論文の内容の要旨

論文題目　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）

氏　　名　李鍾勳

Introduction

DNA methylation, an addition of methyl group on fifth carbon at cytosine, is one of the most important epigenetic modifications. For decades, researchers have focused on methylated CpG dinucleotides (mCpGs) that govern cell type specific functions and cause diseases by regulating transcription [1]. However, recent studies have found that significant amount of methylated CpHs (mCpH; H includes A, C, and T) are observed in pluripotent stem cells and non-dividing cells [2]. The mCpHs pattern over whole genome is associated to cell type specific phenomena such as cell differentiation and brain maturation [3, 4]. In this way, the mCpH, as well as mCpG, emerged as a possible regulator of cell type specific functions, especially in pluripotent stem cells and non-dividing cells such as neuron.

The research of mCpH, however, is limited by two issues. First, detection of mCpH is difficult. The CpHs are generally little methylated (~2%) and distributed genome widely, so that to detect mCpH pattern, the whole genome should be scanned with great accuracy. The Whole genome bisulfite sequencing (WGBS) is an optimized method for detecting genome-wide mCpH pattern; however, it takes great amount of time and cost. Thus, it is limited to secure large number of samples in which mCpHs are accurately detected. Second, it is hard to uncover independent role of mCpH. The DNA methyltransferase 3a and 3b (DNMT3a and DNMT3b, respectively) are responsible for the methylation at both CpG and CpH sites, resulting in spatial correlation between mCpG and mCpH [5]. Thus, it is hard to uncover roles of mCpHs that independent to mCpGs. In this way, the research about mCpH is limited by difficulty on detection and dependency on mCpG.

In this study, we tried to uncover the independent role of mCpH on biological processes by resolving the two difficulties. First, we improved the detection accuracy by developing a novel method for analyzing WGBS data (Section 1). Using the method, we found that DNMT3a and DNMT3b preferentially methylate cytosines at CpHpH and CpHpG context, respectively, which results in differential distribution and function of mCpH in embryonic stem cell (ESC) and neuron (Section 2). Then, we divided the mCpHs into two groups, CpG-proximal and CpG-distal mCpHs, and figured out that CpG-distal mCpH is highly functionally related to brain maturation (Section 3). Altogether, our study sheds light on the distribution and function of mCpH in mammalian cells.

Section 1: An integrative approach for efficient analysis of whole genome bisulfite sequencing data

Whole genome bisulfite sequencing (WGBS) is a high-throughput technique for profiling genome-wide DNA methylation at single nucleotide resolution. However, the applications of WGBS are limited by the high complexity induced by the conversions from bisulfite treatment. Although several computer programs have been developed for accurate detecting, most of the programs have succeeded in improving either quantity or quality of the methylation results. To improve both, we attempted to develop a novel integration of the most widely used bisulfite-read mappers: Bismark [6], BSMAP [7], and BS-seeker2 [8].

Our comprehensive analysis of the three mappers revealed that the mapping results of the mappers were mutually complementary under diverse read conditions, such as read length or sequencing quality. 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 (Figure 1-a). In addition, the amount of detected cytosine was higher than that by Bismark. Furthermore, the integration successfully reduced the fluctuation of detection accuracy induced by read conditions. We applied the integration to real WGBS samples and succeeded in clustering the samples into their originated tissues or cell types by both CpG and CpH methylation patterns (Figure 1-b). It was not clearly distinguished by BS-seeker2.

This section contributes to DNA methylation researches by improving efficiency of methylation detection from

WGBS data and facilitating the comprehensive analysis of public WGBS data. The results have been published in BMC genomics [9].

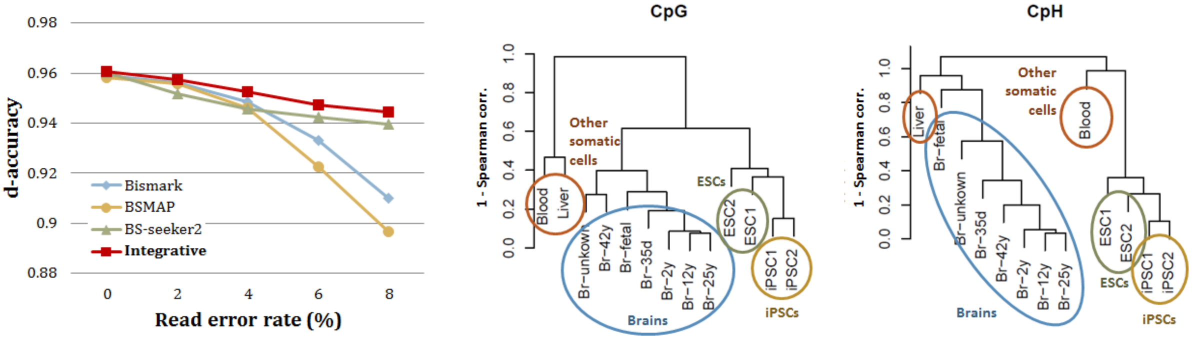


Figure1. a) Methylation detection accuracy in simulated data. b) Hierarchical clustering results of WGBS samples by CpG and CpH methylation pattern

Section 2: Differential activity of DNMT3a and DNMT3b causes distinct distribution and function of non-CpG methylation in embryonic stem cell and neuron.

The methylated non-CpG dinucleotides (mCpH) are significantly abundant in pluripotent stem cells and non-dividing cells such as neuron. Interestingly, distribution and function of the mCpH in those two cell types are highly distinct. The abundant motif at mCpH is "CAG" in ESC, whereas it is "CAC" in neuron [2]. Also, the mCpH level is positively correlated to gene expression in ESC, whereas it is negatively correlated in neuron. These cell type specific features of mCpH have been a mystery among researchers for years.

In this section, we found that the characteristics of mCpH are resulted from differential activity of DNA methyltransferase 3a and 3b (DNMT3a and DNMT3b, respectively). Applying the integrative approach for bisulfite-read aligning (Section 1), we collected whole genome bisulfite sequencing (WGBS) data of 16 human and mouse ESCs, and 8 human brains and neurons, with improved quality. Through comprehensive analysis of DNMT knockout human and mouse ESCs, we confirmed that CpHs are methylated by DNMT3a and DNMT3b, dependently to the methylation at CpG. Interestingly, the DNMT3a tends to methylate cytosine at CpHpH context, whereas DNMT3b at CpHpG context, especially nearby CpGs. Based on their differential expression in ESC and neuron, we concluded that the distinct mCpH motifs in those two cell types are resulted from the differential activity of DNMT3a and DNMT3b. Also, we found that the positive correlation between mCpH level and gene expression in ESC is caused by the preferential interaction between DNMT3b and highly expressed gene-body regions. Collectively, our study revealed that differential contribution of DNMT3a and DNMT3b to CpHpH and CpHpG contexts cause distinct characteristics of mCpH in ESCs and neurons.

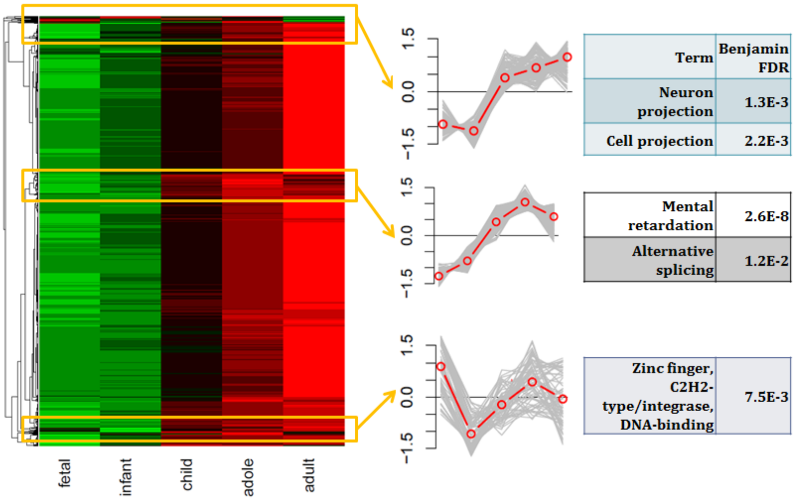
Section 3: Roles of non-CpG methylation on brain maturation

Since the DNMT3a and DNMT3b methylate both CpG and CpH dinucleotides, the mCpH is spatially correlated to mCpG. Based on the correlation and greatly higher affinity of DNMTs on CpGs, some researchers insist that mCpH is merely by-product from over-activity of DNMT3a and 3b that originally target CpG [5]. Subsequent evidences, however, support that mCpH plays roles over cell type specific phenomena such as brain maturation, neuro-diseases, and cell differentiation [3, 4]. Thus, in this section, we tried to uncover function of mCpH that independent to mCpG. Through a comprehensive analysis of correlation between mCpG and mCpH, we divided the mCpHs into two groups, CpG-proximal and CpG-distal mCpHs, based on the distance from proximal CpGs (threshold = 100bp). Interestingly, in brain, mCpHs are abundant at CpG-distal regions, implying that large portion of mCpHs is independently formed to mCpGs. The tendency was double-checked with 224 bisulfite-treated microarray data of human brain and neuron. In further analysis, we clustered genes by the CpG-distal mCpH pattern across brain aging. Interestingly, the genes that shares similar mCpH pattern showed enriched gene ontology terms (GO terms) of "zinc-finger protein activity" and "mental retardation". On the contrary, the genes clustered by mCpG pattern did not show any enriched GO term that related to brain specific activity (Figure 2).

In summary, we uncovered that large number of mCpHs are independently formed to mCpGs in brain, and the mCpHs are highly related to GO terms such as "zinc-finger activity" and "mental retardation".

Conclusion

In this study, we tried to uncover the formation, distribution, and function of mCpHs in mammalian cells. The integrative approach for WGBS data analysis greatly improved both quantity and quality of methylome, and facilitated gathering large number of samples for statistically robustness. Also, by analyzing DNMT knockout samples, we uncovered the mysterious of distinct mCpH characteristics over mammalian cells. Lastly, the functional analysis of CpG-independent mCpHs suggested that mCpH regulate genes related to “zinc-finger activity” and “mental retardation”. Altogether, this study gives insights about the formation, distribution and function of methylated non-CpG in mammalian cells via functional analysis *in silico*.



Term	Benjamin FDR
Neuron projection	1.3E-3
Cell projection	2.2E-3
Mental retardation	2.6E-8
Alternative splicing	1.2E-2
Zinc finger, C2H2-type/integrase, DNA-binding	7.5E-3

Figure2. Genes were clustered by mCpH methylation pattern in gene-body regions across brain aging. The tables on the left side show GO term enrichment in each gene set.

References

1. Holliday, R. and J.E. Pugh, *DNA modification mechanisms and gene activity during development.* Science, 1975. **187**(4173): p. 226-32.
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. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-22.
4. Lister, R., et al., *Global epigenomic reconfiguration during mammalian brain development.* Science, 2013. **341**(6146): p. 1237905.
5. 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.
6. Krueger, F. and S.R. Andrews, *Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.* Bioinformatics, 2011. **27**(11): p. 1571-2.
7. Xi, Y. and W. Li, *BSMAP: whole genome bisulfite sequence MAPping program.* BMC Bioinformatics, 2009. **10**: p. 232.
8. Guo, W., et al., *BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data.* BMC Genomics, 2013. **14**: p. 774.
9. Lee, J.H., S.J. Park, and N. Kenta, *An integrative approach for efficient analysis of whole genome bisulfite sequencing data.* BMC Genomics, 2015. **16 Suppl 12**: p. S14.