Existence and possible roles of independent non-CpG methylation in mammalian brain

Jong-Hun Lee
The University of Tokyo
Yutaka Saito
National Institute of Advanced Industrial Science and Technology
Sung-Joon Park
The University of Tokyo
Kenta Nakai (knakai@ims.u-tokyo.ac.jp)
the University of Tokyo https://orcid.org/0000-0002-8721-8883

Research

Keywords: Non-CpG methylation, Hidden Markov model, Neuro-epigenetics

Posted Date: January 30th, 2020

DOI: https://doi.org/10.21203/rs.2.22195/v1

Version of Record: A version of this preprint was published at DNA Research on August 1st, 2020. See the published version at https://doi.org/10.1093/dnares/dsaa020.

Existence and possible roles of independent non-CpG methylation in mammalian brain

Jong-Hun Lee[1], Yutaka Saito[2,3,4], Sung-Joon Park[1], and Kenta Nakai[1,4]*

[1]Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo 108-8693, Japan.
[2]Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Tokyo 135-0064, Japan.
[3]AIST-Waseda University Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), Tokyo 169-8555, Japan.
[4]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa 277-8568, Japan.

Jong-Hun Lee	christiario@gmail.com
Yutaka Saito	yutaka.saito@aist.go.jp
Sung-Joon Park	sjpark@ims.u-tokyo.ac.jp
Kenta Nakai	knakai@ims.u-tokyo.ac.jp

* To whom correspondence should be addressed
Contact: knakai@ims.u-tokyo.ac.jp

Abstract

Background

DNA methylations at non-CpG sites (mCpHs) yield weak enrichment signals and co-localize with methylated CpGs (mCpGs). Accordingly, they are thought to represent byproducts of hyperactive methyltransferases. However, recent studies revealed that mCpHs are highly cell type-specific and associated with cellular functions in mammal, although their distributional and functional dependency to mCpGs remains to be elucidated.

Results

Through computational analysis of a large publicly available DNA methylation dataset from human and mouse cells, we revealed that mCpHs tend to co-localize with mCpGs in pluripotent stem cells but not in brain cells. A systematic investigation using an empirical hidden Markov model confirmed that there are genomic regions where CpGs and CpHs are differentially methylated in the brain cells (~25%). Remarkably, the regions were frequently located in putative enhancers, and mCpHs within the enhancers were greatly increased as brain aged. We detected 1669 hyper-mCpH enhancers functionally enriched to immune system development, aberration of which can lead to neuro-inflammation and neuro-degeneration.

Conclusions

In mammalian brain cells, there are genomic regions where CpH alone is methylated (or de-methylated). The mCpHs are accumulated in enhancers as brain aged, potentially associated to neuro-inflammation and degeneration. Altogether, this study provides insight into the roles of non-CpG methylation as an epigenetic code in the genomes of mammalian brain cells.

Keywords

Non-CpG methylation, Hidden Markov model, Neuro-epigenetics

Background

For decades, DNA methylation was thought to occur only at CpG (cytosine followed by guanine) sites in mammalian cells. Recently, however, many studies using advanced genome-wide sequencing technology revealed that non-CpG sites are also methylated in several types of mammalian cells [1, 2]. The methylated non-CpGs (mCpHs; H indicates A, C, or T) are highly cell type-specific [1, 3] and associated with transcription of nearby genes [4], and are therefore emerging as important epigenetic markers.

mCpHs are most widely studied within pluripotent stem cells (PSCs) and brain cells [5–8]. In these cell types, mCpHs are more abundant (~5%) than in other kinds of cells (<1%) [9] and exhibit specific patterns such as co-localization with histone modifications [10], depletion in transcription regulatory regions [11], and accumulation in gene-body regions and transposons [6]. These patterns are highly distinct between the two cell types and are potentially linked to cell type-specific functions. For example, the CAG motif tends to be methylated in PSCs, whereas the CAC motif is preferentially methylated in brain tissues [9]. Methylated CAC in the brain is a binding target of methyl-CpG binding protein 2, mutation of which causes Rett syndrome [12]. In addition, enrichment of mCpHs in PSCs is positively correlated with cell differentiation capacity [13].

In general, mCpHs tend to be positioned near mCpGs, although they are methylated at a much lower level [7]. Due to their extensive colocalization with methylated CpGs (mCpGs), it remains unclear whether mCpHs alone have an effect on cellular processes. In mammalian cells, CpHs are methylated by DNA methyltransferases (DNMT) 3a and DNMT3b, which methylate CpGs with much greater efficiency [14]. In PSCs and brain cells, DNMTs are expressed at high levels [9]; consequently, mCpHs have been considered to be byproducts of the hyperactivity of these enzymes [5, 15, 16].

In this study, we attempted to understand the functional relevance of mCpHs in PSCs and brain cells. To this end, we conducted a large-scale analysis using publicly available datasets encompassing 26 whole-genome bisulfite sequencing (WGBS) samples and 368 microarray samples. We also included DNMT-knockout samples to investigate the involvement of DNMTs in methylation processes. Furthermore, we developed a hidden Markov model (HMM) to systematically detect genomic regions where CpG and CpH are differentially methylated, providing an opportunity to infer the functional importance of non-CpG methylation.

Results

Differential distributions of mCpGs and mCpHs in PSCs and brain

To capture the genome-wide methylation status at each CpG and CpH, we prepared human WGBS datasets from PSCs and brain cells, which contain abundant mCpHs, as well as lung and spleen datasets as controls (Table S1). We then analyzed the WGBS reads by applying our statistical method, which integrate outcomes from three tools [17]: Bismark [18], BSMAP [19], and BS-seeker2 [20]. We were able to quantify methylation levels for up to 73% of CpGs and up to 96% of CpHs throughout the genome (Table S1), and to define methylated and unmethylated cytosine bases (see Method).

First, we investigated the positional correlation of mCpHs with mCpGs. As shown in Fig. 1a and Fig. 1b, CpHs in PSCs tended to be hyper-methylated near mCpGs, whereas those in brain cells were hypo-methylated near mCpGs. Given that the mCpG positions in the two cell types may yield the distinctive distributions of mCpHs, we focused only mCpGs present in both PSCs and brain cells, and obtained consistent results (Fig. S1). To further confirm the differential distribution of mCpHs, we analyzed 368 microarray-based DNA methylation datasets and compared mCpH levels in mCpG-proximal sites (within 100 bp) with mCpH levels in CpG-distal sites (more than 100 bp away). As shown in Fig. 1c, mCpG-proximal mCpH levels were significantly lower than mCpG-distal mCpH levels in brain cells but not in PSCs. These results indicate that mCpHs tend to co-localize with mCpGs in PSCs, but not in brain tissues.

In addition, we observed a 180 bp periodicity of mCpH levels in PSCs (Fig. 1a, Fig. S1). This may reflect the positioning of nucleosomes, which regulate the accessibility of the genome to DNMTs [21]. Moreover, when we analyzed only mCpGs common to both cell types, we also detected the periodicity in brain cells (Fig. S1). These results indicate that CpGs and CpHs are methylated simultaneously by DNA ‘walking’ of DNMTs [22]. Notably, this periodicity is alleviated in Fig. 1b, indicating that the brain-specific hypo-methylation at CpHs around mCpGs are mediated to a lesser extent by DNMTs.

Inference of DNMT activities in the methylation processes

DNMT3a and DNMT3b are differentially expressed in PSCs and brain cells: DNMT3a is more strongly expressed in brain cells, whereas DNMT3b is more strongly expressed in PSCs [9]. Therefore, we hypothesized that the patterns of mCpH distribution (Fig. 1a, Fig. 1b) rely on the differential activities of DNMT3a and DNMT3b. To test this hypothesis, we analyzed WGBS data from DNMT1-KO (knockout) mouse embryonic stem cells (ESCs), which primarily harbor DNA methylations mediated by DNMT3a and DNMT3b. As shown in Fig. 1d, mCpH methylation levels

were higher near mCpGs than near unmethylated CpGs, with a clear 180 bp periodicity in both mCpGs and mCpHs, demonstrating the restricted effect of DNMT1 in the mCpH distributions.

Next, to observe methylation by DNMT3a and DNMT3b separately, we analyzed mouse ESCs in which DNMT3a or DNMT3b were re-introduced after knockout of DNMT1, DNMT3a, and DNMT3b [21]. In both DNMT3a- and DNMT3b-reinduced samples, CpH methylation levels were higher near mCpGs (Fig. S2c), indicating that both DNMT3a and DNMT3b mediate co-localization of mCpGs and mCpHs (Fig. S2a). We observed a similar trend in human ESCs in which DNMT3a and DNMT3b were knocked out separately (Fig. S2b). In addition, the genome-wide correlation between mCpGs and mCpHs were similar between the DNMT3a-KO and DNMT3b-KO (Pearson correlation coefficient = 0.38 and 0.32, respectively; Fig. S2b).

These results clearly demonstrated that DNMT3a and DNMT3b can solely mediate mCpGs and mCpHs co-localized. Thus, we concluded that the depletion of mCpH levels near mCpGs in brain cells is not caused by the DNMT3a hyperactivity and may have functional meaning.

Detection of CpG–CpH DMRs

Because both DNMT3a and DNMT3b generated a positive correlation between mCpH and mCpG levels on a genome-wide scale, we assumed that the negative correlation between mCpH and mCpG levels in brain tissue would be highly locus-specific. We designated such genomic regions as CpG–CpH differentially methylated regions (CpG–CpH DMRs). To detect CpG–CpH DMRs, several issues must be considered. First, because average methylation levels differ dramatically between CpGs (~0.8) and CpHs (~0.02), we need to consider the methylation states at CpGs and CpHs in different scopes. Second, because average methylation levels also vary among cell types (Fig. 1a, Fig. 1b), we need to adjust the methylation criteria for each cell type. Finally, the methylation data show positional continuity with the 180 bp periodicity, which is useful for detecting DMRs.

To address these issues, we adopted a HMM approach in which the whole genome was segmented into 180 bp bins, and each bin was categorized into one of three states: P, N, and U (Fig. 2a). In P-state regions, mCpH and mCpG levels correlate positively (P). In N-state regions, the levels correlate negatively (N); i.e., N-state regions are equivalent to CpG–CpH DMRs. In a U-state region, levels are uncorrelated (U), meaning that the mCpH level stays at the average level rather than being affected by the surrounding mCpG level.

We calculated the probability that a bin belongs to any of the three states (emission probability),

and the probability of a transition between two states, which models the positional continuities between two neighboring bins (Fig. 2b). To represent variation of mCpG and mCpH levels among cell types, the transition probability was trained by the expectation-maximization (EM) algorithm for each cell type (Fig. 1a, Fig.1b). After training, our HMM assigns the most probable state for each bin using the Viterbi algorithm (Fig. S3a; see Method). As exemplified in Fig. 2c and Fig. 2d, the state classification of bins using the highest emission probability only (Fig. 2c) and the Viterbi algorithm (Fig. 2d) clearly demonstrated that mCpG and mCpH levels were correlated positively in the P-state and negatively in the N-state. After Viterbi decoding, the U-states were absorbed by P- or N-states, as the probability of transition from U to U is lower (Fig. S3b).

We designed the Viterbi decoding to properly define the bin states by considering methylation information in and around the bin (Fig. S3a). To confirm its performance, we generated an artificially methylated human chromosome 19 with randomly generated N-states, and compared precision and recall rates of the N-state, defined by the emission probability only or by the Viterbi decoding (Fig. S3c). Both precision and recall rates increased when the Viterbi decoding was performed, indicating that CpG–CpH DMRs were more accurately defined by the Viterbi decoding than the emission probability alone.

Functional analysis of brain CpG–CpH DMRs

We applied the HMM to each of the WGBS samples and estimated the states of all genomic regions. Interestingly, 25% of bins in brain samples were detected as CpG–CpH DMRs, a remarkable abundance relative to other cell types (Fig. 3a). Moreover, the DMRs were relatively conserved among brain samples but differed from others (Fig. S4). Overall, the DMRs overlapped significantly with a subset of known genomic elements, including transcribed regions, conserved regions in mammals, and enhancer regions (Fig. 3b). Remarkably, the enhancers marked by DMRs showed high mCpH level, and the mCpH levels increased with age (Fig. 3c).

Next, we collected DMRs containing hyper- and hypomethylation of CpHs, defined as follows: hyper-mCpH-N, >2-fold change relative to the mCpH level in the whole genome; hypo-mCpH-N, <0.5-fold change relative to the mCpH level in the whole genome. The hyper-mCpH-N regions overlapped with enhancers, and hypo-mCpH-N regions were found in the promoter and coding regions (Fig. 3d). Finally, we extracted 1669 hyper- and 946 hypo-mCpH-N enhancers; >80% of an enhancer region were marked by either hyper- or hypo-mCpH-N regions in at least two samples. These enhancers were within 1 Mb of 3184 genes: 1717 genes for hyper-mCpH-N enhancers (Table S2) and 1467 genes for hypo-mCpH-N enhancers (Table S3).

To understand the functional importance of DMRs, we performed Gene Ontology (GO) enrichment analysis on the 3184 putative target genes using the Genomic Regions Enrichment of Annotations Tool (GREAT) [23]. This analysis revealed the possibility that the genes associated with hyper-mCpH-N enhancers encode signal molecules related to the immune system and involved in developmental cell activation (Fig. 4a-c, Table S4). None of the immune-related terms were enriched among genes associated with hypo-mCpH-N enhancers (Fig. 4d). Some of these genes regulate neuroinflammation, which eventually leads to neurodegeneration. For example, the enhancer associated with *PICALM* (phosphatidylinositol binding clathrin assembly protein), a significant genetic risk locus for Alzheimer's disease[24–26], was covered by hyper-mCpG-N regions (Fig. S5a). Similarly, the enhancer associated with *CD28* (cluster of differentiation 28), whose downregulation in T-cells is a hallmark of aging [27] and dementia [28], was also marked by hyper-mCpG-N regions (Fig. S5b).

Taken together, these observations suggest that CpG–CpH DMRs with hyper-methylated CpHs are potentially involved in neurodegeneration via regulation of the immune system.

Discussion

mCpGs and mCpHs tend to be co-localized in mammalian cells, casting doubt on the biological relevance of mCpHs [5, 15, 16]. However, our analyses of large-scale methylation data revealed that the patterns of these two types of methylation are distinct in PSCs and brain cells. In particular, mCpHs were markedly depleted near mCpG sites in brain cells but not in PSCs. Because our analysis of DNMT-KO samples indicated that this characteristic was less introduced by the activity of DNMTs, we hypothesized that the co-methylation of CpHs with CpGs occurs in a cell type-specific manner, and that these methylation processes have an unknown functional meaning.

To systematically detect DMRs based on mCpG and mCpH levels, we designed an HMM that was successfully applied in previous studies [29, 30]. We identified 25% of genome in brain samples as DMRs, a remarkable abundance relative to other tissue types (Fig. 3a). Moreover, these DMRs were conserved in the brain, and their methylation levels increased with age (Fig. 3c). It should be noted that a DMR is a genomic region where either mCpH or mCpG levels are intensified. Thus, to confirm the mCpH depletion near mCpGs, we investigated the DMRs marked by hyper- and hypomethylation of CpHs. Hypo-mCpH DMRs were present in promoter and coding regions (Fig. 3d), consistent with epigenetic regulation of transcription by mCpGs [13].

Unexpectedly, there were hypo-mCpH DMRs largely overlapped with enhancer regions, where methylation levels increased as brain aged. Importantly, the putative target genes of these enhancers are specifically involved in the immune system development, disruption of which increases risk of Alzheimer's disease [31].

Although we found here the existence of functional interplay of non-CpG methylation and brain development, we still do not know the underlying mechanism of the mutually exclusive methylation at CpG and CpH in brain. To understand the mechanism that may involve complex DNA methylation and demethylation processes in mammalian brain, future studies should incorporate comprehensive analyses with experimental validation from various cell types.

Conclusions

Our findings provide evidence of the functional importance of non-CpG methylation as an epigenetic code in the genome of mammalian brain cells.

Methods

Analysis of WGBS data

WGBS data from human PSCs, human tissue cells (brain, lung, spleen), and mouse ESCs were downloaded from Gene Expression Omnibus (Table S1). After filtering out low-quality reads using the FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and duplicated reads using Samtools [32], left were mapped to the reference genomes (hg19 and mm10) using three bisulfite-read mappers: Bismark [18], BSMAP [19], and BS-seeker2 [20]. After collecting cytosine bases where more than 5 reads were mapped by at least two mappers, the rate of unconverted reads over all the mapped reads was calculated at each cytosine [17], and the methylation level Me_i at the i-th cytosine is given as follows:

$$Me_i = \frac{\sum_j m_{i,j}}{\sum_j n_{i,j}} - \mathrm{R}\,,$$

where j stands for one of the three mappers, R is the non-conversion rate, and m and n represent the number of methylated reads and the total number of mapped reads, respectively.

Identification of methylated cytosine

To identify mCpHs, we adopted an approach presented in a previous study [7]. Because average mCpG and mCpH levels are significantly different, we applied different thresholds in the two contexts: in a CpG context, a cytosine base was defined as methylated if Me >0.8 or as unmethylated if Me <0.2. For DNMT1-, DNMT3a-, and DNMT3b-KO samples, we defined the CpGs as methylated if Me >0.5. In a CpH context, the i-th cytosine base is defined as methylated by the significance of the binomial test, given as follows:

$$\Pr(m_i|n_i, R) = \binom{n_i}{m_i} R^{m_i}(1-R)^{n_i-m_i}\,,$$

where m and n represent the number of methylated reads and the total number of mapped reads, respectively, and R stands for the non-conversion rate. To define a P-value threshold, we randomly generated methylated reads from a binomial distribution given by R and n_i, for every i, and defined mC with varied P-values (1.0e-2 to 1.0e-6). Then, we calculated the false discovery rate (FDR) (i.e., $\frac{mCs\ in\ null\ data}{mCs\ in\ real\ data}$) for each mC set, and eventually, set the P-value threshold as 1.0e-5 that allows 1% FDR in all samples.

Analysis of DNA methylation microarray data

Microarray datasets, including 177 samples of human brain and 191 samples of PSCs, were downloaded from BrainSpan (http://www.brainspan.org/) and GSE59091 [33]. Among 3091 CpH sites, 1079 sites had no CpGs within 100 bp (CpG-distal), and 26 sites in average had more than

one mCpG (mCpG-proximal). We then compared the average methylation levels of those CpHs in PSCs and brain cells.

Designing a HMM

After segmenting the whole genome into 180 bp bins, our HMM calculates the emission probability E for the j-th state, j={P, N, U}, at the i-th bin as follows:

$$E_{i,j} = \Pr\left(m_i \middle| n_i, \theta_{i,j}\right) \sim \mathrm{Bin}\left(m_i \middle| n_i, \theta_{i,j}\right) \ ,$$

where m_i and n_i stand for the number of methylated reads and the total reads aligned at i, respectively. P, N, and U represent the correlation state of methylation levels at a bin: P, positive correlation between mCpH and mCpG levels; N, negative correlation between mCpH and mCpG levels; and U, uncorrelated.

We differentiated the probabilities of the three states by optimizing $\theta_{i,j}$ ($\hat{\theta}_{i,j}$) as representing the methylation status of each state. In the U-state, the mCpH level is not correlated to the mCpG level, thus $\hat{\theta}_{i,U}$ is the ratio of all mCs over cytosine bases (Cs) in the whole-genome w as follows:

$$\hat{\theta}_{i,U} = \hat{\theta}_w = \frac{\sum_i m_i}{\sum_i n_i}.$$

Then, $\hat{\theta}_{i,j}$ for the P- and N-states, j={P, N}, are estimated by maximizing the posteriori estimation, where a prior probability Pr is modeled as a beta distribution as follows:

$$\begin{aligned}\hat{\theta}_{i,j} &= argmax_{\theta_{i,j}} Pr\left(\theta_{i,j} \middle| m_i, n_i\right) \propto argmax_{\theta_{i,j}} Pr\left(m_i, n_i \middle| \theta_{i,j}\right) Pr(\theta_{i,j}) \\ &= argmax_{\theta_{i,j}} {\theta_{i,j}}^{(m_i+\alpha_{i,j})}(1-\theta_{i,j})^{(n_i-m_i+\beta_{i,j})} \\ &= \frac{m_i+\alpha_{i,j}}{n_i+\alpha_{i,j}+\beta_{i,j}} \ .\end{aligned}$$

Here, α and β are used as pseudo-counts to reflect the CpG methylation status at each bin. We regulate α and β by adjusting mean M of the distribution at the i-th bin, given as follows:

$$M_{i,P} = \hat{\theta}_U + \mathrm{SD}_{mH} \times \frac{mG_i - MED_{mG}}{\mathrm{SD}_{mG}} \ ,$$

$$M_{i,N} = \hat{\theta}_U - \mathrm{SD}_{mH} \times \frac{mG_i - MED_{mG}}{\mathrm{SD}_{mG}} \ ,$$

where mG_i is average mCpG level at the i-th bin. SD_{mG} and MED_{mG} are the standard deviation and median of mCpG level across the whole genome, respectively. The SD_{mH} is the standard deviation of mCpH level across the whole genome. In brief, if the mCpG level at the i-th bin is higher than the genome-wide average mCpG level, $M_{i,A}$ is higher than $\hat{\theta}_U$, whereas $M_{i,B}$ is lower than $\hat{\theta}_U$. Because $M = \frac{\alpha}{\alpha+\beta}$, we defined α and β as follows:

$$\alpha_{i,j} = n_i \times M_{i,j} \ ,$$
$$\beta_{i,j} = n_i - \alpha_{i,j} \ ,$$

where n_i is the number of reads at all CpGs in the i-th bin.

We calculated $E_{i,j}$ for bins where the number of CpGs and CpHs is over 10, and grouped the genomic regions in each chromosome where continuous undetected bins were not longer than 100,000 bp; the average number of groups in each chromosome was 2.4. The detection of CpG–CpH DMRs was performed separately for each group. We estimated the transition probability by EM algorithm that repeats the estimation-maximization steps until the difference between the previous and current transition probabilities of all state transaction is less than 0.5e-3 [29]. We ran the Viterbi algorithm, which finds an optimal path among the states, and defined CpG–CpH DMRs where the N-state regions continued longer than 1 Kb.

Preparation of simulated dataset

Pseudo-bisulfite–treated reads were uniformly distributed with an average read depth of 10 in cytosine bases of human chromosome 19; total read number was counted by summing the reads at every 180 bp bins. For the i-th bin, the number of unmethylated reads un_m at the CpG or CpH site (j) was randomly chosen from a binomial distribution given by the total read number n and a modeled methylation level Me as follows:

$$\mathrm{Bin}\left(un\_m_{i,j} \middle| n_{i,j}, Me_{i,j}\right) \ .$$

$Me_{i,CpG}$ is a uniform random number distributed between 0 and 1. The $Me_{i,CpH}$ in P-state is Me_{CpG}×0.1, and that in N-state is (1 − Me_{CpG})×0.1, reflecting the lower methylation level at CpHs relative to CpGs.

We generated 500 artificial sets containing 10 to 10,000 randomly distributed N-states, the length of which varied randomly from 1 to 100 bins. Then, we simulated whether our HMM could properly detect the N-state by precision and recall rates given as follows:

$$Precision = \frac{TP}{TP + FP}$$
$$Recall = \frac{TP}{TP + FN} ,$$

where TP is the true positive, FP the false positive, and FN the false negative rate of the detected N-state region.

Preparation of genomic element annotation

Annotation for genomic elements was downloaded from a previous study [34]. The original

source of each annotation is as below: Enhancer_Andersson (FANTOM5 enhancer) from [35]; SuperEnhancer_ and H3K27ac_Hnisz from [36]; weakEnhancer_Hoffman, Transcribed_Hoffman, Repressed_Hoffman, and CTCF_Hoffman from [37]; Conserved_LindbladToh from [38]; H3K4me1_, H3K9ac_, and H3K4me3_Trynka from Roadmap [39]; and post-processing from [40]. Of note, the enhancers are from a broad range of cell types, not specific to brain tissue.

Analysis of GO enrichment

Using the 2615 enhancer regions overlapped with hyper-mCpH-N (>2-fold change of average mCpH level) and hypo-mCpH-N (<0.5 fold change of average mCpH level) in at least two brain samples, GREAT [23] with default parameters detected 3184 putative target genes within 1 Mb of the enhancers. GO enrichment was calculated by performing a binomial FDR test (<0.05).

List of abbreviations

mCpG: Methylated CpG
mCpH: Methylated CpH
DNMT: DNA methyltransferase
HMM: Hidden Markov model
DMR: Differentially methylated region

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All data generated are included in this article and its supplementary information files. The Perl script, example files and description of HMM is available at Github (https://github.com/christiario/HMM).

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by JSPS KAKENHI (grant numbers: 16H01403 and 18H04710). Y.S. was supported by JSPS KAKENHI (grant numbers: 17H06410, 19K20409, 19K06502, and 19K06077).

Authors' contributions

J.-H.L. designed and conducted the data analyses. J.-H.L. and Y.S. built the hidden Markov model. J.-H.L., Y.S., and S.-J.P. analyzed the data and wrote the manuscript. K.N. conceived and designed the study. All authors read and approved the final manuscript.

Acknowledgments

Computational resources were provided by the SHIROKANE supercomputer system at Human Genome Center of the Institute of Medical Science at the University of Tokyo.

References

1. Holliday R, Pugh J. DNA modification mechanisms and gene activity during development. Science (80-). 2005.

2. Riggs AD. X inactivation, differentiation, and DNA methylation. Cytogenet Genome Res. 1975.

3. Baylin SB. DNA methylation and gene silencing in cancer. Nature Clinical Practice Oncology. 2005.

4. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008;454:766–70.

5. Ziller MJ, Müller F, Liao J, Zhang Y, Gu H, Bock C, et al. Genomic Distribution and Inter-Sample Variation of Non-CpG Methylation across Human Cell Types. PLoS Genetics. 2011;7:e1002389.

6. Lister R, Mukamel E a, Nery JR, Urich M, Puddifoot C a, Johnson ND, et al. Global epigenomic reconfiguration during mammalian brain development. Science. 2013;341:1237905. doi:10.1126/science.1237905.

7. Guo JU, Su Y, Shin JH, Shin J, Li H, Xie B, et al. Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain. Nat Neurosci. 2014.

8. Rizzardi LF, Hickey PF, Rodriguez DiBlasi V, Tryggvadóttir R, Callahan CM, Idrizi A, et al. Neuronal brain-region-specific DNA methylation and chromatin accessibility are associated with neuropsychiatric trait heritability. Nat Neurosci. 2019.

9. Lee JH, Park SJ, Nakai K. Differential landscape of non-CpG methylation in embryonic stem cells and neurons caused by DNMT3s. Sci Rep. 2017.

10. Liao J, Karnik R, Gu H, Ziller MJ, Clement K, Tsankov AM, et al. Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. Nat Genet. 2015.

11. Laine VN, Gossmann TI, Schachtschneider KM, Garroway CJ, Madsen O, Verhoeven KJF, et al. Evolutionary signals of selection on cognition from the great tit genome and methylome. Nat Commun. 2016.

12. Lagger S, Connelly JC, Schweikert G, Webb S, Selfridge J, Ramsahoye BH, et al. MeCP2 recognizes cytosine methylated tri-nucleotide and di-nucleotide sequences to tune transcription in the mammalian brain. PLoS Genet. 2017.

13. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009.

14. Arand J, Spieler D, Karius T, Branco MR, Meilinger D, Meissner A, et al. In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. PLoS Genet. 2012.

15. Smith ZD, Meissner A. DNA methylation: Roles in mammalian development. Nature Reviews Genetics. 2013.

16. He Y, Ecker JR. Non-CG Methylation in the Human Genome. Annu Rev Genomics Hum Genet. 2015.

17. Lee JH, Park SJ, Kenta N. An integrative approach for efficient analysis of whole genome bisulfite

sequencing data. BMC Genomics. 2015.
18. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27:1571–2.
19. Xi Y, Li W. BSMAP: Whole genome bisulfite sequence MAPping program. BMC Bioinformatics. 2009.
20. Guo W, Fiziev P, Yan W, Cokus S, Sun X, Zhang MQ, et al. BS-Seeker2: A versatile aligning pipeline for bisulfite sequencing data. BMC Genomics. 2013.
21. Baubec T, Colombo DF, Wirbelauer C, Schmidt J, Burger L, Krebs AR, et al. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature. 2015.
22. Goyal R, Reinhardt R, Jeltsch A. Accuracy of DNA methylation pattern preservation by the Dnmt1 methyltransferase. Nucleic Acids Res. 2006.
23. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010.
24. Parikh I, Fardo DW, Estus S. Genetics of PICALM expression and Alzheimer's disease. PLoS One. 2014.
25. Zhao Z, Sagare AP, Ma Q, Halliday MR, Kong P, Kisler K, et al. Central role for PICALM in amyloid-β blood-brain barrier transcytosis and clearance. Nat Neurosci. 2015.
26. Holmes C. Inflammation in Alzheimer's disease. In: Dementia, Fifth Edition. 2017.
27. Costantini E, D'Angelo C, Reale M. The role of immunosenescence in neurodegenerative diseases. Mediators of Inflammation. 2018.
28. Panossian LA, Porter VR, Valenzuela HF, Zhu X, Reback E, Masterman D, et al. Telomere shortening in T cells correlates with Alzheimer's disease status. Neurobiol Aging. 2003.
29. Saito Y, Tsuji J, Mituyama T. Bisulfighter: Accurate detection of methylated cytosines and differentially methylated regions. Nucleic Acids Res. 2014.
30. Saito Y, Mituyama T. Detection of differentially methylated regions from bisulfite-seq data by hidden Markov models incorporating genome-wide methylation level distributions. BMC Genomics. 2015.
31. Weiner HL, Frenkel D. Immunology and immunotherapy of Alzheimer's disease. Nature Reviews Immunology. 2006.
32. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009.
33. Butcher LM, Ito M, Brimpari M, Morris TJ, Soares FAC, Ährlund-Richter L, et al. Non-CG DNA methylation is a biomarker for assessing endodermal differentiation capacity in pluripotent stem cells. Nat Commun. 2016.
34. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet.

2015.
35. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014.
36. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, et al. XSuper-enhancers in the control of cell identity and disease. Cell. 2013.
37. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 2013.
38. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011.
39. Chadwick LH. The NIH Roadmap Epigenomics Program data resource. Epigenomics. 2012.
40. Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet. 2013.

Figure legends

Figure 1

(a) Distribution of mCpH levels around mCpGs in human PSCs. (b) Distribution of mCpH levels around mCpGs in brain and control tissues. Numbers following underscores identify biological replicates. (c) Difference in mCpH levels at mCpG-proximal (< ±100 bp) and CpG-distal (> ±100 bp) CpH sites. Beta values were used as average methylation levels, and P-value was calculated by the Wilcoxon signed-rank test. (d) Distribution of methylation levels around mCpGs and unmethylated CpGs in DNMT1-knockout mouse ESCs. mCpG, methylated CpG; mCpH, methylated non-CpG; PSC, pluripotent stem cell; Br, brain; y, year; d, day.

Figure 2

(a) Schematic representation of the three states of the HMM. In the P-state, either both CpGs and CpHs are methylated or neither is methylated, resulting in a positive correlation between the mCpG and mCpH levels. In the N-state, only CpGs or CpHs are methylated, resulting in a negative correlation between the levels. In the U-state, the levels are not correlated. The bin size was set to 180 bp because DNMT3a and DNMT3b methylate DNA strands with a ~180 bp periodicity. N-state bins are considered as CpG–CpH differentially methylated regions (CpG–CpH DMRs). (b) Schematic diagram of the state transition in the HMM. The transition probability was estimated using the EM algorithm. (c) Distribution of methylation levels of each bin assigned to one of the three states, using the highest emission probability. (d) Distribution of methylation levels of each bin assigned to one of the three states, using the Viterbi algorithm. (c) and (d) used 180 bp bins in chromosome 19 of sample Br-81y_1.

Figure 3

(a) Fraction of the P-, N-, and U-state regions detected by the highest emission probability or by the Viterbi algorithm. EP, emission probability. (b) Fraction of whole genome and of the known genomic elements overlapped with DMRs. The P-value of Student's t-test between N-state regions in the whole genome, and in each genomic element within the brain samples, was marked as follows: *, <0.01; **, <0.001. (c) Distribution of mCpH levels in P- and N-state regions of brain cells. (d) Fraction of whole genome and of the known genomic elements overlapped with hyper- or hypo-mCpH-Ns.

Figure 4

Results of Gene Ontology (GO) enrichment analysis of 1717 genes associated with hyper-mCpH-N enhancers: (a) Enriched GO biological process terms (FDR<0.05). Top 20 terms were shown.

(b) Enriched GO cellular component terms. (c) Enriched GO molecular function terms. (d) Biological process terms enriched in 1467 genes associated with hypo-mCpH-N enhancers. No cellular component or molecular function terms were significantly enriched for the gene set.

Additional file

Figures

Figure S1

Distribution of mCpH levels around mCpGs in human PSCs and tissues. mCpG, methylated CpG; mCpH, methylated non-CpG; PSC, pluripotent stem cell.

Figure S2

(a) Distribution of mCpH levels around mCpGs in wild-type and DNMTs-KO mouse ESCs. DNMT-KO+3a, KO samples with DNMT3a re-introduced; DNMT-KO+3b, KO samples with DNMT3b re-introduced. (b) Distribution of mCpH levels around mCpGs in wild-type, DNMT3a-KO, and DNMT3b-KO human ESCs. (c) Density plot showing the distribution of mCpG and mCpH levels in 1 kb genomic regions. Regions containing more than 10 CpGs and 10 CpHs were considered. As the density increases, the color gradually changes based on the color scheme grey (lowest density)–blue–red–yellow (highest density). WT, wild type; ESC, embryonic stem cell; KO, knockout; PCC, Pearson correlation coefficient.

Figure S3

(a) Example of the Viterbi algorithm detecting an N-state region. (b) Transition probabilities in human samples. Probability learning was done within a chromosome block in which undetected bins were not longer than 100,000. Error bar indicates standard error among the transition probabilities from all the chromosome blocks (c) Distribution of precision and recall rates of the results obtained by highest emission probability (left) and by the Viterbi algorithm (right). TP, true positive; FP, false positive, FN, false negative; n, the total number of randomly generated N-state regions (length of the regions are randomly set); Highest EP, the highest emission probability.

Figure S4

Clustering dendrogram based on the Jaccard index of DMRs in all the possible pairs of WGBS samples.

Figure S5

Examples for the mCpH enrichment around genes in the brain. (a) *PICALM* (Phosphatidylinositol Binding Clathrin Assembly Protein), which is associated with Alzheimer's disease. (b) *CD28* (Cluster of Differentiation 28), which is associated with aging and dementia.

Tables

Tables S1

Description of WGBS data used in this study

Tables S2

List of hyper-mCpH-N enhancers and their putative target genes

Tables S3

List of hypo-mCpH-N enhancers and their putative target genes

Tables S4

Genes involved in the biological process “Immune system development”

Figures

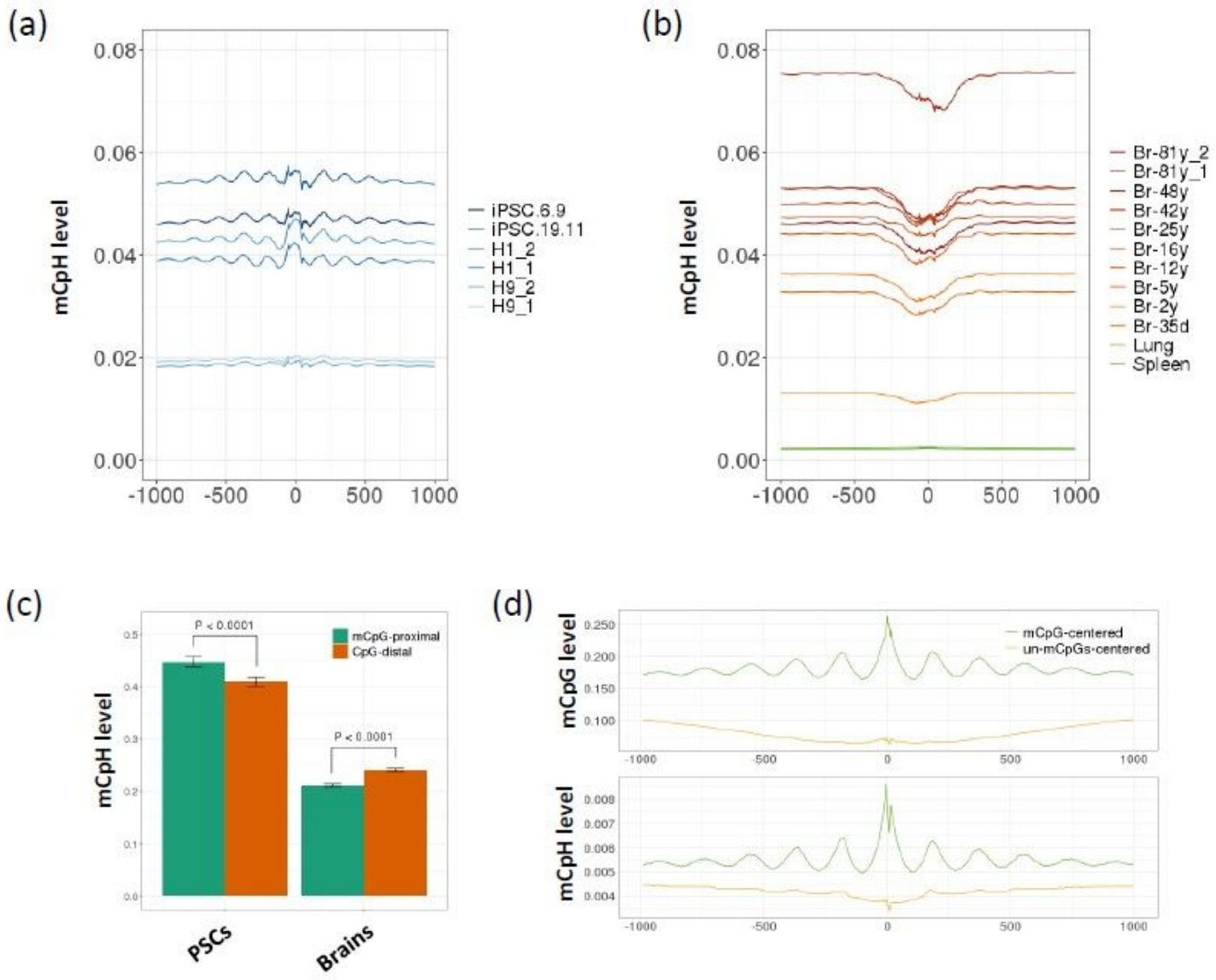

Figure 1

(a) Distribution of mCpH levels around mCpGs in human PSCs. (b) Distribution of mCpH levels around mCpGs in brain and control tissues. Numbers following underscores identify biological replicates. (c) Difference in mCpH levels at mCpG-proximal (< ±100 bp) and CpG-distal (> ±100 bp) CpH sites. Beta values were used as average methylation levels, and P-value was calculated by the Wilcoxon signed-rank test. (d) Distribution of methylation levels around mCpGs and unmethylated CpGs in DNMT1-knockout mouse ESCs. mCpG, methylated CpG; mCpH, methylated non-CpG; PSC, pluripotent stem cell; Br, brain; y, year; d, day.

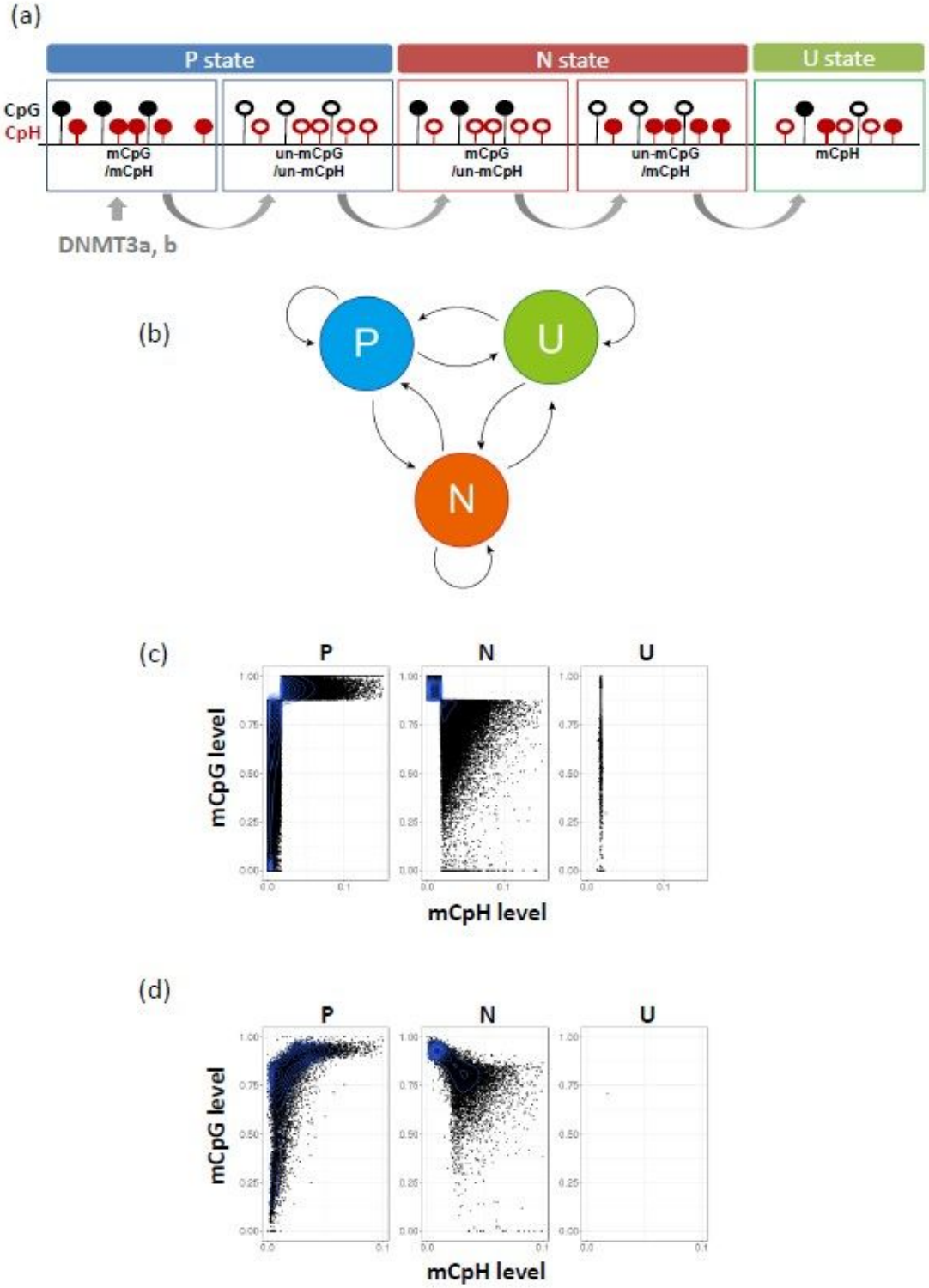

Figure 2

(a) Schematic representation of the three states of the HMM. In the P-state, either both CpGs and CpHs are methylated or neither is methylated, resulting in a positive correlation between the mCpG and mCpH levels. In the N-state, only CpGs or CpHs are methylated, resulting in a negative correlation between the levels. In the U-state, the levels are not correlated. The bin size was set to 180 bp because DNMT3a and DNMT3b methylate DNA strands with a ~180 bp periodicity. N-state bins are considered as CpG–CpH

differentially methylated regions (CpG–CpH DMRs). (b) Schematic diagram of the state transition in the HMM. The transition probability was estimated using the EM algorithm. (c) Distribution of methylation levels of each bin assigned to one of the three states, using the highest emission probability. (d) Distribution of methylation levels of each bin assigned to one of the three states, using the Viterbi algorithm. (c) and (d) used 180 bp bins in chromosome 19 of sample Br-81y_1.

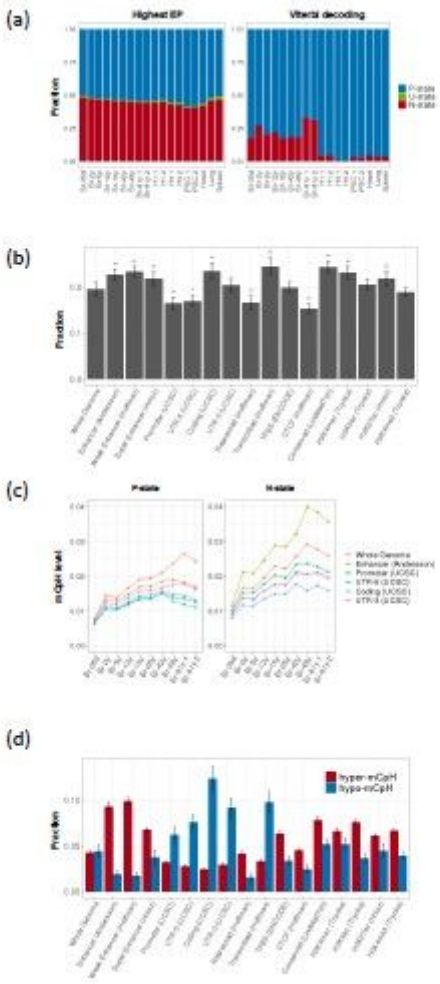

Figure 3

(a) Fraction of the P-, N-, and U-state regions detected by the highest emission probability or by the Viterbi algorithm. EP, emission probability. (b) Fraction of whole genome and of the known genomic elements overlapped with DMRs. The P-value of Student's t-test between N-state regions in the whole genome, and in each genomic element within the brain samples, was marked as follows: *, <0.01; **, <0.001. (c) Distribution of mCpH levels in P- and N-state regions of brain cells. (d) Fraction of whole genome and of the known genomic elements overlapped with hyper- or hypo-mCpH-Ns.

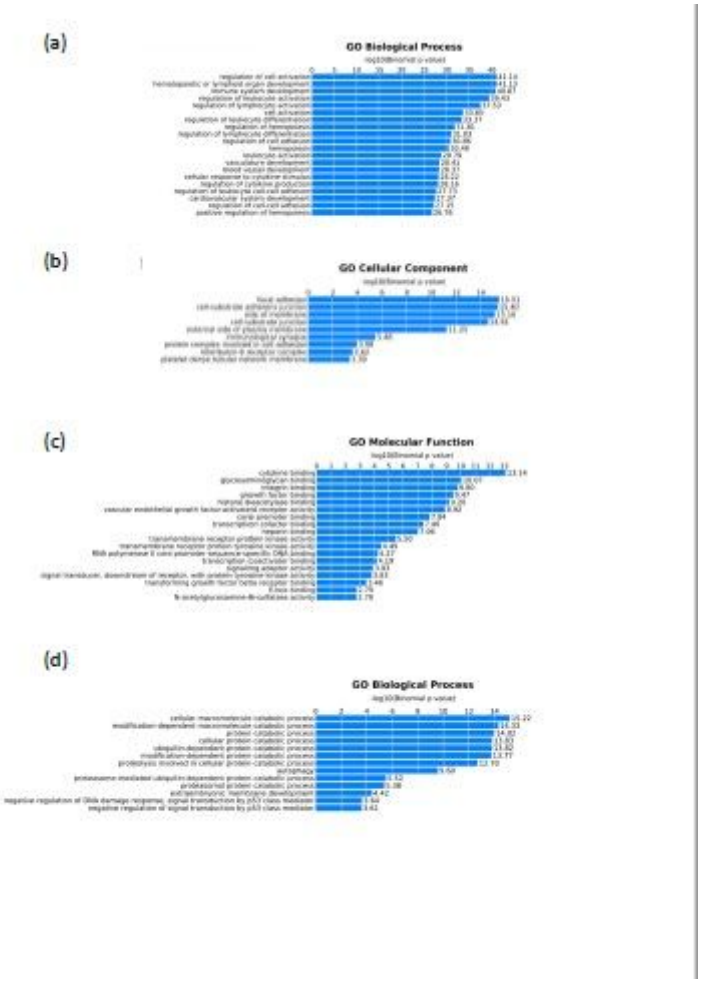

Figure 4

Results of Gene Ontology (GO) enrichment analysis of 1717 genes associated with hyper-mCpH-N enhancers: (a) Enriched GO biological process terms (FDR<0.05). Top 20 terms were shown. (b) Enriched GO cellular component terms. (c) Enriched GO molecular function terms. (d) Biological process terms enriched in 1467 genes associated with hypo-mCpH-N enhancers. No cellular component or molecular function terms were significantly enriched for the gene set.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.xlsx
- SupplementaryFigures.pdf