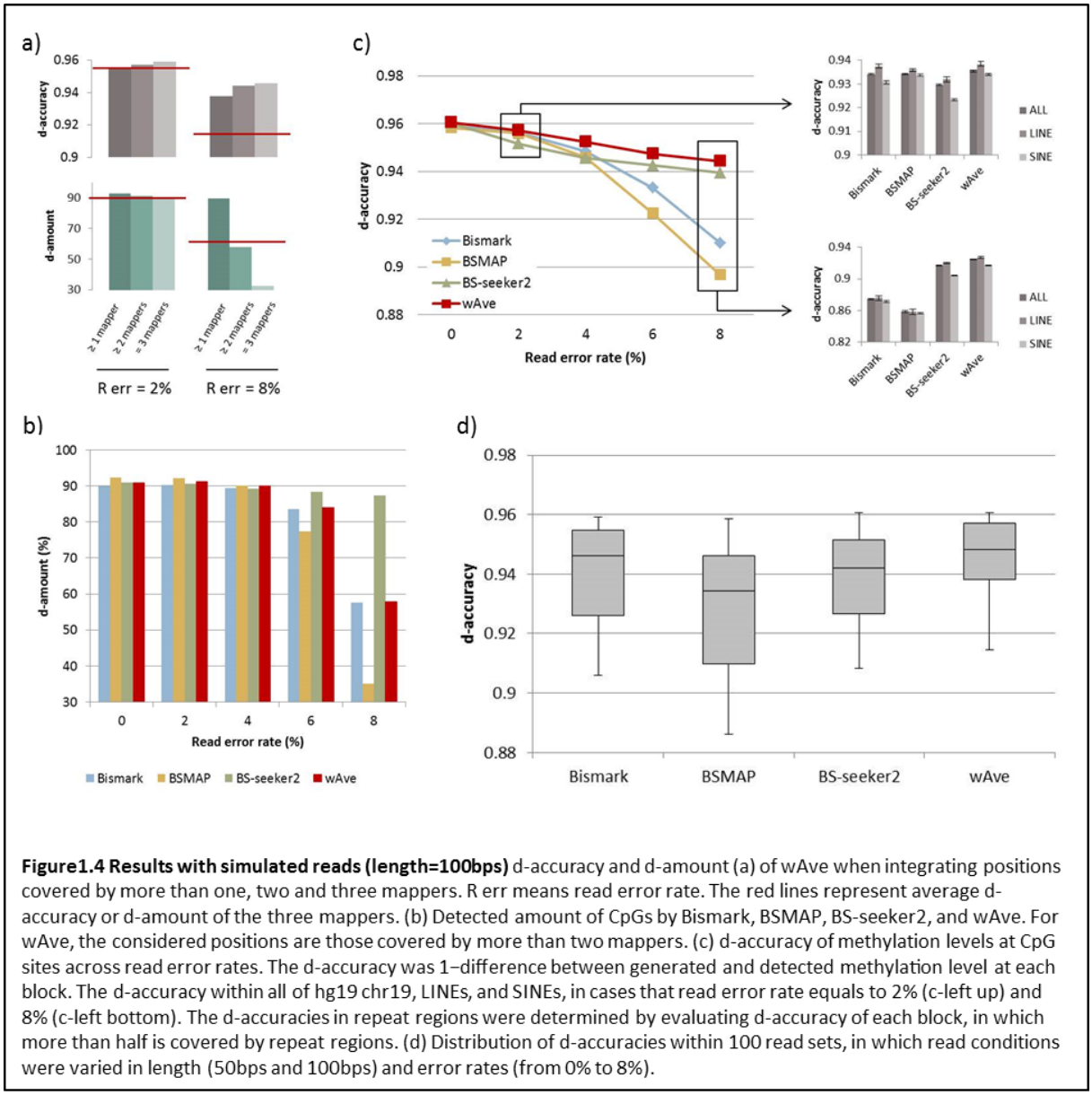


Figure1.4 Results with simulated reads (length=100bps) d-accuracy and d-amount (a) of wAve when integrating positions covered by more than one, two and three mappers. R err means read error rate. The red lines represent average d-accuracy or d-amount of the three mappers. (b) Detected amount of CpGs by Bismark, BSMAP, BS-seeker2, and wAve. For wAve, the considered positions are those covered by more than two mappers. (c) d-accuracy of methylation levels at CpG sites across read error rates. The d-accuracy was 1−difference between generated and detected methylation level at each block. The d-accuracy within all of hg19 chr19, LINEs, and SINEs, in cases that read error rate equals to 2% (c-left up) and 8% (c-left bottom). The d-accuracies in repeat regions were determined by evaluating d-accuracy of each block, in which more than half is covered by repeat regions. (d) Distribution of d-accuracies within 100 read sets, in which read conditions were varied in length (50bps and 100bps) and error rates (from 0% to 8%).

4. The integrative approach facilitated the comprehensive analysis of public WGBS data

Next, we re-analyzed 13 WGBS samples that were generated from various experiments with different read length and quality. In particular, we included six brain samples and four pluripotent stem cells, in which significant amount of CpH methylation has been observed [6, 7, 11, 12].

The mapping rate of WGBS samples was consistent with the mapping results of artificial bisulfite-reads (Supplementary figure 1.3). BSMAP showed the highest mapping rate within nine samples, whereas BS-seeker2 showed the highest mapping rate within the left four samples. Bismark showed the lowest mapping rate in most of the samples. In accordance with mapping rate, the d-amount by BSMAP was the highest among the three mappers within 12 samples, whereas the BS-seeker2 showed the highest d-amount within only one sample. The d-amount by wAve was higher than Bismark within all samples (Figure 1.5-a). Also, it showed higher d-amount than BS-seeker2 within six samples.

We also examined the correlation between samples in terms of CpG and CpH methylation levels detected by the wAve. We found that both methylation levels clearly grouped samples according to their originated tissues (Figure 1.5-b). In particular, while brain samples were produced from three different experiments, they were closely positioned in the dendrogram. Moreover, an unknown-brain sample, a WGBS data from brain of which age was not known, and a sample from liver were produced from the same experiment but were successfully grouped according to each tissue type. These trends were not observed for BS-seeker2 (Supplementary figure 1.4). We could confirm that the wAve clearly reduced false correlation between brain and liver from the same experiment (Figure 1.5-c). We also found that the wAve results in significantly higher correlation of CpG methylation levels in gene-body regions within brain samples compared with that observed by Bismark (Supplementary figure 1.5). Especially, the correlation among brain samples from different experiments were relatively more increased than that from same experiment. Thus, the wAve succeeded in decreasing the false correlation between samples from the same experiment and increasing the correlation among samples from the same tissue.

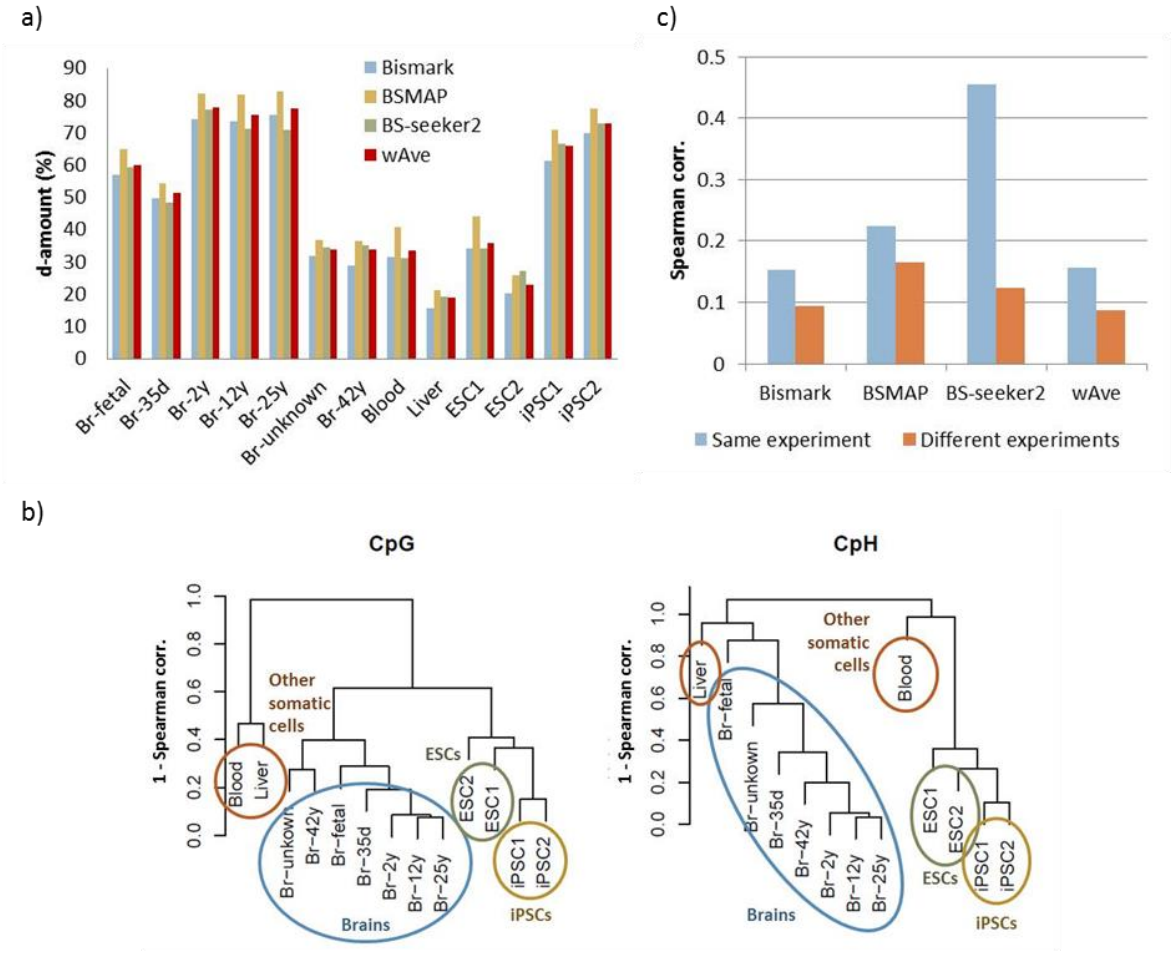


Figure1.5 Results with real WGBS data

(a) d-amount of WGBS samples. ESC=Embryonic stem cell, iPSC=induced pluripotent stem cell, Br=brain, d=day and y=year (e.g. Br-5y means 5 years old brain). (b) Hierarchical clustering results based on CpG and CpH methylation levels of 10kbps blocks across samples. The distance is 1−spearman correlation coefficient. (c) Spearman correlation coefficients of CpH methylation between liver with brain from same experiment (Br-unknown) and with a brain from different experiment (Br-42y).

Methods

Generating artificial sequenced reads

The sequenced reads were generated by Sherman [21]. The Sherman generates virtually bisulfite-treated reads with specific read number, length, error rate, and methylation level. We designated methylation level of CpG and CpH randomly and separately, on every 500bps block of human genome (hg 19) chromosome 19. The human chromosome 19 is short but reveals the highest repeat rate among all chromosomes [32]. Therefore, we could effectively observe the diverse mapping results of the three bisulfite-read mappers with special focus on repeat regions. From the randomly methylated reference genome, we generated long (100bps) and short (50bps) reads separately. The numbers of generated reads were 50 for 100bps reads and 100 for 50bps reads in a block to adjust the average coverage depth equals to 10. Also, we generated reads with designating error rate to 0%, 2%, 4%, 6%, and 8%, in order to determine the dependency of mapping results on read error. We repeatedly generated all the cases of read sets 10 times. In total, we generated 100 read sets (2 read length cases×5 read error cases×10 repeat) for which the read number was 5.4 million and 10.8 million for long and short reads, respectively.

Parameters for read generation

- For short reads

/Sherman -l 50 -n 100 -e [ERROR] --genome_folder [DIR] -CG [mCG] -CH [mCH]

- For long reads

/Sherman -l 100 -n 50 -e [ERROR] --genome_folder [DIR] -CG [mCG] -CH [mCH]

[DIR]: Directory to fasta file that include 500bp-long sequences. Before running Sherman, human chromosome divided into 500bps sequences and saved in separate directories.

[ERROR]: repeatedly set to 0, 2, 4, 6, and 8

[mCG] and [mCH]: randomly and independently set from 0 to 100. After the read generation completed, Sherman reported exact value of the bisulfite-treated rates on CpG and CpH positions. We used the reported values as designated methylation level at each block.

Read mapping and extracting methylation level

We mapped both artificial reads and real WGBS reads with Bismark, BSMAP, and BS-seeker2. In mapping, we unified the maximum mismatch threshold to 4% of the read length, to determine the distinct performances of the three mappers in unified parameters. The command lines for each mapper were as below;

Bismark; we used bowtie2 as an aligner for better performance [21].

perl ./bismark -o [ouput] --bowtie2 [reference genome] [input fastq] --score_min L,0,-0.24

BSMAP; we set the maximum number of equal best hits to one [20].

./bsmap -a fastq file -d [reference genome] -o [output] –w 1 –v 0.04
BS-seeker2; we used bowtie2 as an aligner for better performance. [22]
python ./bs_seeker2-align.py -i [input fastq] -o [output] -g [reference genome] --aligner=bowtie2 –m 0.04

After mapping, we removed duplicates possibly induced by PCR amplification. The duplicated reads from Bismark, BSMAP, and BS-seeker2 were removed by picard [33], samtool rmdup [34], and a program of BS-seeker2 [22], respectively. After removing duplicates, methylation levels of each C were extracted by programs of each mapper. In results with simulated reads, we considered methylation levels at Cs that covered by more than one read. In results with real WGBS data, however, we considered methylation levels at Cs that covered by more than five reads in order to increase the confidence of methylation level at each C [35]. The methylation level at each C was calculated as the ratio of unconverted Cs over the total mapped read number.

Integration of mapping results

The integration of three mappers was conducted at single base resolution. We extracted the number of both converted and un-converted Cs at each cytosine position. The methylation level M_{ij} at the *i*th C position detected by a mapper *j* (=Bismark, BSMAP, BS-seeker2) was calculated as below;

$$M_{ij} = \frac{{N^c}_{ij}}{{N^c}_{ij} + {N^t}_{ij}},$$

where N^c and N^t are the number of Cs and Ts, respectively. If there is no mapped read by the mapper *j*, M_{ij} is set to zero. We integrated M_{ij} using three methods; *Ave* (average), *wAve* (weighted average) and *pwAve* (probabilistic weighted average). *Ave* was given by

$$Ave_i = \frac{\sum_j M_{ij}}{n_i},$$

where *n* is the number of mappers with constrain $M_{ij} > 0$. *wAve* weights M_{ij} by the read depth of mapper *j* with assuming that the methylation level detected by many reads is more confident. This is based on the observation that read-mapping rate and detection accuracy of methylation levels are correlated (Supplementary figure 1.2). The *wAve* was given by

$$wAve_i = \frac{\sum_j W_{ij} M_{ij}}{n_i},$$

$$W_{ij} = \frac{N^d_{ij}}{\sum_j N^d_{ij}},$$

where *W* and N^d is the weight and the read depth, respectively. *pwAve* uses Poisson distribution for weighting M_{ij}. Based on the observations of the performances of the three mappers, we assumed that if a mapper mapped more reads than other mappers, the probability of existing incorrectly mapped reads at each position is also higher than that by other mappers. The *pwAve* was given by

$$pwAve_i = \frac{\sum_j W^p_{ij} M_{ij}}{n_i},$$

$$W^p{}_{ij} = \frac{f(N^d{}_{ij}; \lambda_j)}{\sum_j f(N^d{}_{ij}; \lambda_j)},$$

where W^p is the weight by the probability function f of Poisson distribution with parameter λ that is the average read depth of a mapper over whole genome.

WGBS data preparation

We collected 13 WGBS samples from 5 experiments (Table 1.3). For evaluating the CpH methylation level, seven of human brains and four of pluripotent stem cells, known to have specific CpH methylation patterns [6, 7, 36], were included to the dataset. All the samples were quality-trimmed by fastx toolkit [37], with setting that minimum phred quality score equals to 20 and minimum read length equals to half of the original read length. We mapped all WGBS sample in single mode for the greatest generalization [30].

- Parameters for quality-trimming

fastq_quality_trimmer -t 20 -l [half of read length] -i sample.fastq -o [output directory] -Q [phred score scale]

Tissue	Age	Read #	Length	Type	Experiment	Replica #
frontal cortex	fetal(20-week)	788M	100	Single	GSE47966[6]	9
middle frontal gyrus	35-day	669M	100	Single		12
middle frontal gyrus	2-year	900M	100	Single		12
middle frontal gyrus	12-year	594M	100	Single		12
middle frontal gyrus	25-year	903M	100	Single		12
prefrontal cortex	42-year	706M	100	Paired	GSE46710[38]	2
brain	Unknown	549M	90	Paired	GSE46698[39]	4
liver	Unknown	336M	100	Paired		7
ESC1		2655M	45	Single	GSE40832[40]	1
Blood		1240M	45	Single		1
ESC2		460M	87	Single	GSE16256[41]	5
iPSC1		667M	87	Single		13
iPSC2		837M	87	Single		20

Table 1.3: Description of public WGBS data used in Section 1

Conclusion

To efficiently detect DNA methylation from WGBS data, we analyzed and integrated the three most widely used bisulfite-read mappers, Bismark, BSMAP, and BS-seeker2. The procedure consisted of three steps: mapper analysis, analysis with simulated reads, and analysis with real WGBS dataset.

Firstly, we confirmed that the performances of the three mappers were consistent with the results of former studies of wild-card type (e.g. BSMAP) and three-letter type (e.g. Bismark and BS-seeker2) [19]. In particular, the two types of mappers performed distinctly in SINEs, in which the wild-card type mappers falsely mapped reads, whereas the three-letter type mappers failed to map large number of reads. It should be further investigated what distinction in algorithm induces the difference in mapping results in SINEs. In addition, the performances of Bismark and BSMAP dramatically decreased in case of bed read quality, whereas BS-seeker2 did not affected much by the fluctuation of read error rate. Lastly, the mapping accuracies of BSMAP and BS-seeker2 were found to be dependent on the methylation level, whereas Bismark were not. Based on the complementary performances of the three mappers across varying read conditions, we integrated the mapping results of the three mappers with three methods: average (Ave), read depth-weighted average (wAve), and probabilistically weighted average by Poisson distribution (pwAve).

With the simulated reads, the wAve method resulted in significantly higher detection accuracy than that obtained with individual mappers and other integration methods. On the other hand, pwAve showed decreased accuracy compared with wAve. The superior performance of wAve could be explained by the correlation between d-accuracy and mapping rate. Using read depth as weight, the wAve considered mapping rate as a first element on determining the certainty of the methylation levels from each mapper. On the other hand, pwAve indirectly employed mapping accuracy on weighting by considering the characteristics of the mappers; a mapper that maps larger number of reads compared with other mappers tended to maps reads at incorrect positions. The tendency was clearly revealed within short reads containing low error, so the d-accuracy of pwAve was the highest among the integration methods in those read conditions. Generally, however, d-accuracy was more strongly correlated with mapping rate (Pearson correlation coefficient equals to 0.83) than mapping accuracy (Pearson correlation coefficient equals to 0.64, Supplementary figure 1.6), resulting in the higher d-accuracy of wAve than that of pwAve in most read conditions. It should be further studied what probabilistic methods could improve the detection accuracy compared with read-depth weighting.

In addition, the wAve exhibited higher detection of Cs than Bismark. Indeed, existing bisulfite mappers exhibit smaller increases in either quality or quantity of the methylation results compared with former systems. It is remarkable that the integration improved both the accuracy and amount of methylation detection. Furthermore, the integration reduced the dependency of detection accuracy on read conditions (i.e. error rate and length), proving that our method can facilitate the comprehensive analyses of multiple WGBS samples of which read conditions are heterogeneous.

With real WGBS samples, the wAve reduced the false correlation between WGBS samples generated from same experiments and increased the true correlation between those originated from same tissues. Thus, our method succeeded in facilitating comprehensive analyses of multiple WGBS datasets from various experiments by reducing the dependency of methylation results on read conditions.

In summary, our integrative approach improved both quality and quantity of methylation results from WGBS data, and facilitated the comprehensive analyses of DNA methylation among various read conditions. This study may contribute to researches about methylation patterns among samples in different conditions (e.g. tissue, age, or some diseases) by combining a massive public WGBS data. In addition, this study may give a new clue to algorithmic improvement of bisulfite-read mappers to enhance epigenetic researches.

Section2: Differential activity of DNMT3a and DNMT3b causes distinct distribution and function of non-CpG methylation in embryonic stem cell and neuron

Abstract

DNA methylation is an important epigenetic mark that regulates cellular processing such as cell differentiation, development, and retardation. Although most of the methylation occurs at CpG dinucleotides in mammalian cells, recent studies have reported that considerable amount of methylated non-CpGs (mCpHs; H means A, C, and T) exit in embryonic stem cell (ESC) and neuron. Interestingly, distribution and function of the mCpHs in those two cell types are highly distinct. For example, the methylation preferentially occurs at CAG motif in ESC, whereas it occurs at CAC in neuron. In addition, the CpH methylation level and expression level of genes are positively correlated in ESC but negatively correlated in neuron. These opposite tendencies of mCpHs in the two cell types have been mystery among researchers for years.

In this study, to understand the differential mechanisms of CpH methylation in those two cell types, we conducted a comprehensive computational analysis with public whole genome bisulfite sequencing data of 14 ESCs and neurons incorporated with transcriptome and DNA methyltransferase (DNMT) knockout data. We confirmed that CpHs are methylated by DNMT3a and DNMT3b, dependently to the methylation at flanking CpGs. Remarkably, the DNMT3a and DNMT3b are preferentially methylate CpHpH and CpHpG context, respectively. These DNMTs are differentially expressed in ESCs and brain tissues, resulting in distinct mCpH motifs in those two cell types. In addition, the preferential binding of DNMT3b on H3K36me3 histone marks induces positive correlation between gene expression and CpH methylation in ESCs.

Collectively, our study revealed that DNMT3a and DNMT3b are responsible for the methylation at CpHpH and CpHpG contexts nearby CpGs, respectively, causing distinct distribution and function of mCpH in ESCs and neuron. This result shed light on the importance of cell type specific establishment of both mCpG and mCpH in mammalian cells.

Introduction

DNA methylation, an addition of methyl group on fifth carbon at cytosine, is one of the most important epigenetic modifications. For decades, epigenetic studies have focused on methylated CpG dinucleotides (mCpGs) that govern cell type specific functions and cause diseases by regulating transcription [23, 25, 42]. In contrast, the methylated non-CpG dinucleotides (mCpH; H includes A, C and T) had shown mostly ignorable level in mammalian somatic cells, even though it had been frequently observed in plant [43]. Recently, however, several research groups have uncovered that significant amount of mCpHs are observed in mammalian pluripotent stem cells and non-dividing cells such as neuron, regulating cell type specific regulation such as cell differentiation and neurodevelopment [6, 7, 11-13, 44]. In this way, the mCpH, as well as mCpG, emerged as a key epigenetic factor, especially in pluripotent stem cells and non-dividing cells.

In those cell types, the distribution of mCpHs is depended to that of mCpGs. In contrast to the methylation in plant, where CpG, CpHpH, and CpHpG contexts are methylated by methyltransferases, MET1, DRMs, and CMT3, respectively [45-47], both CpG and CpHpN (N means all kinds of nucleotide) contexts are *de-novo* methylated by DNMT3a and DNMT3b in mammalian cells. It results in spatial correlation between mCpG and mCpH [12]. Since those methyltransferases showed 10 times higher affinity at CpG than CpH, the mCpHs have been considered as by-products from the hyperactivity of DNMTs that originally target CpGs. [2, 11, 12, 48-50].

Despite the spatial dependency of mCpH on mCpG, there are some evidences that mCpHs regulate cell type specific functions that independent to mCpGs. In brain, the mCpHs are gradually increased as aged, in a same pattern with the progress of synaptogenesis [6]; whereas mCpGs are not. In addition, the methyl-CpG binding protein 2 (MeCP2), mutation of which causes Rett syndrome, binds to not only mCpGs but also mCpHs. Considering that postnatal onset of Rett syndrome coincides with the emergence of mCpH, there is a possibility that MeCP2-related neuro-diseases are governed not by mCpG, but by mCpH [12]. Also, there are mega-base mCpH deserts in induced pluripotent stem cells (iPSCs), in which genes are less transcribed compared to those in ESCs. This implies that the failure of epigenetic reprogramming at CpHs leads to genetic aberration in iPSCs. Altogether, even though the mCpHs are spatially correlated to mCpGs, those play important roles on cell type specific processes independently to mCpGs.

One of the underlying mechanisms that mCpHs govern cell type specific phenomena is distinct characteristics of mCpHs across cell types. For example, the abundant DNA motif at mCpHs is “CAG” in ESC [5, 7]; whereas it is “CAC” in neuron [6, 10]. Also, the mCpHs tend to be abundant at intragenic regions in ESC, whereas those are abundant at intergenic regions in neuron [6, 10]. Especially, the CpHs are hyper-methylated at actively expressed gene bodies in ESC, whereas those in neuron are hypo-methylated [6, 7]. Since DNA methylation is generally negatively correlated to gene expression in most somatic cells [9], the positive correlation between mCpH and gene expression in ESC has been mysterious for years. In this way, the mCpH shows distinct distribution and potential function to

gene expression in ESC and neuron. However, it is still unknown what causes the distributional and functional difference among cell types.

In this study, to uncover the mechanism of cell type specific CpH methylation, we analyzed the whole genome bisulfite sequencing (WGBS) data of human ESCs, neurons, and brain tissues. In addition, we included WGBS data of DNMT knock out ESCs to trace the role of DNMTs on cell type specific methylation [6, 35, 38, 49, 51-54]. Through comprehensive analysis, we confirmed that DNMT3a and DNMT3b are responsible for both mCpG and mCpH, resulting in spatial dependency of mCpHs on mCpGs. Interestingly, the two DNMTs differentially methylate cytosines at CpHpH and CpHpG contexts, resulting in distinct pattern of mCpHs in ESCs and neurons. In addition, we found that in ESC, the preferential affinity of DNMT3b on histone mark H3k36me3 causes positive correlation between mCpH level in gene bodies and gene expression level. Based on the results, we suggested a differential CpH methylation model that can explain the mystery of distinct mCpH characteristics in ESC and neuron. Altogether, our results give an insight for understanding cell type specific formation, distribution and function of mCpHs in mammalian cells.

Results

1. The integrative approach for WGBS read aligning successfully reproduced known characteristics of mCpH in ESC and neuron.

To analyze methylation at both CpG and CpH contexts, we re-aligned WGBS reads that generated from various experiments (Table 2.1). Three bisulfite read mappers, Bismark [21], BSMAP [20], and BS-seeker2 [22] were used for read aligning, and the outputs were integrated by previously introduced strategy (Method section) [55]. Also, we adjusted the methylation level by subtracting bisulfite non-conversion rate based on a statistical model [6].

Statistic features of the regenerated data were coincident with those in previous studies (Figure 2.1-a) [2, 6, 7, 11-13]. In human samples, although most of the CpGs were hyper-methylated (75~85%), the average methylation levels at CpH (mCpH level) were distinct across cell types. It was mostly abundant in neurons (>5%), abundant in ESCs derived from male (H1) and matured brain tissues (>1%), detectable in ESCs derived from female (H9), early-passage ESCs (HUES64), and immature brain tissues (0 to 5 year-old; ~1%), and mostly ignorable in other tissues (Heart, Spleen, and Lung; <0.5%). Also, we observed the increase of mCpH level along with brain aging [6], and lower mCpH level in H9 than in H1 [5, 7, 56]. Remarkably, the WGBS samples were clearly clustered into their originated tissues (or cell types) by both mCpG and mCpH patterns (Figure 2.1-b). With the methylation patterns that extracted from each bisulfite read mapper, however, it was not clearly grouped into their originated tissues (Supplementary Figure 2.1). Lastly, we confirmed the motif abundancy at hyper-methylated CpHs (beta value>50%). It was “CAG” in all ESC samples, whereas it was “CAC” in all neuron and brain samples (Supplementary Figure 2.2) [5-7, 10, 12]. Altogether, our integrated methylome successfully reproduced the known characteristics of CpH methylation in mammalian cells. Since the mCpH levels in early brains and somatic cells are extremely low, we excluded those cells from further analysis, and focused on methylation in ESCs, neurons, and adult brains.

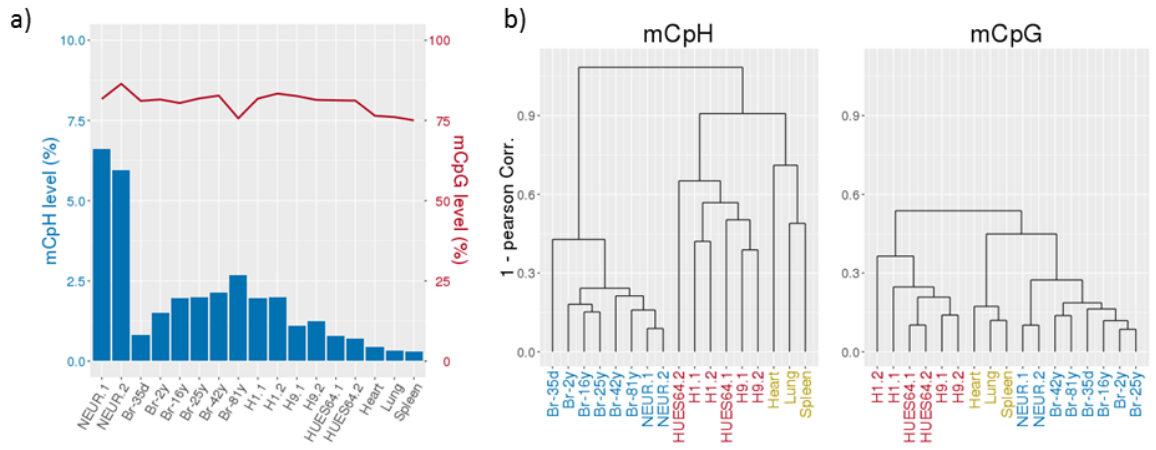


Figure2.1 Dataset used in section 2
(a) Genome-wide average methylation level at CpGs (red line) and CpHs (blue bar). Labels in x-axis are cell and tissue types of WGBS data. NEUR and Br represent neuron and brain tissue, respectively. The numbers followed by Br- means age that the brain tissue is sequenced (d and y mean day and year, respectively). H1, H9 and HUES64 are cell lines of human embryonic stem cell. Numbers back in the label represents biological replicates (NEUR.1 means neuron biological replicate 1). (b) The hierarchical clustering has been conducted by CpH and CpG methylation pattern across whole genome. Colors in x-axis represents same cell (or tissue) types (blue: brain and neuron, red: ESCs, and yellow: other tissues).

2. There are cell-type specific CpH methylation pattern around mCpGs (±100bp).

With the WGBS dataset, we analyzed the correlation between CpG and CpH methylation. The methylation levels of CpG and CpH across 1k-bp blocks were positively correlated to each other, which coincident with results from previous studies [10-12, 48] (Supplementary Figure 2.3; Pearson Corr. ~0.44). In further analysis, we found that among blocks where CpGs are hemi-methylated (difference of mCpG level between strands > 0.5), the mCpH levels are significantly higher at the same strand with highly methylated CpG, compared to those at opposite strand. In addition, the mCpH levels were significantly highly correlated to the mCpG levels in same strand than those in opposite strand, implying that the CpG and CpH methylations are correlated in strand-specific way (Supplementary Figure 2.4-a, b). Also, the correlation between mCpG and mCpH level was significantly higher at exon and promoter regions (Supplementary Figure 2.4-c).

Next, we checked the spatial correlation between mCpG and mCpH by measuring the probability that mCpHs exist at a distance from CpGs. Interestingly, the probability was greatly high within ±100bp distance from mCpG (Figure2.2-a). In addition, the correlation between mCpG and mCpH levels was greatly high when those are within ±100bp, supporting that methylations at CpGs and CpHs are highly correlated especially when those are within ±100bp distance (Supplementary Figure 2.5). The correlation pattern has been observed in all of ESCs, neurons and brain samples. Based on the spatial correlation, we grouped CpHs within 100-bp from mCpGs (beta value > 0.8) as mCpG-correlated CpHs and analyzed the methylation pattern at those CpHs. The mCpG-correlated CpHs showed conserved methylation pattern among same cell types (Supplementary Figure 2.6). In ESCs, a clear CpH methylation peak was observed at -4bp from mCpG that had reported in previous study [14]. Also, in neurons and brains, the 8-10bp periodicity among CpH methylation peaks was

observed, which has been suggested as a potential mark for methylation by DNMT3a-DNMT3l enzyme complex [57]. Remarkably, the mCpG-correlated CpH methylation patterns were clearly distinguishable between ESC and neuron (Figure 2.2-b), implying the differential mechanisms of mCpG-correlated CpH methylation between the two cell types exist. Altogether, the methylation pattern at mCpG-correlated CpHs is highly conserved in same cell types but distinguishable between ESCs and neurons.

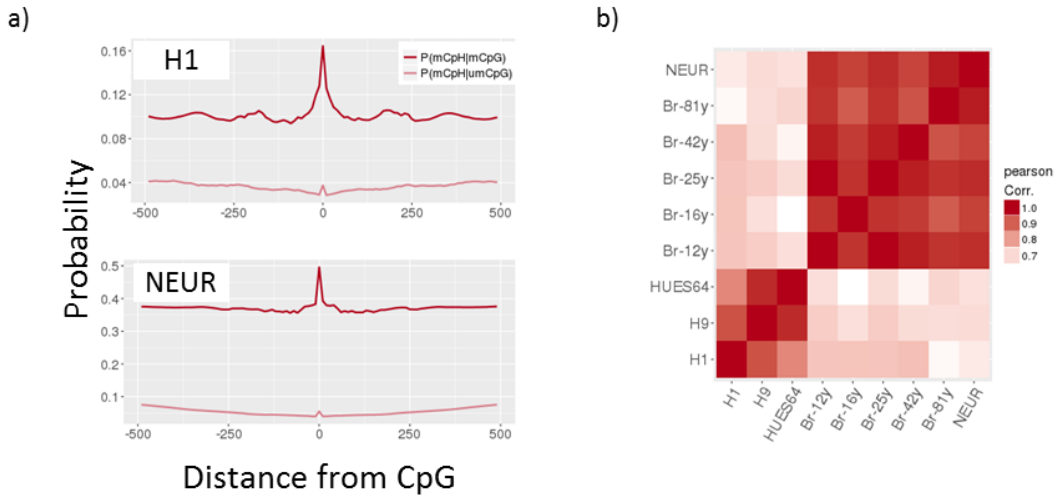


Figure2.2
(a) Shown is the probability that mCpH exist around methylated CpGs (deep red), and un-methylated CpGs(pale red). We divided the surrounding (± 500bp) of every CpG into 10bp-long blocks, and calculated the probability of mCpH by counting how many blocks contain mCpHs in the condition of CpGs (methylated or un-methylated). (b) The heatmap shows Pearson correlation of mCpG-correlated (± 100bp) CpH methylation among samples. NEUR, and Br represents neuron, and brain, respectively.

3. The DNMT3a and DNMT3b preferentially methylate CpHpH and CpHpG contexts, respectively.

To uncover the differential mechanism of mCpG-correlated CpH methylation in ESC and neuron, we analyzed DNMT knock out human and mouse ESCs (mESCs).

As a first step, we analyzed mouse ESCs in which DNMT1, and both DNMT3a and 3b are knocked out (DNMT1-KO mESC and DNMT3-DKO mESC, respectively) [52]. The mCpH level was little decreased in DNMT1-KO mESC compared to wild type, whereas it was dramatically decreased in DNMT3-DKO mESC, implying that DNMT3a and DNMT3b are mainly responsible for CpH methylation (Supplementary Figure 2.7-a). In addition, the correlation between CpG and CpH methylation levels was higher in DNMT1-KO mESC and lower in DNMT3-DKO mESC, comparing to wild type (Supplementary Figure 2.7-b; Pearson correlation coefficient= 0.4, 0.07, and 0.2, respectively). Furthermore, we observed that mCpH level is greatly high nearby mCpGs in DNMT1-KO sample

(Figure 2.3-a). A clear periodicity of ~180bp between mCpH level peaks was also observed in DNMT1KO mESC, implying that CpG and CpH could be simultaneously methylated by DNA walking of methyl-transferases. Altogether, the mCpG-correlated CpH methylation is introduced not by DNMT1, but by DNMT3a and DNMT3b.

Next, we analyzed the contribution of DNMT3a and DNMT3b on CpH methylation in human ESC (HUES64 cell line). Interestingly, in DNMT3a knock out sample, the methylation level at CpHpH contexts decreased more than that at CpHpG contexts, whereas it was opposite in DNMT3b knock out sample (Figure 2.3-b). Also, in H9 cells [51], the methylation level at CpHpG contexts was more decreased than that at CpHpH contexts as DNMT3b knocked out (Supplementary Figure 2.7-c). We observed the same tendency in mouse ESC WGBS data that generated by Dr. Tuncay Baubec's group [53]. They knocked out all of DNMT1, DNMT3a and DNMT3b from mouse ESC, then induced DNMT3a and DNMT3b, respectively, for measuring *de novo* methylation by each methyltransferase. With this sample set, we found that *de novo* CpH methylation is greatly focused at CpG-correlated CpHs (Figure 2.3-c). Remarkably, among the CpHs, the CpHpH context is more methylated by DNMT3a than by DNMT3b, whereas the CpHpG context is mostly methylated by DNMT3b. In summary, we found that both DNMT3a and DNMT3b are key factors of mCpG-correlated CpH methylation, but, DNMT3a is more responsible for the methylation at CpHpH contexts, whereas DNMT3b is at CpHpG contexts.

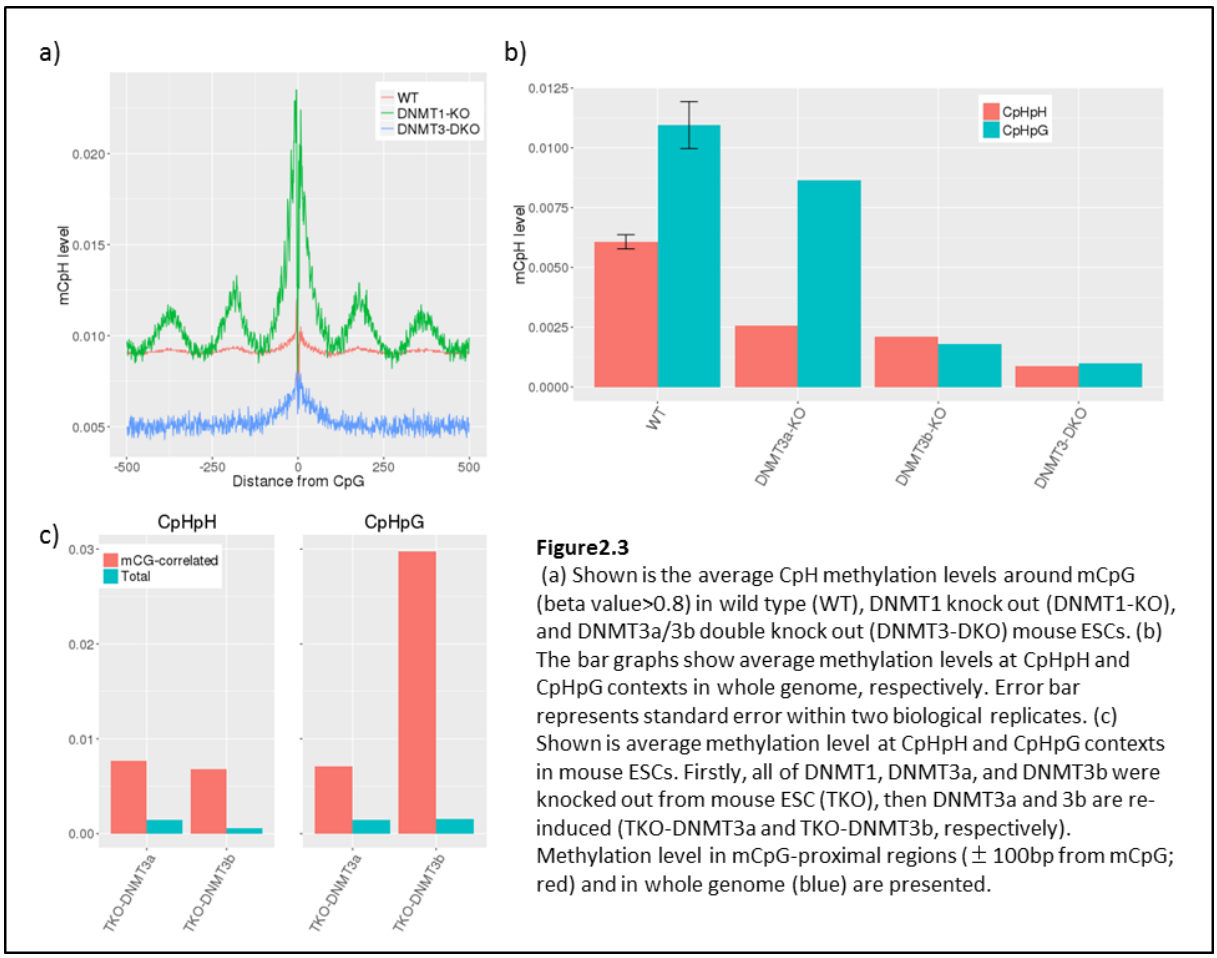


Figure2.3
(a) Shown is the average CpH methylation levels around mCpG (beta value>0.8) in wild type (WT), DNMT1 knock out (DNMT1-KO), and DNMT3a/3b double knock out (DNMT3-DKO) mouse ESCs. (b) The bar graphs show average methylation levels at CpHpH and CpHpG contexts in whole genome, respectively. Error bar represents standard error within two biological replicates. (c) Shown is average methylation level at CpHpH and CpHpG contexts in mouse ESCs. Firstly, all of DNMT1, DNMT3a, and DNMT3b were knocked out from mouse ESC (TKO), then DNMT3a and 3b are re-induced (TKO-DNMT3a and TKO-DNMT3b, respectively). Methylation level in mCpG-proximal regions (± 100bp from mCpG; red) and in whole genome (blue) are presented.

4. Distinct characteristics of mCpH between ESC and neuron are resulted from differential activity of DNMT3a and DNMT3b.

Through comprehensive analysis of WGBS data and RNA-seq data, we found some evidences that the distinct mCpH patterns in ESC and neuron are resulted from differential activity of DNMT3a and DNMT3b on CpHpH and CpHpG methylation.

First, the DNMT3a is highly expressed in neuron, whereas DNMT3b is highly expressed in ESC (Figure 2.4-a). Through analyzing an expression dataset that contains 143 human adult brains and 36 human ESCs (Table 2.2), we confirmed that DNMT3a is significantly highly expressed in brains, whereas DNMT3b is in ESCs. Following that, in all the human ESC WGBSs, the methylation level at CpHpG context (mCpHpG level) was higher than that at CpHpH context (mCpHpH level), whereas it was opposite in all the neuron and brain samples (Figure 4-b). Remarkably, the abundant motif at mCpH in DNMT3b knock out ESC was not "CAG", but "CAC" which is the abundant mCpH motif in neurons (Supplementary Figure 2.2). It implies that the abundant mCpH motif, "CAG", is induced by hyperactivity of DNMT3b in ESC.

Meanwhile, the cell type specific mCpG-correlated mCpH patterns are also caused by differential activity of DNMT3a and DNMT3b. The methylation pattern at mCpG-correlated CpHs in DNMT3b-knock out ESC (HUES64) was more similar to that in brains, than that in wild type ESCs, implying that the DNMT3b contributes to the distinct mCpG-correlated CpH methylation pattern in ESC (Figure 2.4-c).

Lastly, we found that the hyper-methylation at highly expressed gene-body regions in ESC is selectively occurs at CpHpG context (Supplementary Figure 2.8). Considering the preferential affinity of DNMT3b on H3k36me3 histone marks that positioned at highly expressed gene-body regions [25], it could be inferred that the hyper-activity of DNMT3b, by preferentially methylating CpHpGs in highly expressed gene-bodies, causes the positive correlation between mCpH level and gene expression in ESC.

To prove that, we confirmed the relationship between mCpHpGs and histone mark H3K36me3s. Among 13 histone marks, the H3K36me3 showed significantly highly positive correlation with mCpHpG in both H1 and H9 cell lines. Also, it is known that the H3K36me3 is highly abundant at highly expressed gene-body regions [53]. Therefore, we compared the mCpH level in gene body regions between wild type and DNMT-3b knock out HUES64 cell lines. Remarkably, the difference of mCpH levels in highly/lowly expressed gene-body regions almost disappeared when DNMT3b is knocked out, whereas it was maintained when DNMT3a is knocked out (Supplementary Figure 2.8). Lastly, we analyzed the mCpH levels at gene-body regions in a DNMTs and SETD2 knocked out mouse ESC [53]. In that sample, the three DNMTs (DNMT1, DNMT3a, and DNMT3b) were knocked out and DNMT3b was re-induced. In addition, the SETD2 enzyme, known as catalyzer of H3K36me3 marks, was knocked out, resulting in absent of h3k36me3 marks [53]. Interestingly, the positive correlation between mCpHpG and gene expression level disappeared in the Setd2 knock out sample (Figure 4-e), implying that the positive correlation between mCpH and gene expression is resulted from the

interaction between H3K36me3 and DNMT3b that responsible for methylation at CpHpG.

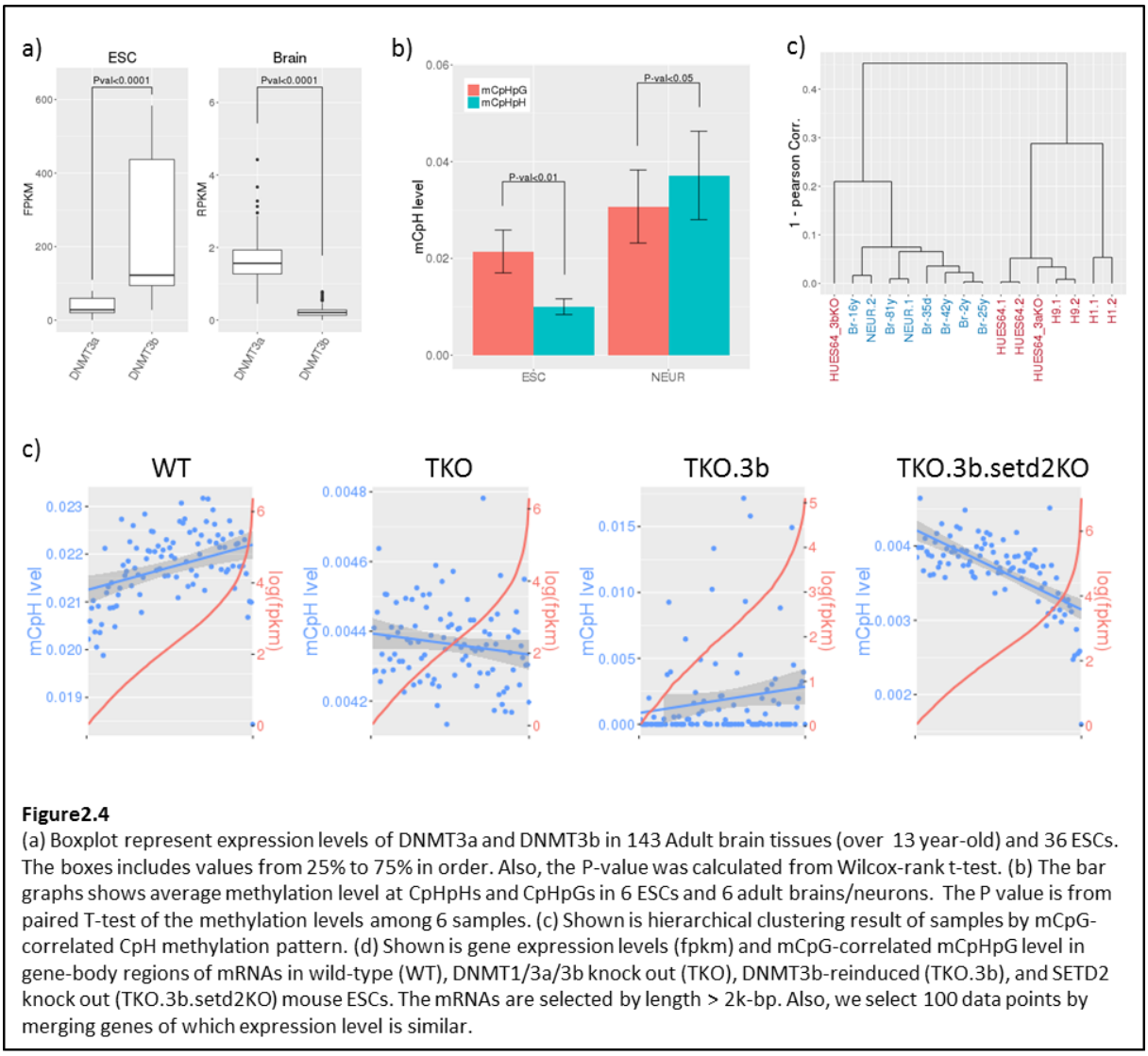


Figure2.4
(a) Boxplot represent expression levels of DNMT3a and DNMT3b in 143 Adult brain tissues (over 13 year-old) and 36 ESCs. The boxes includes values from 25% to 75% in order. Also, the P-value was calculated from Wilcox-rank t-test. (b) The bar graphs shows average methylation level at CpHpHs and CpHpGs in 6 ESCs and 6 adult brains/neurons. The P value is from paired T-test of the methylation levels among 6 samples. (c) Shown is hierarchical clustering result of samples by mCpG-correlated CpH methylation pattern. (d) Shown is gene expression levels (fpkm) and mCpG-correlated mCpHpG level in gene-body regions of mRNAs in wild-type (WT), DNMT1/3a/3b knock out (TKO), DNMT3b-reinduced (TKO.3b), and SETD2 knock out (TKO.3b.setd2KO) mouse ESCs. The mRNAs are selected by length > 2k-bp. Also, we select 100 data points by merging genes of which expression level is similar.

Method

WGBS data analysis

The sra type WGBS data (Table 2.1) were downloaded from Gene Expression Omnibus (GEO) database and converted into fastq type by using fatq-dump, a part of SRA Toolkit provided by NCBI. Then, the fastq files were quality-controlled by fastqx-toolkit tools.

The high-quality bisulfite sequenced reads were mapped into refseq reference genome (hg19 for human samples and mm10 for mouse samples) by three bisulfite-read aligners, Bismark, BSMAP, and BS-seeker2. Then, we removed possible duplicated reads by using picard (for the output from Bismark and BS-seeker2) and samtools (for the output of BSMAP). Then, we extracted cytosine positions covered by more than 5 reads in outputs of more than 2 bisulfite-read aligners. Methylation levels at the cytosines were calculated as read-depth weighted average of those from individual aligner, and the bisulfite non-conversion rate was subtracted from the value, based on the statistical model suggested in previous study [6]. The scripts for used software tools and calculation for methylation is below.

Preprocessing

> fastq-dump --split-3 INPUT FILE –O OUTPUT DIR

>fastq_quality_trimmer -t 20 -l [half of original read length] -i input.fastq -o OUTPUT

>fastq_quality_filter -q 20 -p 50 -i input.fastq -o OUTPUT

Read aligning (Single type read case)

>perl bismark --bowtie2 pre_built_bismark_reference -o OUTDIR input.fastq

>bsmap -a input.fastq –d reference_fasta_file -o OUTDIR -w 1

>python bs_seeker2-align.py -i input.fastq -o OUTFILE -g BS-seeker2_reference --aligner=bowtie2

Methylation detection

$$Me_i = \frac{\sum_j M_{ij}}{\sum_j t_{ij}} - non\ conversion\ rate$$

Where Me_i is methylation level at cytosine i, t_{ij} is aligned read number at cytosine i by bisulfite-read aligner j, and M_{ij} is unconverted read number at cytosine i by bisulfite-read aligner j. In case the Me_i is under 0, we set $Me_i = 0$.

Identification of methylated cytosines

Although average methylation level was used for most of the analysis, the identification of mCpH was necessary for measuring probability of mCpH existing around CpGs (Result2). We used binomial distribution to detect the methylated cytosine loci. At every detected cytosine loci(i), we calculated the probability that methylated reads (k_i) occurs out of total read number (n_i) with set the success probability (p) as bisulfite non-conversion rate. If the probability is under certain threshold, we identified the cytosine loci as methylated.

$$f(k_i; n_i, p) = Pr(X_i = k_i) = \binom{n_i}{k_i} p^{k_i}(1-p)^{n_i - k_i}$$

To set the probability threshold, we made artificial methylome for every WGBS sample, in which methylated read number is randomly generated following binomial distribution with setting the trial number as real read depth at each location, and the success probability as bisulfite non-conversion rate. Then, we calculated FDR by measuring how much of the false methylation loci occur by certain probability threshold. Finally we changed the threshold as 0.00001, by which the FDR was under 0.01 within all of the WGBS samples. This method was used in previous study [12].

Correlation analysis

We divided whole-genome into 1 kilo base-pair blocks to compare the methylation pattern at CpG and CpH. Then, we extracted blocks that contain more than 10 points of CpG and CpH to secure accurate methylation level at each block. Also, to compare the mCpHs in Cis/Trans-strand to mCpG, we extracted blocks of which the difference of mCpG level is over 0.3. Lastly, to measure the correlation between mCpG and mCpH in genic regions, we extracted blocks that more than 500bp in the block is covered to genic regions. As meaning of genic regions, promoter defined as transcription start site ±5000bp, intragenic as all the regions of transcription start site (TSS) to transcription termination site (TTS; the strand of each gene bodies were considered), and intergenic as regions that uncovered by intragenic. The position information of TSS, TTS, and exon regions is from refseq data.

Gene expression

The sra type RNA-seq data were downloaded from GEO database, and preprocess had been conducted in same process with WGBS data processing. The high-quality sequenced reads were aligned by Tophat2 [58], and the fpkms were extracted by Cufflink [59]. Follows are command line for running Tophat2 and Cufflink

>tophat2 -g 1 --b2-sensitive -o OUT_DIR -G refseq_annotation.gtf --no-novel-juncs bowtie2_reference input.fastq
>cufflinks -G refseqannotation.gtf -o OUT_DIR tophat_output.bam

In addition, we used the processed data of gene expression levels in brains [60] and ESCs [61]. Specifics for the data are described in Supplementary Table 2.2.

Histone marks

The processed Chip-seq data for histone marks were downloaded from as bed files. The bed file included loci that more than one read mapped into hg19. Specifics for the data are described in

Supplementary Table 3.

Species	Sample name	Cell type	sex	age	experiment	Bisulfite conv. rate
Human	H1.1	H1 cell line	M	Passage:30	GSE16256[41]	0.996
	H1.2	H1 cell line	M	passasge:25		0.996
	H9.1	H9 cell line	F	passage:42		0.995
	H9.2	H9 cell line	F	passage:40		0.995
	H9.3bko	DNMT3b knocked out H9	F		GSE32268[51]	0.995
	HUES64.1	HUES64 cell line	M	passage:27	GSE17312[39]	0.995
	HUES64.2	HUES64 cell line	M	passage:23		0.995
	HUES64_3ako	DNMT3a knocked out HUES64	M	passage:24+22	GSE63278[49]	0.996
	HUES64_3bko	DNMT3b knocked out HUES64	M	passage:24+22		0.998
	HUES64_dko	DNMT3a,b knocked out HUES64	M	passage:24+5+7		0.999
	Br-35d	tissue (middle frontal gyrus)	M	35-day	GSE47966[6]	0.9962
	Br-2y	tissue (middle frontal gyrus)	M	2-year		0.9963
	Br-16y	tissue (middle frontal gyrus)	M	16-year		0.9966
	Br-25y	tissue (middle frontal gyrus)	M	25-year		0.9963
	Br-42y	tissue (frontal cortex)	F	42-year	GSE46710[38]	0.9961
	Br-81y	tissue (frontal cortex)	F	81-year	GSE46644[35]	0.9732
	NEUR.1	neuron	F	53-year	GSE47966[6]	0.9933
	NEUR.2	neuron	M	55-year		0.996
	heart	heart	M	34-year	GSE16256[41]	0.9914
	lung	lung	F	30-year		0.992
	spleen	spleen	M	34-year		0.9921
Mouse	mESC_wt	mouse ESC			GSE61457[52]	0.99

	mESC_1ko	DNMT1 knocked out mouse ESC				0.99
	mESC_dko	DNMT3a,3b knocked out mouse ESC				0.99
	mESC_tko	DNMT1,3a,3b knocked out mouse ESC				0.99
	mESC.tko.3a	DNMT1,3a,3b knocked out and 3a induced mouse ESC			GSE57413[53]	0.995
	mESC.tko.3b	DNMT1,3a,3b knocked out and 3b induced mouse ESC				0.995
	mESC.tko.3b.setd2	mESC.tko.3b + setd2 knock out				0.995

Table 2.1: Description of public WGBS data used in Section 2

Type	**Specifics**	**Number**	**Source**	**Analyzer**
ESCs	H1 and H9	8	GSE30567 [61]	Illumina Genome Analyzer II
		6	GSE24399 [62]	
		22	GSE75748 [63]	Illumina HiSeq 2500
brain	Brain tissues from 13 to 40year-old	144	Brainspan [60]	Illumina Genome Analyzer II

Table 2.2: Description of public RNA-seq data

Conclusion

Through comprehensive analysis of WGBS data, we uncovered the differential mechanism of CpH methylation in ESC and neuron.

The WGBS has been considered as the only way for extracting reliable information about mCpH, since other experiments, such as microarray data [15] or RRBS [17], are mainly targeting CpG dinucleotide [19]. It, however, is financially and timely consuming so that researchers had to deduce results from insufficient number of samples. The integrative read aligning strategy solved this problem by facilitating re-using of public data with improved accuracy and reduced experimental bias. Through the integrative approach, we employed 16 of ESCs, 2 of neurons into dataset. In addition, we added 6 brain tissues for detecting characteristics of mCpH in neuron, since most of mCpH are positioned in neuron among the cells included in brain tissue. The quality for the dataset has been verified by reproducing known characteristics of mCpH in each cell type, and clustering samples into their originated cell types by methylation pattern. The enlarged sample set with great quality contributed to the robustness of the following results.

Through analyzing the methylation pattern on both CpG and CpH dinucleotides, we confirmed the positive spatial correlation between CpG and CpH methylation. Especially, the correlation greatly increased when CpG and CpH are in same strand, and those are in promoter or exon region. It implies that the CpG and CpH could be methylated in correlated way by accessing of methyltransferases on open chromosome structure. In addition, we found the correlation between CpG and CpH methylation is significantly high in CpG-proximal regions (CpG±100bp). It supports the insistence that mCpH is a by-product from the hyperactivity of DNA methyltransferases that originally targets CpGs [11]. Remarkably, however, the mCpHs in CpG-proximal region (±100bp) showed differential distribution in ESC and neuron. In addition, even though the mCpG and mCpH in genic regions showed high correlation (Pearson correlation coefficient ~0.6), the relation with gene expression showed exactly opposite tendency in ESC; the mCpG is negatively correlated to gene expression, whereas the mCpH is positively correlated to that [7]. Furthermore, sufficient evidence exist that mCpHs play roles on cell type specific biological processes such as cell differentiation or synaptogenesis [6, 13]. Collectively, even though the mCpH is spatially correlated to mCpG, it plays roles on cell type specific biological processes, driven by cell type specific distribution.

The differential genome-wide distribution of mCpHs in ESC and neuron was explained by differential activity of DNMT3a and 3b. By comparing genome-wide mCpH level in wild type and DNMTs knock out mouse ESCs, we found that the mCpH is formed not by DNMT1 but by DNMT3a and 3b. Remarkably, we found that DNMT3a preferentially methylates CpHpH context, whereas DNMT3b methylates CpHpG context. The differential targeting of DNMT3a and 3b combined with differential expression of those in ESC and brain, emerged as the main reason of distinct mCpH distribution in the two cell types. In ESC, the hyperactivity of DNMT3b results in preferential methylation at CpHpGs. Considering the affinity of DNMT3-series enzymes on CpA among CpH contexts [64], it is reasonable

that the motif at mCpH is "CAG". On the other hand, since DNMT3b is almost not expressed in neuron, the DNMT3a is responsible for most of mCpH that results in higher methylation at CAC than CAG. Decisively, the motif abundant at mCpH was not CAG but CAC in DNMT3b knock out ESC. Furthermore, the mCpG-proximal CpH methylation pattern in DNMT3b knocked out HUES64 was more similar to that in brain tissues or neurons, than wild type ESCs. It implies the distinct mCpH pattern in ESC is induced by the hyperactivity of DNMT3b.

Finally, the positive correlation between CpH methylation level on gene body and expression level in ESC was explained by the activity of DNMT3b. In recent study [53], DNMT3b showed preferential interaction with histone mark H3k36me3 in highly expressed gene bodies. In our analysis, the preferential methylation on highly expressed gene bodies were greatly focused on CpHpG contexts, and it disappeared as DNMT3b knocked out. In addition, the positive correlation between mCpHs and gene expression disappeared in either DNMT3b or SETD2 (enzyme for catalyzing histone mark, H3k36me3) is knocked out. Collectively, the DNMT3b, by selectively methylates CpHpGs upon histone mark H3k36me3, causes positive correlation between CpH methylation and gene expression level in ESC.

Based on the results above, we suggest distinct CpH methylation mechanism by DNMT3a and DNMT3b in ESCs and neurons (Figure2.5). Still, the specifics for the differential methylation ability of DNMT3a and DNMT3b are remained as further research subject. Altogether, this study uncovered the reason of differential distributions and functions of mCpH in ESCs and neuron, and gives a crucial hint for uncovering the cell-type specific CpH methylation mechanism.

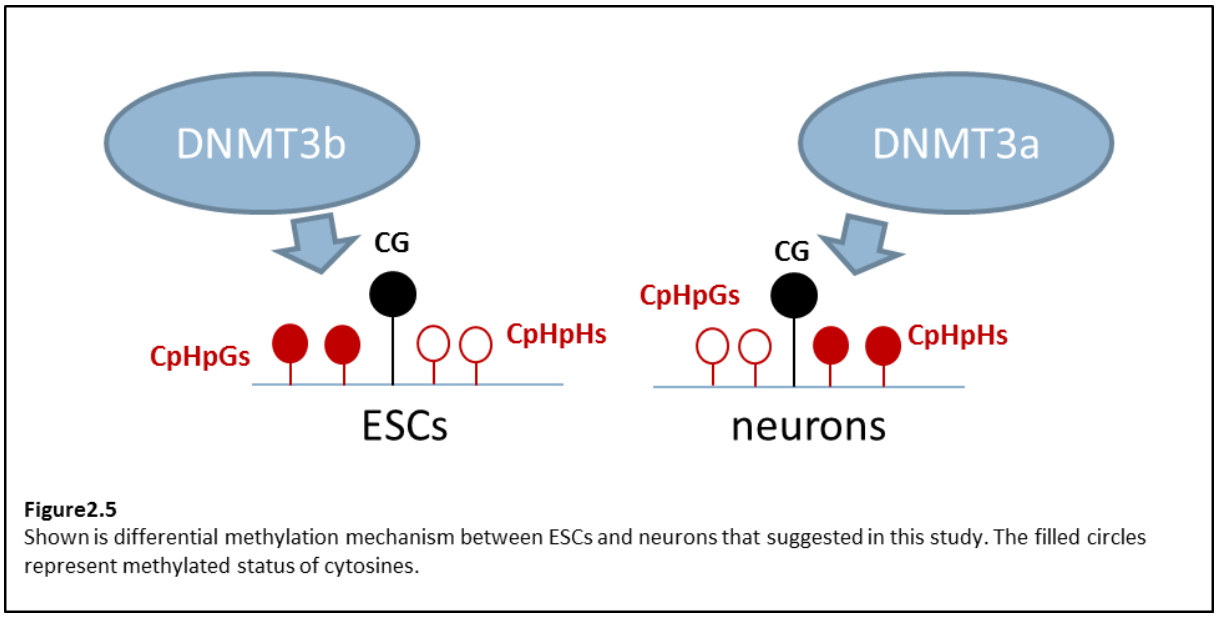


Figure2.5
Shown is differential methylation mechanism between ESCs and neurons that suggested in this study. The filled circles represent methylated status of cytosines.

Section3: Role of mCpH on brain maturation

Abstract

The methylated non-CpGs (CpH; H means A, C, and T) are emerged as a key epigenetic mark in mammalian cell. Despite of the small quantity, those regulate cell type specific functions such as embryonic cell differentiation, or synaptogenesis. Especially, those are gradually accumulated in brain tissues as mammalian aged, and show negative correlation with gene expression level, implying possible epigenetic regulation on brain maturation. However, the specifics of it, such as what kinds of genes or functions are regulated by mCpH are still unclear.

In this study, we attempted to uncover the role of mCpHs on brain maturation. To do that, we analyzed whole genome bisulfite sequencing (WGBS) data of 10 brain tissues with diverse age, combined with that of other tissues and hundreds of microarray data. First, we found that the mCpHs within ±100bp from CpGs are highly depended to the methylation at CpG. Interestingly, in brain, the mCpH is much abundant outside of the regions, implying large portion of mCpHs are independently induced to mCpGs. With the mCpG-independent CpHs, we succeeded in clustering genes sharing gene ontology (GO) terms related to brain specific functions such as “mental retardation”. Altogether, this research shed light on the connection between mCpHs with brain maturation.

Introduction

DNA methylation, an addition of methyl groups on 5th carbon of cytosine, is one of the most well-known epigenetic marks. Since the DNA methylation pattern over genome is highly distinct among cell types, those are considered as identification of cells governing cell development and maintenance [3]. In mammalian cells, the DNA methylation mainly occurs at CpG sites. The methylated CpGs (mCpG) affects cell processes such as genomic imprinting or X-chromosome inactivation [65], and causes disease such as cancer by regulating transcription of genes [66]. However, recent studies have found that significant amount of methylated non-CpGs (CpHs; H means A, C, and T) exist in several cell types. Those are related to cell type specific functions such as cell differentiation [7] and synaptogenesis [6]. In this way, both mCpHs, as well as mCpGs are emerged as important epigenetic mark.

One of the tissues that contains significant amount of mCpHs is brain [2]. Remarkably, the mCpHs are gradually increased as brain aged [6]; whereas mCpGs are remained as stable. Since the increasing pattern is analogous with synaptogenesis, it is assumed that the mCpHs may have roles on brain development. In addition, the average methylation level at CpHs in gene-body regions is negatively correlated with gene expression levels [6, 12]. The negative correlation is more obvious that that between mCpGs and gene expression, even though the quantity of mCpHs is much smaller than that of mCpGs. It implies that the mCpHs, rather than mCpGs, are effectively repressing gene expression in brain. Since the mCpHs exist as mostly ignorable level in other somatic tissues, it is expected that mCpHs are responsible for brain specific cell processes.

However, there is a limitation on discovering the cellular function of mCpHs. Since both mCpGs and mCpHs are induced by common enzymes, DNA methyltransferase 3a and 3b (DNMT3a and DNMT3b), those are spatially correlated to each other [12]. In addition, the average methylation level is greatly higher at CpGs than CpHs, so that it is hard to uncover the independent role of mCpHs. In fact, some researchers insist that mCpHs are stochastic results of mis-matching of DNMTs because of the spatial correlation and great difference of amount between mCpGs and mCpHs [2, 11, 48]. However, substantial evidences, such as the similar increasing pattern with synaptogenesis, points the possible independent role of mCpHs over brain specific cell processes. In this way, the role of mCpH over brain functions is rather controversial, and the specifics of it, such as what kind of genes are affected by mCpHs, is still unclear.

Thus, in this study, we attempted to discover the role of mCpHs over brain functions by extracting mCpG-independent mCpHs. To do that, we collected whole genome bisulfite sequencing data (WGBS) of 10 brain samples and 9 other cells. Through comprehensive analysis, we found that the methylation at CpHs that proximal (±100bp) to CpGs are highly affected by that at CpGs. Therefore, we focused on the CpHs that outside of the regions, named CpG-distal CpHs. Interestingly, the methylation occurs more preferentially at CpG-distal CpHs, whereas it was opposite in control samples such as PSCs, implying that large portion of mCpHs are accumulated independently to

mCpGs in brain. Further analysis found that genes sharing common CpH methylation pattern over ages are related to brain specific functions. The enriched Gene ontology terms (GO terms) of genes clustered by the CpH methylation levels at their gene bodies included brain-function related terms such as “zinc-finger protein activity” or “mental retardation”; whereas the clustering results by CpG methylation pattern did not contain any brain specific terms. It implies that the mCpH affects to brain development or degeneration by regulating genes related zinc-finger proteins or mental retardation. Altogether, our results discovered that mCpHs affects brain functions independently to mCpGs. The study give an insight on discovering mechanisms that CpH methylation regulate cell type specific functions.

Results

1. Large portion of mCpHs are distal from CpGs in brain tissues

To understand the brain specific mCpH pattern, we compared the characteristics of mCpHs in brains with that of other tissues. Since PSCs are another cell type in which mCpHs are abundant , and the average CpH methylation level in PSCs are relatively comparable with that in brains (Supplementary Figure 3.1-a), we collected WGBS of 4 embryonic stem cells (ESCs) and 2 induced pluripotent stem cells (iPSCs) for control sample set. The WGBS reads of 10 brain tissues and 8 PSCs were aligned to human reference genome (hg19) by the integrative approach (Section 1), and the sample clustering result showed high reproducibility of the known characteristics of mCpH (Supplementary Figure 3.1-b).

With the sample set, we confirmed spatial correlation between mCpGs and mCpHs. Similar with the results in Section2, the methylation at CpG and CpH are highly correlated to each other when those are in ±100bp distance (Figure 3.1-a). Also, it was clearly shown that ~180bp nucleosome positioning pattern exist around CpG, implying that CpG and CpH could be simultaneously formed by DNA walking of enzymes such as DNA methyltransferases. Since the significantly high correlation around ±100bp from CpG was common in both brains and PSCs, We divided CpHs into two groups, CpG-proximal and CpG-distal CpHs, based on the threshold of ±100bp from CpG, representing CpG-dependently and CpG-independently formed CpHs, respectively.

Next, we tried to measure the rate of CpG-dependently and independently induced mCpHs. Since the methylation level at CpHs in CpG-proximal regions showed high correlation with that at CpGs, we assumed that the mCpHs in mCpG-proximal regions (beta value of mCpG > 0.8) are induced dependently to the centered mCpGs. Thus, we compared the mCpH level in mCpG-proximal regions with that in CpG-distal regions. Since the mCpHs proximal to un-methylated CpGs are hard to distinguish whether CpG-dependent or independent, we counted those into neither of the two groups (Figure 3.1-b).

Interestingly, the mCpH level on CpG-distal region is higher than that on mCpG-proximal region in brain, whereas it was opposite in PSCs (Supplementary Figure 3.2). We confirmed the tendency with microarray data of 177 brains with 161 PSCs, and found that the brain showed significant abundance of mCpHs at CpG-distal region, whereas PSCs showed at mCpG-proximal region. The mCpH pattern around mCpGs in WGBS data set also showed that mCpHs are concentrated on CpG-proximal region in PSCs, whereas it is reduced in brains (Figure 3.1-d), implying that large portion of CpHs is spatially independent to mCpGs in brain.

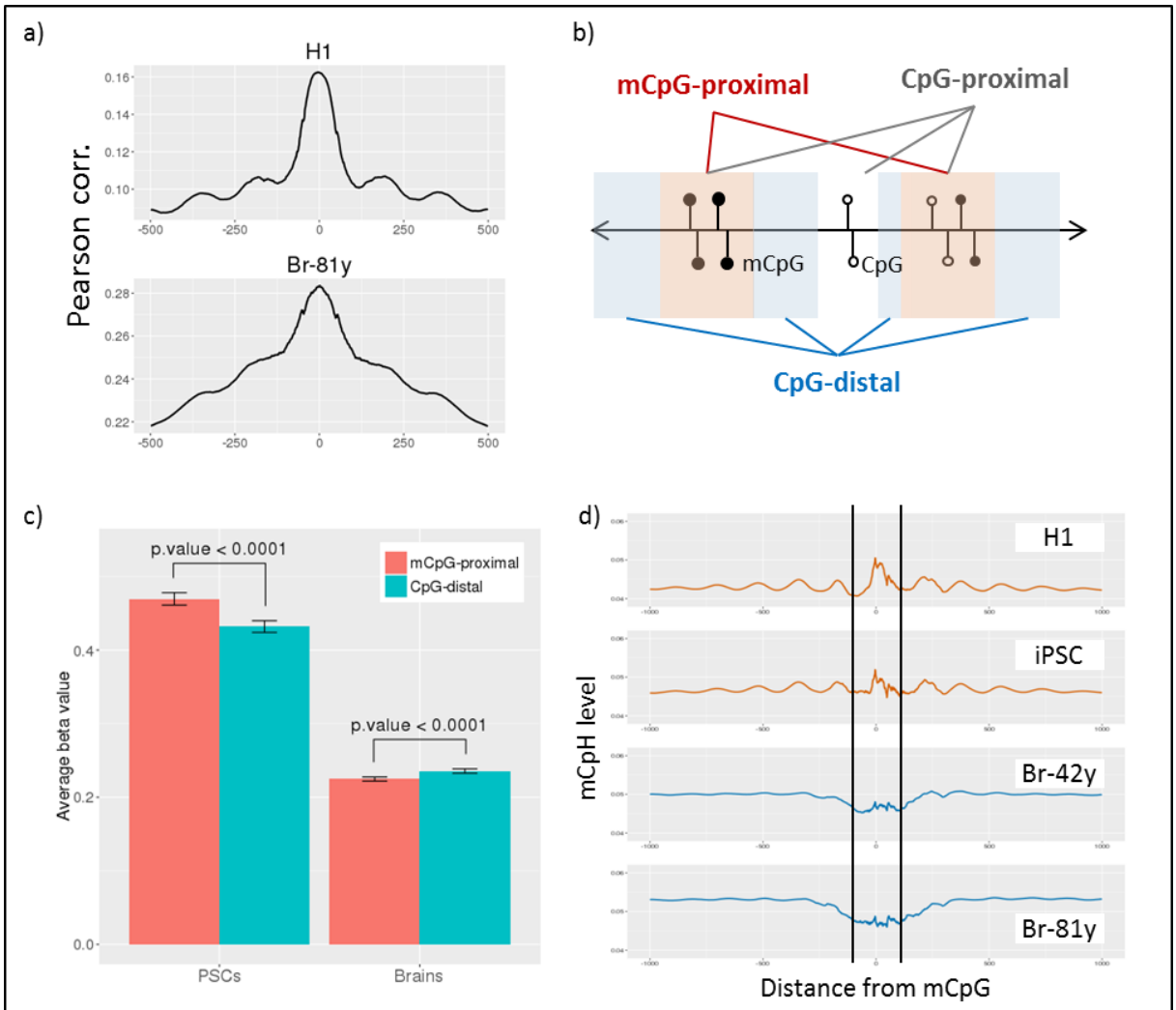


Figure3.1
(a) Shown is Pearson correlation values between methylation levels at CpGs and around CpHs. Since the CpHs are not exist every points around CpGs, we calculated the average methylation level of CpHs within ± 100bp from each distance to CpGs (x-axis). (b) Shown is concept of CpG-proximal, CpG-distal, and mCpG-proximal regions. Here, mCpG means the CpGs at which methylation level is over 0.8. (c) The bar graph represents average CpH methylation level in mCpG-proximal and CpG-distal regions in 177 brains and 161 PSCs. p-value is calculated from paired Wilcox-rank t-test. (d) Shown is the average methylation level at CpHs around mCpG.

2. Genes related to brain specific functions is clustered by the CpG-distal mCpH pattern.

We attempted to discover the function of mCpG-distal mCpHs in brain. Since the methylation at CpH is gradually accumulated as brain aged, and the mCpH level at gene-body region is negatively correlated to the gene expression level, we tried to cluster genes by the pattern of mCpH level in gene bodies across ages. We divided the brain samples into five stages, fetal (Br-fetal), infant (Br-35d), child (Br-2y and 5y), adolescent (Br-12y, 16y, and 25y), and adult (Br-42y, 81y.1, and 81y.2), and analyzed the mCpH pattern across the stages (Method section). Coincident with the genome-wide mCpH levels, most of genes show increasing pattern as brain aged. However, some genes showed partially decreasing pattern, such as decreasing from fetal to infant stages, or from adolescent to adult (Figure 3.2). Remarkably, the genes sharing partial decreasing stages were highly related to

brain specific functions. For example, the GO terms of the genes having decreasing mCpH level from fetal to infant were enriched in “Zinc finger” and “Neuron projection”. In addition, the GO terms of the genes having decreasing mCpH pattern from adolescent to adult were highly enriched with “mental retardation”. However, the gene clusters generated by mCpG pattern were not shown any enriched GO terms related to brain specific functions (Supplementary Figure 3.3).

For removing artifacts on analysis, we did the gene clustering by automated dynamic three cut after hierarchical clustering, then did hierarchical clustering again with GO-terms enriched in any of the cluster (Method section). Although not much of the GO-terms were shared by multiple clusters, there were some GO-terms that enriched in clusters in similar pattern. For example, The GO terms related to embryonic development were enriched in the gene clusters showing gradually increasing pattern of mCpGs (Supplementary Figure 3.4). It is reasonable that the genes related to embryonic development shows low methylation level at their gene bodies in early stage since the methylation normally works as inhibitor of gene expression. It implies that the GO term analysis results well represents overall role of methylation over cell processes. However, there were no enriched GO-terms related to brain specific functions on the genes clustered by mCpG pattern. Meanwhile, the GO term analysis by mCpH showed some enriched terms related to brain specific functions (Figure 3.3). For example, the GO-term of “mental retardation” was enriched in two clusters, and both cluster showed decreasing pattern of mCpH from adolescent to adult stage. In addition, the GO terms related to “zinc finger“, were shared in two clusters, and both reveals decreasing pattern of mCpHs from fetal to infant stage. The zinc finger activity is known to be crucial for brain development [67, 68]. The results imply that the mCpHs, rather than mCpGs, are more responsible for the brain specific functions, especially on brain development and retardation. Altogether, our results revealed that the mCpG-distal mCpHs are related to zinc finger activity and mental retardation.

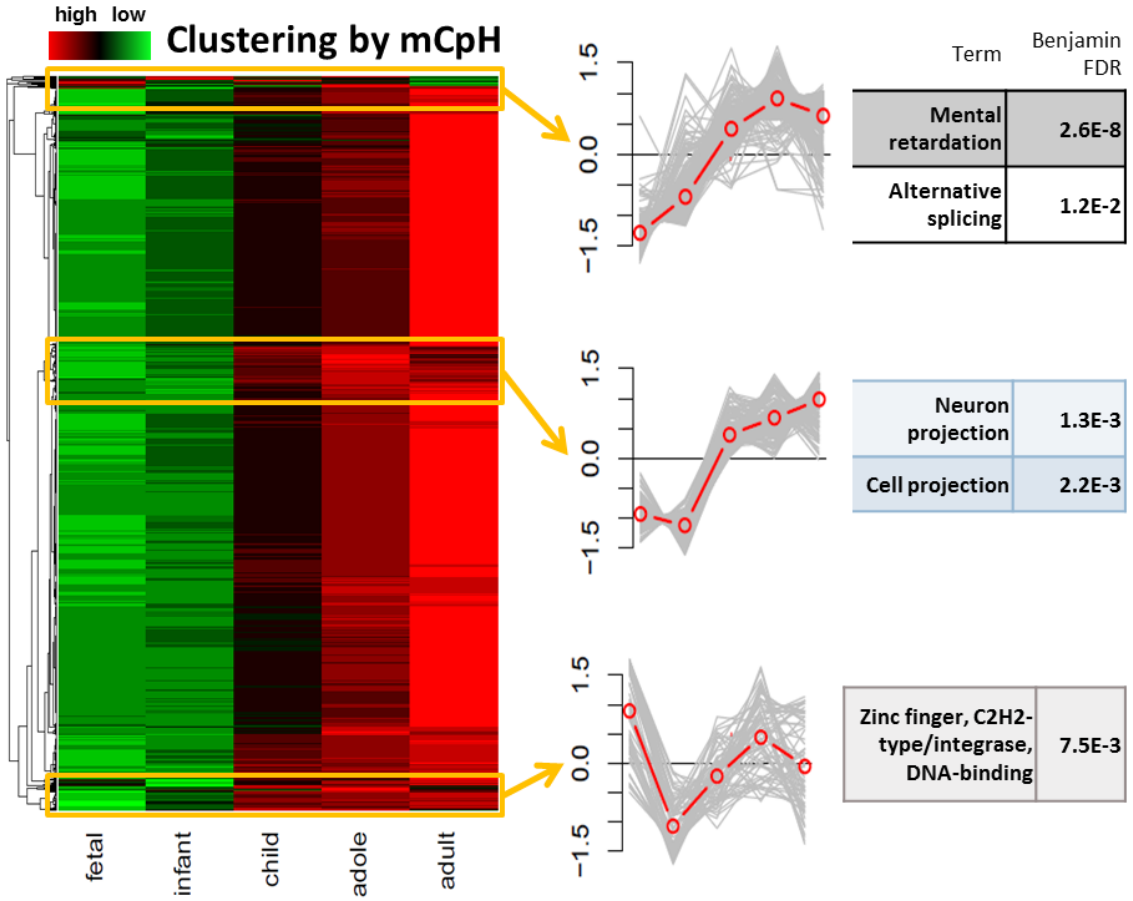


Figure3.2
The heat map shows hierarchical clustering of genes by mCpH pattern across stages. The line graph in the middle shows normalized mCpH levels at each gene bodies (gray lines) and averages of those (red line). Also, the enriched GO term with Benjamin FDR values are shown on the right side.

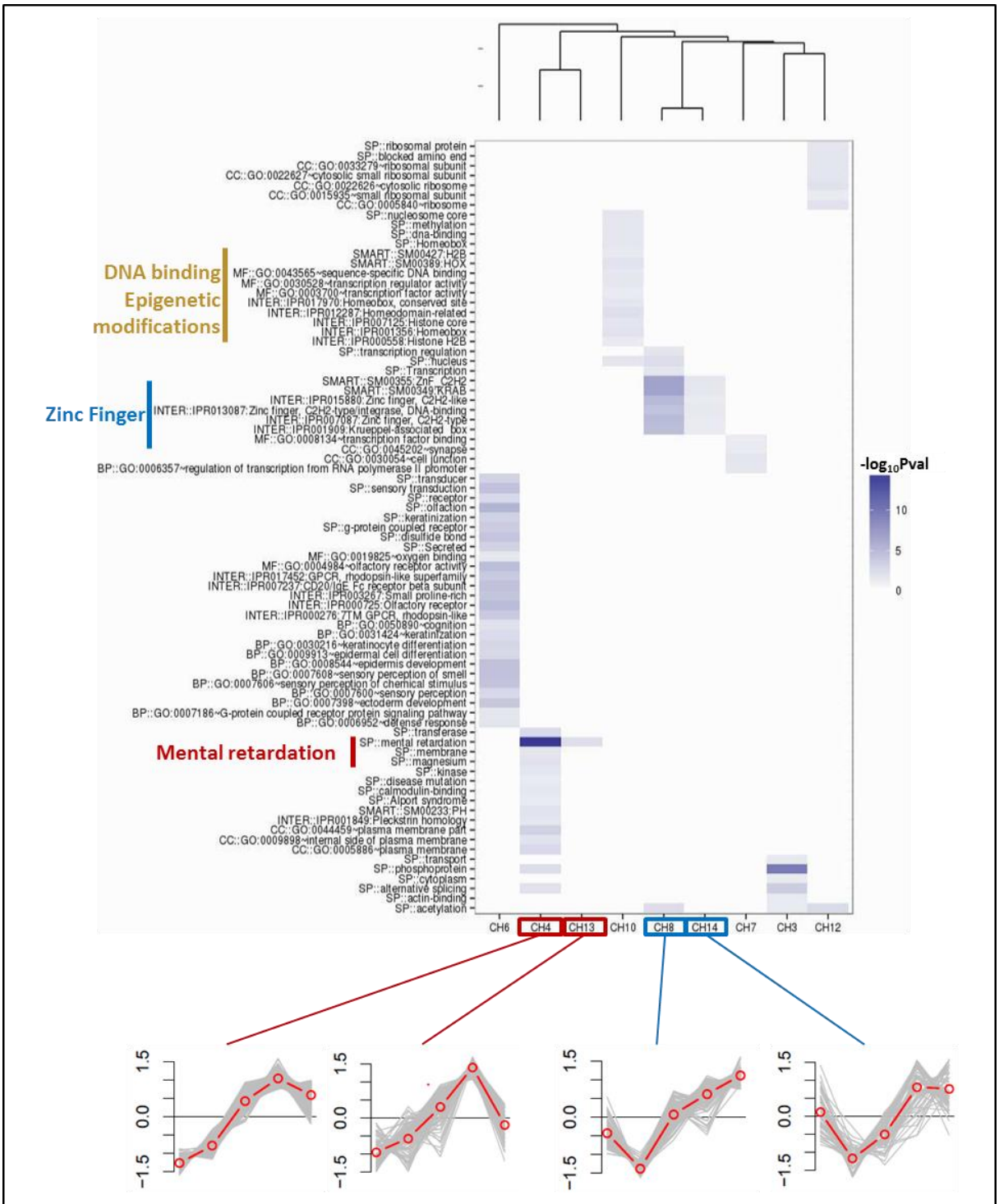


Figure3.3
The heat map shows hierarchical clustering of gene clusters and GO terms. The gene clusters (x-axis) are generated by hierarchical clustering by mCpH level in gene-bodies (Figure3.2). Also, the GO terms are those appeared more than once in the enriched GO term list (Benjamini FDR<0.01). The line graphs at bottom shows the normalized mCpH levels in gene bodies included in each cluster at the 5 age stages (Fetal, Infant, Child, Adolescent, and Adult). Red circle and lines show average value of the grays

Method

WGBS data analysis

The WGBS data of 10 brains and 6 PSCs were analyzed in same way with the previous sections (Section1, and Section2). The public WGBS data used in this section is described in Table 3.1. Also, the identification of the methylated CpH was done in the same way with Section 2. In addition, we used Infinium 450K bisulfite-microarray data for confirming the findings in WGBS dataset (Table 3.2)

Correlation analysis

The spatial correlation between mCpG and mCpH were analyzed as following. First, we re-aligned detected CpHs based on the distance from flanking CpGs. Then, we calculated Pearson correlation coefficient between methylation levels at CpGs with the representative methylation level at CpHs against all the CpGs. Since the CpHs are not always exist at the specified distance from CpG, we calculated the representative methylation level of the CpH as the average methylation level of CpHs within ±100bp from the specified distance to CpG. For example, the Pearson correlation coefficient between methylation levels of CpGs and of CpHs at -4bp from the CpGs was calculated as the correlation between methylation levels at CpGs with average CpH methylation level between -104bp to 96bp from the CpGs.

Gene clustering by methylation pattern

The gene clustering by methylation pattern was conducted as following. First, we calculated the average methylation level at each gene body. The information about the genes such as in which strand the gene exist or the position information about transcript start site (TSS) and transcript termination site (TTS) were from NCBI Reference Sequence Database (Refseq) [69]. In addition, to reduce falsely detected methylation level, we counted genes in which more than 10% of cytosines are detected. Then, we collected age-specifically methylated genes using ANOVA. Through the processes, we extracted 2408 genes for CpG methylation and 14684 for CpH methylation. Then, the genes were hierarchically clustered by z-normalized average methylation values. The Pearson correlation coefficient was used for measuring distance between pairs. Then, the clusters were defined automatically by dynamic cut tree with setting the minimum cluster size as 30, and values for scatter and gab as defaults in R package [70].

Gene ontology clustering

The gene ontology term (GO term) analysis was done by DAVID [71]. Following databases were used for the analysis.

SP: SP_PIR_KEYWORDS

CC: GOTERM_CC_FAT

MF: GOTERM_MF_FAT

BP: GOTERM_BP_FAT

SMART

INTER: INTERPRO

First, we collected all of the enriched terms with setting threshold as Benjamin–Hochberg FDR < 0.05. Then, the hierarchical clustering for the gene sets have conducted against the enriched term list, scored by the Benjamin-Hochberg FDR for each GO term against gene set. Last, we checked whether gene sets sharing similar methylation pattern are properly clustered by the GO terms.

Sample name	**Cell type**	**sex**	**age**	**experiment**	**Bisulfite conv. rate**
H1.1	H1 cell line	M	Passage:30	GSE16256[41]	0.996
H1.2	H1 cell line	M	passasge:25		0.996
H9.1	H9 cell line	F	passage:42		0.995
H9.2	H9 cell line	F	passage:40		0.995
iPSC.1	iPSC 19.11	M	passage:32		0.995
iPSC.2	iPSC 06.09	M	passage:33		0.995
Br-35d	middle frontal gyrus	M	35-day	GSE47966[6]	0.9962
Br-2y	middle frontal gyrus	M	2-year		0.9963
Br-5y	middle frontal gyrus	M	5-year		0.9961
Br-12y	middle frontal gyrus	M	12-year		0.9965
Br-16y	middle frontal gyrus	M	16-year		0.9966
Br-25y	middle frontal gyrus	M	25-year		0.9963
Br-42y	tissue (frontal cortex)	F	42-year	GSE46710[38]	0.9961
Br-81y.1	tissue (frontal cortex)	F	81-year	GSE46644[35]	0.9732
Br-81y.2	tissue (frontal cortex)	F	81-year		0.9742
heart	heart	M	34-year	GSE16256[41]	0.9914
lung	lung	F	30-year		0.992
spleen	spleen	M	34-year		0.9921

Table 3.1: Description of public WGBS data used in Section3

Cell-type	Experiment	Derived cell	# of samples
iPS	GSE59091[72]	dermal fibroblast	109
		erythroblast	22
		endothelial progenitor	12
		foreskin fibroblast	6
ESC		male	3
		female	5
		H9	5
Brain	Brainspan [60]	various age, sex, and structure	177

Table 3.2: Description of public Infinium 450K data

Conclusion

In this paper, we uncovered the function of mCpHs on brain maturation. The process was conducted with largely two steps.

First, we attempted to extract mCpHs that independently induced to mCpGs. The comprehensive analysis about the spatial correlation between mCpGs and mCpHs found that the methylation at CpHs is highly correlated to that at flanking CpGs, especially when those are within ±100bp distance. Although there were positive correlation between mCpGs and mCpHs positioned out of ±100bp, the Pearson coefficient value was ignorable compared to that between in ±100bp. Based on the observed correlation, we divided the whole genome into two regions, CpG-proximal and CpG-distal regions, in which the methylation at CpHs occurs (in)dependently to that at flanking CpGs.

Interestingly, mCpHs were abundant at CpG-distal regions in brains, whereas those in PSCs were abundant at mCpG-proximal regions, implying that the mCpHs in brain may play roles on cell functions independently to mCpGs. This results supports against the insistence of some researchers that the mCpHs are merely stochastic by-product of DNMT's activity that originally targets CpGs. In fact, most of their evidences were from PSCs, such as distributional correlation between mCpGs and mCpHs over whole genome, or higher DNMT3 activities compared to other somatic cells[11, 48]. In this study, however, we weighted on functional independence of mCpHs by uncovering the distributional independency of mCpHs on mCpGs in brain.

Next, we tried to uncover the functional relationship between brain specific cell process and mCpH pattern across varied age. The hierarchical clustering of genes by mCpH pattern among age groups showed that decreasing of mCpH is associated to brain specific functions on that age. For example, the genes showing decreasing mCpH pattern on fetal-to-infant stages were highly related to zinc-finger protein activity, known as key factor of neurogenesis[67, 68]. In addition, the genes sharing decreasing pattern on adolescent-to-adult stages showed high GO term enrichment on "mental retardation". These results implies that the mCpHs in gene-bodies regulates (or are regulated by) the transcription of genes related to brain development or maturation. Meanwhile, the mCpG pattern could not cluster any brain specifically functional genes, implying that the role of mCpHs on brain development is independent to mCpGs.

In summary, we uncovered the functional relationship between mCpHs and brain specific cell process by hierarchical clustering of genes by mCpG-distal mCpHs. The study shed light on research about distribution and function of CpH methylation in brain.

General conclusion

In this study, we attempted to uncover the distribution and function of CpH methylation in mammalian cells via functional analysis *in silico*. The progress consists of three steps. First, we developed a tool for analyzing WGBS data. Second, we discovered the reasons of cell type specific mCpH features in neuron and ESC. Third, we uncovered the role of mCpHs on brain maturation.

In Section 1, we described the limitation of WGBS data analysis and how we improved it. To get the high quality and quantity of methylome from WGBS data, we developed a methylation detection tool. It successfully improved both accuracy of detected methylation level at each cytosine and amount of detected cytosine by integrating outputs of three most widely used bisulfite-read mappers; Bismark2, BSMAP, and BS-seeker2. Through comprehensive analysis of bisulfite read mappers, we found that the three mappers could be mutually complementary against false detection. In addition, the weighting by read depth greatly improved the accuracy of the methylation detection. We confirmed the improvement with both simulated data and real WGBS data.

Through the results, we contributed to methylation study by improving both quantity and quality of methylome from WGBS data, and facilitate reusing of public WGBS data by reducing experimental bias. Especially, the second part, reusing of the public data, is crucial for analyzing CpH methylation. Since WGBS is almost the only way for detecting genome-wide CpH methylation pattern, generating WGBS data is essential step for CpH methylation study. The generation of WGBS data, however, consumes great amount of resources so that many of the previous researches used little number of WGBS samples for deducing their conclusions [6, 13, 38]. We facilitated integration of public WGBS data into one's dataset by reducing experimental bias generated by read heterogeneity. As a result, the following results in Section2 and Section3 were statistically validated by more number of samples than those used in previous studies. Altogether, the study in section 1 sheds light on CpH methylation study by improving the quality and quantity of methylome from WGBS data.

In section 2, we attempted to uncover the reason of distinct characteristics of mCpHs in neuron and ESC. The two cell types are representatives among mammalian cells having significant amount of mCpHs. However, the distribution of mCpHs and potential function of those to transcription are highly distinct in those cells. Through comprehensive analysis, we found that DNMT3a and DNMT3b preferentially methylate CpHpH and CpHpG contexts, respectively, resulting in distinct characteristics of mCpHs in ESCs and neurons. For example, the abundant motif at mCpHs in brain is "CAC", whereas that is "CAG" in ESCs, because of the differential activity of DNMT3a and DNMT3b in those two cell types. It may give hints for understanding cell type specific mCpH regulation. For example, the MeCP2 is known to interact with CpAs, as well as CpGs, and damaging on MeCP2 induces brain diseases such as Rett syndrome [12]. The fact that DNMT3a preferentially targets CpHpHs may give any hints for researching MeCP2 binding or related diseases. In addition, we found cell type specific mCpH patterns around CpGs by focusing on CpG-proximal regions. In addition, we deduced significant difference of methylation level at CpHpH and CpHpG contexts in DNMT3a- and DNMT3b-reinduced

mouse by focusing on the CpG-proximal regions. The results were not reported in the paper [53] produced the WGBS data since the difference is not shown as significant if measuring over whole genome rather than focusing on CpG-proximal regions. In addition, the main focus of the previous study was not CpH methylation, but CpG methylation. In summary, we succeeded in capturing significant difference of CpH methylation in ESCs and neurons by applying the integrative approach, introduced in section 1, and by focusing on CpG-proximal regions.

Another important found was the mechanism of the positive correlation between mCpHs with gene expression level in ESCs. Since the mCpHs shows negative correlation with gene expression in general, the positive correlation was mysterious among researchers. By analyzing wild type, DNMT1/3a/3b-knocked out (DNMT-TKO), that with DNMT3b-reinduced (DNMT-TKO-3b), and that with SETD2 knock out (DNMT-TKO-3b-SETD2KO) mouse ESCs, we found that the DNMT3b and H3k36me3 histone mark, induced by SETD2, is crucial for the positive correlation between mCpHs and gene expression. Altogether, we concluded that the preferential interaction between DNMT3b with H3k36me3 histone mark in highly expressed gene bodies, by methylating CpHpGs on those, causes the positive correlation between CpH methylation level and gene expression level. The results in section 2 contributed to CpH methylation study by uncovering the mechanism of cell type specific CpH methylation and potential function to gene expression.

Lastly, in section 3, we unearthed the role of mCpHs on brain maturation. To do that, we firstly attempted to extract CpHs that methylated independently to the methylation at CpGs. Through comprehensive analysis, we found that the methylation at CpHs was highly correlated to that at CpGs when those are within ±100bp-distance. Interestingly, the methylation occurs at CpHs positioned out of the ±100bp from CpGs in brain, implying that large portion of CpHs is independently methylated to CpGs.

With the CpG-distal mCpHs in gene bodies, we succeeded in clustering genes related to brain specific processes. The genes sharing common decreasing pattern across ages showed highly enriched GO terms of “zinc finger protein activity” and “mental retardation”. Especially, the zinc-finger protein activity is known as crucial for neurogenesis in early brains. The result implies that the mCpHs are affecting (or being affected by) the brain development and retardation. Meanwhile, the clustering results by mCpGs did not contain any brain-specific GO terms, implying that the mCpHs are solely affecting (or being affected by) brain specific functions. In fact, the role of CpH methylation over cell functions has been suspicious because of the spatial correlation between mCpGs and mCpHs. Since the mCpHs are abundant at nearby mCpGs, some researchers insisted that the CpH methylation is merely stochastic event of mis-capturing by DNMTs. However, in this study, we weighed on the role of mCpHs over cell functions by extracting CpG-distal mCpHs and uncovering the brain specific GO terms enriched in genes clustered by the CpG-distal mCpH pattern across ages. Altogether, the results in section 3 shed light on the roles of CpH methylation level over cell type specific functions.

In summary, this study attempted to understand the function of CpH methylation over cell type specific processes by improving quantity and quality of methylome, uncovering the mechanism of cell

type specific mCpH characteristics, and revealing role of mCpHs over brain maturation. The analytic methods and results will greatly contribute to understand the mechanism of mCpHs governing cell processes.

References

1. Ehrlich, M., et al., *Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells.* Nucleic Acids Res, 1982. **10**(8): p. 2709-21.
2. He, Y. and J.R. Ecker, *Non-CG Methylation in the Human Genome.* Annu Rev Genomics Hum Genet, 2015. **16**: p. 55-77.
3. Varley, K.E., et al., *Dynamic DNA methylation across diverse human cell lines and tissues.* Genome Res, 2013. **23**(3): p. 555-67.
4. Meissner, A., et al., *Genome-scale DNA methylation maps of pluripotent and differentiated cells.* Nature, 2008. **454**(7205): p. 766-70.
5. Laurent, L., et al., *Dynamic changes in the human methylome during differentiation.* Genome Res, 2010. **20**(3): p. 320-31.
6. Lister, R., et al., *Global epigenomic reconfiguration during mammalian brain development.* Science, 2013. **341**(6146): p. 1237905.
7. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-22.
8. Zhang, J. and Y.G. Zheng, *SAM/SAH Analogs as Versatile Tools for SAM-Dependent Methyltransferases.* ACS Chem Biol, 2016. **11**(3): p. 583-97.
9. Smith, Z.D. and A. Meissner, *DNA methylation: roles in mammalian development.* Nat Rev Genet, 2013. **14**(3): p. 204-20.
10. Xie, W., et al., *Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome.* Cell, 2012. **148**(4): p. 816-31.
11. Ziller, M.J., et al., *Genomic distribution and inter-sample variation of non-CpG methylation across human cell types.* PLoS Genet, 2011. **7**(12): p. e1002389.
12. Guo, J.U., et al., *Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain.* Nat Neurosci, 2014. **17**(2): p. 215-22.
13. Lister, R., et al., *Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells.* Nature, 2011. **471**(7336): p. 68-73.
14. Shirane, K., et al., *Mouse oocyte methylomes at base resolution reveal genome-wide accumulation of non-CpG methylation and role of DNA methyltransferases.* PLoS Genet, 2013. **9**(4): p. e1003439.
15. Morris, T.J. and S. Beck, *Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data.* Methods, 2015. **72**: p. 3-8.
16. Weber, M., et al., *Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells.* Nat Genet, 2005. **37**(8): p. 853-62.
17. Meissner, A., et al., *Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis.* Nucleic Acids Res, 2005. **33**(18): p. 5868-77.

18. Comfort, N., *Essay reviews. [Review of: Rabinow P. Making PCR: a story of biotechnology. University of Chicago Press, 1996; and Fujimura J. Crafting science: a sociohistory of the quest for the genetics of cancer. Harvard University Press, 1996].* Oral Hist Rev, 1999. **26**(2): p. 181-6.
19. Bock, C., *Analysing and interpreting DNA methylation data.* Nat Rev Genet, 2012. **13**(10): p. 705-19.
20. Xi, Y. and W. Li, *BSMAP: whole genome bisulfite sequence MAPping program.* BMC Bioinformatics, 2009. **10**: p. 232.
21. Krueger, F. and S.R. Andrews, *Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.* Bioinformatics, 2011. **27**(11): p. 1571-2.
22. Guo, W., et al., *BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data.* BMC Genomics, 2013. **14**: p. 774.
23. Day, K., et al., *Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape.* Genome Biol, 2013. **14**(9): p. R102.
24. Horvath, S., *DNA methylation age of human tissues and cell types.* Genome Biol, 2013. **14**(10): p. R115.
25. Baylin, S.B., *DNA methylation and gene silencing in cancer.* Nat Clin Pract Oncol, 2005. **2 Suppl 1**: p. S4-11.
26. Alper, B.S., *SOAP: solutions to often asked problems. Choice of antihistamines for urticaria.* Arch Fam Med, 2000. **9**(8): p. 748-51.
27. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
28. Chatterjee, A., et al., *Comparison of alignment software for genome-wide bisulphite sequence data.* Nucleic Acids Res, 2012. **40**(10): p. e79.
29. Tran, H., et al., *Objective and comprehensive evaluation of bisulfite short read mapping tools.* Adv Bioinformatics, 2014. **2014**: p. 472045.
30. Kunde-Ramamoorthy, G., et al., *Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing.* Nucleic Acids Res, 2014. **42**(6): p. e43.
31. Sherman - bisulfite-treated Read FastQ Simulator: http://www.bioinformatics.babraham.ac.uk/projects/sherman/
32. Grover, D., et al., *Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition.* Bioinformatics, 2004. **20**(6): p. 813-7.
33. Picard: http://broadinstitute.github.io/picard/
34. Samtools rmdup: http://samtools.sourceforge.net/
35. Ziller, M.J., et al., *Charting a dynamic DNA methylation landscape of the human genome.* Nature, 2013. **500**(7463): p. 477-81.
36. Guo, W., et al., *Characterizing the strand-specific distribution of non-CpG methylation in*

human pluripotent cells. Nucleic Acids Res, 2014. **42**(5): p. 3009-16.

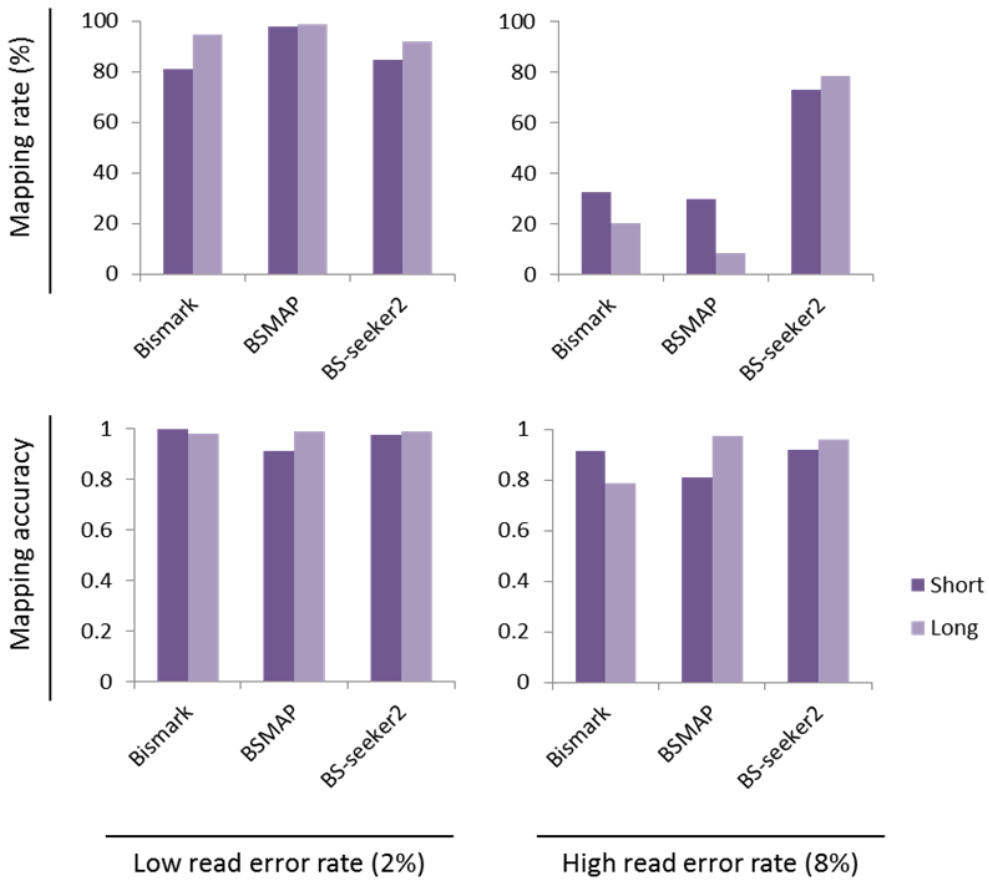
37. http://hannonlab.cshl.edu/fastx_toolkit/index.html
38. Wen, L., et al., *Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain.* Genome Biol, 2014. **15**(3): p. R49.
39. Court, F., et al., *Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment.* Genome Res, 2014. **24**(4): p. 554-69.
40. Whole Genome Bisulfite Sequencing by ENCODE/HAIB [http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40832]
41. UCSD Human Reference Epigenome Mapping Project [http://www.roadmapepigenomics.org/]
42. Holliday, R. and J.E. Pugh, D*NA modification mechanisms and gene activity during development.* Science, 1975. 1**87**(4173): p. 226-32.
43. Vanyushin, B.F. and V.V. Ashapkin, D*NA methylation in higher plants: past, present and future.* Biochim Biophys Acta, 2011. 1**809**(8): p. 360-8.
44. Haines, T.R., D.I. Rodenhiser, and P.J. Ainsworth, A*llele-specific non-CpG methylation of the Nf1 gene during early mouse development.* Dev Biol, 2001. 2**40**(2): p. 585-98.
45. Cao, X., et al., R*ole of the DRM and CMT3 methyltransferases in RNA-directed DNA methylation.* Curr Biol, 2003. 1**3**(24): p. 2212-7.
46. Cao, X. and S.E. Jacobsen, L*ocus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes.* Proc Natl Acad Sci U S A, 2002. 9**9 Suppl 4:** p. 16491-8.
47. Wada, Y., et al., P*referential de novo methylation of cytosine residues in non-CpG sequences by a domains rearranged DNA methyltransferase from tobacco plants.* J Biol Chem, 2003. 2**78**(43): p. 42386-93.
48. Arand, J., et al., I*n vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases.* PLoS Genet, 2012. 8(6): p. e1002750.
49. Liao, J., et al., T*argeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells.* Nat Genet, 2015. 4**7**(5): p. 469-78.
50. Holz-Schietinger, C. and N.O. Reich, T*he inherent processivity of the human de novo methyltransferase 3A (DNMT3A) is enhanced by DNMT3L.* J Biol Chem, 2010. 2**85**(38): p. 29091-100.
51. Martins-Taylor, K., et al., R*ole of DNMT3B in the regulation of early neural and neural crest specifiers.* Epigenetics, 2012. 7(1): p. 71-82.
52. Li, Z., et al., D*istinct roles of DNMT1-dependent and DNMT1-independent methylation patterns in the genome of mouse embryonic stem cells.* Genome Biol, 2015. 1**6:** p. 115.
53. Baubec, T., et al., G*enomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation.* Nature, 2015. 5**20**(7546): p. 243-7.

54. Chadwick, L.H., T*he NIH Roadmap Epigenomics Program data resource.* Epigenomics, 2012. 4(3): p. 317-24.
55. Lee, J.H., S.J. Park, and N. Kenta, A*n integrative approach for efficient analysis of whole genome bisulfite sequencing data.* BMC Genomics, 2015. 1**6 Suppl 12:** p. S14.
56. Zvetkova, I., et al., G*lobal hypomethylation of the genome in XX embryonic stem cells.* Nat Genet, 2005. 3**7**(11): p. 1274-9.
57. Ferguson-Smith, A.C. and J.M. Greally, E*pigenetics: perceptive enzymes.* Nature, 2007. 4**49**(7159): p. 148-9.
58. Kim, D., et al., T*opHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.* Genome Biol, 2013. 1**4**(4): p. R36.
59. Trapnell, C., et al., T*ranscript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. 2**8**(5): p. 511-5.
60. BrainSpan: Atlas of the Developing Human Brain [Internet]. Funded by ARRA Awards 1RC2MH089921-01, 1RC2MH090047-01, and 1RC2MH089929-01. © 2011. Available from: http://developinghumanbrain.org.
61. Consortium, E.P., An *integrated encyclopedia of DNA elements in the human genome. Na*ture, 2012. 489(**74**14): p. 57-74.
62. Yang, L., et al., Gen*omewide characterization of non-polyadenylated RNAs. Ge*nome Biol, 2011. 12(**2**): p. R16.
63. Chu, L.F., et al., Sin*gle-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. Ge*nome Biol, 2016. 17(**1**): p. 173.
64. Suetake, I., et al., Dis*tinct enzymatic properties of recombinant mouse DNA methyltransferases Dnmt3a and Dnmt3b. J* Biochem, 2003. 133(**6**): p. 737-44.
65. Sharp, A.J., et al., DNA *methylation profiles of human active and inactive X chromosomes. Ge*nome Res, 2011. 21(**10**): p. 1592-600.
66. Li, W. and B.F. Chen, Abe*rrant DNA methylation in human cancers. J* Huazhong Univ Sci Technolog Med Sci, 2013. 33(**6**): p. 798-804.
67. Bellefroid, E.J., et al., X-M*yT1, a Xenopus C2HC-type zinc finger protein with a regulatory function in neuronal differentiation. Ce*ll, 1996. 87(**7**): p. 1191-202.
68. Morris, D.R. and C.W. Levenson, Zin*c regulation of transcriptional activity during retinoic acid-induced neuronal differentiation. J* Nutr Biochem, 2013. 24(**11**): p. 1940-4.
69. O'Leary, N.A., et al., Ref*erence sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nu*cleic Acids Res, 2016. 44(**D1**): p. D733-45.
70. Langfelder, P., B. Zhang, and S. Horvath, Def*ining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bi*oinformatics, 2008. 24(**5**): p. 719-20.
71. Huang da, W., B.T. Sherman, and R.A. Lempicki, Sys*tematic and integrative analysis of*

*large gene lists using DAVID bioinformatics resources. Na*t Protoc, 2009. 4(1): p. 44-57.

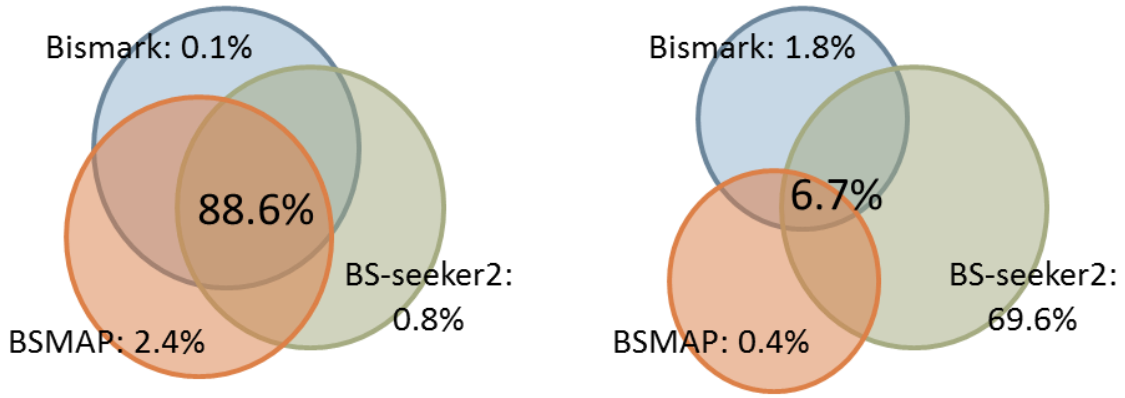
72. Butcher, L.M., et al., Non-*CG DNA methylation is a biomarker for assessing endodermal differentiation capacity in pluripotent stem cells. Na*t Commun, 2016. 7: **p**. 10458.

Supplementary Figures

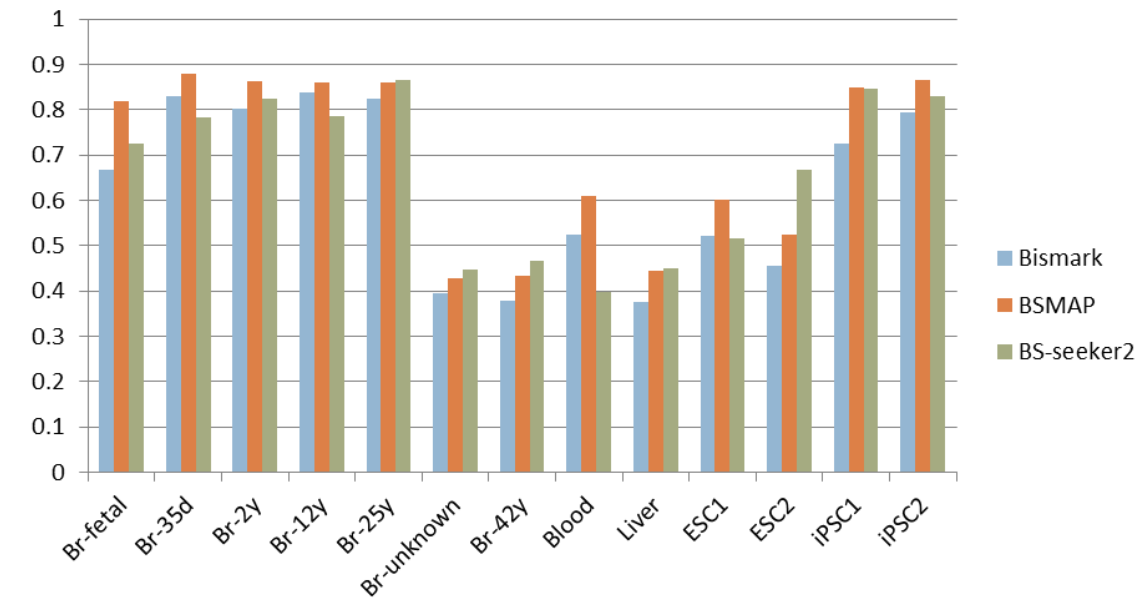


Supplementary Figure 1.1 Mapping results of the three mappers with long (100bp) and short (50bp) reads

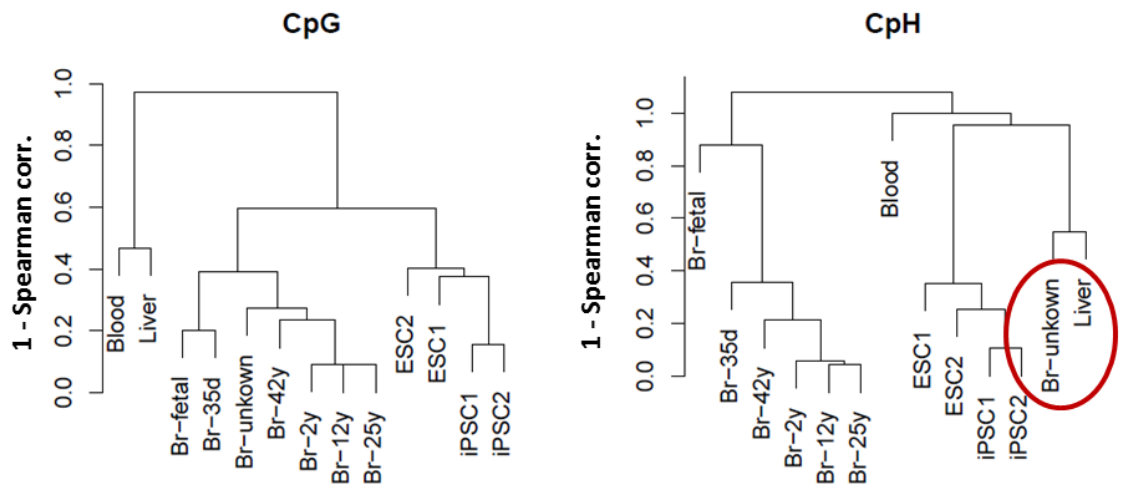
The bars show mapping rate and mapping accuracy of the three mappers with reads that contains 2% error (a and c for mapping rate and mapping accuracy, respectively) and 8% error (b and d for mapping rate and mapping accuracy, respectively). The dark purple bars represent results with short (50bp) reads and the light purple bars represent results with long (100bp) reads.



Supplementary Figure 1.2 Rate of correctly mapped reads by three mappers
The numbers show rate of reads that correctly mapped by each three mapper over total read number, when read error rate equals to 2% (a) and 8% (b). The numbers in middle reveal rate of reads that correctly mapped by all three mappers. Also the numbers followed by each mapper shows rate of reads that correctly mapped only by the mapper.

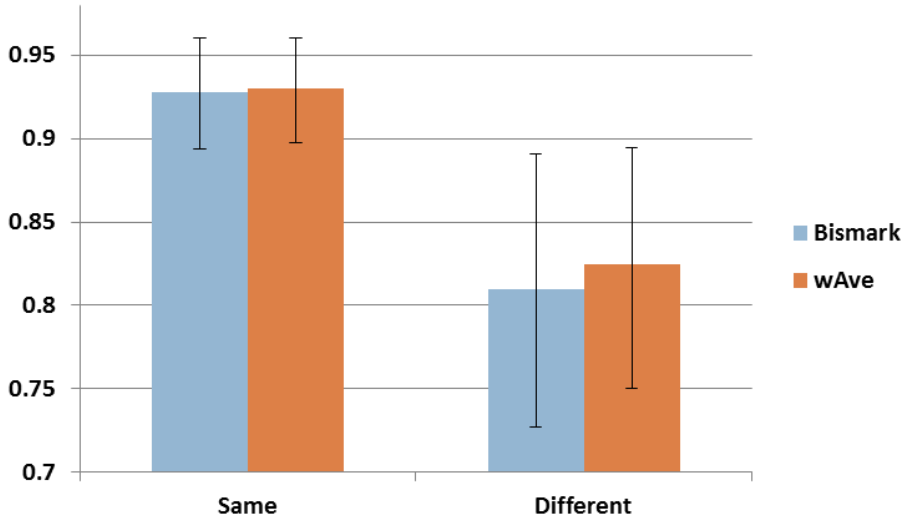


Supplementary Figure 1.3 Mapping rate of whole genome bisulfite sequencing data
It shows mapping rates of WGBS samples by the three mappers; ESC=Embryonic stem cell, iPSC=induced pluripotent stem cell, Br=brain, d=day and y=year (cf. Br-5y means 5 years old brain)



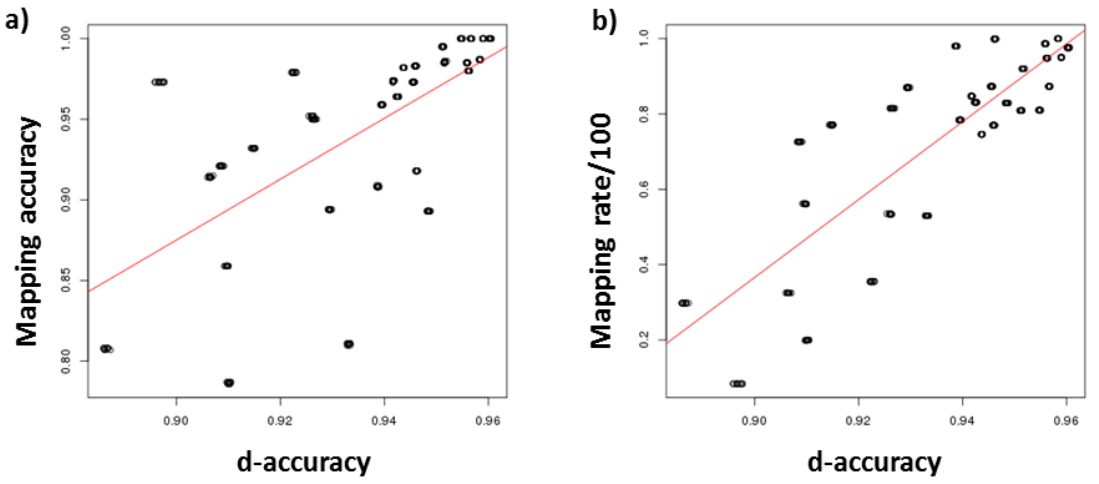
Supplementary Figure 1.4 Hierarchical clustering results base on CpG and CpH methylation levels extracted by BS-seeker2

Hierarchical clustering results base on CpG and CpH methylation levels extracted by BS-seeker2; Distance is 1−spearman correlation coefficient. ESC=Embryonic stem cell, iPSC=induced pluripotent stem cell, Br=brain, d=day and y=year (cf. Br-5y means 5 years old brain). Also, the red circle groups the two samples that produced by same experiment.

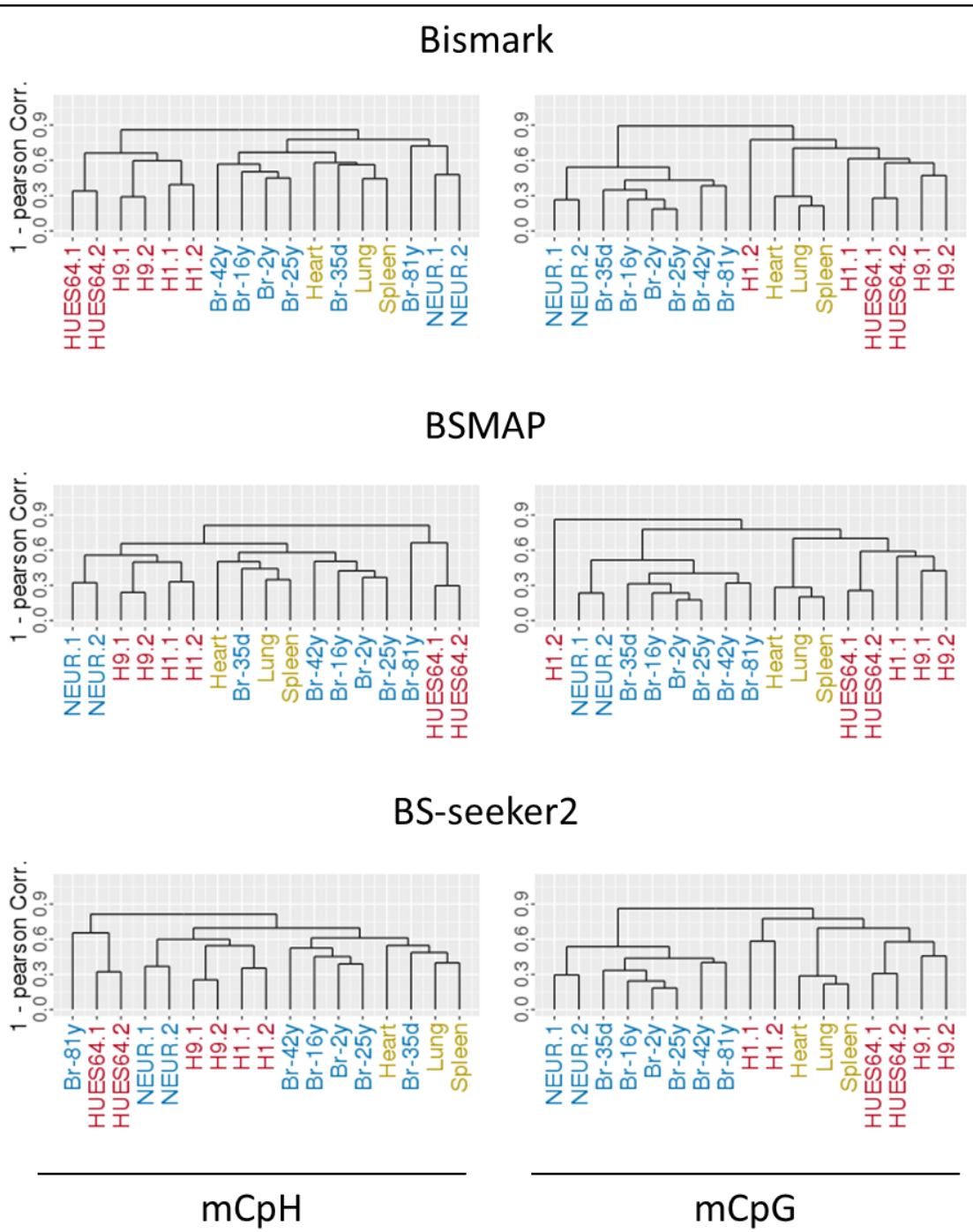


Supplementary Figure 1.5 Correlation of CpG methylation levels among brain samples that produced from same experiment and different experiments

Spearman correlation of CpG methylation in gene-body regions between brain samples that produced from same experiment and multiple experiments (7 samples). Error bars represent maximum and minimum correlation value between samples. The information of gene-body regions was downloaded from refseq database.

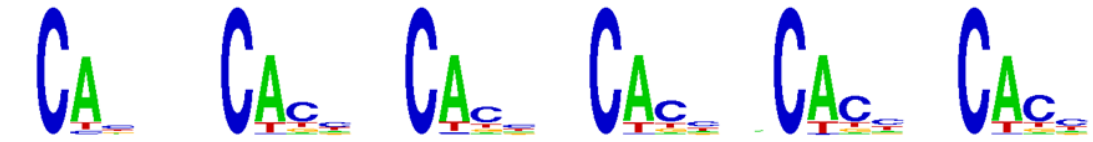


Supplementary Figure 1.6 Correlation of detection accuracy with mapping accuracy and mapping read It shows proportional relationship between detection accuracy with mapping accuracy (a), and detection accuracy with mapping rate (b). Each point represents mapping results by Bismark, BSMAP and BS-seeker2 with read sets in which the read error rates equal to 0%, 2%, 4%, 6% and 8%, and read lengths equal to 50bp and 100bp.



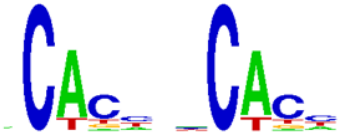
Supplementary Figure 2.1 Hierarchical clustering results by mCpG and mCpH patterns extracted by bisulfite mappers, Bismark, BSMAP, and BS-seeker2.
The hg19 refseq genome reference was divided into 1k-bp blocks, and the Pearson correlation coefficients was calculated by comparing the average mCpG and mCpH levels at each blocks among samples. The coverage depth for all three read mappers is ≧5． Colors in x-axis represents same cell (or tissue) types (blue: brain and neuron, red: ESCs, and yellow: other tissues). The numbers followed by Br- means age that the brain tissue is sequenced (d and y mean day and year, respectively). Also, NEUR represents neuron. The numbers following sample names represents biological replica

Brain tissues



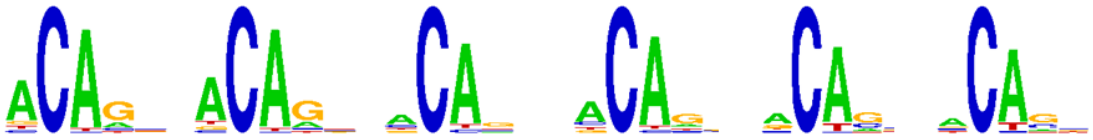
Br-35d Br-2y Br-16y Br-25y Br-42y Br-81y

Neurons



NEUR.1 NEUR.2

ES cell lines



H1.1 H1.2 H9.1 H9.2 HUES64.1 HUES64.2

Other tissues

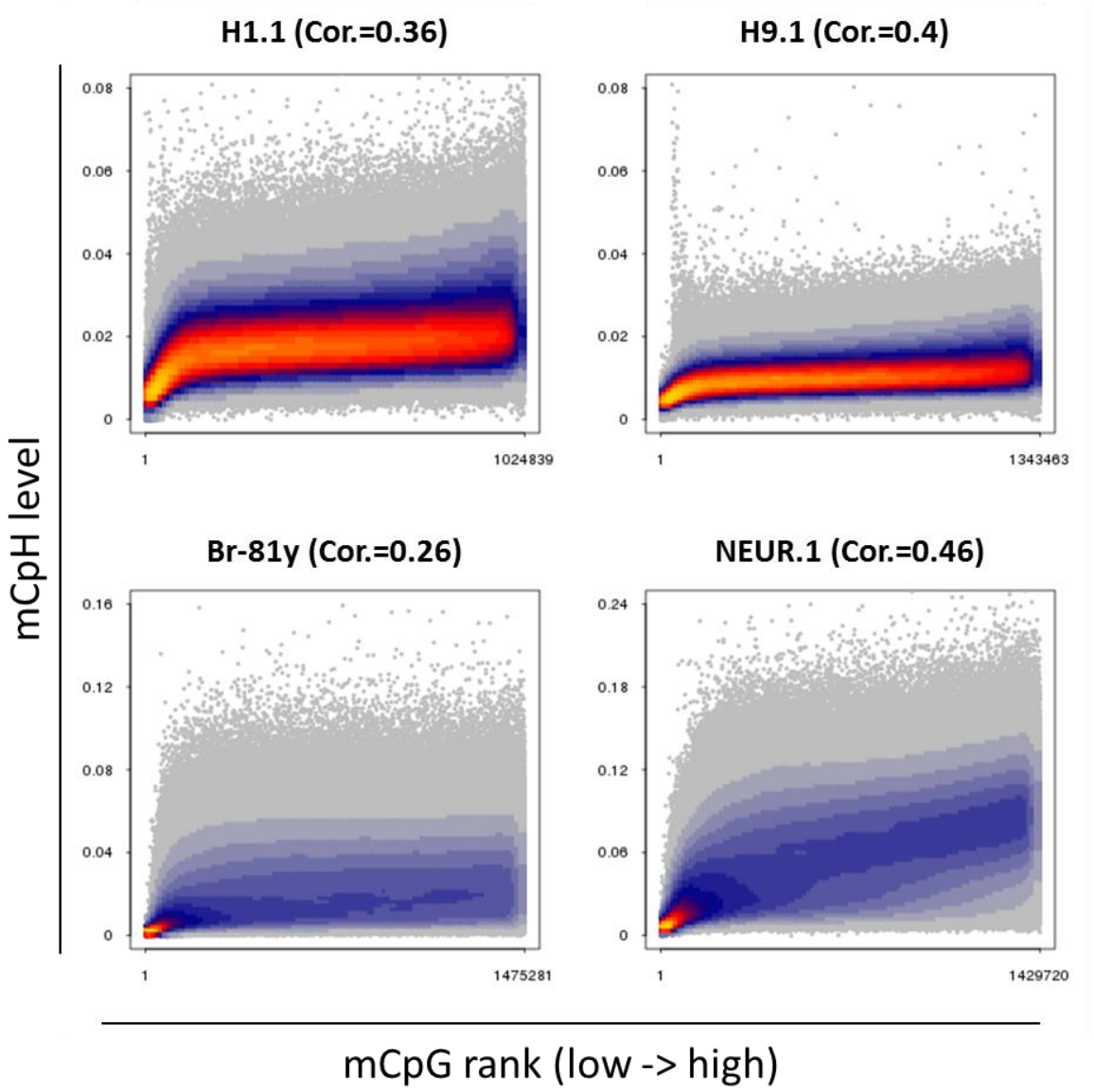
Spleen Lung Heart

DNMT knock out HUES64s

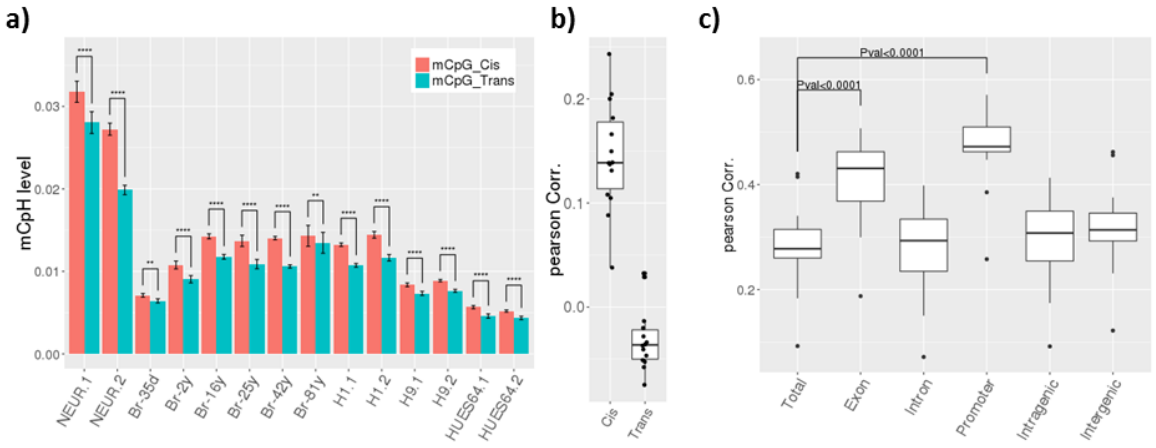
DNMT3a-KO DNMT3b-KO DNMT-DKO

Supplementary Figure 2.2 DNA motifs around mCpHs

Shown DNA motifs are abundant nearby hyper-methylated CpHs (-1bp to +3bp from cytosines at CpH contexts; beta value>0.5). The software "weblogo3", in which the height of the characters are defined as the posterior mean relative entropy through Bayesian calculation. The average number of CpHs for detecting the motifs is 3,290,487.

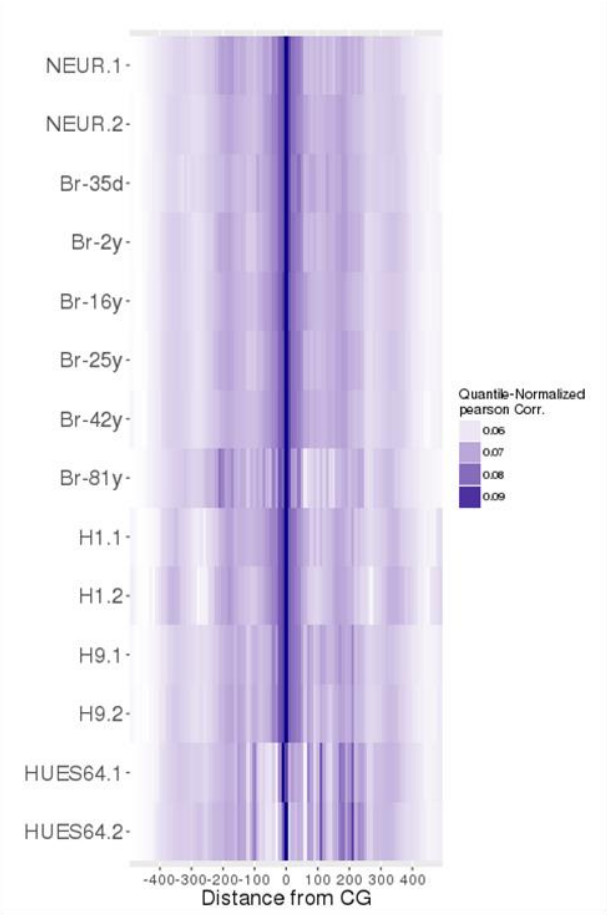


Supplementary Figure 2.3 The correlation between mCpG and mCpH
The scatter plots shows average CpH methylation levels (mCpH level) in 1k-bp blocks that ranked by average CpG methylation level . For robust methylation values, blocks in which more than 10 CpGs and CpHs exist were counted. X-axis shows the number of counted blocks. Colors represent density of points; yellow represents the highest density. Also, Cor. represents Pearson correlation coefficients between mCpH and mCpG levels across blocks. This graph has been drawn by R package, 'LSD'.



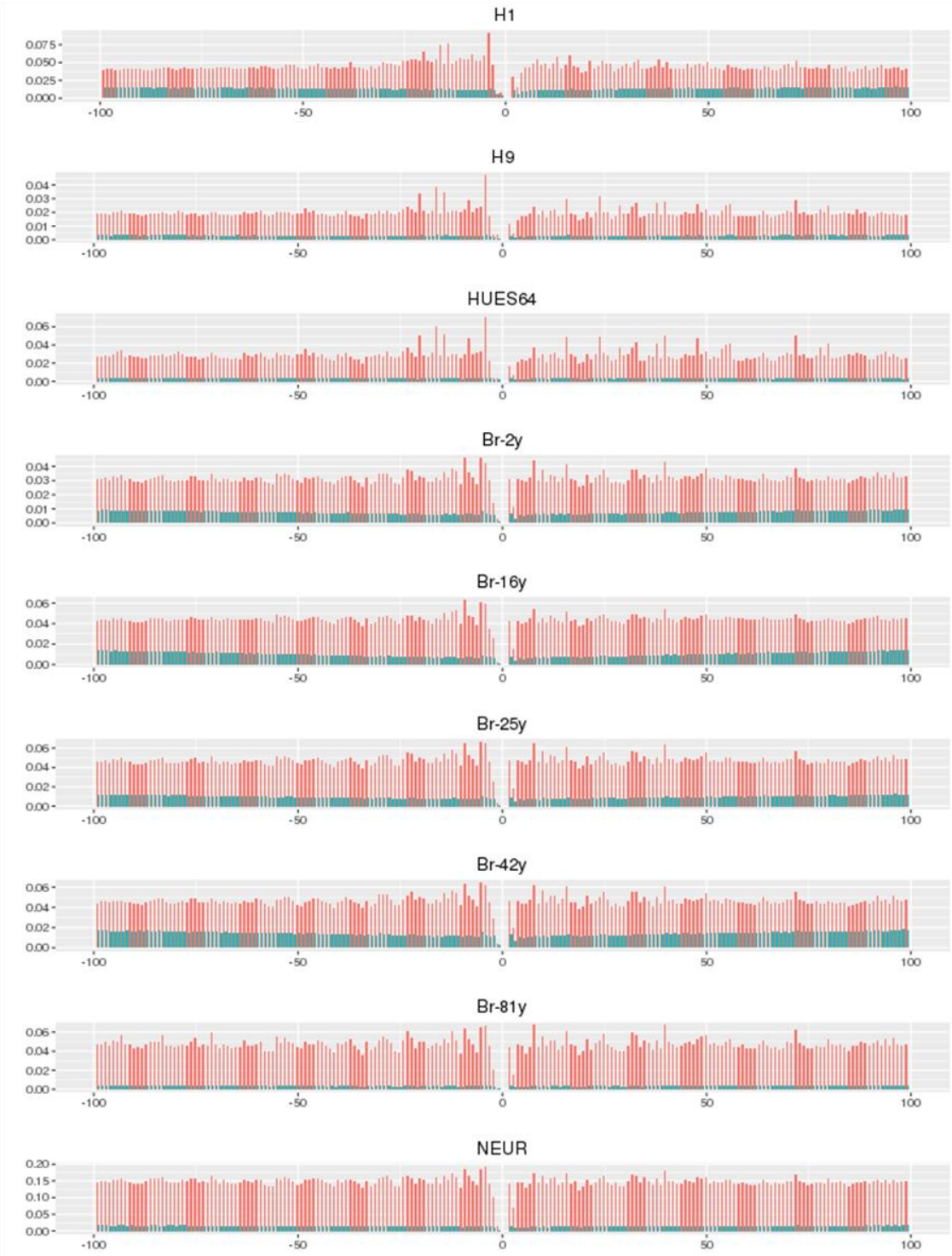
Supplementary Figure 2.4

(a) Bar graphs show average mCpH levels in blocks where CpGs are hemi-methylated (difference mCpG levels between strands ≧0.5). Error bars represents standard error among the blocks. Stars up on the bars represents p-values from Wilcoxon Rank Sum test; *,**,***,and **** represents that p-value is under 0.1, 0.01, 0.001, and 0.0001, respectively. (b) Box plot represents Pearson correlation coefficient between mCpG and mCpH levels across CpG-hemi-methylated blocks in all samples listed up at x-axis in graph (a). Cis represents the correlation between mCpG and mCpH in same strand, and Trans represents that in opposite strands. The upside and underside of the boxes represents 25% and 75% of the sample set, respectively. (c) The Boxplots represents the Pearson correlation between mCpG and mCpH levels across blocks in each genic regions. The annotation of the regions are from 'refseq', especially, the promoter region means transcription start sites (TSS) ±5000bp.

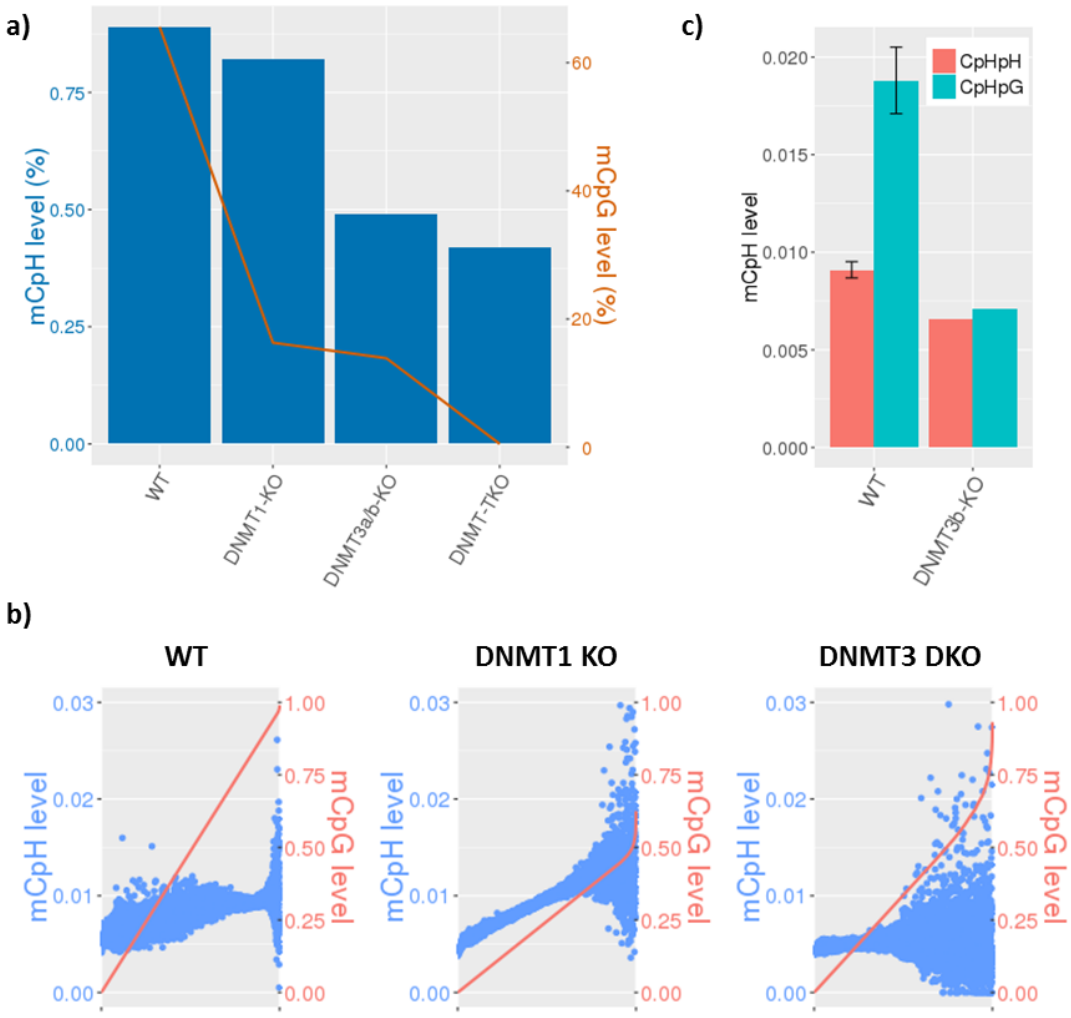


Supplementary Figure 2.5

Shown is the correlation between mCpG level and mCpH level at the specified distance from CpG. To calculate the Pearson correlation coefficients, the CpHs around CpGs (±500bp) were allocated into 10bp-long blocks according to the distance from CpG, and compared the mCpG level at the center with average mCpH level at each block. To show clear tendency, the correlation coefficients were quantile-normalized.

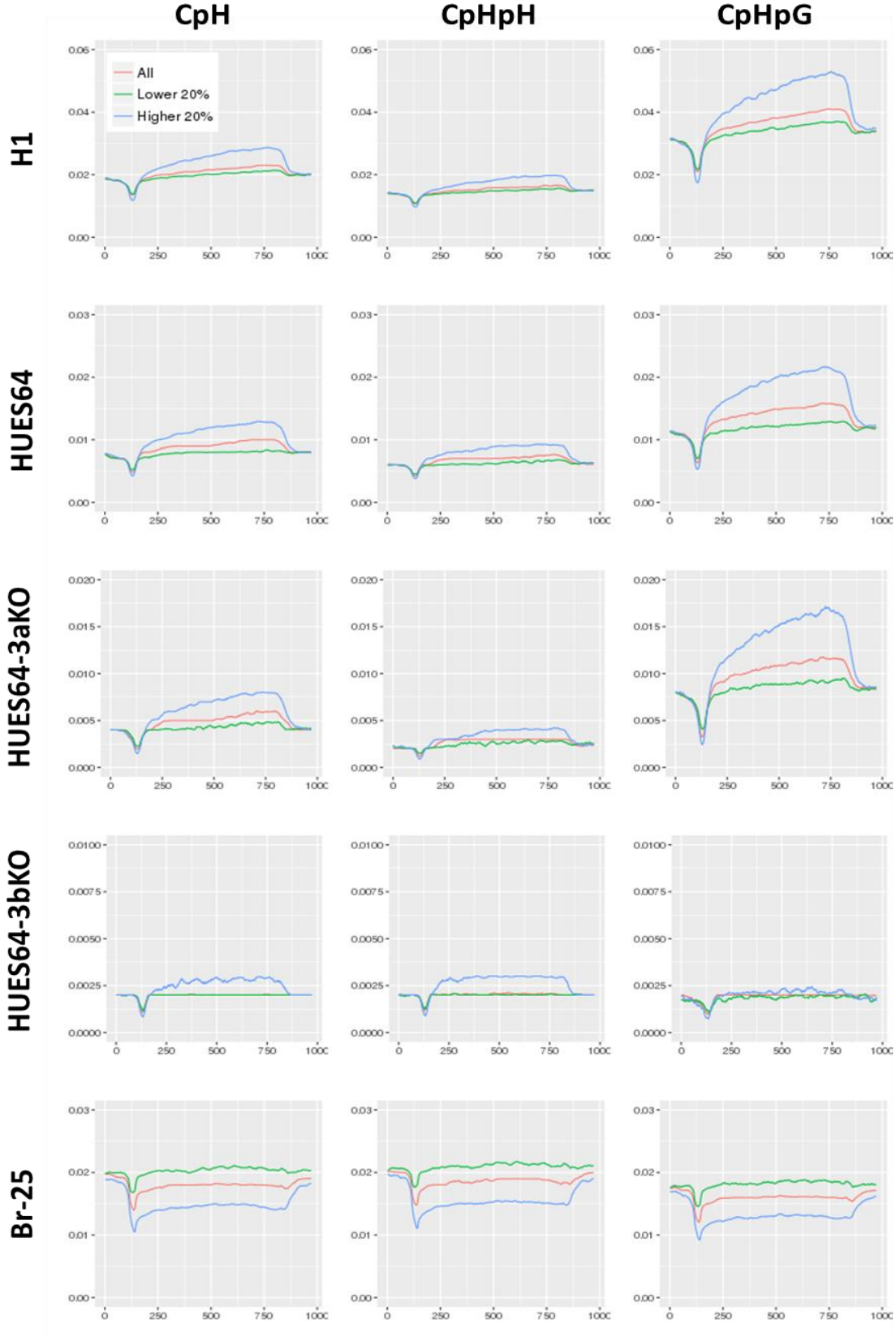


Supplementary Figure 2.6
Shown is the average mCpH levels at distance from mCpGs (red), and un-methylated CpGs (blue).



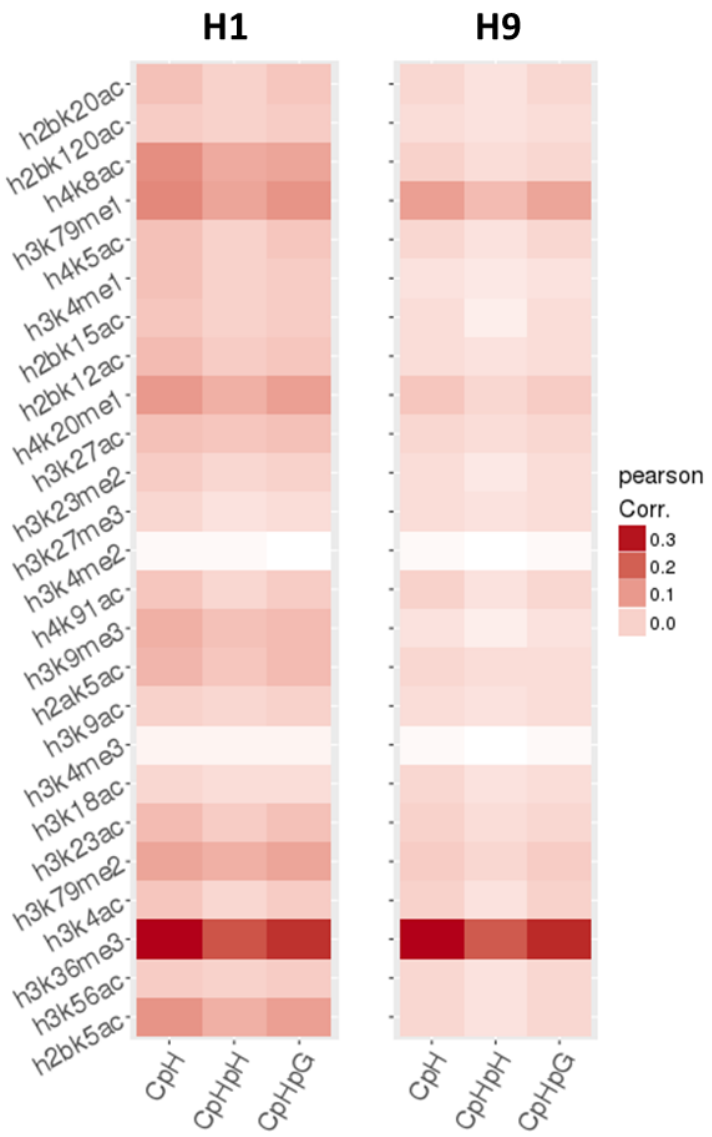
Supplementary Figure 2.7

(a) Shown is the average mCpG (red line) and mCpH (blue bar) levels in wild type, DNMT1 knock out, DNMT3a/b double knock out, and DNMT1/3a/3b triple knock out mouse ESC. (b) Shown is average mCpG and mCpH levels in wild type (WT), DNMT1 knock out(DNMT1 KO), and DNMT3a/b knock out (DNMT3 DKO) mouse ESCs. We divided mouse whole genome into 1k-bp blocks and ordered those by average mCpG level. Then, we combined blocks that shares same mCpG level for visualization. The blue points represents average mCpH level of the combined blocks and red line shows mCpG level of those. For data reliability, we counted blocks that contain more than 10 CpG and CpH points. The Pearson correlation coefficients for WT, DNMT1 KO, and DNMT3 DKO is 0.2, 0.4, and 0.07, respectively. (c) Shown is average CpH methylation levels at CpHpH (red) and CpHpG (green) contexts in wild type (WT) and DNMT3b knock out (DNMT3b-KO) H9 cell line. Error bars represents standard error within 2 biological replica.



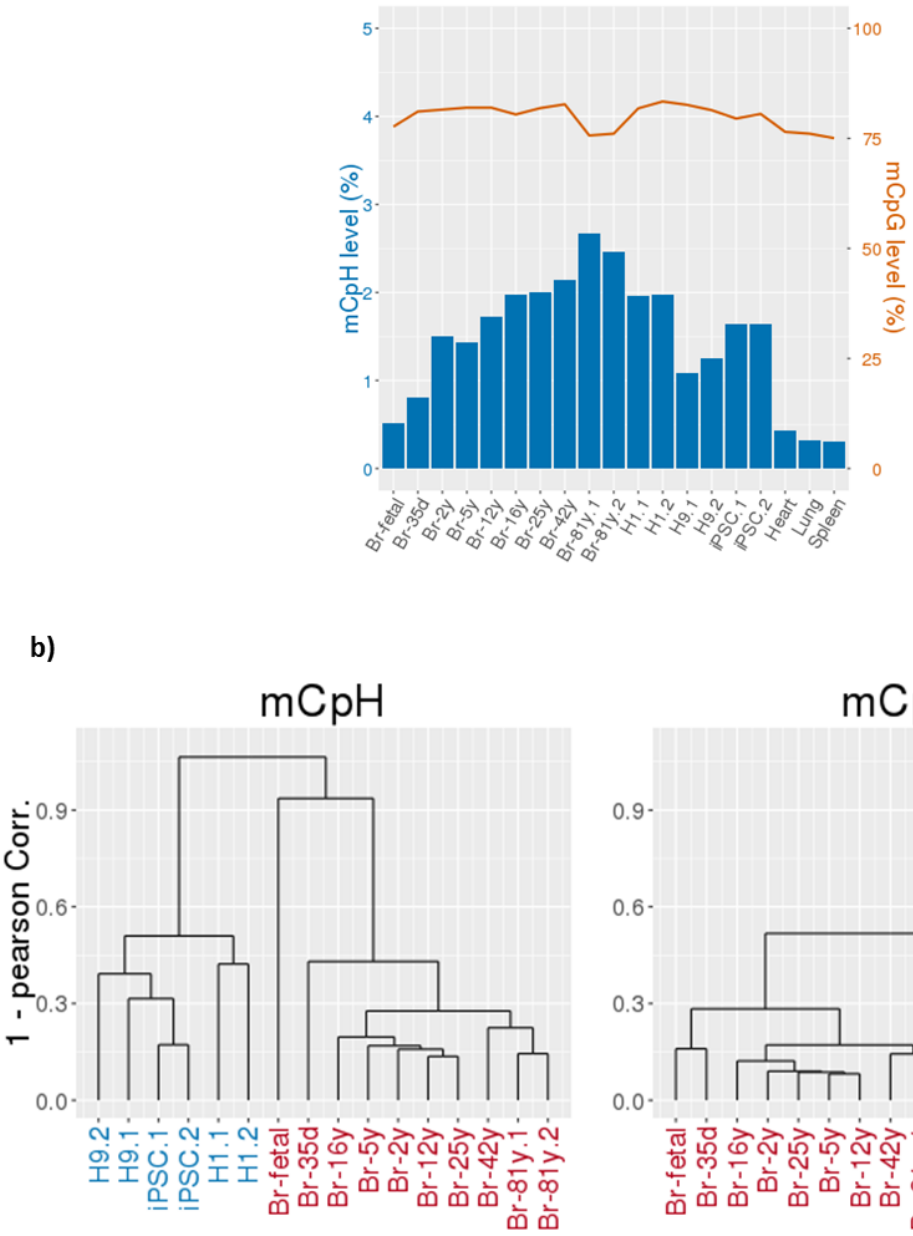
Supplementary Figure 2.8

Shown is methylation patterns around gene-body regions. The regions of gene-bodies ± 20% of those was normalized into 1000 bins and draw average methylation levels at each bin. X-axis is bins and y axis is average methylation level. The transcription star site (TSS) and transcription terminate site (TTS) are in 142'th and 572'th bins, respectively.



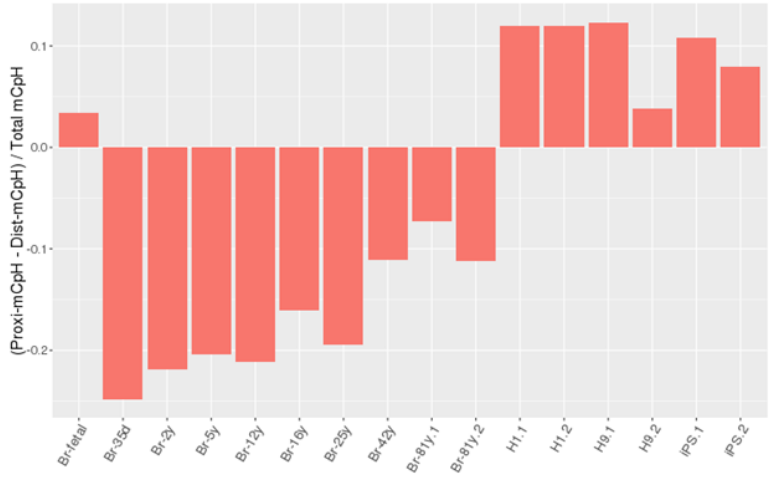
Supplementary Figure 2.9

Shown is the correlation between methylation levels and histone marks in H1(left) and H9 (right). We divided human whole genome into 1k-bp blocks and measured Pearson correlation coefficient by comparing the methylation patterns with Chip-seq peaks piled up at each block.

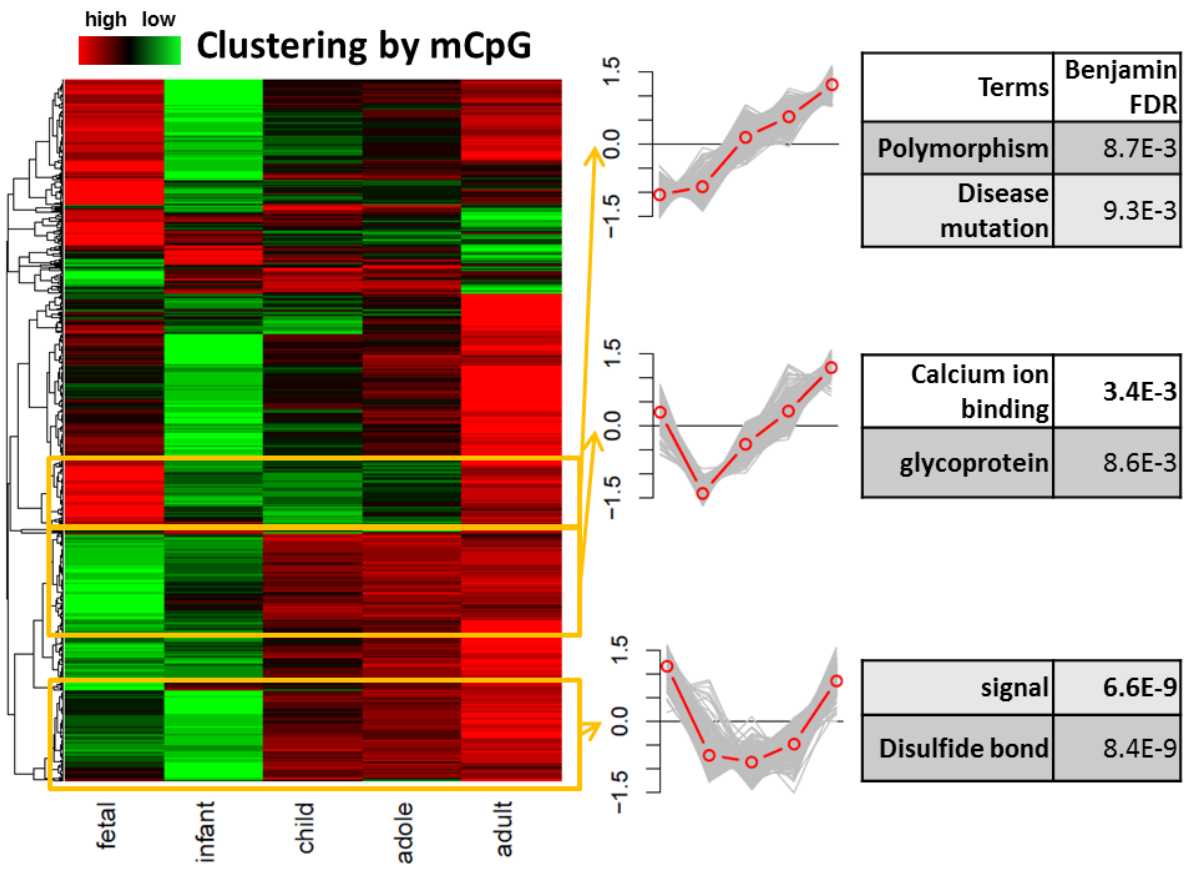


Supplementary Figure 3.1

(a) Genome-wide average methylation level at CpGs (red line) and CpHs (blue bar). Labels in x-axis are cell and tissue types of WGBS data. NEUR and Br represent neuron and brain tissue, respectively. The numbers followed by Br- means age that the brain tissue is sequenced (d and y mean day and year, respectively). H1 and H9 are cell lines of human embryonic stem cell. Numbers back in the label represents biological replicates (iPSC.1 means biological replicate 1 of iPSC). (b) The hierarchical clustering has been conducted by CpH and CpG methylation pattern across whole genome. Colors in x-axis represents same cell (or tissue) types (blue: brain and neuron, red: ESCs, and yellow: other tissues).

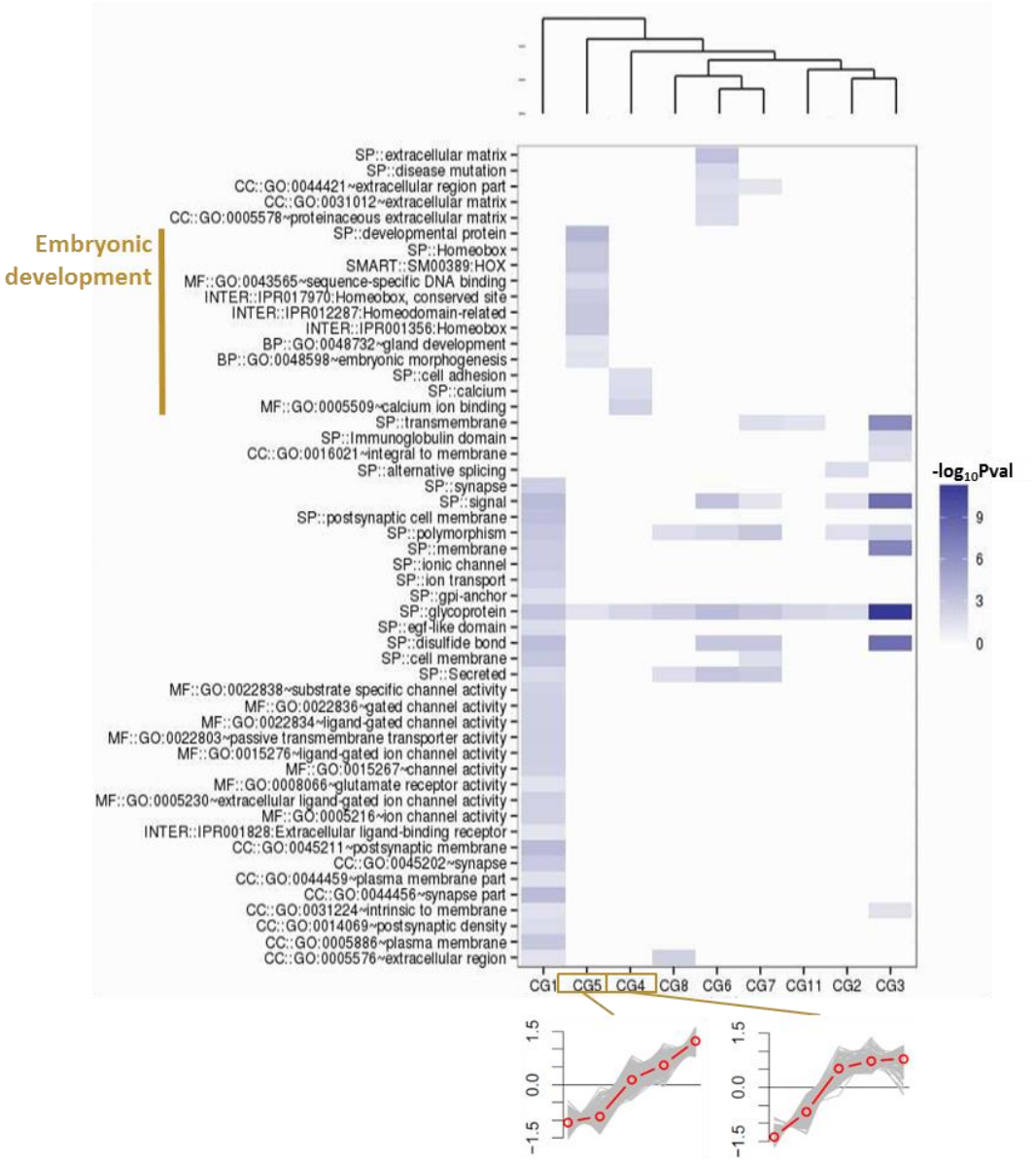


Supplementary Figure 3.2
Shown is the difference between mCpH level in mCpG-proximal and CpG-distal regions. The value was divided by total mCpH level for normalization.



Supplementary Figure 3.3

The heat map shows hierarchical clustering of genes by mCpG pattern across stages. The line graph in the middle shows normalized mCpH levels at each gene bodies (gray lines) and averages of those (red line). Also, the enriched GO term with Benjamin FDR values are shown on the right side.



Supplementary Figure3.4

The heat map shows hierarchical clustering of gene clusters and GO terms. The gene clusters (x-axis) are generated by hierarchical clustering by mCpG level in gene-bodies (Figure3.2). Also, the GO terms are those appeared more than once in the enriched GO term list (Benjamini FDR<0.01). The line graphs at bottom shows the normalized mCpH levels in gene bodies included in each cluster at the 5 age stages (Fetal, Infant, Child, Adolescent, and Adult). Red circle and lines show average value of the grays

Supplementary Table

P-value for mCpG with 100bps-long reads				P-value for mCpH with 100bps-long reads			
R err (%)	Bismark	BSMAP	BSseeker2	R err (%)	Bismark	BSMAP	BSseeker2
0	2.9E-21	5.3E-09	3.7E-01	0	7.7E-05	2.1E-04	8.8E-01
2	2.3E-14	6.4E-03	4.0E-140	2	1.6E-07	8.6E-07	2.2E-57
4	2.8E-96	5.0E-157	6.3E-103	4	5.9E-64	1.0E-152	1.9E-66
6	0	0	8.7E-46	6	0	0	1.16E-28
8	0	0	1.5E-62	8	0	0	1.12E-85
P-value for mCpG with 50bps-long reads				**P-value for mCpH with 50bps-long reads**			
R err (%)	Bismark	BSMAP	BSseeker2	R err (%)	Bismark	BSMAP	BSseeker2
0	7.4E-63	8.0E-302	2.5E-07	0	1.0E-38	8.6E-152	2.6E-06
2	1.0E+00	1.4E-184	5.9E-129	2	1.6E-01	3.7E-131	9.2E-80
4	1.0E+00	1.3E-193	9.0E-296	4	9.6E-01	3.1E-148	4.0E-150
6	1.99E-54	0	6.12E-152	6	7.75E-92	0	4.72E-87
8	0	0	2.36E-54	8	0	0	4.08E-34

Supplementary Table 1 One-side Wilcoxon single-rank test between d-accuracy by wAve and d-accuracy by three mappers

The tables show P-values by one-side Wilcoxon single-rank test between d-accuracy by wAve and d-accuracy by three mappers across 500-long blocks of hg19 chr19. If the P-value is lower than 0.05, it means the d-accuracy by wAve is significantly lower than the d-accuracy by each mapper across blocks. The bold types are values that lower than 0.05. R err means read error rate.

Cell-type.histone_mark	source	passage/age	sex
H1.h3k27me3	GSM605308	20-40	
H9.h3k27me3	GSM706066	30-50	
H1.h3k36me3	GSM605309	20-40	
H9.h3k36me3	GSM605310	30-50	
H9.h3k4me1	GSM667626	30-50	
H1.h3k4me1	GSM605312	27	
H9.h3k4me3	GSM616128	30-50	
H1.h3k4me3	GSM605315	20-40	
H9.h3k9ac	GSM616129	30-50	
H1.h3k9ac	GSM605323	20-40	
H9.h3k9me3	GSM667633	30-50	
H1.h3k9me3	GSM605328	20-40	
H9.h3k27ac	GSM665037	30-50	
H1.h3k27ac	GSM466732	25-45	
H1.h2ak5ac	GSM602257	20-40	
H9.h2ak5ac	GSM667609	30-50	
H1.h2bk120ac	GSM789281	32	
H9.h2bk120ac	GSM752962	42	
H1.h2bk12ac	GSM605297	54	
H9.h2bk12ac	GSM667610	30-50	
H1.h2bk15ac	GSM605298	51	
H9.h2bk15ac	GSM864034	30-50	
H9.h2bk20ac	GSM752963	42	
H1.h2bk20ac	GSM605300	51	
H9.h3k18ac	GSM667616	30-50	
H1.h3k18ac	GSM605304	25-45	
H1.h3k23ac	GSM667618	26	
H9.h3k23ac	GSM667620	30-50	
H1.h3k4ac	GSM667624	32	
H9.h3k4ac	GSM667625	30-50	
H1.h3k4me2	GSM602260	25-45	
H9.h3k4me2	GSM616127	30-50	
H1.h3k56ac	GSM667627	32	
H9.h3k56ac	GSM706076	30-50	
H1.h3k79me1	GSM605319	20-40	
H9.h3k79me1	GSM667629	30-50	

H1.h3k79me2	GSM605321	20-40	
H9.h3k79me2	GSM706078	30-50	
H1.h4k20me1	GSM789284	32	
H9.h4k20me1	GSM667634	30-50	
H1.h4k5ac	GSM752990	32	
H9.h4k5ac	GSM667636	30-50	
H1.h4k8ac	GSM908966	32	
H9.h4k8ac	GSM667638	30-50	
H1.h4k91ac	GSM752991	32	
H9.h4k91ac	GSM667640	30-50	
H9.h3k23me2	GSM667621	30-50	
H1.h3k23me2	GSM605305	52	
H1.h2bk5ac	GSM605302	20-40	
H1.h2bk5ac	GSM667613	30-50	
brain.1.h3k36me3	GSM669982	75y	
brain.1.h3k27ac	GSM1112810	75y	
brain.1.h3k4me1	GSM670015	75y	Female
brain.1.h3k4me3	GSM670016	75y	
brain.1.h3k9ac	GSM670021	75y	
brain.1.h3k9me3	GSM669965	75y	
brain.2.h3k27ac	GSM773015	81y	
brain.2.h3k27me3	GSM772833	81y	
brain.2.h3k36me3	GSM7730113	81y	Male
brain.2.h3k4me1	GSM773014	81y	
brain.2.h3k4me3	GSM773012	81y	
brain.2.h3k9me3	GSM772834	81y	

Supplementary Table 3: Source for histone marks

Acknowledgement

This study could be completed by all of the supports from those below. First, I would like to thank Professor Kenta Nakai, who has instructed me for 5 years as a supervisor. For his great leadership, I could step into the research field and could dream of being a good researcher. Also, I express my gratitude to my beloved wife, Jung-Min Park, who has supported me with great patience and wisdom. Also, I thanks to my family in Korea for their strong love and deep prayer. In addition, I appreciate for all the financial supports by Japanese Government (MEXT), and computational capability supported by IMSUT. Last, and most importantly, I thank and honor to Jesus Christ, my Lord and best friend forever.