이학박사학위논문

A Statistical Method on DNA Methylation Calling and its Application with Next generation sequencing technique

차세대 시퀀싱 기술을 이용한 메틸화 판정 및 그 응용에 관한 통계적 방법론 연구

2015년 8월

서울대학교 대학원

통계학과

허 익 수

A Statistical Method on DNA Methylation Calling and its Application with Next generation sequencing technique

지도교수 박태성

이 논문을 이학박사 학위논문으로 제출함

2015 년 08 월

서울대학교 대학원

통계학과

허 익 수

허익수의 이학박사 학위논문을 인준함

2015 년 06 월

위 원 장 조 신섭 (인)

부위원장 박 태성 (인)

위 원 오 희석 (인)

위 원 임 요한 (인)

위 원 이 수진 (인)

A Statistical Method on DNA Methylation Calling and its Application with Next generation sequencing technique

by

Iksoo Huh

A thesis
submitted in fulfillment of the requirement
for the degree of Doctor of Philosophy
in Statistics

Department of Statistics

College of Natural Sciences

Seoul National University

Aug, 2015

Abstract

A Statistical Method on DNA Methylation Calling and its Application with Next generation sequencing technique

Iksoo Huh

Department of Statistics

The Graduate School

Seoul National University

Epigenomics is the study of biological factors that induce mitotically and/or meiotically heritable changes in gene functions, except by changes in deoxyribonucleic acid (DNA) sequence. Representative mechanisms that make such changes are DNA methylation and histone modification. These two mechanisms control gene expression via changing affinities of transcription factor and/or altering patterns of DNA packing, rather than altering the underlying DNA sequence. Those epigenetic processes play roles in imprinting, gene silencing, X chromosome inactivation, position effect, maternal effects, and the progress of carcinogenesis. Therefore, the importance of the epigenetic process is rapidly increasing.

Especially, DNA methylation is one of the mostly interesting and vigorously studied phenomena. Methylation process is defined as addition of a methyl group

to a substrate, or the substitution of an atom or molecular group by a methyl group. In DNA methylation process, methylation arises in the cytosine, which is one component of the four nucleotides: guanine (G), adenine (A), thymine (T), and cytosine (C). After methylation process on cytosines, those would be 5-Methylcytosine. This process especially actively occurs in cytosines that are located next to a guanine nucleotide in the DNA sequence. It is defined as CpG sites.

There have been many biochemical techniques to measure level of methylation. Firstly developed techniques are based on immunoprecipitation techniques. The techniques uses phenomenon of immunoprecipitation using florescent antibodies that combines to 5-Methylcytosines (Methylated DNA immunoprecipitation: MeDIP). Whole genomic region are divided into many small region such as genes and those divided DNA sequences are located in their own location in microarray panel. After hybridization of the sequences with florescent antibodies to reference sequences in the microarray panel, we measure light intensities of all microarray spots and the intensities would be regarded as methylation intensities of genomic regions (MeDIP-chip technique).

The MeDip-chip techniques are widely used to get methylation level information of genomic regions. However, the microarray based approaches have limitation because they use pre-defined sequences. To overcome the problem, Next-Generation sequencing (NGS) techniques were combined to MeDIP approach (MeDIP-seq). After DNA sequence fragments with florescent antibodies are sequenced and mapped to the reference whole genome, we can obtain methylated CpG cytosine regions and its intensities by sequencing read depth. Then we can tell whether some genomic regions are methylated or non-methylated. However, the

MeDIP-seq techniques also have low resolution (several dozen base pairs) because they only count numbers of mapped DNA fragments that are methylated as methylation intensities.

To overcome these limits of MeDIP, a new method is recently developed based on NGS and Bisulfite treatment. In order to discriminate non-methylated cytosines and methylated cytosines, the technique included bisulfite treatment which converts non-methylated cytosine into thymine and we then estimate methylation level by measuring ratio between numbers of cytosines and thymines in each CpG cytosine site. Because we can obtain information of base-pair resolution methylation from the technique, new statistical methods are needed to handle this new type data. The first issue is to develop method which classifies binary methylation status (methylation calling) and the second issue is to develop method which detects differentially methylated region (DMR).

For those two issues, we proposed two statistical approaches. For the binary methylation binary calling issue, we propose a new classification tool using bayes classifier and local information (Bis-Class). This method used the biological phenomenon that methylation status are spatially correlated, therefore the method performs better than binomial test using false discovery rate (FDR) especially on the condition of low coverage depth and low methylation level. We showed advantages of our method through simulation and real data analysis using honeybee dataset. For the differential methylated region (DMR) detecting method, we proposed a modified Cochran–Mantel-Haenszel (CMH) statistic.

The original CMH statistic was proposed to test conditional independence between two variables with stratification. However, because there is substantial and

spatial correlation of methylation level between adjacent CpG cytosine sites, we additionally included spatial correlation structure and impose biological importance weights on the binary called base-pair resolution information. Moreover, the method has advantages that it can be applied to more various situations such as analysis of ordinal or multinomial response. We compared our method with Fisher′s exact test that has been used for binary called bisulfite sequencing data. Using the modified CMH test, we can avoid type 1 error inflation and handle multiple biological replicated samples in each experimental group. We also conducted simulation study and real data analysis using honeybee bisulfite sequencing dataset to detect differential methylated region.

We expect that our proposed methods to handle bisulfite sequencing data via NGS techniques are widely used to elucidate biological relationships between epigenetic data and many biological endpoints such as cancers, aging, gene silencing, etc.

Keywords: DNA Methylation, Next Generation Sequencing (NGS), Bisulfite treatment, Methylation binary calling, Differential Methylation Region (DMG) test

Student number: 2008-20272

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background of Deoxyribonucleic acid (DNA) Methylation process

1.1.1 Definition of Methylation process

Epigenomics is the study of biological factors that induce mitotically and/or meiotically heritable changes in gene functions, except by changes in deoxyribonucleic acid (DNA) sequence. Representative mechanisms that make such changes are DNA methylation and histone modification. These two mechanisms control gene expression via changing affinities of transcription factor and/or altering patterns of DNA packing, rather than altering the underlying DNA sequence. Those epigenetic processes play roles in imprinting [1], gene silencing [2], X chromosome inactivation [3], maternal effects [4], and the progress of

carcinogenesis [5]. Therefore, the importance of the epigenetic process is rapidly increasing.

Especially, DNA methylation is one of the mostly interesting and vigorously studied phenomena. Methylation is the process that a methyl group is added to a substrate or the substitution of an atom or group by a methyl group. In DNA methylation process, methylation arises in the cytosine, which is one component of the four nucleotides: guanine (G), adenine (A), thymine (T), and cytosine (C). After methylation process on cytosines, those would be 5-Methylcytosine. This process especially actively occurs in cytosines that are located next to a guanine in the DNA sequence. It is defined as CpG sites. Some genomic regions where there are dense CpGs are names as CpG islands. It is very interesting that CpGs in CpG islands are generally lower methylated than CpGs that are not located in CpG islands. This phenomenon is thought to be due to that genetic mutation are more likely to happen in methylated CpG cytosines. Therefore, CpG islands are being depleted as time flows and it may affect to genomic sequences indirectly.

1.1.2 Basic Biological functions of Methylation process

DNA methylation process generally affects to biological phenomenon by controlling gene expression level. Many important biological processes have been known to be related or induced by the DNA methylation.

One of the most important functions of the DNA methylation is to develop cancers [6]. DNA methylation is known to be an important regulator of gene transcription by many studies, and those studies have shown that highly methylated genes especially in their promoter region would be transcriptionally silent. Compared with normal tissues, there are two aberrant statuses: hypermethylation

and hypomethylation, that means higher density of methylation level and lower density of methylation level, respectively. These two abnormal statuses can induce cancers via opposite mechanisms because hypermethylation of genes are generally silenced while hypomethylation of genes are over-expressed. Therefore, hypermethylation induces cancers through silencing tumor suppressor genes while hypomethylation induces cancers through activating oncogenes.

The other functions are related to aging process. Although there are controversial and contradictory results of methylation effect on aging, many specific CpG cytosines are known to be statistically associated with the aging process. [7]. Those found CpG cytosines provide well prediction of aging, and therefore they can be used as biological, or epigenetic clock that promise biomarkers of aging.

The DNA methylation is also related to development process of organisms. When female organisms are growing, one of two X−chromosomes is inactivated to avoid overdose of gene expression in X−chromosomes [8]. In that process, hypermethylation of inactivated X-chromosomes silences gene expression. Genomic imprinting that is the epigenetic phenomenon by which certain genes are expressed in a parent-of-origin-specific manner is also associated with the DNA methylation process. As a result of the DNA methylation, genes are expressed only in non-imprinted region. Therefore, if a gene from father is imprinted, it would be silenced while the gene from mother, which is not imprinted, would be expressed.

1.2 Review of methylation measuring techniques

1.2.1 Immunoprecipitation-based methylation measuring technique (MeDIP)

The firstly proposed technique to measure methylation level is based on immunoprecipitation. Immonoprecipitation is the reaction of hybridizing antigens and antibodies (″immunity″) and sinking of hybridized compound (″precipitation″).

For methylated CpG cytosines, there are complementary antibodies and we used it to separate from the non-methlyted CpG cytosines. Before the process, DNA sequences are divided into many several kb length fragments via sonification. Then, via the immunoprecipitation, only DNA fragments which include methylated CpGs are selected. These sequential process are named as Methylated DNA immunoprecipitation (MeDIP), and they are further applied to microarray-chip based technique (MeDIP-chip) [9] or sequencing technique (MeDIP-seq) [10]

In the MeDIP-chip, DNA fragments including methylated CpGs labeled with cyanine-3 (Cy3; green), while those including non-methylated CpGs labeled with cyanine-5 (Cy5; red). Then both are hybridized to the genomic sequences in a microarray pannel. There are several hundreds of thousands spots containing predefined genomic sequences in the pannel. Therefore, if some of genomic regions are highly methylated, more DNA fragment labeled with green fluorescent light than those with red fluorescent light can be hybridized to the according spots. After hybridization, we eventually measure light intensities from both fluorescent colors (Figure 1-1A).

In the MeDIP-seq, instead of labeling, it sequences methylated DNA fragments by various sequencing techniques including NGS technique. Then it

maps the fragments to the whole genome reference and evaluates how many fragments are attached to some specific regions. If some genomic regions are highly methylated, amount of the DNA fragments would be large and density of the fragment to the region also would be high while no fragment would be attached to the non-methylated region (Figure 1-1B).

Those two representative MeDIP based technique are widely used to obtain information of methylation level in genomic region. However, they have technical limitations. In the MeDIP-chip technique, methylation level of only predefined regions can be measured because all spots are prepared before analysis with fixed length. Therefore, it is not possible to measure methylation level of unprepared regions. Although the MeDIP-seq technique can measure any region in a whole genome because it sequenced all methylated fragments and mapped them to the original reference, it has also limitation of resolution of detection. Generally, lengths of the fragments are from 300bp to 1000bp used in MeDIP-chip technique, and from 40bp to 400bp used in MeDIP-seq technique. Therefore, we only measure regional methylation level instead of one base-pair resolution. It restricts obtaining detailed information and conducting analysis.

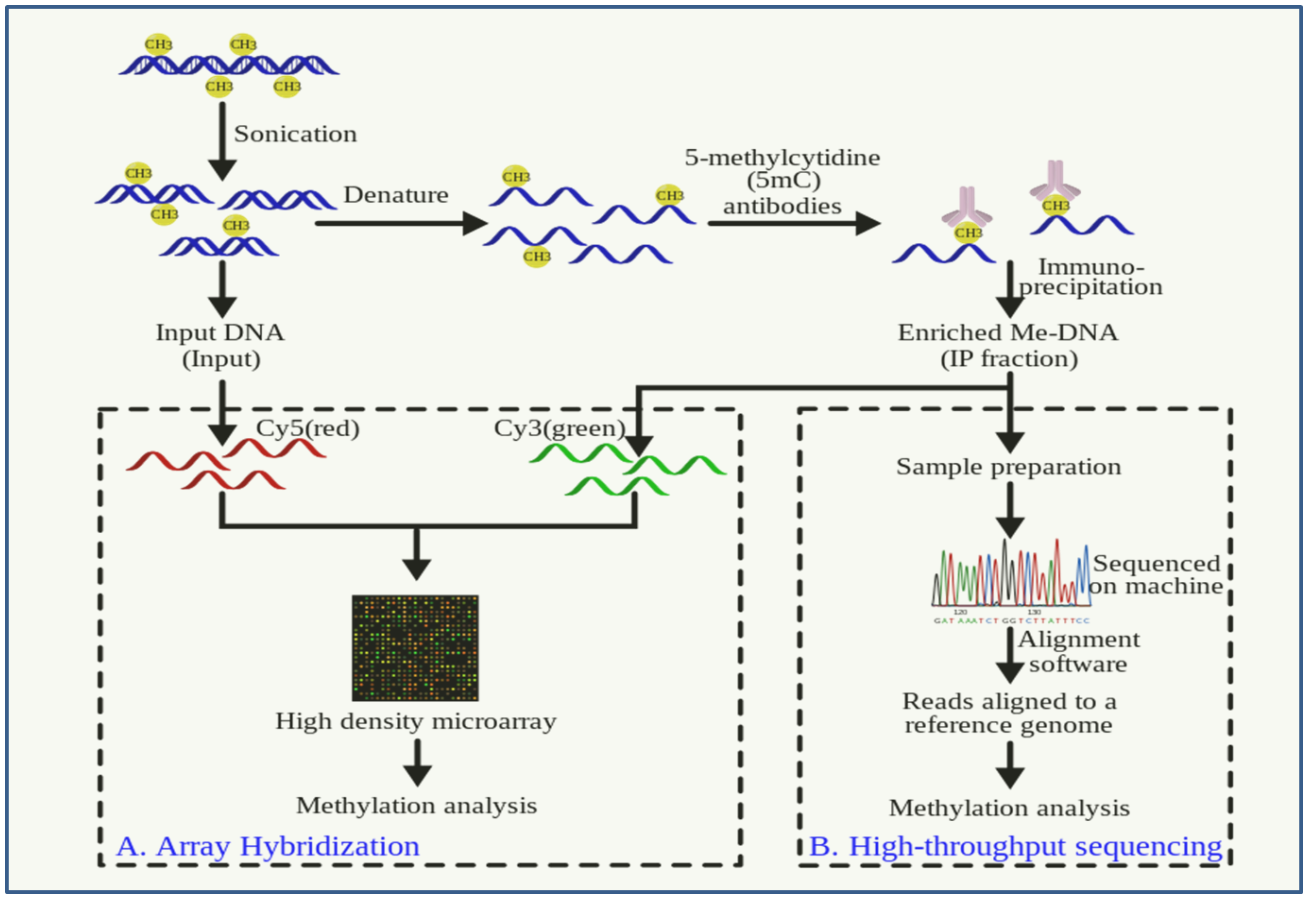

Figure 1-1. Schematization of the MeDIP-based techniques (referred to https://en.wikipedia.org/wiki/Methylated_DNA_immunoprecipitation#/media/File:MeDIP.svg)

1.2.2 Bisulfite-sequencing based methylation level measuring technique (BS-seq)

In order to overcome the limitations of immunoprecipitation based methylation measuring method, bisulfite treatment using NGS sequencing technique was developed. It is also called bisulfite sequencing (BS-seq) [11]. Bisulfite means the HSO_3^- ion. When bisulfite treatment is applied to genomic sequences, methylated cytosines are not affected while non-methylated cytosines are converted to uraciles. Uracil is one of four elements in mRNA sequence, and it is alternative form of thymine in DNA sequence. Therefore, after polymerase chain reaction (PCR) process, methylated cytosine remains cytosine while non-methylated cytosines are converted to thymines. Then DNA fragments are sequenced and mapped to reference. Based on the assumption that sequencing error can be ignorable, DNA fragments (it is also named as sequencing reads) having thymine in the cytosine location of reference genome are regarded as non-methylated in that site (Figure 1-2). After sequencing and mapping are finished, numbers of cytosines and thymines in each CpG cytosine site are counted. If a CpG cytosine site has a lot of mapped reads having cytosine and few reads having thymine can be classified as methylated site. In opposite case, the site can be classified as non-methylated site, vice versa. Even though we consider experimental errors during bisulfite treatments, the properties are still valid to determine the status of methylation on each cytosine.

This technique provides information of methylation via count of cytosine reads and thymine reads to all CpG cytosine sites. Therefore, we can make decisions about whether some CpG cytosine sites are methylated at base-pair resolution and the technique is regarded as gold standard of measuring methylation.

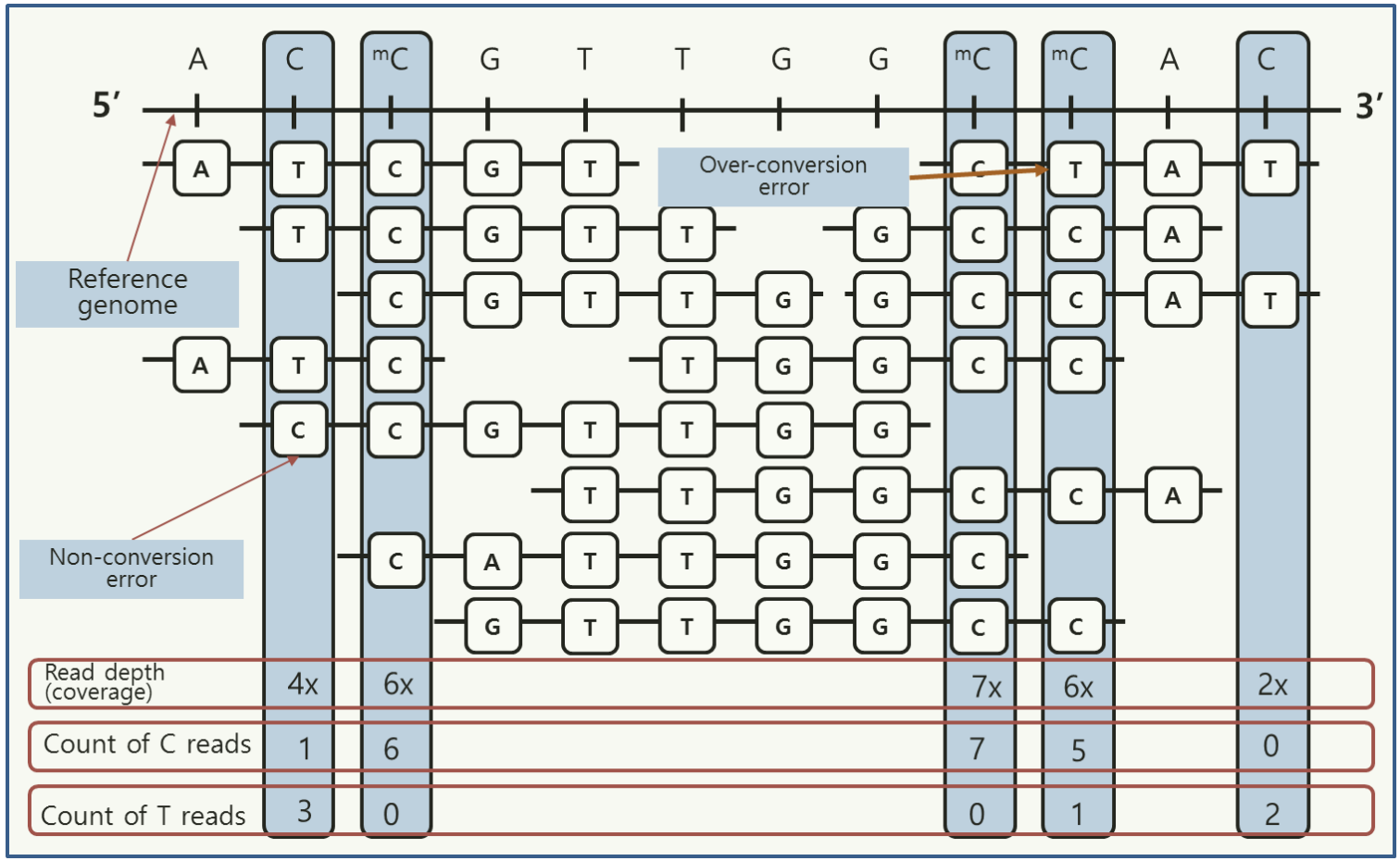

Figure 1-2. Schematization of sequencing reads alignment on a reference genome using the BS-seq technique

Because the cost of producing BS-seq data was expensive, a few samples have been produced, and used to validate data from other technique in initial stage of the BS-seq technique. However, usage of the BS-seq data is increasing, as the cost is gradually decreasing.

1.3. Purpose of this study

The main purpose of this thesis is to develop the statistical methods to classify binary status of methylation, and detect differentially methylated regions using categorically classified information from BS-seq dataset.

In the previous Binomial test using FDR adjustment which have been used widely to detect methylated sites, is not powerful to detect methylated locus when it is applied to BS-seq dataset of low methylation level and low read coverages. To overcome the limitations of the Binomial test, we used Bayes classifier method, and include local information of methylation level into the method (Bis-Class). Using the biological fact that methylation levels of CpG cytosines are spatially correlated, some locus having small amount of information that are not statistically significant in the Binomial method can be classified as methylated sites in our proposed method when it is located in highly methylated region. We validated better performance of our method via simulation study and real data analysis using Honeybee dataset.

In the later part, using binary information obtained from Bis-Class, we proposed modified CMH test to detect differentially methylated region (DMR) in the BS-seq dataset. Detecting DMRs is one of ultimate goals in DNA methylation analysis, therefore it is very important to develop statistical method for the DMR analysis. In the previous Fisher′s exact test used in many researches, there is type 1

error inflation because it can′t consider positive correlation between adjacent cytosine sites within several kilobytes. Therefore, we employed CMH test which uses stratification information that is regarded as locus information in our analysis and modified variance part via including correlation part from permutation test. In the simulation study, we found that our method preserves type 1 error rate. Then we applied our method to real dataset to detect DMRs using same dataset used in the previous part.

1.4 Outline of the thesis

This thesis is organized as follows. Chapter 1 is an introduction of basic knowledge including definition of DNA methylation process, its biological functions, and techniques of measuring methylation intensities. Chapter 2 includes overview of methylation measuring method in aspect of regional and base-pair resolution level, respectively. Chapter 3 is the study of a new binary methylation calling method (Bis-Class) in base-pair resolution. Chapter 4 is the study of detecting differentially methylated region (DMR) method of BS-seq dataset. Introduction of statistical method, conducting simulation studies, and real data analysis are included in both 3 and 4 chapters. Finally, all the results and conclusions are summarized in Chapter 5.

Chapter 2

Overview of methylation level measuring methods

2.1 Regional measuring methods

After data processing and handling such as hybridizing DNA fragments to microarray spots or mapping methylated sequencing reads to reference genome, it is needed to develop method that converts methylation level into quantitative values for additional analysis.

For the MeDIP-based technique, ratio between intensity of green fluorescent light and red fluorescent light is used to show methylation level. If the ratio is big, high level of methylation exists in the DNA sequence. However, there are many factors which may affects the ratio such as DNA sequence pattern, CpG density, and fragment length etc. Therefore, some normalization methods of fluorescent light intensity were proposed. One of the approaches is direct normalizing method comparing with sequence known to be fully methylated. The other method is normalizing from regression using CpG density as explanatory variable.

For BS-seq based technique, easier evaluation is used because informations of all CpG sites are given. If C_i and T_i denote counts of cytosine reads methylated read) and thymine reads (non-methylated read) for i^{th} CpG site in a region, methylation level of the region can be calculated as $\sum_i C_i / (\sum_i C_i + \sum_i T_i)$ for i=1,2,...,I [12]. However, this estimation can be biased when read number of each site (=coverage) are not equally distributed and either independent with methylation level. Therefore, $\sum_i C_i / (C_i + T_i)$ or $\sum_i \sqrt{n_i} \times C_i / (C_i + T_i)$ can be alternative form for regional methylation level estimation. In chapter 2.1.1, we introduced an article which used BS-seq dataset for estimation of gene-wise methylation level.

2.1.1 Application to explore relationship between transcriptional noise and DNA methylation

We introduced an article which uses gene methylation level calculated from BS-seq dataset. The paper is the publication that investigates the relationship between transcriptional noise and methylation level of gene bodies. Gene region is divided by two parts: Gene body, and promoter region. In particular, DNA methylation in transcription units (′gene bodies′) is highly conserved across diverse taxa. However, the functional role of gene body methylation is not yet fully understood.

A long-standing hypothesis posits that gene body methylation reduces transcriptional noise associated with spurious transcription of genes. Despite the plausibility of this hypothesis, an explicit test of this hypothesis has not been performed until now. Therefore, using nucleotide-resolution data on genomic DNA methylation and abundant microarray data, here we investigate the relationship

between DNA methylation and transcriptional noise. Transcriptional noise measured from microarrays scales down with expression abundance, confirming findings from single-cell studies. We show that gene body methylation is significantly negatively associated with transcriptional noise when examined in the context of other biological factors.

Levels of gene expression vary between cells even with the same genetic materials and under the same biological conditions [13-15]. Understanding the nature and mechanism of such variability, which is commonly referred to as 'transcriptional noise', has manifold functional consequences [16]. Recently, there have been significant improvements in experimental methods to measure transcriptional noise, as well as in the theoretical understanding of transcriptional noise. These studies indicate that transcriptional noise may occur due to transcriptional bursting of promoters, as well as spurious transcription within coding sequences [17-20]. Transcriptional noise in multicellular organisms, such as mammals, cannot be easily dissected using experimental means. However, they can be approximated using abundant expression datasets, for example utilizing normalized variation among microarray assays between replicates of populations [21-22]. For example, Yin et al. [21] compared the transcriptional noise measured from microarrays to those measured from single-cell experiments.

The two results correspond remarkably well [21]. Similar results were seen in another study, comparing expression variation among populations to experimentally measured transcriptional noise [22]. Following these approaches, in this study we approximated transcriptional noise of human genes as the coefficient of variation of transcriptional abundance, assayed between replicates of

populations of the same tissue samples under normal conditions.

There have been significant recent technical improvements in analysis of genomic DNA methylation. In particular, researchers have begun to generate whole-genome maps of DNA methylation at the nucleotide level, via whole-genome sequencing of bisulfite-converted genomic DNA [23-25]. This method quantifies the methylation level of each CpG dinucleotide across the whole genome, enabling us to discern gene body methylation levels for individual genes. In this study, we analyzed DNA methylation and transcriptional noise of the prefrontal cortex (brain) and the peripheral blood mononuclear cells (blood). We chose these two tissues for the following reasons. First, we decided to analyze 'normal' tissues (as opposed to cell lines). While there exists vast information on transcriptional variation of cell lines, gene expression profiles of cell lines are known to have significantly diverged from those of normal tissues [26]. Consequently, we chose not to consider cell lines in the current study. Second, we chose tissues whose genome-wide methylation maps are currently available. Finally, large numbers of microarray data in the 'control' (as opposed to disease) conditions exist for these tissues, thereby enabling us to measure transcriptional noise with confidence. We used rigorous quality control processes to curate microarray data from these tissues (see Methods). The resulting data are from the same technical platforms, and exhibit high correlation levels among experiments.

Gene expression data was obtained from National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov /geo/). Because there are considerable technical variations between platforms, we restricted platforms to only the Affymetrix Human Genome U133 series. After

quality control, we obtained a total of 52 datasets (12 datasets for the prefrontal cortex and 40 datasets for blood). Gene lengths were determined based upon the RefSeq annotation provided by the UCSC genome browser. Nucleotide resolution whole DNA methylation maps of the human prefrontal cortex (brain) were obtained from a recent study ([27], data available at NCBI Gene Omnibus under the record number GSE37202). DNA methylation maps of mature peripheral blood mononuclear cells were from Li et al. [24], generated using a similar method.

To obtain gene body methylation levels of non-repetitive portions of genes, we used the annotation of TEs from the RepeatMasker database (http://www.repeatmasker.org). A custom Perl script was used to mask the TEs in gene bodies. For each mapped cytosine, the fractional methylation value was calculated as: total number of ′C′ reads/(total number of ′C′ reads + total number of ′T′ reads), following previous studies [23,24,28]. We then calculated the fractional methylation level of each transcription unit, using the RefSeq database of hg18. Gene body methylation level for each gene was estimated as the mean fractional methylation value for all the mapped cytosines within each transcription unit. When alternative transcripts were present, we chose the longest transcript for each gene. The promoter methylation level for each gene was estimated as fractional methylation for regions spanning 1,500 bp upstream and 500 bp downstream of the transcription start site (TSS), similar to Zeng et al. [27].

Microarray raw data files were first processed using raw intensity using the MAS5.0 method [29]. Using other normalization methods provided similar results. We used the median probe intensities assigned to each gene as gene expression levels. We then analyzed correlation between pairwise samples, to assess

similarities between datasets from the same tissue. Datasets within the same tissues exhibiting correlation coefficient greater than 0.8 are included in this study (Figure 2.1). Quantile normalization using the R package 'preprocesscore' [30] was conducted within each tissue. Transcriptional noise was defined as the coefficient of variation (CV: standard deviation/mean) of transcriptional abundance within each tissue, following Yin et al. [21].

We performed multiple linear regression analyses to elucidate relationships between transcriptional noise and several biological factors (gene expression abundance, gene body methylation, promoter methylation, and gene lengths) simultaneously. CV and gene length were log transformed to improve normality. The actual equation for the multiple linear regression is below.

$$\log(\mathrm{CV}) = \beta_0 + \beta_1 \times \mathrm{expr} + \beta_2 \times \mathrm{GBM} + \beta_3 \times \mathrm{PRM} + \beta_3 \times \log(\text{gene length}) + \mathrm{e}$$

In this equation, expr, GBM, and PRM denotes gene expression, gene body methylation, and promoter methylation, respectively.

We found that, in both tissues, gene body methylation shows significant negative relations to transcriptional noise (Table 2.1). This is in accord with the hypothesis that gene body DNA methylation suppresses transcriptional noise [31]. Before the regression analysis, we confirmed that the transcriptional noise was negatively associated with expression abundance in both studied human tissues (Figure 2.2), and the fact were also observed in previous studies [18,19].

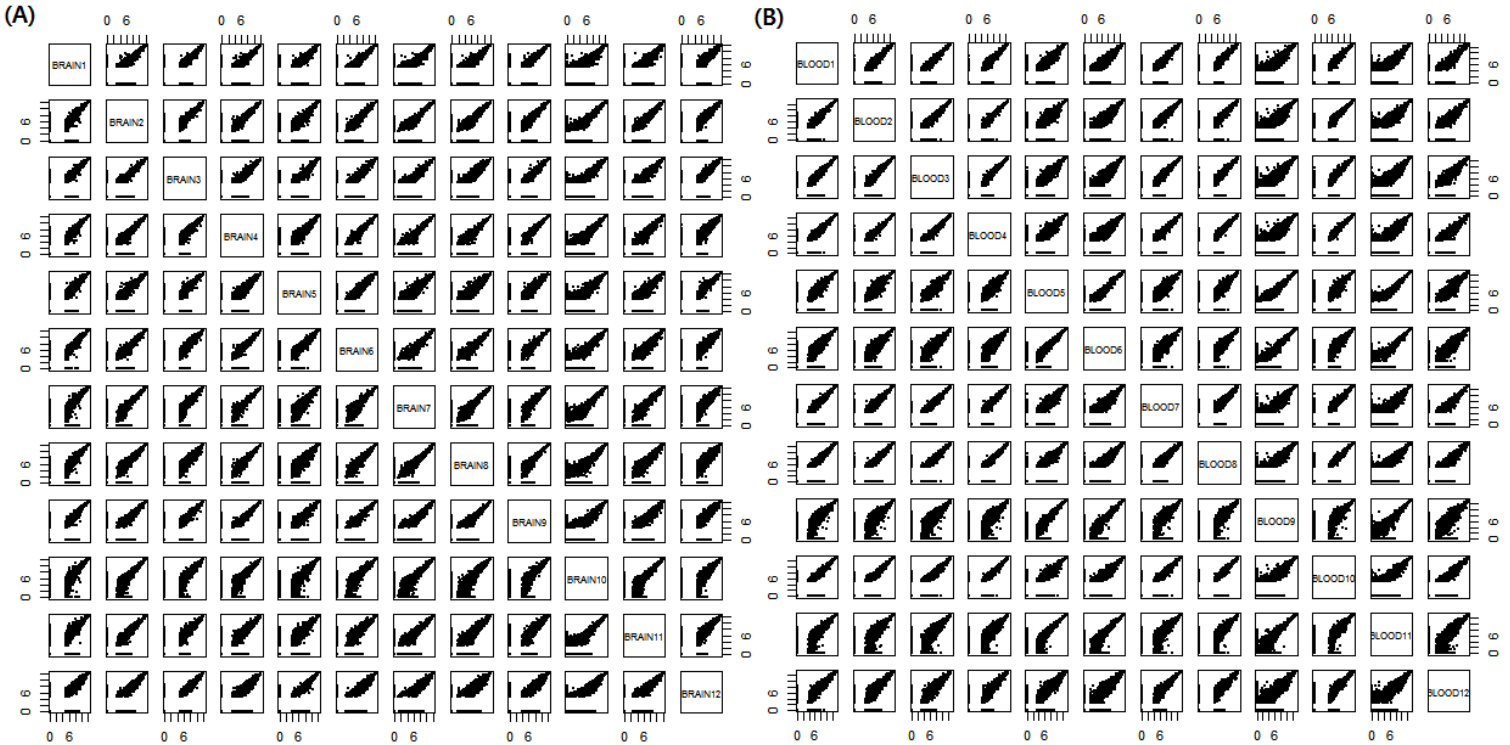

Figure 2.1 Pairwise scatterplots of microarray gene expression data between samples in each tissue (A) Plots of 12 brain samples (B) Plots of 12 blood samples among total 40 blood samples

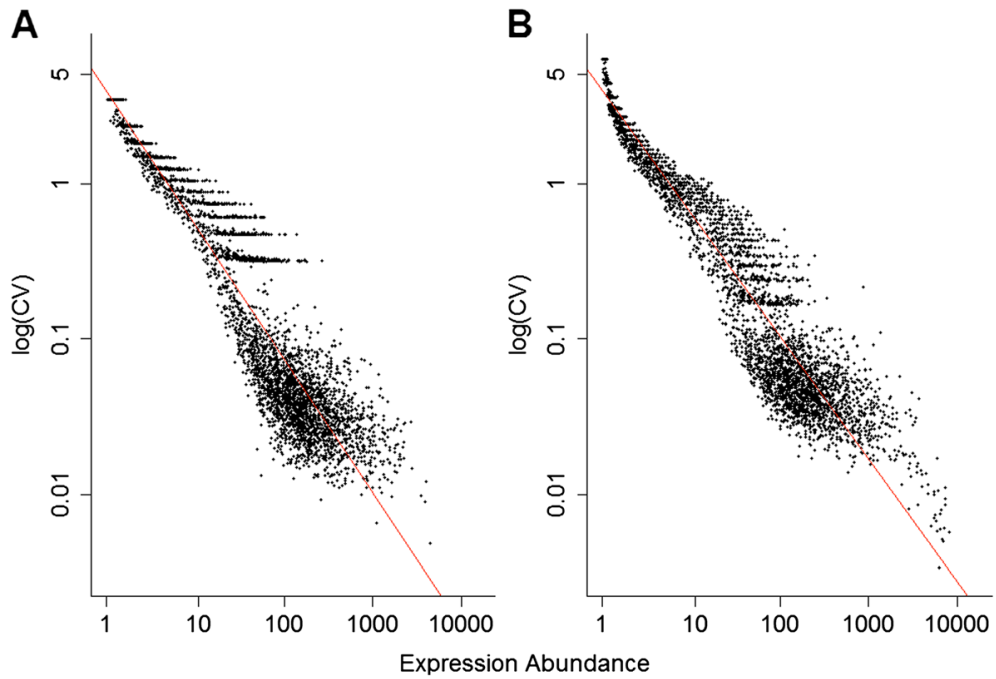

Figure 2.2. Transcriptional noise and expression abundance are significantly and negatively correlated in (A) brain, and (B) blood. Transcriptional noise is measured as the coefficient of variation of transcriptional abundance (see Methods section). The regression coefficients between these variables are −0.60 ($P < 0.001$) and −0.55 ($P < 0.001$) for brain and blood, respectively.

As gene length increases, there may be more opportunities for spurious transcription. In other words, gene length may be positively correlated with transcriptional noise. According to our multiple linear regression analysis, however, the effect of gene length on transcriptional noise, while controlling for other factors, was negligible in the brain data, but significantly negative in the blood data (Table 2.1). Analyzing more tissue samples would clarify the effect of gene length on transcriptional noise. Interestingly, promoter methylation again exhibited strong positive relations with the transcriptional noise in a multiple linear regression setting (Table 2.1).

Especially, variation between microarray expression data can be divided by technical and biological variation. Biological variation comes from differences between individuals when technical variation originates in the noise of technical experiments. Therefore we separate the transcriptional noise into biological variation and technical variation and include them into the regression model. Analysis result from the new regression model is below (Table 2.2).

We confirmed that the new analysis produced similar result compared with that without separation of variations. We explored the relationship between transcriptional noise and DNA methylation, using gene expression variability among different populations of cells as a proxy for transcriptional noise. Our analysis confirms the inverse relationship between gene expression abundance and transcriptional noise, while revealing novel relationships between DNA methylation and transcriptional noise. In particular gene body DNA methylation exhibits a negative correlation with transcriptional noise. This observation supports

Table 2.1 Multiple linear regression results explaining variation of transcriptional noise in different tissues

Tissue	Variables	Beta	S.E	P	VIF
Brain	Intercept	1.47	19.51	$<10^{-4}$	
	Expression abundance	−0.59	−180.50	$<10^{-4}$	1.21
	Gene body methylation	−0.28	−4.74	$<10^{-4}$	1.96
	Promoter methylation	0.2	4.94	$<10^{-4}$	1.27
	Log (gene length)	0.00092	0.099	0.921	2.19
	Adjusted R2		0.87		
Blood	Intercept	1.89	28.92		
	Expression abundance	−0.55	−237.24	$<10^{-4}$	
	Gene body methylation	−0.37	−6.68	$<10^{-4}$	1.11
	Promoter methylation	0.29	7.36	$<10^{-4}$	1.27
	Log (gene length)1	−0.038	−5.09	$<10^{-4}$	1.65
	Adjusted R2	0.92		$<10^{-4}$	1.79

Table 2.2 Multiple linear regression results in which technical versus biological components of transcriptional noise are separately analyzed

Tissue	Variables	Beta	t	P	VIF
Model1	Intercept	1.201	14.12	$<10^{-4}$	
	Expression abundance	−0.442	−78.19	$<10^{-4}$	1.06
	Gene body methylation	−0.797	−7.33	$<10^{-4}$	1.07
	Promoter methylation	0.613	6.17	$<10^{-4}$	1.06
	Adjusted R^2				0.53
Model2	Intercept	0.769	11.157	$<10^{-4}$	
	Expression abundance	−0.337	−61.354	$<10^{-4}$	3.39
	Gene body methylation	−0.566	−9.463	$<10^{-4}$	1.10
	Promoter methylation	0.431	7.969	$<10^{-4}$	1.07
	Technical noise	0.608	32.467	$<10^{-4}$	3.30
	Adjusted R^2				0.82

a longstanding hypothesis that gene body DNA methylation may reduce transcriptional noise. In light of evolutionary findings that gene body methylation is a widespread, conserved form of DNA methylation, the ancestral role of DNA methylation may have been related to the reduction of transcriptional noise. On the other hand, promoter DNA methylation is positively related to transcriptional noise, raising the possibility that epigenetic status of promoters may affect transcriptional bursts.

2.2 Base-pair resolution measuring method

2.2.1 Experimental errors considered in Statistical test and Binomial test using False discovery rate (FDR)

In the previous chapter, we reviewed regional methylation measuring method from the various measuring techniques. However, only BS-seq technique can measure base-pair resolution methylation level directly and more accurately than other techniques. BS-seq datasets constitutes sequencing read attached to each CpG cytosine site and the sequenced base pair are mainly cytosine (C) or thymine (T). When we use NGS sequencing technique for measuring gene expression, we can observe C read in most CpG sites. However, we observed C and T read owing to bisulfite treatment in the BS-seq and we should further consider errors from the technique.

The first error in the technique is non-conversion error which is the probability of un-converging from C read to T read in a non-methylated site. Vice versa, the second error is over-conversion error which is the probability of over-converging from T read to C read in a methylated site. These two errors are firstly discussed in Grunau's article [32]. Among two errors, non-conversion error rate is

widely used to test existence of methylation in CpG sites. Based on the assumption of consistent error rate across whole genome, the error rate have been estimate from the regions that are biologically well known as absolutely non-methylated region such as non-methylated spiked in DNA [33] or the mammalian mitochondrial sequences [34]. The best known test using the non-conversion error is Binomial test and false discovery rate (FDR) adjustment.

To explain how the error rate is related to the Binomial test, we first describe the method in some details. In this method, the probability that a non-methylated C remains as C, or the 'non-conversion' error (which we will refer to as p_0), is used to infer whether the observed methylation signal is more likely to have arisen by chance. Specifically, the methylation status of a site is determined under a binomial distribution where p_0 is used as the success rate. The null hypothesis is that the site is not methylated, and the P-value for this null hypothesis is:

$$\mathrm{P}(\mathrm{X} \geq \mathrm{k}) = 1 - \mathrm{P}(\mathrm{X} < k) = 1 - \sum_{0 \leq i \leq k-1} \binom{N}{i} {p_0}^i (1 - p_0)^{N-i} \quad (1)$$

where k is the number of methylated reads at the site of interest, and N is the total number of reads at this site. The resulting P-values are further corrected for multiple testing, typically by the false discovery rate (FDR) [35]. The main parameter p_0 is determined either by examining non-methylated portions of the genome (such as repetitive regions in insect genomes or chloroplast genomes in plants, e.g., [36,37]) or from 'spiked-in' lambda genomic DNA (e.g., [28]). This approach is intuitive and straightforward. Therefore it has been used in many methylome researches. This error depends on the condition of experiments, therefore its reported values are from 0.001 to 0.04.

The other error in the BS treatment, over-conversion rate is confirmed in the

same research. However it has not been relatively highlighted compared with non-conversion rate. It is because researchers are more interested in detecting methylated site based on the null hypothesis that is the locus is not methylated than detecting non-methylated site. It is thought that methylation process is generally changing process from non-methylated sites to methylated sites. However, we used it for development of new method and improving classification accuracies in special cases (see chapter 3). We will explain how we include the error in our new model later.

Finally, there are sequencing errors that happen in sequencing process. The error is the probability of mis-reading a site as other nucleotide. Although the error exists apparently, its value is relatively small compared with errors from BS-treatment (generally less than 0.001). Therefore, that can be ignored or included in BS treatment related errors. We also include those errors into our new model and estimate it with BS-seq errors as the confounded form.

Chapter 3

A new classification tool of methylation status using bayes classifier and local methylation information

3.1 Introduction

DNA methylation is a prevalent epigenetic modification of genomic DNA with fundamental functional consequences on developmental processes, regulation of gene expression and diseases [33,38]. Accurately inferring the level of DNA methylation at a specific nucleotide in the genome is a critical step toward elucidating the molecular mechanisms of regulation via DNA methylation. A method widely gaining popularity is the whole genome sequencing of bisulfite converted genomic DNA, often referred to as ′methylC-seq′ (also referred to as BS-seq′ elsewhere). This method is based upon the particular chemical properties of DNA methylation to ′protect′ cytosines from converting to uracils by sodium

bisulfite [32]. Specifically, during the sodium bisulfite conversion process, non-methylated cytosines are changed to uracils, which then change to thymine after PCR amplification. Consequently, following a sodium bisulfite treatment, non-methylated cytosines should be read as thymines while methylated cytosines should remain as cytosines.

Compared to microarray-based methods, the methylCseq method is powerful in a multitude of ways. In addition to the fact that it can provide information on every nucleotide in the genome, a notable strength of this method is that it can be applied to analyses of non-model species where pre-defined microarrays (such as beadchip) are not readily available. For this reason, methylC-seq has been instrumental in the recent surge of genomic DNA methylation analyses from diverse taxa, in particular from many invertebrates (e.g., [39,40,41]). These studies show that invertebrate genomes generally exhibit very different patterns of DNA methylation compared to those of mammalian genomes. The most significant difference is that invertebrate genomes tend to be much more sparsely methylated than mammalian genomes. For example, the mean level of DNA methylation for individual CpGs in the honey bee genome is <1% [36,42], which is far lower than that in the human genome (60 ~ 90% [43,44]). Even relatively heavily methylated genomes of some aquatic species such as the freshwater snail Biomphalaria glabrata or the pacific oyster Crassostrea gigas harbor only a few percent of methylated cytosines [45]. Similarly, plant genomes appear to be generally much more sparsely methylated than mammalian genomes. For example, only a few percent of cytosines are methylated during the early stages of Populous floral development [46]. Analyzing such sparsely methylated genomes presents unique technical challenges. In heavily methylated species such as mammals, the measure

of interest is usually the fraction of methylated reads (′C′ reads) in the total number of reads per site, the so-called ′fractional DNA methylation′ [23,24,28]. In sparsely methylated genomes, these values are typically very small. Moreover, these values are heavily influenced by errors associated with the conversion and sequencing processes (see below). For these reasons, it is often important to determine whether a specific position has any methylation or not. In other words, a binary classification of methylated versus non-methylated cytosines is critical to evaluate the distribution of DNA methylation and different levels of DNA methylation [36,37,39,47]. In principle, this should be simple: cytosines covered by any number of ′C′ reads should be considered methylated.

However, in reality, this step is not straightforward due to the nature of chemistry underlying the MethylC-seq method. Specifically, the sodium bisulfite conversion step is not perfect, and includes both i) the possibility of non-conversion (non-methylated C is not properly converted to U/T), leading to an overestimation of actual DNA methylation (Figure 3.1A), as well as ii) overconversion (methylated C is also converted), leading to an underestimation of actual DNA methylation (Figure 3.1A) [32]. Consequently it is necessary to take into account these technical errors for a binary classification of a specific nucleotide. In particular, these errors can occur at rates comparable to the actual methylation levels in some genomes. Despite these well-known and substantial technical issues, methods to efficiently account for these imperfections are surprisingly rare. The most widely used method is the so-called binomial method [36,37]. However, this method has some shortcomings when the genomic

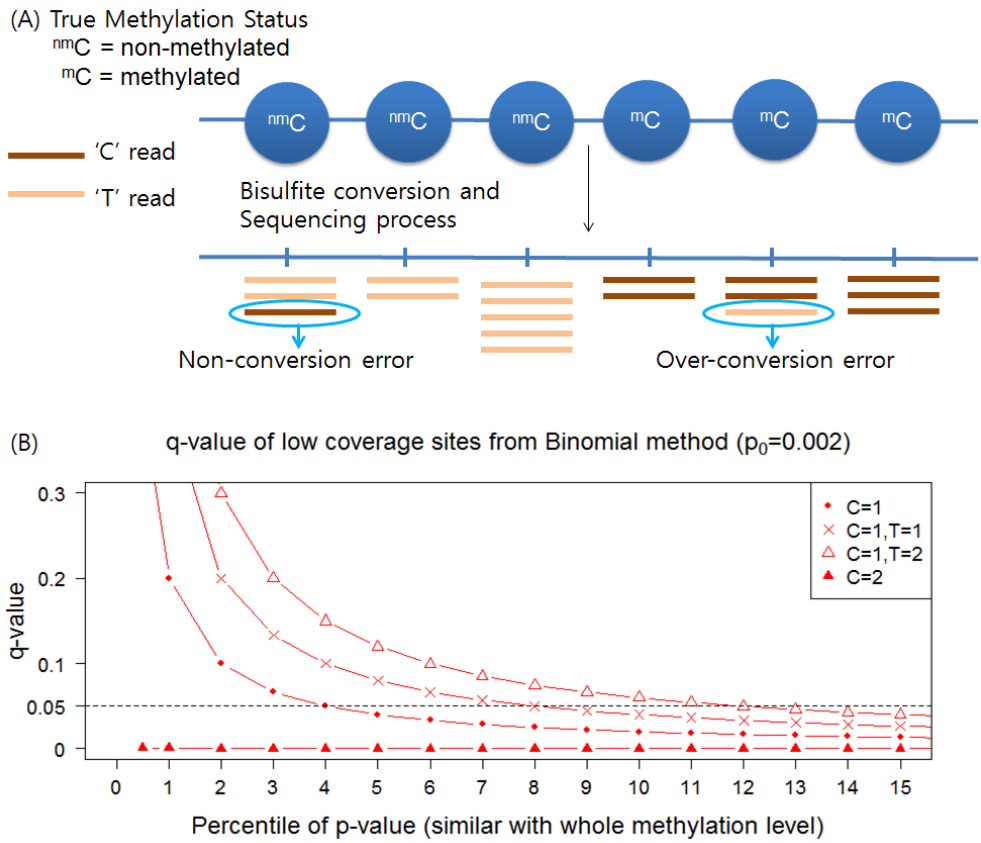

Figure 3.1 Potential errors and biases of methylC-seq and binomial method. **(A)** Errors associated with the methylC-seq method. Non-methylated Cs may not be completely converted (non-conversion error, non-methylated C remains as C). In addition, methylated Cs may undergo conversion (over-conversion error, methylated C converts to T). **(B)** Reduced power of the binomial test in sparsely methylated genomes and low coverage. The Y-axis indicates FDR-corrected q-values from the binomial test, calculated following the equation (2) in the main text. Four cases are shown, including when a specific cytosine is covered by a single 'C' read (filled circles), one 'C' and one 'T' reads (crosses), one 'C' and two 'T' reads (open triangles) and two 'C' reads (filled triangles)

methylation levels and the coverage of specific site are low (see below). Here, we propose a new method, the Bisulfite-sequencing data classification method (Bis-Class). This method takes the prior methylation distribution into account to infer methylation status in the framework of Bayesian probabilistic models, which is known to minimize classification errors in the presence of a known alternative hypothesis [48]. In addition to utilizing a Bayesian classification scheme, we take into account the fact that DNA methylation levels of adjacent sites are correlated (Figure 3.2). Consequently, including information on DNA methylation levels of the genomic neighborhood improves our ability to correctly infer the DNA methylation status. We demonstrate that Bis-Class alleviates the problems of the binomial method and improve sensitivity and accuracy using extensive simulations as well as analyses of actual methylC-seq data.

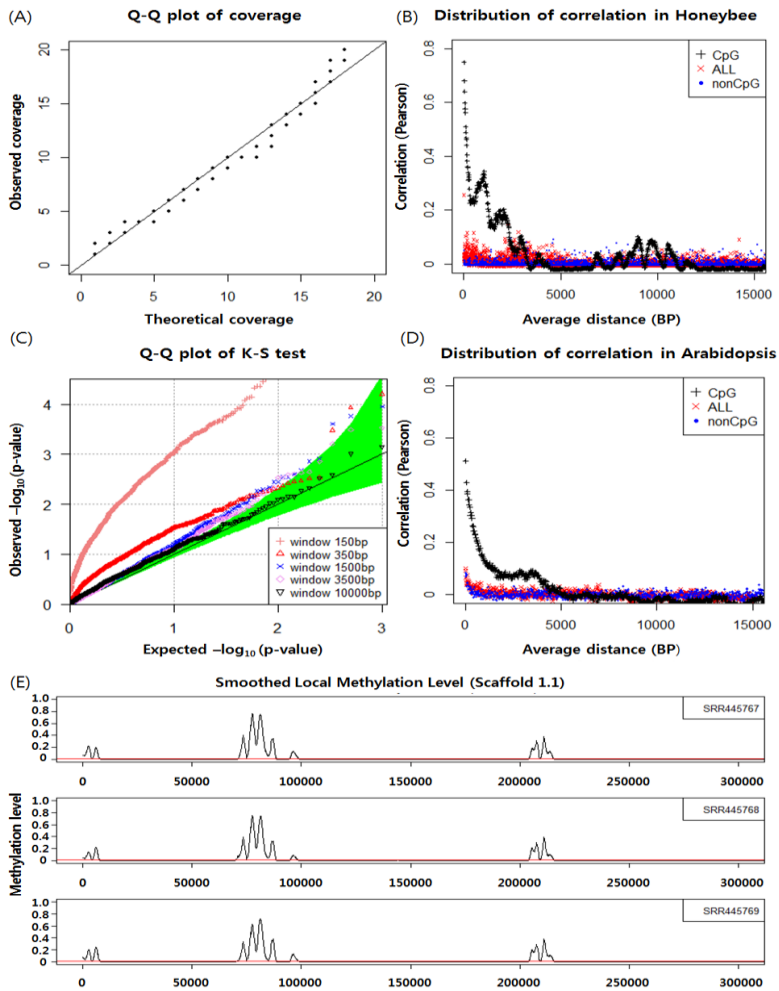

Figure 3.2 Properties of methylC-seq coverage and spatial correlation of CpG methylation level. (A) Quantile-quantile (Q-Q) plot between observed coverage and theoretical coverage which is from a shifted negative binomial distribution. (B) Spatial correlation plot of a honeybee methylome from Herb et al. [42]. (C) Q-Q plot between observed p-values from Kolmogorov-Smirnov (K-S) test and theoretical p-values from a null distribution. (D) Spatial correlation plot of an Arabidopsis methylome (methylC-seq data from GSM276809, [49]). (E) Smoothed methylation level using triangle weight in scaffold 1.1, for three samples. X-axis is physical location and Y-axis is methylation level. Red lines represent average methylation fractions calculated from whole CpG methylomes.

3.2 Methods

3.2.1 Binomial test using FDR and its limit

We overviewed Binomial test using FDR in the previous chapter. Although it has been widely used, the power of the binomial method is weak when the number of reads (N) is small (i.e., equation (1) above). Moreover, when combined with the correction for multiple testing with FDR, the power of the binomial method is particularly reduced at low coverage sites in sparsely methylated genomes. This reduction is because the FDR-corrected q-value of the ith site is calculated as

$$p_i \times \frac{\#\ of\ total\ sites}{rank\ of\ p_i} \qquad (2)$$

where pi is the binomial P-value for a specific site i. Since the binomial P-values are limited by the number of reads (N) for that site, P-values from low-coverage sites (low N) will have moderate ranks at most.

Consequently, in sparsely methylated genomes, even if a site is truly methylated, if the number of reads is small, the P-value of such site will not be ranked sufficiently low to be classified as methylated after FDR correction. Figure 3.1B demonstrates this phenomenon using specific examples. For example, a site covered by a single ‘C’ read (the line with filled circles, Figure 3.1B) will not be classified. as ‘methylated’ with the binomial method in genomes with the overall methylation levels typical of hymenopteran insects (i.e., < 4%). Likewise, a site with 50% methylation with coverage of two (the line with crosses, Figure 3.1B) will also be classified as non-methylated in sparsely methylated genomes. Consequently, the binomial method may produce a high number of false negatives

in lowly methylated genomes. To avoid this problem, some studies suggest using sites that are covered by at least two [50], or four [37] reads. However, since the number of reads typically has a large variance, even with a moderate coverage sequence data, substantial numbers of cytosines are covered by single reads (Table 3.1), making it impractical to remove positions with a small number of reads. For example, if we remove sites with fewer than four reads, almost 50%. of data are discarded in representative methylC-seq datasets in honey bee (Table 3.1).

Table 3.1. Properties of the methylC-seq data sets used in this study.

Species	Subtype	Sample ID	Coverage	Variance	Variance/Coverage	% of under 3 reads (% of 1 read)
Honey Bee	**Worker**	SRR039814	5.86	22.96	3.918	0.4255 (0.1547)
	Queen	SRR039815	7.24	27.521	3.801	0.3728 (0.1118)
	Forager	SRR445767	3.86	10.28	2.663	0.5608 (0.1892)
		SRR445768	4.17	13.27	3.182	0.5345 (0.1825)
		SRR445769	3.86	9.951	2.578	0.5620 (0.1944)
		SRR445770	4.04	12.838	3.177	0.5521 (0.1907)
		SRR445771	4.51	14.510	3.217	0.4823 (0.1522)
		SRR445773	5.86	18.275	3.119	0.3111 (0.0798)
	Nurse	SRR445774	3.13	6.081	1.943	0.6709 (0.2512)
		SRR445775	4.49	14.812	3.299	0.4868 (0.1552)
		SRR445776	3.84	12.203	3.178	0.5802 (0.2032)
		SRR445777	4.51	14.978	3.321	0.4856 (0.1553)
		SRR445778	2.65	6.846	2.583	0.7801 (0.359)
		SRR445799	4.05	12.014	2.966	0.5434 (0.1875)
Human (Brain)		HS1570_0731	8.59	37.157	4.32	0.2874 (0.0945)

3.2.2 Bis-Class

To overcome the aforementioned problems in the binomial method, here we propose to use a Bayesian probabilistic model to infer methylation status. The posterior probability of methylation status is determined based upon the product of prior distribution of methylation and the likelihood of specific reads aligned to a site. Specifically, the posterior distribution of methylation is given as.

$$\mathrm{P(M|R)} = \frac{\pi(M)\times P(R|M)}{P(R)}$$

where M is a random variable representing methylation status (m for methylated, nm for non-methylated). R = {R_1, R_2,.., R_N} is the set of sequence reads mapped to a site. If a sample consists of N number of CpGs and ith CpG has ni reads, R_i denotes a set of reads assigned in ith CpG and R_{ij} denotes jth read of ith CpG (i= 1, ..., N and j =1, ..., n_i). In addition, likelihood P (R_i|M) is given as the product of P(R_{ij}|M)s. π(M) is the prior information on DNA methylation. The likelihood P (R_i|M) can be calculated by explicitly incorporating the errors associated with the inference of methylation.

The main source of errors for non-methylated sites is the non-conversion rate (denoted as p_0, Figure 3.1A). If there is no additional error, the probability of obtaining a C read in non-methylated site is equivalent to the non-conversion rate p_0. Likewise, the probability of obtaining a C read in methylated site is 1- (over-conversion rate), which we denote as p_1 (Figure 3.1A). There may be additional errors occurring during sequencing process. We define the sequencing error (ε) as the probability of being misread from other nucleotide (For example, reading C read as T read or vice versa). Consequently, our observation likelihood P(R_i|M)

consists of the following distributions according to the methylation status.

$$\mathrm{P}\left(\mathrm{R}_{ij}\middle|\mathrm{M}=\mathrm{nm}\right)=\begin{cases} p_0{}' = p_0\times(1-\varepsilon)+(1-p_0)\times\varepsilon & if\ \mathrm{R}_{ij}=C \\ 1-p_0{}' = (1-p_0)\times(1-\varepsilon)+p_0\times\varepsilon & if\ \mathrm{R}_{ij}=T \end{cases} \quad (3)$$

$$\mathrm{P}\left(\mathrm{R}_{ij}\middle|\mathrm{M}=\mathrm{m}\right)=\begin{cases} p_1{}' = p_1\times(1-\varepsilon)+(1-p_1)\times\varepsilon & if\ \mathrm{R}_{ij}=C \\ 1-p_1{}' = (1-p_1)\times(1-\varepsilon)+p_1\times\varepsilon & if\ \mathrm{R}_{ij}=T \end{cases} \quad (4)$$

Since sequencing errors are confounded with p_0 or p_1 in reality, we will regard $p_0{}'$ and $p_1{}'$ as p_0 and p_1 in this article, respectively. The parameters p_0 or p_1 are inferred from data using the Expectation-Maximization (EM) algorithm [51]. Because p_0 and p_1 are probabilities of emerging 'C' read in a non-methylated site and a methylated site, respectively; latent status for them are given as non-methylated (M=*nm*) or methylated (M=*m*). For convenience, 0 and 1 denote *nm* and *m* for methylation status, respectively. Expected complete log-likelihood for N' selected samples is then given as

$$\mathrm{Q}(\theta|\hat{\theta}_t)=E\left[log\prod_{i=1}^{N'}\left[\hat{\pi}_{1,t}\times\hat{p}_{1,t}{}^{C_i}\times\left(1-\hat{p}_{1,t}\right)^{T_i}\right]^{M_i}\left[\hat{\pi}_{0,t}\times\hat{p}_{0,t}{}^{C_i}\times\left(1-\hat{p}_{0,t}\right)^{T_i}\right]^{1-M_i}\right]. \quad (5)$$

$\hat{\theta}_t$ is a set of parameter obtatined t^{th} iteration and $\hat{\pi}_{0t}$, $\hat{p}_{0t}$ and $\hat{p}_{1t}$ are its components. From this equation, we can calculate probability of being methylated as

$$\hat{\mu}_{i,t}=E\left[M_i\middle|C_i,T_i,\hat{\theta}_t\right]=\mathrm{p}\left(M_i=1\middle|C_i,T_i,\hat{\theta}_t\right)=\frac{\hat{\pi}_{1,t}\times\hat{p}_{1,t}{}^{C_i}\times\left(1-\hat{p}_{1,t}\right)^{T_i}}{\hat{\pi}_{1,t}\times\hat{p}_{1,t}{}^{C_i}\times\left(1-\hat{p}_{1,t}\right)^{T_i}+\hat{\pi}_{0,t}\times\hat{p}_{0,t}{}^{C_i}\times\left(1-\hat{p}_{0,t}\right)^{T_i}}. \quad (6)$$

In the maximization step, $\hat{\theta}_{t+1}$ can be obtained by maximizing $\mathrm{Q}\left(\theta\middle|\hat{\theta}_t\right)$ and the solutions are given as

$$\hat{\pi}_{1,t+1}=\frac{1}{N'}\sum_i\hat{\mu}_{i,t}\ ,\ \hat{p}_{0,t+1}=\frac{\sum_i\hat{\mu}_{i,t}\times C_i}{\sum_i\hat{\mu}_{i,t}},\ \hat{p}_{0,t+1}=\frac{\sum_i(1-\hat{\mu}_{i,t})\times C_i}{\sum_i(1-\hat{\mu}_{i,t})}. \quad (7)$$

We iterated this process until change of $\hat{\theta}_t$ is small enough to be regarded as conversion. Then the converged estimates are maximum likelihood estimate of p_0 and p_1. In order to obtain independent samples for constructing likelihood function

used in EM, we divided the whole genome by 10kb windows and select a site for each window by the result from Figure 2C. We repeated this process 100 times and calculated median of the results to obtain final estimates of p_0 and p_1. If we already know $\hat{p}_0$ from the experiment, we may fix p_0 in the EM algorithm and maximize likelihood only to estimate of p_1.

We determined window size which assures independence between selected samples. Figure 3.2C shows a quantile-quantile (Q-Q) plot to determine the physical distance required for all CpGs to be mutually uncorrelated. We selected a CpG for each window and calculated C/(C+T) for all selected CpGs. The methylation fraction of i^{th} selected CpG is denoted as F_i and i= 1, 2, …, N. Then we calculated $\tilde{\mathrm{F}}_i = \mathrm{F}_i - \frac{1}{N}\sum \mathrm{F}_i$, which denotes residual of F_i. Since methylation level does not follow normal distribution, the Durbin-Watson test, which is popularly used in regression analysis for test of independence of residuals, is impractical. Instead, we used nonparametric procedures. First, we extracted the sign of $\tilde{\mathrm{F}}_i$ and then tested whether positive loci are uniformly distributed via the Kolmogorov-Smirnov test. For a fixed window size, we repeated the procedure 1,000 times and compared the obtained p-values with the theoretical p-values, based on the null distribution. The distribution of observed p-values becomes concordant with the distribution of theoretical p-values as window size increases, and we can determine the window size which imposes mutual independence on all selected CpGs from these results.

We previously demonstrate that methylated cytosines are heterogeneously distributed and locally clustered in different species (Figure 3.2). For example, in the honeybee genome, some regions exhibit >70-fold higher methylation levels

compared to other regions (Figure 3.2E). We take this observation into account and incorporate local methylation levels into the methylation prior to improve classification accuracy. Since methylated cytosines are heterogeneously distributed and locally clustered, the use of local methylation information would be useful. Since some regions may have extreme methylation values, it might be also useful to include information on the global methylation level. Here, we propose using a weighted average of local and global methylation levels to produce a more robust estimation of posterior odds. Specifically, if we denote the global methylation level as $\hat{\pi}_1{}^G$ and the local methylation level as $\hat{\pi}_1{}^L$, the combined methylated prior, $\hat{\pi}_1{}^C$, can be represented as $\hat{\pi}_1{}^G \times (1 - \mathrm{w}) + \hat{\pi}_1{}^L \times \mathrm{w}$. The weight parameter, w, can decide how much local versus global methylation levels can be included in the prior. This factor can have any value between 0 and 1. In our analyses we used 0.5, to treat local and global information equally. In our experience, using the weight factor of 0.5 produced good AUC (Area Under Curve), sensitivity and low error rate compared with other weight factor values for honey bee data (See Chapter 3.5). Nevertheless, in this implementation of Bis-Class the users can choose any arbitrary value of the weight factor.

We describe the estimation process for the global methylation levels, $\hat{\pi}_1{}^G$ and $\hat{\pi}_0{}^G$ which are the estimates of proportion of methylated and non-methylated sites in the whole methylome. We define C_i as number of C reads assigned to an i^{th} site, T_i as the number of T reads assigned to the i_{th} site. The total number of reads in the i^{th}site of a sample L_i is then C_i+T_i. Then the proportion of cytosine read in i^{th} site, $F_i = C_i/L_i$, is equivalent to the widely used 'fractional methylation' measure [23,24,28]. Therefore, for any *i*, expectation of F_i can be

estimated as:

$$\mathrm{E}(F_i) = \pi_1 \times \mathrm{E}(F_i|M = m) + \pi_0 \times \mathrm{E}(F_i|M = nm) = \pi_1 \times p_1 + \pi_0 \times p_0. \quad (8)$$

Using the method of moments and adjusting for the total read count L_i to impose more confidence to deep coverage sites, the estimate of $\mathrm{E}(F)$ across whole genome will be $(1/\sum_{\mathrm{i}=1}^{\mathrm{N}}\sqrt{L_i}) \times (\sum_{\mathrm{i}=1}^{\mathrm{N}}\sqrt{L_i} \times F_i)$. From equation (8) and $\pi_1 = 1 - \pi_0$, $\hat{\pi}_1$ for global methylome, denoted by ${\hat{\pi}_1}^G$, is as follows:

$$(\widehat{\mathrm{E}(\mathrm{F})} - \hat{p}_0)/(\hat{p}_1 - \hat{p}_0). \quad (9)$$

In case of local methylation level, $\mathrm{E}(F)$ is estimated as $(1/\sum_{\mathrm{i}=1}^{\mathrm{N}}\sqrt{L_\mathrm{i}} \times K(d_k)) \times (\sum_{\mathrm{i}=1}^{\mathrm{N}}\sqrt{L_i} \times F_i \times K(d_k))$ which additionally adopts weight function $\mathrm{K}(d_k)$ which adjusts the weight of a specific function to consider distance from the site which is to be determined. k = 1, 2, ..., K denotes the index of CpGs in a window. For a weight function K(d), d is the physical distance from a site which is of interest. Then ${\hat{\pi}_1}^L$ can be estimated as weighted average through weight function, as follows:

$${\hat{\pi}_1}^L = \frac{1}{\sum_k \sqrt{L_k} \times K(d_k)} \left(\sum_{\mathrm{k}=1}^{\mathrm{K}} \sqrt{L_k} \times K(d_k) \times (F_k - \hat{p}_0)\right)/(\hat{p}_1 - \hat{p}_0) \quad (10)$$

$\hat{p}_0$and $\hat{p}_1$ are estimates of p_0 and p_1, respectively. The weight function $K(d)$ can be many types of functions which decrease as d increases. In our analyses, we chose to use the triangle weight which decreases linearly for $\mathrm{d} \leq \mathrm{d}_0$ and zero for $\mathrm{d} > \mathrm{d}_0$. In addition, we define a window around the considered site as the region whose weights are not zero. The window size, d_0, can be arbitrarily selected. We also define $K(0) = 0$ to exclude the focal cytosine. Our approach is very flexible, as the width and the weight function type can be easily changed according to the properties of each dataset. We selected the triangle weight function because it is

similar to be observed patterns of spatial correlation between methylation levels of adjacent CpG sites (Figure 3.2B). Applying alternative functions such as Gaussian or Laplace provided similar results (Figure 3.3). The width of region to which the weight is applied in our analyses was determined as the point where the spatial correlation decreases to below 0.2, which is approximately 1.5 kb in the honey bee data (Figure 3.2B).

After following the above steps, the posterior odds for ith CpG can be constructed as:

$$\text{Posterior odds} = \frac{\mathrm{P}(M=m|R_i)}{\mathrm{P}(M=nm|R_i)} = \frac{P(M=m)\times P(R_i|M=m)}{P(M=nm)\times P(R_i|M=nm)} = \frac{\pi(M=m)\times\prod_j P(R_{ij}|M=m)}{\pi(M=nm)\times\prod_j P(R_{ij}|M=nm)} \qquad (11)$$

If the value of a specific site is larger than some criteria, it will be classified as methylated. We propose using 19 as the criterion (meaning that the probability of being classified as methylated is 19 times bigger than that of being classified as non-methylated). This criterion also means that the probability of being falsely classified as methylated is smaller than 0.05 at the site [52]. Consequently this is equivalent to the FDR-corrected q-value < 0.05, as typically used in the binomial test.

3.3 Material and its description: Honeybee dataset

In this section we present analyses of actual bisulfite sequencing data that are pertinent to our proposed method. Honey bee is one of the first invertebrates for which the methylC-seq method has been applied. The usage of the methylC-seq method has been crucial to elucidating the importance of DNA methylation on gene regulation in honey bee, including its role in the differentiation of castes [36], behavioral differentiation of worker bees [42], and alternative splicing [53,54]. We examined two recent methylC-seq datasets of honey bee, one from the brains of

worker and queen bees [36], the other from brains of six forager and six nurse bees [42]. All data have been mapped to the assembly 2.0 using BSmap [55]. As reported previously in the original studies, mean fractional methylation levels are extremely low, between 0.3 ~ 0.5% for all cytosines, and 0.5% ~ 0.9% for CpGs ($E(\hat{F})$ in Table 2). The mean coverage in these data sets ranges between 2.65 and 7.24 (Table 1) and the variance of read depths is quite high (Table 1). The distribution of coverage follows a shifted negative binomial distribution with similar mean and variance as observed (Figure 2A). An important consequence of this is that most of the data (~50%) are covered by fewer than four reads, and substantial portion of the data are covered by only a single read (Table 3.1).

Methylated cytosines are not randomly distributed along the genome. DNA methylation levels of nearby cytosines are correlated; for example, in a forager sample from Herb et al. [42], the correlation coefficient between two CpGs 100 bps apart is 0.5 (Figure 3.2B). The correlation decreases as the distance between two cytosines increases, and this pattern is more pronounced for CpGs than non- CpGs (Figure 3.2B). Co-variation of DNA methylation of adjacent cytosines extends to several kilobases (Figure 3.2B and 3.2C). We observed similar trends in multiple species analyzed. For example, in Arabidopsis, a similar pattern is observed (Figure 3.2D, also see [56]). A similar level of spatial correlation has been also observed in the human genome [57]. When examined in detail, methylated cytosines in the honey bee are locally clustered in the genome (Figure 3.2E), with several regions in the chromosome exhibiting elevated levels of DNA methylation (Figure 3.2E).

Importantly, this pattern and the locations of methylated clusters are consistent across different biological replicates (Figure 3.2E), indicating that the spatial correlation is not caused by technical noises, but reflects the inherent pattern in the

genomic distributions of DNA methylation in these species. Together with the results in the above section, we show that a substantial portion of the genome is covered by very few reads, the overall level of methylation is low, and that local methylation levels are correlated. As discussed above and seen in the Figure 3.1, such aspects of data render the binomial method prone to high false negative rates. Consequently, we propose Bis-Class as a practical alternative to the commonly used binomial method of classification. In the next section, we show comprehensive simulation results based upon the observed parameters of the data, indicating that Bis-Class outperforms the binomial method.

3.4 Simulation study

We performed extensive simulation to compare the performance of Bis-Class to the binomial method. We generated methylC-seq data for a genome of 100,000 cytosines, with the mean coverage ranging between 3× to 9×. The numbers of total reads at each site were generated from a shifted negative binomial distribution with the whole genome coverage as the mean and three times the mean as the variance, similar to the typical methylC-seq dataset in honey bee (Table 3.1, Figure 3.2A). The selected parameters p0 and p1, as well as the total methylation levels are also similar to those observed in the empirical data (Table 3.2).

Table 3.2. Methylation Classification results using the Binomial and Bis-Class methods.

Species	Subtype	Sample ID	$\hat{p}_0$ (experiment)	$\hat{p}_0$ (EM)	$\hat{p}_1$ (EM)	$\widehat{E(F)}$ (CpG only)	$\hat{\pi}_1^G$ (CpG only)	$\hat{\pi}_1^G$ ($p_1 = 0.95$) (CpG only)	# of mCpG (binomial)	# of mCpG (Bis-Class)
Hobey Bee	**Worker**	SRR039814	0.0029	0.0021	0.7081	0.004985 (0.005414)	0.004086 (0.004694)	0.002727 (0.003496)	83509	111900
	Queen	SRR039815	0.0024	0.0018	0.7273	0.004009 (0.005750)	0.003044 (0.005472)	0.00232 (0.003872)	98769	117829
	Forager	SRR445767	0.0012	0.0015	0.6669	0.003271 (0.007931)	0.002662 (0.009665)	0.001867 (0.006780)	97220	106204
		SRR445768	0.0013	0.0015	0.6479	0.003309 (0.008171)	0.002799 (0.01032)	0.001907 (0.007033)	101628	108382
		SRR445769	0.0011	0.0014	0.6608	0.0032 (0.007830)	0.002730 (0.009752)	0.001897 (0.006778)	96934	105128
		SRR445770	0.0012	0.0014	0.6461	0.00322 (0.008059)	0.002823 (0.01033)	0.001918 (0.007019)	100173	107755
		SRR445771	0.0012	0.0014	0.6682	0.003269 (0.008434)	0.002803 (0.01055)	0.001970 (0.007415)	107688	111434
		SRR445773	0.0013	0.0014	0.6574	0.003087 (0.007448)	0.002572 (0.009221)	0.001778 (0.006375)	117971	116825
	Nurse	SRR445774	0.0012	0.0015	0.6888	0.003115 (0.007391)	0.002350 (0.008571)	0.001702 (0.006201)	75340	95191
		SRR445775	0.0011	0.0015	0.6539	0.003209 (0.007849)	0.002620 (0.009733)	0.001801 (0.006693)	103257	109181
		SRR445776	0.0012	0.0015	0.6506	0.003113 (0.007717)	0.002485 (0.009579)	0.001700 (0.006554)	92131	103492
		SRR445777	0.0011	0.0015	0.6571	0.003228 (0.007936)	0.002636 (0.009818)	0.001821 (0.006785)	105435	110196
		SRR445778	0.0013	0.0013	0.6563	0.003252 (0.008262)	0.002980 (0.01063)	0.002057 (0.007338)	63375	90489
		SRR445799	0.0012	0.0015	0.6574	0.003314 (0.008137)	0.002766 (0.01012)	0.001912 (0.006997)	99063	106836
Human (Brain)		HS1570_0731	0.0013	0.01	0.94	(0.8064)	(0.8563)	(0.8472)	882351[a]	901228[a]

[a]Calling only on 10^6 CpGs in Chr 10.

We also examined the effects of each parameter when they were slightly greater than the observed values. The weight parameter we used is 0.5, to consider global information and local information equally. To examine the effect of DNA methylation clustering, we generated two types of genomes. The first is a genome where methylated CpGs are uniformly distributed. In the second type, DNA methylation is concentrated in 1/10 of the genome in a 10× intensity compared to whole genome methylation level. We generated 100 replicates for each parameter combination. Local information was obtained from the 200 nearest cytosines (which is equivalent to considering CpGs with 3000 bps of a specific site in the honey bee methylation data).

We then compared classification results with the true status and calculated the sensitivity as the proportion of sites classified as methylated when they are truly methylated (Figure 3.3). The higher the sensitivity, the lower the rate of false negatives. In genomes where DNA methylation occurs uniformly (′homogeneous′), both the binomial method and Bis-Class provide similar results across almost all settings (purple and green bars filled with diagonal lines in Figure 3.3). We note that the binomial methods in clustered genomes and homogenous genomes are statistically equivalent, which is apparent in the simulation results. Sensitivities are low when the sequence coverage is low, and increase with sequence coverage. In the non-homogenous, clustered genomes, Bis-Class (solid green bar) outperforms the binomial method and exhibits much higher sensitivity (therefore lower false negatives) than the binomial method (solid purple bar, Figure 3.3). While Bis-Class displays higher sensitivities compared to the binomial method in a variety of settings, the difference is most pronounced when the coverage is low. In addition,

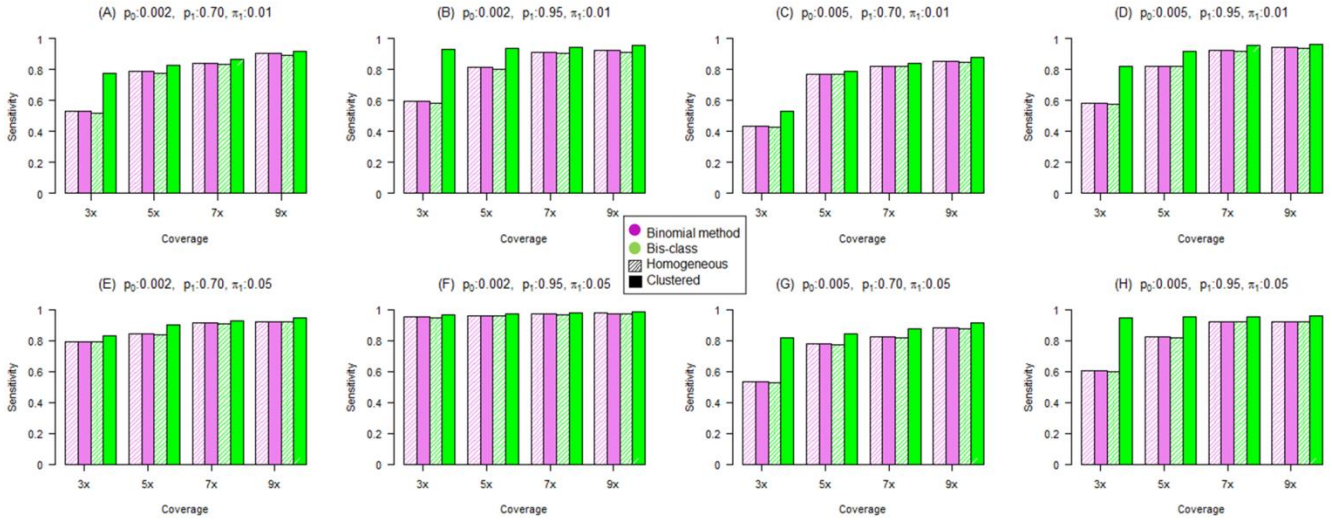

Figure 3.3 Comparison of sensitivities of Bis-Class and the binomial method using simulated data. Sensitivities are evaluated in a variety of parameter settings and plotted in (A)-(H). Purple bars and green bars indicate the results from the binomial method and the Bis-Class, respectively. Bars with diagonal lines indicate the results from homogeneous methylomes and solid bars indicate those from regionally clustered methylomes.

the difference between Bis-Class and the binomial method is large when the ratios between the two error rates (p0 and p1) are high and the whole genome methylation level is low.

We also examined the incidences of mis-classification. Because the proportions of methylation and non-methylated sites are not balanced, a direct comparison between accuracy measures is difficult to perform. For this, we define ′1-specificity′ as the ratio of the number of mis-classified non-methylated sites to the number of true methylated sites. The resulting plots (Figure 3.4) show that all methods have acceptably low error rates (less than five percent of true methylated sites). These simulation results demonstrate that, with the cutoff comparable to FDR-corrected q-value < 0.05, Bis-Class exhibits a greater sensitivity and a comparable specificity compared to the binomial method. Overall, Bis-Class has a greater accuracy (calculated as the sum of (proportion of methylated sites) × sensitivity and (proportion of non-methylated sites) × specificity) than the binomial method. To illustrate this further we evaluated the Area Under Curve (AUC) measure of the ROC (Receiver operating characteristic) under identical simulation settings, which is expected to provide a comprehensive comparison because it summarizes both sensitivity and specificity across all possible cutoff values [58]. This analysis (Figure 3.5) demonstrates that the AUC values of Bis-Class are larger than those of the binomial method, especially in settings where the sequence coverage is low and DNA methylation occurs heterogeneously, i.e., settings closely resembling the observed patterns in the actual methylC-seq data (Tables 3.1 and 3.2, Figure 3.2).

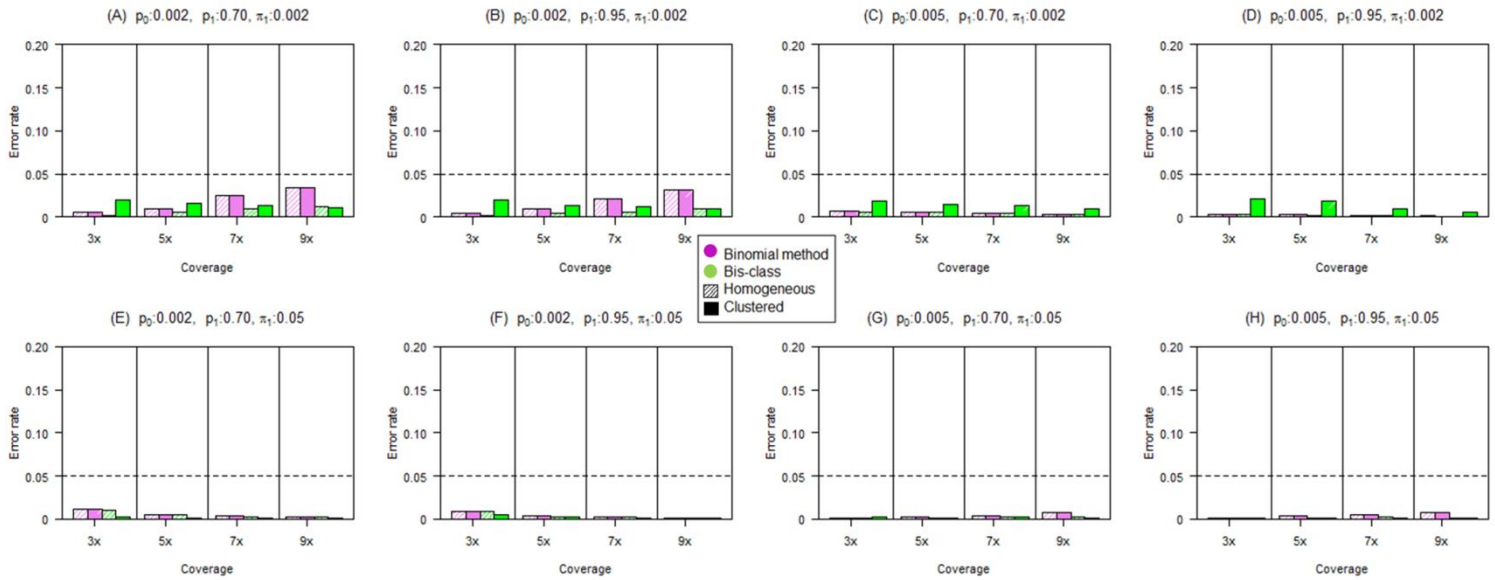

Figure 3.4 Comparison of misclassification rates of non-methylated CpGs via the Bis-Class and the binomial method using simulated data. 1-specificities are evaluated in a variety of parameter settings and plotted in (A)-(H). The Y-axis indicates the ratio of the number of misclassified non-methylated CpGs to the total number of methylated CpGs. Purple bars and green bars are the results from the binomial method and the Bis-Class, respectively. Bars with diagonal lines imply the results from homogeneous methylomes and solid bars imply those from regionally clustered methylomes.

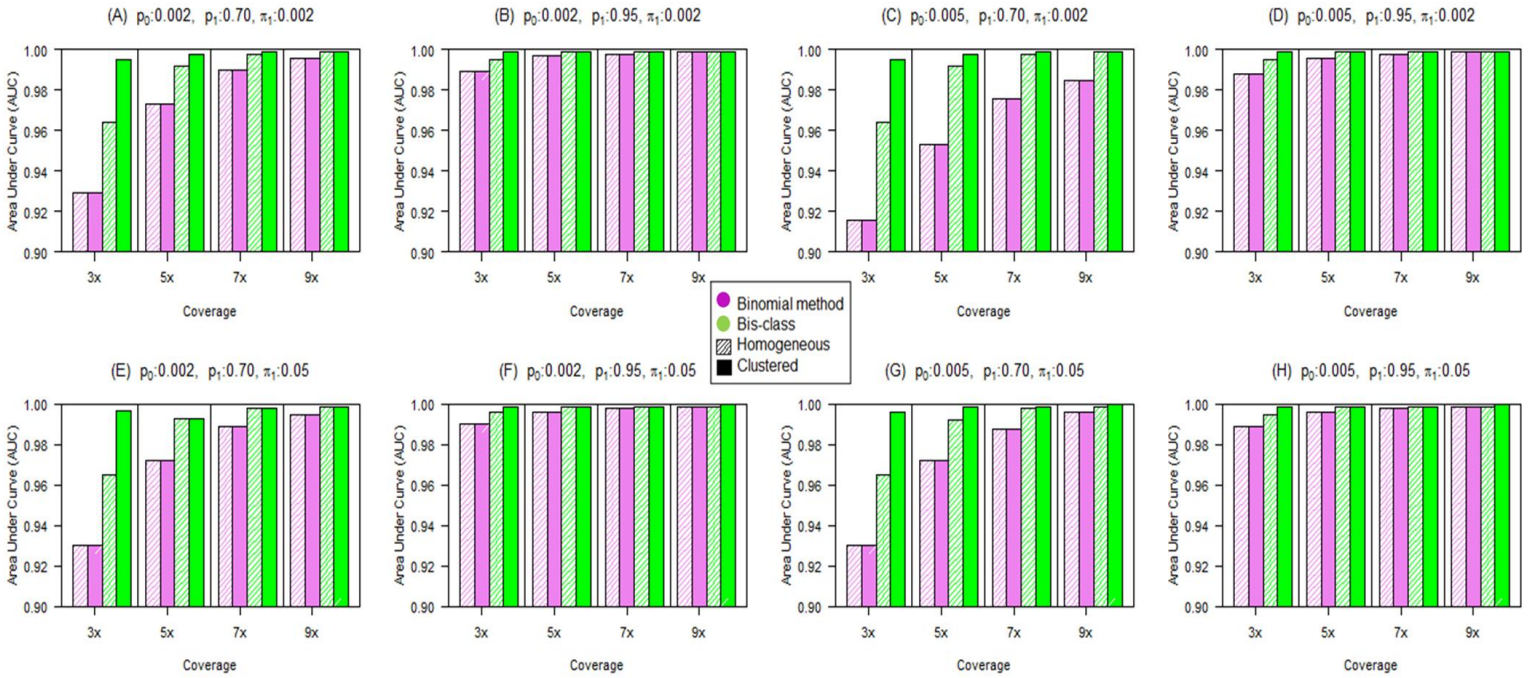

Figure 3.5 Comparison of the AUC measures in simulated data sets. Parameter settings of the simulation are identical with those in the Figures 3 and 4 in the main text. AUC is generally higher for the Bis-Class compared to the Binomial method.

Together these results indicate that Bis-Class provides superior results compared to the binomial method. When we applied our method to intermediately (=30%) or densely methylated (=80%) regions that are artificially generated, it produced similar performances with the Binomial method (Figure 3.6, 3.7). Although there are small losses of sensitivities in Bis-Class of homogenous settings, compared with that of the Binomial method, there are small gains of specificities in Bis-class method. These simulation results correspond with the real analysis result in Table 3.2. In honeybee dataset, almost samples are found to have more methylated CpG cytosines from Bis-Class than the Binomial method, while more CpG cytosines are classified to be methylated from Bis-Class method than the Binomial method.

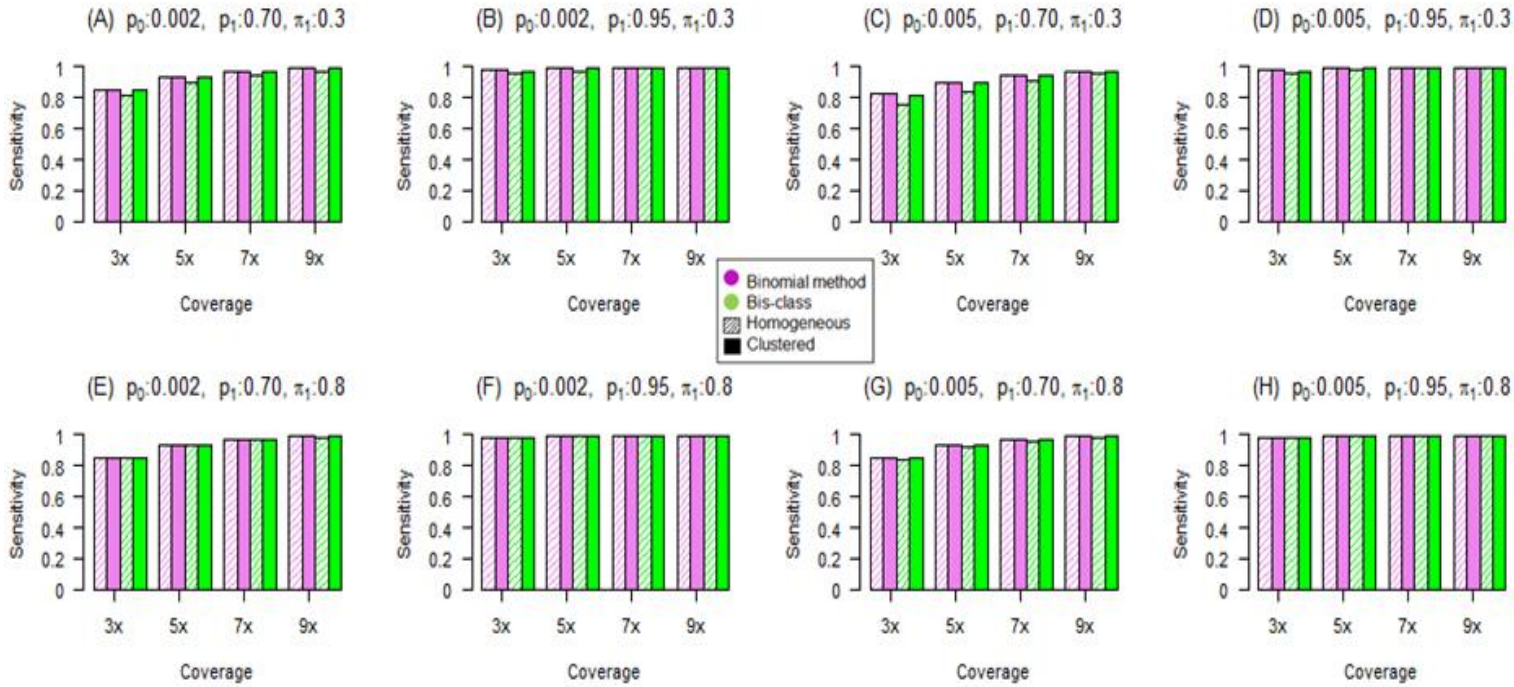

Figure 3.6 Comparison of sensitivities of Bis-Class and the binomial method using simulated data (intermediately or densely methylated region). We plotted sensitivities for intermediately or densely methylated region. Compositions of the eight figures are similar with those of figure 3.3.

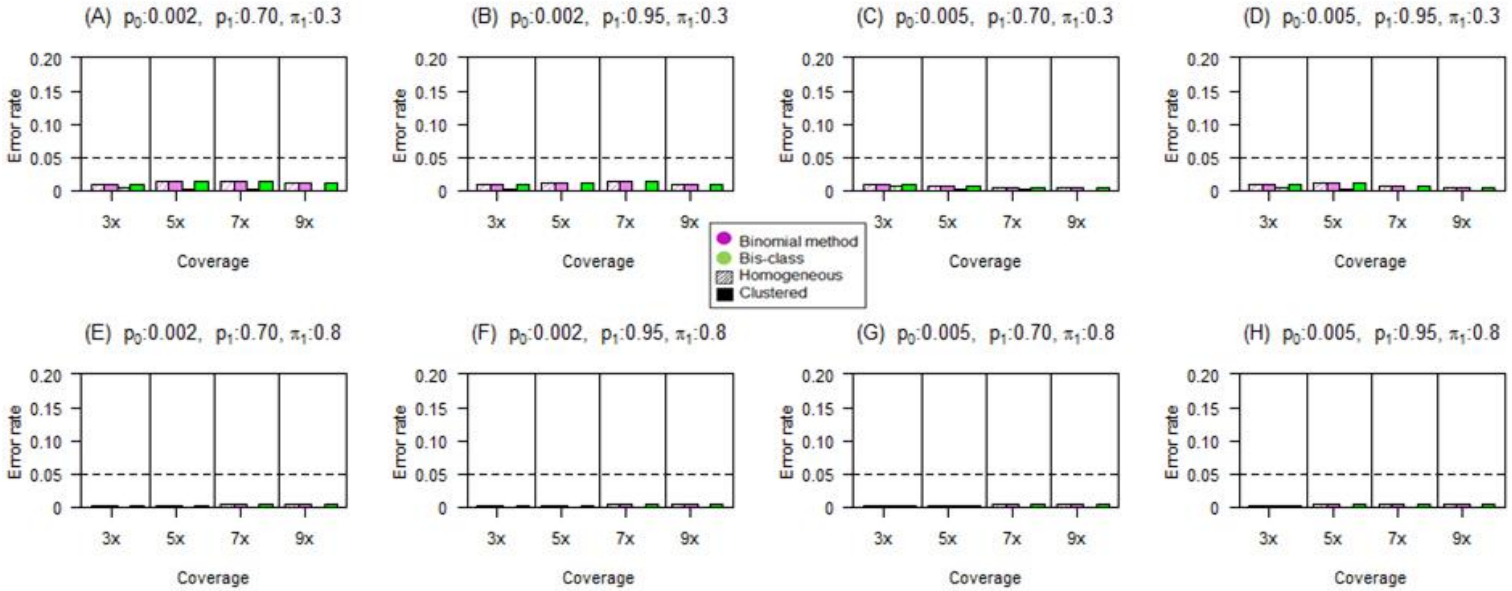

Figure 3.7 Comparison of misclassification rates of non-methylated CpGs via the Bis-Class and the binomial method using simulated data (intermediately or densely methylated region). We plotted 1-specificities for intermediately or densely methylated region. Compositions of the eight figures are similar with those of figure 3.4

3.5 Application to real dataset

3.5.1 Calling of honeybee (Insect) dataset and validation of our method

We applied the Bis-Class to the aforementioned honey bee data sets. We first estimated the parameters p_0 and p_1 using the EM algorithm. The results are shown in Table 2; all data had very low p_0, indicating that the error rates due to non-conversion are small. Importantly, the p_0 estimated from EM are highly similar to the values provided by the authors using experimental methods (Table 3.2). The estimates of p_1 values are around 70% for honey bee datasets. These are much lower than the estimate from the human genome (Table 3.2). The underlying cause for this discrepancy needs to be studied in future experiments. The genome-wide mean DNA methylation levels $\hat{\pi}_1{}^G$ are inferred from the estimated p_0 and p_1. These are highly similar to, but slightly lower than, the fractional methylation levels ($E(\hat{F})$ in Table 3.2). Intuitively, because the non-conversion rate (p_0) is substantial and on par with the mean methylation levels (Table 3.2), the fractional methylation levels at the face value could over-estimate the actual methylation levels. On the other hand, the fact that there may exist substantial levels of over-conversion (1-p_1) indicates that ignoring the effect of over-conversion can lead to under-estimate the overall methylation levels. For instance, if we assume p_1=0.95 (near perfect conversion), the estimated global methylation level $\hat{\pi}_1{}^G$ is much lower than fractional methylation (Table 3.2).

It is interesting to note that in the human data, the rate of over-conversion (1-p_1) is much lower than in the honey bee data. Nevertheless, due to the over-

conversion effectively under-estimating the actual methylation levels, the observed fractional methylation levels may be underestimates of the true methylation levels in the human genome. Again, if we assume a better over-conversion rate, the estimated global methylation level is closer to the observed fractional methylation levels (Table 3.2). The estimate of global methylation level is, as shown in the equation (9), affected by both i) the numerator term, which is the difference between the global methylation level and the non-conversion rate (false positives) and ii) the denominator term, which is the difference between over-conversion (false negatives) and non-conversion. In sparsely methylated genomes, the non-conversion rate, which affects the numerator significantly, will have a larger influence. In heavily methylated genomes, because the non-conversion rate is negligible compared to the overall methylation levels, the numerator will not change much by the correction, and the overall methylation levels will be more influenced by the over-conversion rate in the denominator.

For example, In the case of honey bee data in Table 3.2, mean methylation level $\widehat{\mathrm{E(F)}}$ (=0.0033) is only about twice that of $\hat{p}_0$ (=0.0015) and therefore the numerator $\widehat{\mathrm{E(F)}} - \hat{p}_0$ is reduced almost by half to 0.0018. Further dividing this further by $\hat{p}_1 - \hat{p}_0$(=0.65) leads to a smaller estimate of global methylation level compared to the $\widehat{\mathrm{E(F)}}$, especially when $\hat{p}_1$ is close to one. However, in the human data, $\widehat{\mathrm{E(F)}}$(=0.8064) is about 80 times bigger than $\hat{p}_0$ (=0.01), and subtracting $\hat{p}_0$ from $\widehat{\mathrm{E(F)}}$ does not make big difference. Further dividing this number further by $\hat{p}_1 - \hat{p}_0$ (=0.93) makes $\hat{\pi}_1$ even bigger than $\widehat{\mathrm{E(F)}}$. In summary, if $\widehat{\mathrm{E(F)}}$ is big compared to $\hat{p}_0$ and $\hat{p}_1 - \hat{p}_0$ is small, $\hat{\pi}_1$ will be generally bigger than $\widehat{\mathrm{E(F)}}$.

Biologically speaking, in the honey bee data, the non-conversion error rate is

similar to the global methylation level, thus overall inflate the methylation levels. Correcting for the errors will thus reduce the estimated methylation level. In the human data however, the non-conversion rate error is generally negligible compared to the actual methylation levels. The effect of over-conversion, even though lower than for the honey bee data, generally deflate the global methylation levels. Correcting for errors will thus increase the methylation levels.

We then evaluated posterior odds of each site to classify each site as methylated or non-methylated. Local information is obtained from 3 kb adjacent to the focal CpG site (1.5 kb on each side), and the weight parameter used is 0.5. As mentioned in Chapter 3.2, we found that the weight factor of 0.5 and the weight function of triangle are good from validation study (Figure 3.8). The numbers of methylated and non-methylated CpGs are shown in Table 2. In honeybee samples, Bis-Class detects on average 10% more methylated CpGs compared to the binomial method (Table 3.2). To determine whether this increase is due to high false positives or due to the improved inference, we investigated the difference between callings provided by the binomial method and the Bis-class methods further by several different approaches.

First, we found that many of these mCpGs detected by Bis-class are sites that are covered by a single C read that occur in highly methylated regions (Figure 3.9). This improved detection is because while the binomial method cannot recognize any mCpGs covered by only a single read (Figure 3.9), Bis-Class can provide methylation calling if that position occurs near other methylated CpGs. We demonstrate this property using two examples recovered from the data.

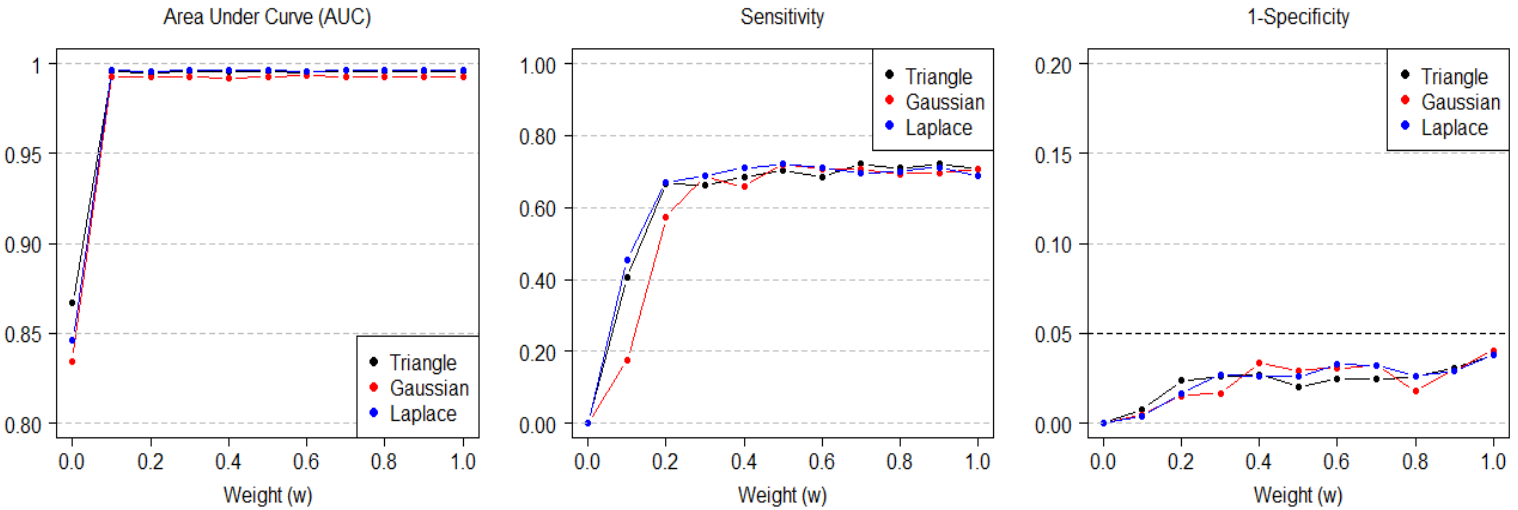

Figure 3.8 Using high-confidence CpG sites (coverage ≥7) and sampling one read for each site, we examined the AUC, sensitivity, and 1-specificity of different weight functions and weight factors. The results indicate that the three weight functions tried (Triangle, Gaussian, and Laplace) provide similarly high sensitivity and acceptable 1-specificity. Gaussian

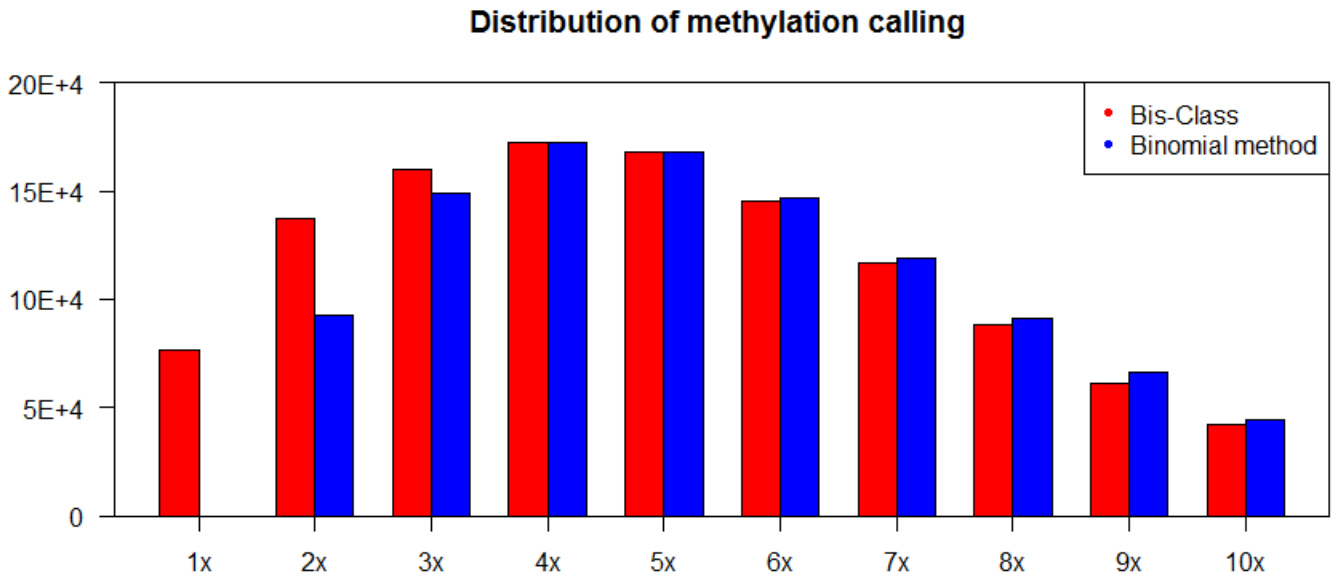

Figure 3.9 Histogram of mCpG counts detected using the Bis-Class and the Binomial method. Red and blue bars are the results from the Bis-Class and the Binomial method, respectively. X-axis indicates the coverage of each site and the Y-axis indicates the sum of methylated CpG counts in the 12 samples in Herb et al.

The first example is the gene (GB-16479) from two honey bee MethylC-seq data sets (Figure 3.10). In this data, four cytosines cluster in a region with high overall methylation levels (the fractional methylation level of a 1000 bps encompassing these four sites is ~ 0.9 in both samples). In sample A, the four cytosines were covered by only single reads, all 'C's. In sample B, the same four cytosines were covered by two 'C' reads. The binomial method calls all cytosines in the first sample as ′non-methylated', while calling all four cytosines in the second sample as 'methylated'. This example demonstrates the pitfalls of the binomial method clearly: two samples with exactly same qualitative information (100% 'C' reads in both cases) are classified as opposite directions due to the low sample size. Bis-Class, on the other hand, classified all four cytosines as 'methylated' for both cases. In the second example (Figure 3.10), we show the distribution of reads mapped to the locus GB 13135 in Herb et al. [42]. There are twelve samples in this data (six forager bees and six nurse bees). In the Forager 1 sample, the third and fifth positions are covered by single C reads. The Binomial method will call these as non- methylated (Figure 3.11B). However, since these sites occur in a heavily methylated region, Bis- Class calls both of these sites as methylated (Figure 3.11B). In other samples, these sites are covered by more than one read.

For example, in the Forager 6 sample, both positions third and five are covered by seven C reads, and consequently called as methylated CpGs. The similarity between different biological replicates indicates that using local information improves methylation- calling accuracy. FDR corrected q-values and posterior odds for each position of this locus are provided in the Table 3.3.

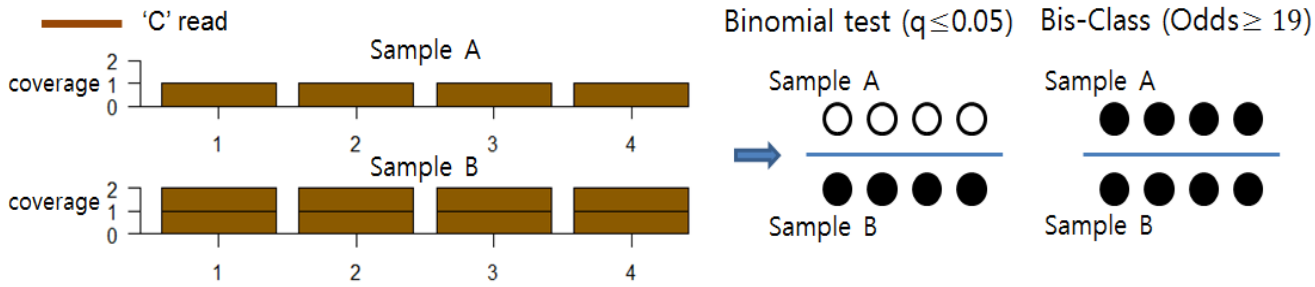

Figure 3.10 The GB 16479 locus exhibits qualitatively identical information yet opposite methylation calling under the binomial method. Data are from unpublished methylC-seq experiments of two honey bee individuals from the Yi lab, and are available upon request. All reads mapped to the four CpGs are 'C' reads (indicating 100% methylation). However, the binomial method provides a different methylation calls for these two samples. Specifically, the binomial calls all CpGs in the sample A as non-methylated (white dots), and all CpGs in the sample B as methylated (black dots). Bis-Class correctly identifies identical methylation features in the two replicates.

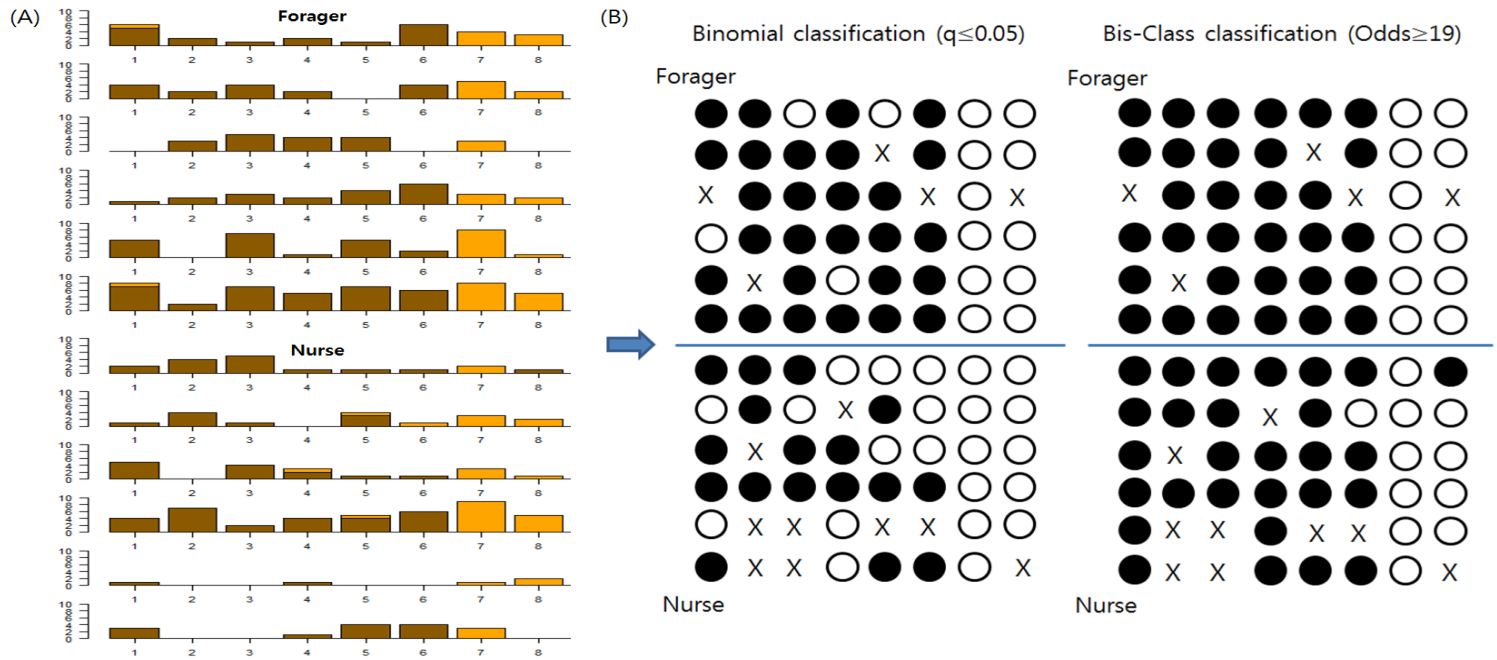

Figure 3.11 Contrasting methylation-calling results of the GB 13135 locus in Herb et al. [42] data by the two methods. (A) The numbers of ′C' reads (brown) and 'T' reads (orange) in 8 CpG positions of GB-13135. (B) Classification results following the binomial method (q-value < 0.05) and the Bis-Class method (Odds ≥ 19). CpGs classified as methylated are shown as black dots and those classified as non-methylated are shown as white dots. Sites with no read are marked as X. Bis-Class provides results that are more consistent across the biological replicates.

Table 3.3. q-values and odds of 12 honeybee samples in GB-13135 that are displayed in Figure 6.

Sample \ Position		1	2	3	4	5	6	7	8
SRR445767	q	2.63E-11	4.59E-4	0.22	4.59E-4	0.22	0	1	1
	Odds	3.01E+12	1.04E+05	2.20E+02	9.54E+04	1.97E+02	3.15E+15	5.86E-03	1.74E-02
SRR445768	q	1.77E-9	4.00E-4	1.77E-9	0.0004	NA	1.77E-9	1	1
	Odds	1.25E+10	7.07E+04	1.08E+10	6.09E+04	NA	8.99E+09	1.65E-03	3.66E-02
SRR445769	q	NA	7.74E-7	3.47E-12	1.6E-9	1.6E-9	NA	1	NA
	Odds	NA	6.34E+7	1.23E+13	2.64E+10	2.33E+10	NA	0.0207	NA
SRR445770	q	0.200	3.82E-4	7.23E-07	3.82e-4	1.45E-09	0	1	1
	Odds	216.13	9.66E+4	3.84E+7	8.52E+4	1.50E+10	3.25E+15	0.0167	0.0467
SRR445771	q	2.27E-12	NA	0	0.197775	2.27E-12	0.000357	1	1
	Odds	1.34E+13	NA	2.79E+18	255.2	1.13E+13	108110	7.77E-05	0.1639
SRR445773	q	0	3.38E-4	0	1.75E-12	0	0	1	1
	Odds	2.50E+18	1E+5	2.25E+18	1.04E+13	1.97E+18	4.23E+15	8.46E-05	2.06E-3
SRR445774	q	5.8E-4	3.73E-09	1.05E-11	0.240	0.240	0.240	1	0.240
	Odds	1.26E+5	2.55E+10	1.03E+13	252.28	220	219.81	0.0499	206.43
SRR445775	q	0.216	1.77E-9	0.216	NA	3.18E-6	1	1	1
	Odds	149.29	1.11E+10	124.91	NA	6.88E+6	0.097517	1.13E-2	3.23E-2
SRR445776	q	5.71E-12	NA	2.37E-9	0.001332	0.223	0.223	1	1
	Odds	5.06E+12	NA	1.06E+10	2.18E+4	122.49	122.37	0.0129	0.103
SRR445777	q	1.67E-09	0	4.22E-4	1.67E-09	7.66E-9	0	1	1
	Odds	1.56E+10	1.26E+18	7.35E+4	1.36E+10	4.16E+09	2.21E+15	2.40E-05	0.00169
SRR445778	q	0.190	NA	NA	0.190	NA	NA	1	1
	Odds	169.12	NA	NA	137.63	NA	NA	0.0899	0.0316
SRR445799	q	8.98E-7	NA	NA	0.215	1.92E-9	1.92E-9	1	NA
	Odds	3.82E+7	NA	NA	182.67	1.21E+10	1.21E+10	0.01557	NA

Second, we did the following experiments to directly assess the difference between the binomial method and Bis-Class when the numbers of reads is reduced. We assumed that we could distinguish methylated and non-methylated positions in coverage-rich CpG sites. We selected CpGs with over coverage of 7 in the honey bee scaffold 1.1. There were 9300 CpGs that satisfied this criterion. For these coverage-rich sites, we considered those with < 10% 'C' reads as non-methylated, and > 30% 'C' reads as methylated. We then generated a new methyl-seq data set by randomly selecting only a single read from these sites, thereby artificially reducing the coverage. We then used the binomial method and Bis-Class for methylation calling. Since we have information on the true methylation status, we can directly assess the false positives and false negatives from this experiment. We also performed the same experiments for the coverages of two and three reads. Each experiment was repeated 100 times. The results of these analyses, shown in Figure 3.12, demonstrate that Bis-Class is superior in these low coverage sites in the real data.

Third, we examined biological consistency across different methylC-seq data sets. We compared the calling results across the biological replicates offered by Herb et al. [42]. Bis-Class yields methylation callings that are more consistent among biological replicates. First, the coefficient of variation (CV) of methylated CpG counts in 12 samples from Bis-Class (0.067) is less than half of the CV from the binomial method (0.150). Second, we calculated pairwise correlations of gene methylation levels between samples in each subtype (foragers and nurses). Correlations between individuals are much greater for Bis-Class than those via the binomial method (Figure 3.13). Based on the biological facts that methylation

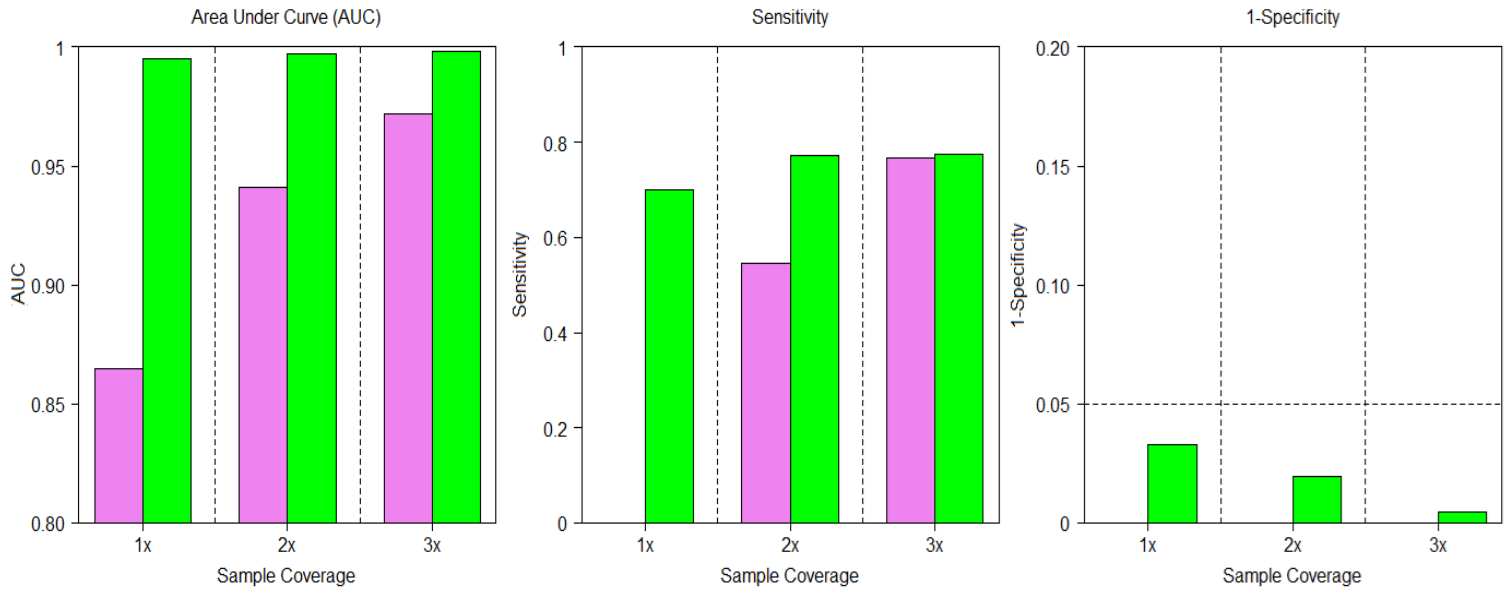

Figure 3.12 Validation results using real dataset. Violet bars are results from Binomial method and green bars are results from Bis-class

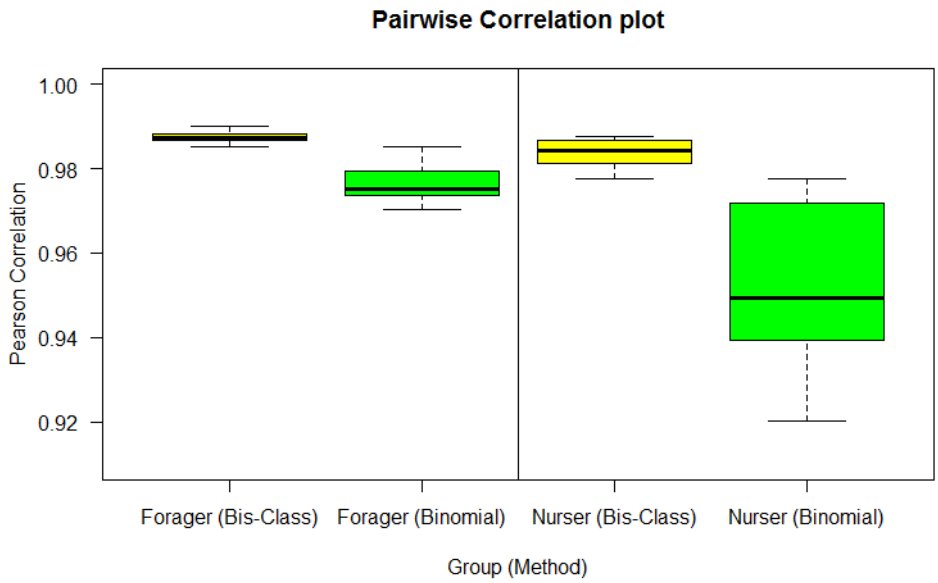

Figure 3.13 Correlations between biological replicates are higher in the Bis-Class calling compared to the binomial calling. The left panel represents the pairwise correlations between the methylation statuses of biological replicates in the forager samples from Herb et al. [42] data. The right panel represents the pairwise correlations between the nurse samples from the same study.

patterns are similar between individuals in the same species (e.g., Figure 3.2E), the observed higher correlations implies more realistic classification of DNA methylation via Bis- Class. We also note that in the binomial method, pairwise correlations are highly sensitive to the coverage levels. Specifically, nurse samples have more variable coverage than forager samples (Table 3.1), and the calling via method is highly variable, in contrast to the more stable methylation calling from Bis-Class.

3.6 Conclusion

The development of the methylC-seq method has propelled genome-wide evaluation of DNA methylation in diverse genomes across the tree of life. Due to the next generation sequencing nature of methylC-seq, the information content at each position varies greatly. Given such constraints, statistical methods that can perform robustly, even when sequence coverage is low, are desired. The existing binomial method is prone to errors in low coverage sites, particularly in sparsely methylated genomes. Our approach solves this problem by explicitly incorporating local DNA methylation levels in a Bayesian framework. This is based upon the observation that methylated sites are locally clustered. By utilizing both global and local methylation information, we can obtain more biologically consistent and relevant information. Bis-Class is particularly well-suited in the analyses of sparsely methylated genomes such as insect genomes.

Chapter 4

Application to real dataset and detecting differentially methylated region (DMR) analysis.

4.1 Introduction of DMR method

4.1.1 Fisher's exact test using binary calling dataset and its limit

In previous three chapters, we overviewed biological functions of methylation process, measuring techniques of methylation process, and calling method of the BS-seq technique. Based on the developed techniques and methods, researchers began to detect regions which are differentially methylated between individuals in different biological conditions because it can be strong evidences of many biologically important phenomena such as oncogenesis, aging process, onset of many diseases through controlling gene expressions. Therefore, there are needs for

developing of differential methylated region (DMR). In the initial stage, researchers use empirical criteria to detect DMR such as absolute difference between regions [59]. For example, if differences of methylation levels in some regions between two groups are over 0.3, they are decided to be differentially methylated regions. However, this is not statistically strict criteria and it can not control some important concepts of decision theory such as type 1 error.

In order to handle BS-seq data, firstly used statistical method for detecting DMRs is the Fisher's exact test [37,50]. When there are two groups distinguished by experimental condition or biological differences, we can construct a 2×2 contingency table (Table 4.1). The table constitutes of count of methylated status: methylated sites and non-methylated sites, and count of group status: case and control.

Then p-value of the test for independence would be given as sum of extreme probabilities calculated from hypergeometric distribution. Although the test is easy to conduct and intuitive, there is a problem in controlling type 1 error. The Fisher's exact test assumes that samples should be independent with each other. However, methylation levels of CpG cytosines within a distance are positively correlated as shown in the chapter 3. As a result, effective sample size would be smaller than actual number of CpG cytosine sites.

To demonstrate the inflation of type 1 error in Fisher's exact test, we artificially generated binary classified methylation dataset. We assumed that there is only one sample in each group, and number of groups is two. In each sample, numbers of CpG cytosine sites are decided to be 30, 50, 100, respectively. Proportions of methylated sites in each sample based on the null hypothesis are set to be 0.1, 0.3, and 0.5. In addition, correlation structures are set to be three statuses

Table 4.1 Data structure to conduct Fisher's exact test of a gene region for detecting DMR.

	Group A	**Group B**	**Total**
Methylated site	$\mathbf{n_{11}}$	$\mathbf{n_{12}}$	$\mathbf{n_{1+}}$
Non- Methylated	$\mathbf{n_{21}}$	$\mathbf{n_{22}}$	$\mathbf{n_{2+}}$
Total	$\mathbf{n_{+1}}$	$\mathbf{n_{+2}}$	$\mathbf{n_{++}}$

considering various situations: no correlation, weak correlation, and strong correlation. In order to generate correlated binary samples, we used R package "bindata".

For each correlation structure, we used following matrices. $\boldsymbol{R}$ denotes matrix of correlation structure with $N \times N$ size. N denotes number of CpG cytosine sites in each sample. $\boldsymbol{R_{ij}}$ is the (i,j) element of the matrix $\boldsymbol{R}$, and means that the correlation between i and j^{th} CpG cytosine sites. Detailed descriptions are below.

For no correlation setting, we used identity matrix as $\boldsymbol{R}$. However, $\boldsymbol{R_{ij}}$ s are linearly decreasing as the distance between CpG cytosines are increasing for two substantial correlation settings. For weak correlation, $\boldsymbol{R_{ij}}$ is represented as $\max(0.3-d_{ij}/3000,0)$ for $i \neq j$, and 1 for i=j. d_{ij}s are physical distances between i and j^{th} CpG cytosine sites. In this simulation, d_{ij} between the nearest CpG sytosines is set to be 15 and those are assumed to be uniformly distributed across whole genome. For strong correlation, $\boldsymbol{R_{ij}}$ is represented as $\max(0.7-d_{ij}/3000,0)$ for $i \neq j$, and 1 for i=j.

For each simulation setting, we iterated 1000 times and calculate proportion of tests whose p-value is below 0.05 and 0.005, respectively. The result of simulation is summarized in table 1. As we can see in the Table 4.2, patterns of type 1 error inflation become clear, as number of CpG sites, size of correlation strength, and proportion of methylated CpG sites are increasing. In the worst case, over 80% of tests are rejected at $\alpha = 0.05$. As seen in the Figure 3.2, correlations are generally positive and it can be similar with the patterns of strong correlation in our simulation. Therefore, this property may result in overestimation of number of differentially methylated genes.

Table 4.2 Estimated type 1 error rate from simulation via Fisher's exact test. The first column means correlation structure, and the second column is proportion of methylated CpG cytosines. First row is number of CpG cytosine sites in a gene. Values except headers are estimated type 1 error rates (values out of the parenthesis are estimated at α=0.05, and values in the parenthesis are estimated at α=0.005)

		30	**50**	**100**
No correlation	**0.1**	0.011 (0.001)	0.020 (0.002)	0.025 (0.002)
	0.3	0.028 (0.002)	0.032 (0.003)	0.033 (0.003)
	0.5	0.027 (0.006)	0.034 (0.004)	0.042 (0.003)
Weak Correlation	**0.1**	0.239 (0.155)	0.349 (0.236)	0.503 (0.366)
	0.3	0.482 (0.328)	0.558 (0.416)	0.609 (0.471)
	0.5	0.489 (0.342)	0.554 (0.445)	0.625 (0.493)
Strong Correlation	**0.1**	0.264 (0.23)	0.313 (0.268)	0.411 (0.360)
	0.3	0.566 (0.505)	0.637 (0.566)	0.739 (0.677)
	0.5	0.648 (0.569)	0.735 (0.662)	0.815 (0.751)

4.2 Methods

4.2.1 Overview of the Cochran-Mantel-Haenszel (CMH) test

In order to resolve the problem which the Fisher's exact test has, we considered the Cochran-Mantel-Haenszel (CMH) test because it can use information of location as stratification variables. In this chapter, we overview the original CMH test, and we will extend it to slightly modified version in chapter 4.2.2.

The original CMH test proposed by Mantel and Haenszel is the method to tests conditional independence of $2 \times 2 \times K$ contingency tables [60,61], meaning that the method is commonly used to test for conditional independence between two binary variables, after adjusting for the effect of confounding variables with *K* strata. Statistics of the test follow chi-square distributions with one degree of freedom and perform best when the associations of two binary variables have the same directions in each partial table.

Situational application of this approach has been generalized by Birch [62], Landis [63], and Mantel [64] to an $I \times J \times K$ table in which the predictor variable and the response variable have *I* and *J* levels, respectively, that can be treated as not only nominal but also as ordinal. Therefore, the generalized CMH method consists of two more tests, in addition to the conditional independence test for two nominal variables. One test examines the mean score difference when one variable is ordinal, and the other test evaluates the correlation when both variables are ordinal [64]. The generalized CMH statistics is given as

$$L^2 = [\sum_k B_k(n_k - \mu_k)]'[\sum_k B_k V_k B_k']^{-1}[\sum_k B_k(n_k - \mu_k)] \qquad (1)$$

In the above equation, $\boldsymbol{B}_k$ is the Kronecker product between the row score $\mathbf{u_k}$ and the column score $\mathbf{v_k}$, $\mathbf{n_k}$ and $\boldsymbol{\mu}_\mathbf{k}$ are vectors of observed and expected counts of length of $I \times J$ in the k^{th} strata, respectively. $\boldsymbol{V}_k$ is an $(I \times J) \times (I \times J)$ variance matrix of $\mathbf{n_k}$ **,** evaluated under an assumed hypergeometric distribution. Therefore, $\mathbf{n_k}$ and $\boldsymbol{\mu}_\mathbf{k}$ is represented as $(n_{11k}, n_{12k}, ..., n_{IJk})$ and $(n_{1+k} \times n_{+1k}, n_{1+k} \times n_{+2k}, ..., n_{I+k} \times n_{+Jk}) / n_{++k}$ respectively. Moreover, elements of $\boldsymbol{V}_k$ consist of covariance terms between n_{ijk} and $n_{i'j'k}$, and are represented as $n_{i+k}(\omega_{ii'} n_{++k} - n_{i'+k}) n_{+jk} (\omega_{jj'} n_{++k} - n_{+j'k}) / (n_{++k}^2 (n_{++k} - 1))$ where $\omega_{ab} = 1$, when *a* = *b* and $\omega_{ab} = 0$ otherwise.

Three types of tests can be derived by imposing ordinal or nominal weights on $\mathbf{u_k}$ and $\mathbf{v_k}$. When $\mathbf{u_k}$ is used as the nominal variable, it is described as a $(I-1) \times I$ matrix (***I***, **-1)**, where ***I*** is an identity matrix of size $I-1$, and **1** denotes a column vector of $I-1$ ones. When $\mathbf{u_k}$ is used as the ordinal variable, it is given as $(u_1, u_2, ..., u_I)$, with an ordered score vector given to each level of predictor. $\mathbf{v_k}$ is constructed similarly with $\mathbf{u_k}$. Therefore, the general association test is conducted if both variables are nominal, the mean score test is conducted if only one variable is ordinal, and the correlation test is conducted if both variables are ordinal. The degrees of freedom are given as $(I-1) \times (J-1)$ for the general association test, $I-1$ or $J-1$ for the mean score test, and 1 for the correlation test [65].

4.2.2 Application of the CMH Method to BS-seq Data.

Because the original CMH method used information of strata for conditional independent test, we applied it to BS-seq data and regarded locus information for stratification. However, as seen in the chapter 3, there are substantial correlations between neighbor CpG cytosine sites. Therefore, if we ignore the correlation part of BS-seq data, type 1-error rate can be inflated as in the case of Fisher's exact test. In order to resolve this problem we additionally included covariant part to the statistics of CMH, especially on the denominator. The resulting statistics including covariance part can be represented as below.

$$L^2 = [\sum_k B(n_k - \mu_k)]'[\sum_k BV_kB' + \sum_{k,k'} BV_{k,k'}B']^{-1}[\sum_k B(n_k - \mu_k)] \quad (2)$$

In this equation, $V_{k,k'}$ is the covariance matrix between k^{th} and k'^{th} locus. For estimating $\sum_{k,k'} BV_{k,k'}B'$, we proposed two schemes. The one of two schemes is derivation from underlying distribution such as multivariate hypergeometric because variance part of CMH test is estimated from hypergeometric distribution. This scheme used all possible pairwise combination of methylation sites, make a new 2×4 contingency table (Figure 4.1) for each combination, and finally calculate covariance part. Although this scheme is basic and conserves type 1 error rate, it did not consider physical distance between CpG cytosine sites. Therefore, the other scheme includes covariance part estimated from spatial correlation (Figure 4.2). We illustrated two methods using mathematical formula below.

Y	m_k	$m_{k'}$
1	0	1
1	1	1
1	1	1
1	1	0
1	1	0
0	0	0
0	0	0
0	0	0
0	0	1
0	1	1

K^{th} loci	Methylated	Non-methylated	Total
Case	n_{11k}(=4)	n_{12k} (=1)	n_{1+k} (=5)
Control	n_{21k} (=1)	n_{22k} (=4)	n_{2+k} (=5)
Total	n_{+1k} (=5)	n_{+2k} (=5)	n_{++k}(=10)

K'^{th} loci	Methylated	Non-methylated	Total
Case	$n_{11k'}$ (=3)	$n_{12k'}$ (=2)	$n_{1+k'}$ (=5)
Control	$n_{21k'}$(=2)	$n_{22k'}$(=3)	$n_{2+k'}$ (=5)
Total	$n_{+1k'}$ (=5)	$n_{+2k'}$ (=5)	$n_{++k'}$(=10)

(K, K')	(m,m)	(m,nm)	(nm,n)	(nm,nm)	Total
Case	$n_{11kk'}$ (=2)	$n_{12kk'}$ (=2)	$n_{13kk'}$ (=1)	$n_{14kk'}$ (=0)	$n_{1+kk'}$ (=5)
Control	$n_{21kk'}$ (=1)	$n_{22kk'}$ (=0)	$n_{23kk'}$ (=1)	$n_{24kk'}$ (=3)	$n_{2+kk'}$ (=5)
Total	$n_{+1kk'}$ (=3)	$n_{+2kk'}$ (=2)	$n_{+3kk'}$ (=2)	$n_{+4kk'}$ (=3)	$n_{++kk'}$(=10)

Figure 4.1 A new 2×4 contingency table for estimating covariance between k^{th} and k'^{th} sites. m is an abbreviation of methylated and nm is that of non-methylated.

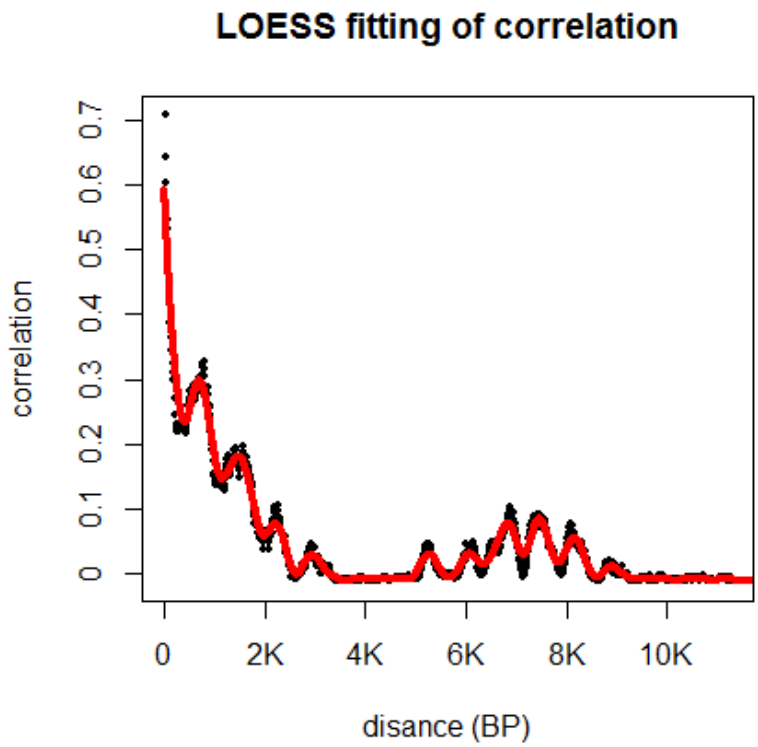

Figure 4.2 LOESS fitting of spatial correlation in honeybee dataset. Black dots represent Pearson correlations calculated in all possible distances. The red line represents fitted curve of black dots via LOESS method [66].

For the first scheme, covariance term between k and k'^{th} CpG cytosine sites can be calculated as follows, depicted in Figure 4.1, and finally plugged into $V_{k,k'}$covariance matrix.

$$V_{k,k'(1,1)} = \mathrm{cov}(n_{11k}, n_{11k'})$$
$$= \mathrm{cov}(n_{11kk'} + n_{12kk'}, n_{11kk'} + n_{13kk'})$$
$$= \mathrm{var}(n_{11kk'}) + \mathrm{cov}(n_{11kk'}, n_{13kk'}) + \mathrm{cov}(n_{12kk'}, n_{11kk'}) + \mathrm{cov}(n_{12kk'}, n_{13kk'}) \quad (3)$$

$\mathrm{Cov}(n_{1ikk'}, n_{1jkk'})$ is estimated via simple calculation and the resulting value is $-\frac{n_{1+}n_{+i}n_{+j}}{{n_{++}}^2} \times \frac{n_{++}-n_{1+}}{n_{++}-1}$. Therefore, $\mathrm{cov}(n_{11k}, n_{11k'})$ can be summarized as $\frac{n_{1+}\times(n_{++}-n_{1+})\times n_{+1}}{n_{++}(n_{++}-1)} - \frac{n_{1+}\times(n_{++}-n_{1+})}{{n_{++}}^2(n_{++}-1)} \times ({n_{+1}}^2 + n_{+1} \times n_{+2} + n_{+1} \times n_{+3} + n_{+2} \times n_{+3})$.

In this scheme, the number of terms to be calculated increases in quadratic order of the number of CpG cytosines in a region. Therefore, permutation can be used to estimate denominator part of the statistic when there are too many CpG cytosines in some genes.

For the second scheme, covariance term between k and k'^{th} CpG cytosine sites can be estimated from empirical spatial correlation calculated using real dataset. Firstly, spatial correlation according to distance between CpG cytosines is estimated following steps. If we assume that we have N CpG cytosines in hand and N is sufficiently large (e.g. 10000), we can index them by physical location. For correlation of *d* distance, it can be calculated as average of all correlations of vectors in a group having *d* distance. When there are small numbers of pairs in some distances, we use LOESS (locally weighted scatterplot smoothing) [66] to stably estimate correlations (Figure 4.2, 4.3). Parameters for fitting complexity can be chosen by researcher's decision. With the fitted LOESS curve, we predict correlation value for any positive distances.

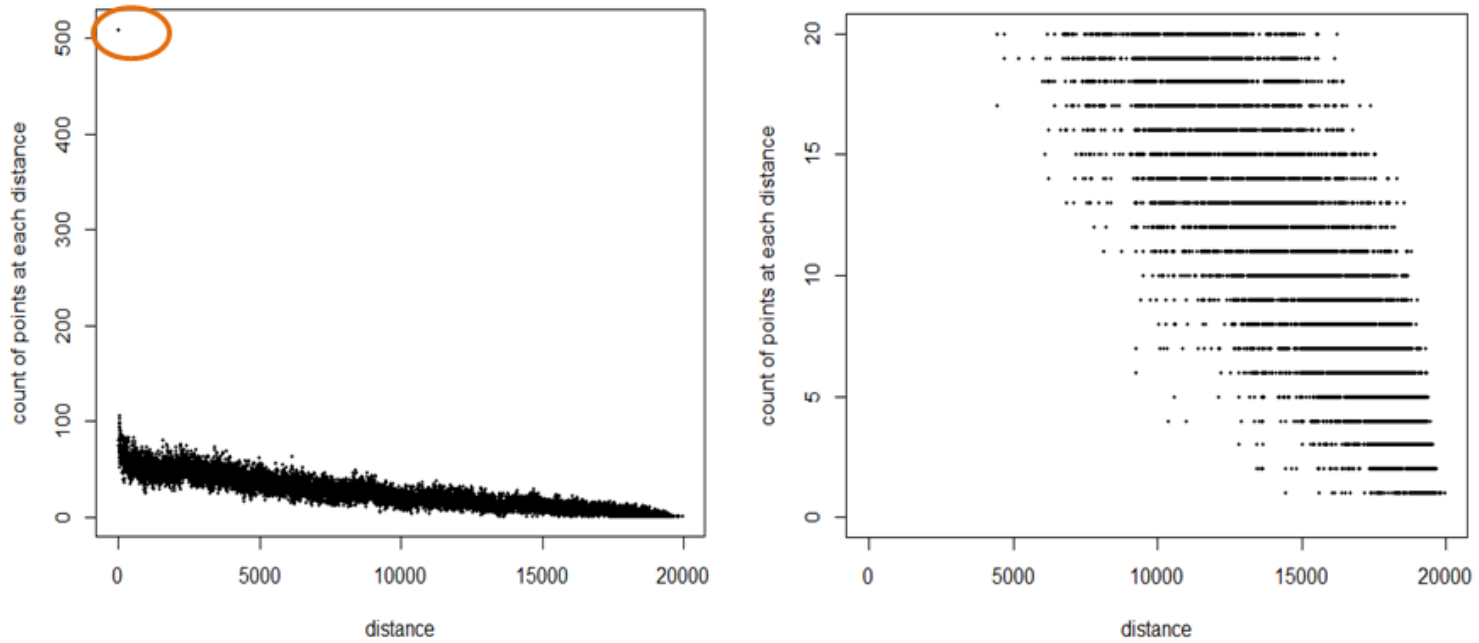

Figure 4.3 Plots of number of pairs for each distance using a region of 20 KB. (A) Plot of count within 20KB from honeybee dataset. A maximum value is over 500 in distance of 1 base pair. (B) Detailed plot of (A) whose range of y-axis is from 0 to 20.

Finally, using the formula of correlation, we obtain estimated covariance as below.

$$V_{k,k'(1,1)} = \mathrm{cov}(n_{11k}, n_{11k'}) = r_{k,k'} \times \mathrm{sd}(n_{11k}) \times \mathrm{sd}(n_{11k'})$$
$$= r_{k,k'} \times \mathrm{var}(n_{11k})^{1/2} \times \mathrm{var}(n_{11k'})^{1/2} \quad (4)$$

$r_{k,k'}$ denotes estimated correlation value from all pairs of vectors having same distance with the distance between k and k'^{th} CpG cytosines. It means that any pairs of CpG sites having same distance are assumed to have same correlation coefficients.

In order to confirm the our proposed methods conserve type 1 error inflation, we conducted simulation studies and used same correlation structure used in validation of the Fisher's exact test. As a result, we obtained estimated type 1 errors and those are summarized in Table 4.3 and 4.4. In this simulation studies, we used multiple samples in each group such as 5 or 10. As seen in the tables, we can conclude that the our proposed modified CMH method preserves type 1 error rate compared with Fisher's exact test, although there are some settings which exceed type 1 error rates with small amounts. As number of samples in a group increases, or there are balanced methylated CpGs in a group (i.e methylated proportion reaches 0.5), estimated type 1 error rates reach the nominal significant level. Although there are slight conservativeness in our proposed method, we expect that selected regions (genes) from our method can be regarded as truly differentially methylated regions.

Table 4.3 Simulation of type 1 error rates for five of sample size in a group. Values without parenthesis are type 1 error rate for 0.05, and values in parenthesis are type 1error for 0.005.

		Estimation from underlying distribution			Estimation from spatial correlation		
		30	**50**	**100**	**30**	**50**	**100**
No correlation	**0.1**	0.037 (0.001)	0.051 (0.000)	0.046 (0.002)	0.048 (0.005)	0.047 (0.004)	0.052 (0.006)
	0.3	0.042 (0.001)	0.047 (0.000)	0.042 (0.000)	0.052 (0.004)	0.051 (0.002)	0.044 (0.004)
	0.5	0.034 (0.001)	0.044 (0.001)	0.045 (0.002)	0.045 (0.002)	0.052 (0.007)	0.061 (0.003)
Weak Correlation	**0.1**	0.033 (0.000)	0.023 (0.000)	0.020 (0.001)	0.058 (0.001)	0.07 (0.002)	0.076 (0.004)
	0.3	0.046 (0.000)	0.041 (0.002)	0.040 (0.000)	0.054 (0.004)	0.05 (0.004)	0.056 (0.005)
	0.5	0.049 (0.002)	0.038 (0.001)	0.042 (0.000)	0.046 (0.005)	0.04 (0.004)	0.05 (0.004)
Strong Correlation	**0.1**	0.006 (0.001)	0.010 (0.000)	0.012 (0.000)	0.009 (0.000)	0.011 (0.000)	0.017 (0.000)
	0.3	0.037 (0.001)	0.043 (0.000)	0.04 (0.001)	0.035 (0.000)	0.037 (0.000)	0.045 (0.002)
	0.5	0.049 (0.003)	0.046 (0.001)	0.052 (0.002)	0.044 (0.002)	0.047 (0.005)	0.064 (0.003)

Table 4.4 Simulation of type 1 error rates for ten of sample size in a group. Values without parenthesis are type 1 error rate for 0.05, and values in parenthesis are type 1error for 0.005.

		Estimation from underlying distribution			Estimation from spatial correlation		
		30	**50**	**100**	**30**	**50**	**100**
No correlation	**0.1**	0.047 (0.002)	0.065 (0.001)	0.052 (0.003)	0.045 (0.007)	0.041 (0.003)	0.053 (0.005)
	0.3	0.056 (0.005)	0.051 (0.006)	0.054 (0.003)	0.053 (0.006)	0.049 (0.007)	0.052 (0.009)
	0.5	0.041 (0.004)	0.042 (0.002)	0.035 (0.001)	0.046 (0.002)	0.050 (0.006)	0.042 (0.005)
Weak Correlation	**0.1**	0.034 (0.001)	0.041 (0.001)	0.037 (0.001)	0.047 (0.002)	0.049 (0.004)	0.051 (0.004)
	0.3	0.041 (0.004)	0.050 (0.006)	0.051 (0.004)	0.061 (0.006)	0.056 (0.005)	0.049 (0.005)
	0.5	0.060 (0.005)	0.055 (0.003)	0.05 (0.003)	0.049 (0.006)	0.053 (0.008)	0.051 (0.007)
Strong Correlation	**0.1**	0.023 (0.000)	0.029 (0.000)	0.026 (0.000)	0.02 (0.001)	0.027 (0.001)	0.028 (0.001)
	0.3	0.052 (0.004)	0.052 (0.003)	0.049 (0.002)	0.047 (0.003)	0.049 (0.003)	0.047 (0.003)
	0.5	0.057 (0.003)	0.051 (0.003)	0.054 (0.002)	0.053 (0.003)	0.048 (0.004)	0.038 (0.004)

4.3 Application to real dataset

4.3.1 Detecting DMRs using honeybee dataset and validation of our method

We applied our modified CMH method to real dataset which is used in chapter 3 to validate our binary calling method. Using our proposed DMR method, we planned to detect DMR regions between forager and nurse groups. There are about 8800 genes in our honeybee dataset, especially SRR445767 sample. As seen in the Figure 4.4, median of length of 8800 genes is 2576 base-pair, and median of number of CpG cytosines is 112. Both histograms are shown in $\log_{10}$ transformed scale in X-axis. For the honeybee dataset of six forager samples and nurse samples [42], we conducted DMR analysis using four approaches. The first approach is applying Fisher's exact test. The second approach is applying original CMH test without correlation part. The third approach is applying modified CMH test with correlation part estimated from underlying distribution. The fourth approach is applying modified CMH test with correlation part estimated from spatial correlation and LOESS fitting.

Before the DMR analysis, we set CpG cytosines to be non-methylated when the fraction of C reads are smaller than 30%, otherwise they are set to be methylated. This is because the fraction of C reads of methylated cytosines in honeybee samples are around 70%. Then we applied above four approaches to conduct binary classified dataset. For the first sample of forager subtype, only 0.36% of total CpG cytosines are classified as methylated cytosines.

After conducting CMH test, we plotted quantile-quantile (Q-Q) plot after

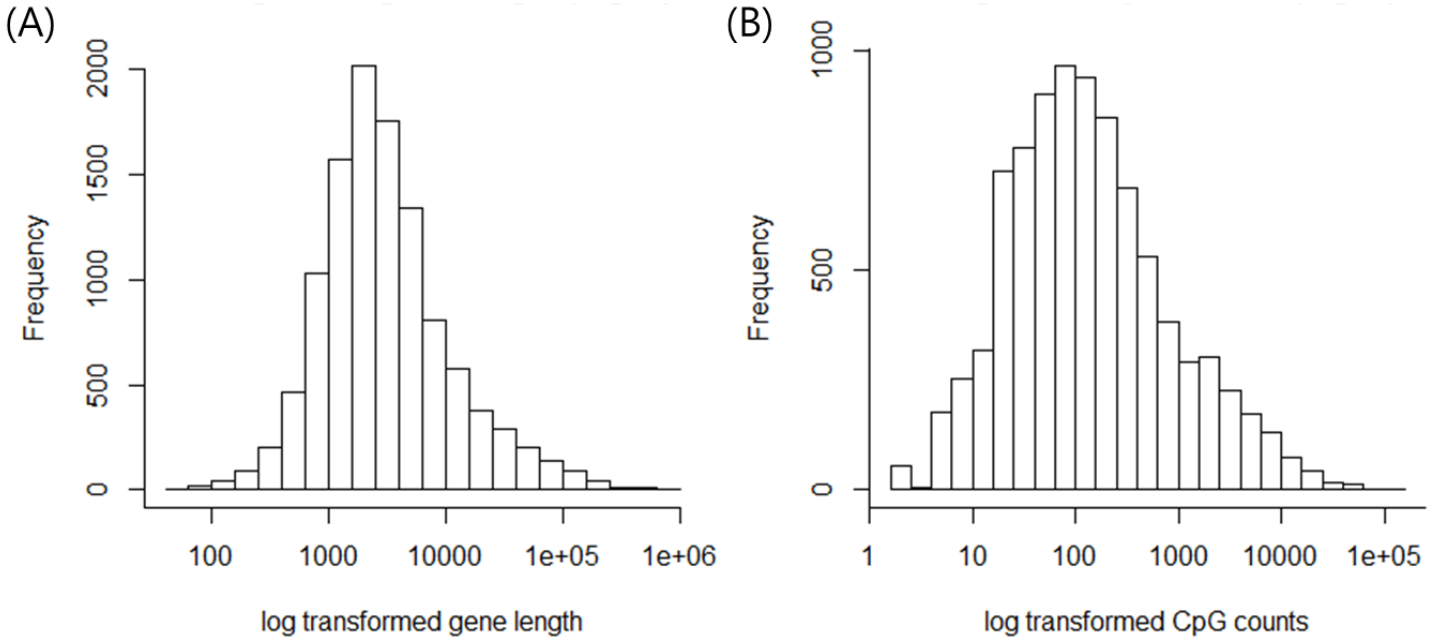

Figure 4.4 Histograms of gene length and CpG counts in honeybee dataset. X-axis implies $\log_{10}$-transformed counts of gene length (A) and CpGs (B)

converting p-values into negative value of $\log_{10}$ transformed p-values ($-\log_{10}$(p-values)) in order to see patterns of the obtained p-values (Figure 4.5). In the Q-Q plot, we can see that the p-values from the first approach are very inflated because large portion of dots are located out of confidence interval of p-values. Second approach has also inflated p-values, although we can't distinguish them with true DMRs. Third and fourth approaches show very similar results, and their resulting dots are almost located in the confidence interval. From the Figure 4.5, we can conclude that there are generally positive correlations in the honeybee dataset rather than negative correlations.

Although there are no significant genes from the third and fourth approaches after FDR adjustment, we also plotted Ben diagram of selected genes (Figure 4.6) having p-values under 0.05. We only plotted a result of the modified CMH method using hypergeometric distribution for correlation because two modified CMH methods are produced very similar results. As seen in the figure, we can see that large portion of significant gene at nominal level from the original CMH test is also included in the result of the Fisher's exact test. This may imply that the original CMH test can violate type 1 error rate, but the amount of the violation is slightly smaller than the Fisher's exact test. However, 40% regions detected the modified CMH method is not overlapped with the Fisher's exact test. It may be because those are negatively correlated regions.

We further constructed Table 4.5 to list top 20 significant regions from the modified CMH method. We also included p-values from the Fisher's exact test and the original CMH test to the corresponding regions. There are some regions whose p-value from the Fisher's exact test and the original CMH test are over 0.05. This also implies that there are some negative correlated regions in the honeybee dataset.

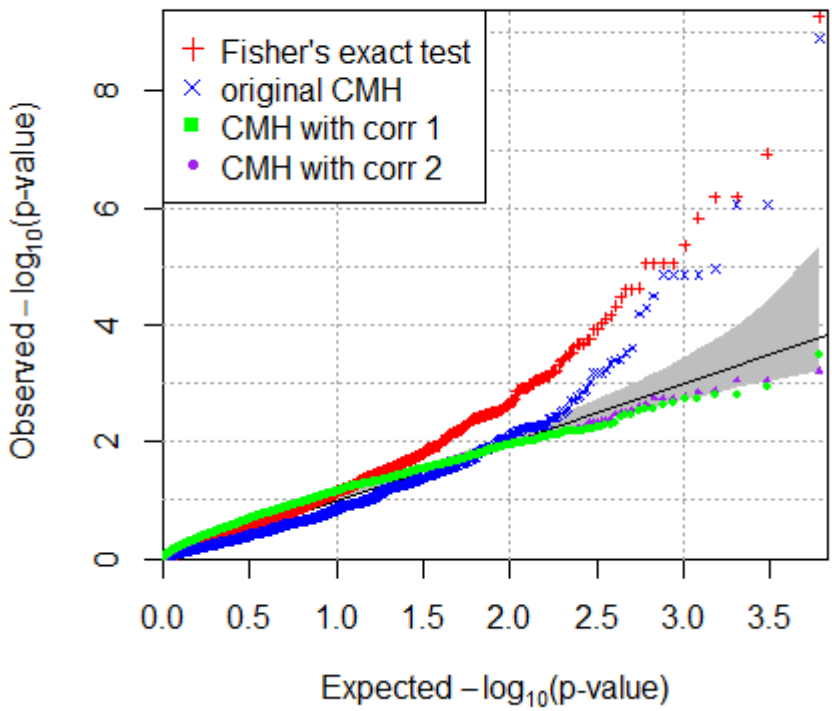

Figure 4.5 Q-Q plots of p-values obtained from four approaches of honeybee data analysis. Red, blue, green, violet dots are from the Fisher's exact test, the original CMH test without correlation part, the modified CMH test with correlation part estimated from hypergeometric distribution (CMH with corr 1), and the modified CMH test with correlation part estimated from distance-wise correlation (CMH with corr 2), respectively.

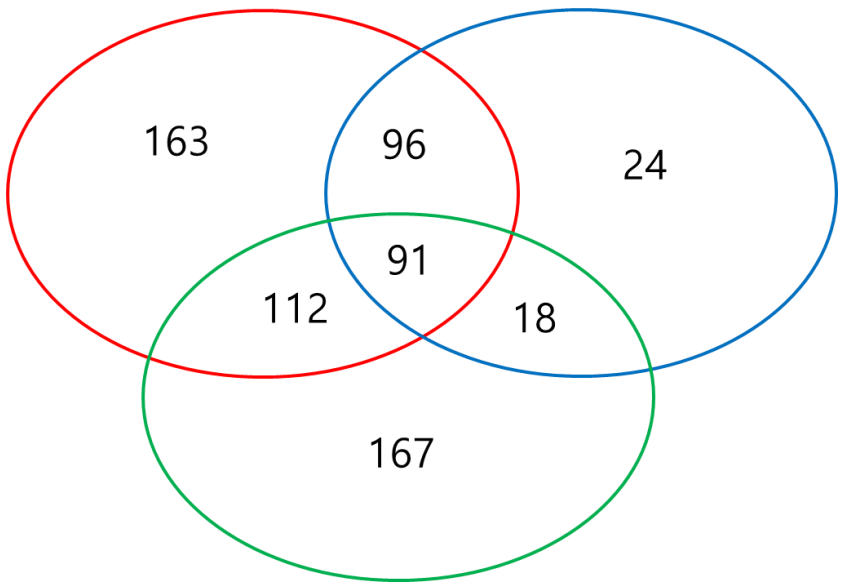

Figure 4.6 Ben diagram of selected genes having p-values under 0.05. Red, blue, green circles imply results from the Fisher's exact test, original CMH test, and the modified CMH test using hypergeometric distribution for correlation, respectively.

Table 4.5. Top 20 significant results from the modified CMH test. We listed 20 genes having smallest p-values among whole genes. We also listed p-values from other approaches. All selected region produce mRNAs.

name	Scaffold	Start site	End site	P-value (modified CMH)	P-value (original CMH)	P-value (Fisher's exact test)
GB17492	GroupUn.2	579690	582581	0.00032	0.24	0.0097
GB17128	Group1.28	135479	139349	0.0012	0.033	0.043
GB15134	Group8.13	180976	187205	0.0015	0.16	2.39E-05
GB10567	Group3.8	206318	233781	0.0016	0.019	0.019
GB11307	Group10.20	82592	83215	0.0018	0.084	0.0062
GB11197	Group9.3	56101	95598	0.0019	0.023	0.021
GB19139	Group10.3	32281	33184	0.0021	0.20	0.0095
GB13565	Group9.3	357850	360508	0.0023	0.029	0.067
GB10850	GroupUn.221	1285	4440	0.0027	0.087	0.039
GB17882	Group1.3	1072	10143	0.0028	0.24	0.050
GB19462	Group3.2	89820	113604	0.0030	0.037	0.0031
GB14929	Group11.15	5327	10595	0.0034	0.12	0.18
GB18951	Group7.29	276958	277629	0.0035	0.50	0.15
GB19156	Group5.24	194774	197456	0.0039	0.65	0.026
GB15614	Group8.17	49585	53354	0.0049	0.047	0.00017
GB16762	Group9.6	61748	65008	0.0052	0.0086	0.021
GB30047	Group15.8	637770	641350	0.0053	0.0057	0.013
GB15175	Group9.10	248530	250094	0.0057	0.066	0.0086
GB16499	GroupUn.77	63276	65144	0.0058	8.74E-07	6.48E-07
GB30202	Group4.18	123163	124187	0.0059	0.0053	0.013

4.4 Conclusion

The BS-seq dataset needs to have a new statistical method to consider the spatial correlation property of methylation status, because there are non-ignorable and substantial amount of positive (generally) correlations among adjacent CpG cytosines. Therefore we used the CMH test and modified it by including correlation structure to analyze BS-seq dataset. From simulation studies, we confirmed that our proposed method conserves type 1errors, while the Fisher's exact test can't achieve it. However, our proposed method also reduces number of statistical significant regions in compensation.

Using real data analysis, we found that our method did not detect statistically significant regions after applying multiple hypothesis correction method. Therefore, we may need to improve the property of the modified CMH method although forager and nurse are not very different sub-species, and whole methylation levels of two subcastes are very similar (Table 3.2). As a solution, only considering correlation within a certain window may work in order to exclude correlations of far distance that may not be meaningful in biological aspect.

In addition, ordinal and multinomial conversion of methylation status can be valid in our modified CMH method, and it can improve power because it uses more information that binary dataset analysis. Therefore, we will further study those approaches as a future work. Because read coverages of CpG cytosines have large variance and substantial proportion of CpG cytosines have very low coverages, categorical variable analysis can be more reasonable than the continuous variable analysis. As one of the representatively categorical approaches, CMH-based approach of methylation analysis is expected to be used in many biological studies.

Chapter 5

Summary and conclusion

Recently, research of the epigenetics receives a lot of attentions because many important biological functions were not explained enough via only genetic analysis. Especially, the DNA methylation process is one of the most highlighted topics among epigenetic phenomena. The DNA methylation process generally affects gene expressions by silencing when it is highly methylated. Therefore, it can result in onset of cancer, X-chromosome inactivation, aging, and genomic printing, etc.

Therefore, methylation measuring techniques have been developed together with needs of methylation studies. By the appearance of BS-seq technology, we can get information of methylation with base-pair resolution and new corresponding statistical methods to handle the BS-seq dataset are needed.

In Chapter 3, we proposed a new binary classification method using local information and Bayes classifier. Using the fact that methylation status of adjacent CpG cytosines are correlated, we improved detection power of methylated CpG cytosines compared with the binomial test using FDR especially in the dataset of low methylation level and low sequencing read coverages. Unlike human

methylation dataset which are generally highly methylated, some species of insects or plants are very sparsely methylated and methylated region are generally clustered. However, in order to get enough information, increasing average sequencing read depth would be very costly. Therefore, there are many BS-seq datasets whose methylated CpG cytosine sites are not well detected via the existing binomial method.

To resolve this problem, we additionally employed overconversion rate in the methylated CpG cytosines, and constructed Bayes classifier to include local information. Our proposed classification method is confirmed to be better than the binomial method from simulation study and real data analysis of insect dataset.

In Chapter 4, we proposed a modified CMH test which detects DMRs in BS-seq data analysis. Instead of only using variance part of original CMH test, we additionally included covariance part between adjacent CpG cytosines to control type 1 error. For adjusting covariance part, we proposed several method of estimation. Firstly, we derive covariance part from hypergeometric distribution. When the numbers of CpG cytosine sites are too large, we instead estimated it from permutation. However, this test did not consider location information directly. Therefore, we also used local correlation information estimated in each physical distance, and additionally introduced LOESS fitting method for sparsely counted pairs of some distances. This method provides an approximate estimate of spatial correlation according to each of physical distances between adjacent CpG cytosines, and has an advantage in testing small number of samples in a group or relatively narrow regions having a few CpG cytosines. From simulations, we confirmed that the modified CMH test conserved type 1 error rates, which is not conserved in the Fisher's exact test. Finally, we applied our proposed method to real data analysis

with honeybee methylation dataset, and discussed analysis results.

In summary, the bisulfite sequencing technique helps us to get the most precise information of methylation status, but it needs some considerations of spatial correlation between adjacent CpG cytosines. One of the reasons can be that enzymes involved in the DNA methylation process make several CpG cytosines to be methylated simultaneously. We considered this phenomenon in the proposed statistical models, and validated improved properties via simulation study and real data analysis.

References

1. Li E, Beard C, Jaenisch R: Role for DNA methylation in genomic imprinting. Nature, 1993, 366(6453): 362-5.
2. Newell-Price J, Clark AJ, King P: DNA methylation and silencing of gene expression, Trends Endocrinol Metab, 2000, 11(4): 142-8.
3. Sharp AJ, Stathaki E, Migliavacca E, Brahmachary M, Montgomery SB, Dupre Y, Antonarakis SE: DNA methylation profiles of human active and inactive X chromosomes, Genome Res, 2011, 21(10): 1592–1600.
4. Pickard B, Dean W, Engemann S, Bergmann K, Fuermann M, Jung M, Reis A, Allen N, Reik W, Walter J: Epigenetic targeting in the mouse zygote marks DNA for later methylation: a mechanism for maternal effects in development, Nature, 2001, 103: 35-47.
5. Widschwendter M, Jones PA.: DNA methylation and breast carcinogenesis, Oncogene, 2002, 21(35): 5462-82.
6. Daura-Oller E, Cabre M, Montero M.A, Paternain J.L, Romeu A: Specific gene hypomethylation and cancer: New insights into coding region feature trends, Bioinformation. 2009, 3(8): 340–343.
7. Horvath S: DNA methylation age of human tissues and cell types. Genome Biol 2013, 14(10): R115.
8. Chow J, Yen Z, Ziesche S, Brown C: Silencing of the mammalian X chromosome. Annu Rev Genomics Hum Genet 2005, 6:69–92.
9. Wilson IM, Davies JJ, Weber M, Brown CJ, Alvarez CE, MacAulay C, Schübeler D, Lam WL: Epigenomics: Mapping the Methylome. Cell Cycle 2005,

5(2):155–8.

10. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Gräf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Bäckdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJ, Durbin R, Tavaré S, Beck S: A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol 2008, 26(7):779-85.
11. Chatterjee A, Stockwell PA, Rodger EJ and Morison IM: Comparison of alignment software for genome-wide bisulphite sequence data. Nucleic Acids Research 2012, 40(10): e79.
12. Huh IS, Zeng J, Park T, Yi SV: DNA methylation and transcriptional noise. Epigenetics & Chromatin 2013, 6:9.
13. Bahar R, Hartmann CH, Rodriguez KA, Denny AD, Busuttil RA, Dolle MET, Calder RB, Chisholm GB, Pollock BH, Klein CA, Vijg J: Increased cell-to-cell variation in gene expression in ageing mouse heart. Nature 2006, 441:1011–1014.
14. Novick A, Weiner M: Enzyme induction as an all-or-none phenomenon. Proc Nat Acad Sci USA 1957, 43:553–566.
15. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A: Regulation of noise in the expression of a single gene. Nat Genet 2002, 31:69–73.
16. Raj A, van Oudenaarden A: Nature, nurture, or chance: stochastic gene expression and its consequences. Cell 2008, 135:216–226.
17. Struhl K: Transcriptional noise and the fidelity of initiation by RNA polymerase II. Nat Struct Mol Biol 2007, 14:103–105.

18. Bar-Even A, Paulsson J, Maheshri N, Carmi M, O'Shea E, Pilpel Y, Barkai N: Noise in protein expression scales with natural protein abundance. Nat Genet 2006, 38:636–643.

19. Newman JRS, Ghaemmaghami S, Ihmels J, Breslow DK, Noble M, DeRisi JL, Weissman JS: Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature 2006, 441:840–846.

20. Raser JM, O'Shea EK: Noise in gene expression: origins, consequences, and control. Science 2005, 309:2010–2013.

21. Yin S, Wang P, Deng W, Zheng H, Hu L, Hurst L, Kong X: Dosage compensation on the active X chromosome minimizes transcriptional noise of X-linked genes in mammals. Genome Biol 2009, 10:R74.

22. Dong D, Shao X, Deng N, Zhang Z: Gene expression variations are predictive for stochastic noise. Nucleic Acids Res 2011, 39:403–413.

23. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J, Wei C-L: Dynamic changes in the human methylome during differentiation. Genome Res 2010, 20:320–331.

24. Li Y, Zhu J, Tian G, Li N, Li Q, Ye M, Zheng H, Yu J, Wu H, Sun J, Zhang H, Chen Q, Luo R, Chen M, He Y, Jin X, Zhang Q, Yu C, Zhou G, Sun J, Huang Y, Zheng H, Cao H, Zhou X, Guo S, Hu X, Li X, Kristiansen K, Bolund L, Xu J, Wang W, Yang H, Wang J, Li R, Beck S, Wang J, Zhang X: The DNA methylome of human peripheral blood mononuclear cells. PLoS Biol 2010, 8:e1000533.

25. Lister R, Ecker JR: Finding the fifth base: genome-wide sequencing of cytosine methylation. Genome Res 2009, 19:959–966.

26. Lukk M, Kapushesky M, Nikkila J, Parkinson H, Goncalves A, Huber W, Ukkonen E, Brazma A: A global map of human gene expression. Nat Biotech 2010, 28:322–324.

27. Zeng J, Konopka G, Hunt BG, Preuss TM, Geschwind D, Yi SV: Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. Am J Hum Genet 2012, 91:455–465

28. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR: Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 2009, 462:315–322.

29. Hubbell E, Liu W-M, Mei R: Robust estimators for expression analysis. Bioinformatics 2002, 18:1585–1592.

30. Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003, 19:185–193.

31. Bird A: Gene number, noise reduction and biological complexity. Trends Genet 1995, 11:94–100.

32. Grunau C, Clark SJ, Rosenthal A: Bisulfite genomic sequencing: systematic investigation of critical experimental parameters. Nucleic Acids Res 2001, 29(13):e65.

33. Jones PA: Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 2012, 13:484–492.

34. Jaenisch R, Bird A: Epigenetic regulation of gene expression: how the genome

integrates intrinsic and environmental signals. Nat Genet 2003, 33:245–254.

35. Benjamini Y, Hochberg Y: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B 1995, 57:289–300.

36. Lyko F, Foret S, Wolf S, Falckenhayn C, Maleszka R: The honey bee epigenomes: differential methylation of brain DNA in queens and workers. PLoS Biol 2010, 8:e1000506.

37. Calarco JP, Borges F, Donoghue MT, Van Ex F, Jullien PE, Lopes T, Gardner R, Berger F, Feijo JA, Becker JD, Martienssen RA: Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. Cell 2012, 151(1):194–205.

38. Suzuki MM, Bird A: DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 2008, 9:465–476.

39. Gao F, Liu XS, Wu X-P, Wang X-L, Gong D, Lu H, Song Y, Wang J, Du J, Liu S, Han X, Tang Y, Yang H, Jin Q, Zhang X, Liu M: Differential DNA methylation in discrete developmental stages of the parasitic nematode Trichinella spiralis. Genome Biol 2012, 13:R100.

40. Hunt BG, Glastad K, Yi SV, Goodisman MAD: Patterning and regulatory associations of DNA methylation are mirrored by histone modifications in insects. Genome Biol Evol 2013, 5:591–598.

41. Wang X, Wheeler D, Avery A, Rago A, Choi J-H, Colbourne JK, Clark AG, Werren JH: Function and Evolution of DNA Methylation in Nasonia vitripennis. PLoS Genet 2013, 9(10):e1003872.

42. Herb BR, Wolschin F, Hansen KD, Aryee MJ, Langmead B, Irizarry R, Amdam GV, Feinberg AP: Reversible switching between epigenetic states in honeybee

behavioral subcastes. Nat Neurosci 2012, 15:1371–1373.

43. Zeng J, Nagrajan HK, Yi SV: Fundamental diversity of human CpG islands at multiple biological levels. Epigenetics 2014, 9(4):483–491.

44. ZillerMJ, Gu H, Muller F, Donaghey J, Tsai LTY, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, Gnirke A, Meissner A: Charting a dynamic DNA methylation landscape of the human genome. Nature 2013, 500(7463): 477–481.

45. Gavery MR, Roberts SB: Predominant intragenic methylation is associated with gene expression characteristics in a bivalve mollusc. PeerJ 2013, 1:e215.

46.Vining KJ, Pomraning KR, Wilhelm LJ, Priest HD, Pellegrini M, Mockler TC, Freitag M, Strauss SH: Dynamic DNA cytosine methylation in the Populus trichocarpa genome: tissue-level variation and relationship to gene expression. BMC Genomics 2012, 13:27.

47. Becker C, Hagmann J, Muller J, Koenig D, Stegle O, Borgwardt K, Weigel D: Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature 2011, 480(7376):245–249.

48. Devroye L, Györfi L, Lugosi G: A probabilistic theory of pattern recognition. New York: Springer-Verlab; 1996.

49. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. Cell 2008, 133(3):523–536.

50. Dinh HQ, Dubin M, Sedlazeck FJ, Lettner N, Mittelsten Scheid O, von Haeseler A: Advanced methylome analysis after bisulfite deep sequencing: an example in Arabidopsis. PLoS One 2012, 7(7):e41528.

51. Dempster AP, Laird NM, Rubin DB: Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Ser B 1977, 39(1):1–38.

52. Storey JD: The positive false discovery rate: a bayesian interpretation and the q-value. Ann Stat 2003, 31(6):2013–2035.

53. Foret S, Kucharski R, Pellegrini M, Feng S, Jacobsen SE, Robinson GE, Maleszka R: DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. Proc Natl Acad Sci 2012, 109(13):4968–4973.

54. Li-Byarlay H, Li Y, Stroud H, Feng S, Newman TC, Kaneda M, Hou KK, Worley KC, Elsik CG, Wickline SA, Jacobsen SE, Ma J, Robinson GE: RNA interference knockdown of DNA methyl-transferase 3 affects gene alternative splicing in the honey bee. Proc Natl Acad Sci 2013, 110(31):12750–12755.

55. Xi Y, Li W: BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics 2009, 10(1):232.

56. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE: Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 2008, 452(7184):215–219.

57. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, Haefliger C, Horton R, Howe K, Jackson DK, Kunde J, Koenig C, Liddle J, Niblett D, Otto T, Pettett R, Seemann S, Thompson C, West T, Rogers J, Olek A, Berlin K, Beck S: DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet 2006, 38(12):1378–1385.

58. Bradley AP: The use of the area under the ROC curve in the evaluation of

machine learning algorithms. Pattern Recogn 1997, 30(7):1145–1159.

59. Rakyan, VK; Down, TA; Thorne, NP; Flicek, P; Kulesha, E; Gräf, S; Tomazou, EM; Bäckdahl, L; Johnson, N; Herberth, M; Howe, KL; Jackson, DK; Miretti, MM; Fiegler, H; Marioni, JC; Birney, E; Hubbard, TJ; Carter, NP; Tavaré, S; Beck, S: An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). Genome Research 2008, 18 (9): 1518–29.

60. Mantel N, Haenszel W: Statistical aspect of the analysis of data from retrospective studies of disease, Journal of National Cancer Institute 1959, 22(4): 719-748.

61. Cochran WG: Some methods of Strengthening the common $\chi 2$ tests, Biometrics 1954, 10(4): 417-451.

62. Birch MW: The detection of partial association II:The general case, J R Stat Soc Series B 1965, 27(1): 111-124.

63. Landis JR, Heyman ER, Koch GG: Average partial association in three-way contingency tables: A review and discussion of alternative tests, INT STATIST REV 1978, 46(3): 237-254.

64. Mantel N: Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel Procedure, J AM STAT ASSOC 1963, 58(303): 690-700.

65. Alan A: An Introduction to Categorical Data Analysis 2nd edition, Wiley, 2007.

66. Cleveland WS, Grosse E, Shyu WM: Local regression models. Chapter 8 of Statistical Models in S, Chapman and Hall/CRC, 1991

국 문 초 록

후성유전학(epigenetics)은 DNA의 염기서열이 변화하지 않는 상태에서 이루어지는 유전자 기능의 조절인 후생유전적 유전자 발현 조절을 연구하는 유전학의 하위 학문이다. 후성유전학적인 현상으로는 대표적으로 DNA 메틸화(methylation)와 histone 단백질의 변화가 있다. 이들은 DNA의 염기서열을 직접적으로 변화시키지 않으면서 유전자 발현을 조절한다는 특징이 있다. 이러한 후생유전학적 현상들은, 유전자 각인, 모성 유전, 유전자 침묵화 및 암 발병에 관여한다. 따라서 후성유전학 과정에 대한 연구의 중요성은 날로 늘어가는 추세이다.

이 중에서, DNA 메틸화는 최근 가장 활발히 연구되고 있는 현상이다. 메틸화는 원자나 분자에 메틸기가 붙는 과정을 의미하는 화학적 용어이다. DNA 에서 일어나는 메틸화는, 사이토신이라는 염기에서 작용하고 이는 구아닌, 아데닌, 티민과 함께 염기서열을 구성하는 네가지 원소 중 하나이다. 메틸화가 이루어진 사이토신은 5-메틸사이토신이 되며, 특히 구아닌과 연접한 사이토신 영역을 일컫는 CpG site에 대해서 활발하게 일어나는 것이 알려져 있다.

이러한 메틸레이션이 일어나는 정도를 파악하기 위해서, 다양한 방법들이 개발되어 왔다. 첫번째로 개발된 기술들은 면역침강원리에 기반한 것들이다 (MeDIP). 이 원리를 이용하여, 5-메틸사이토신에 형광물질이 첨가된 항체가 붙은 뒤, 침강된 DNA 서열 조각들이 패널에 마련된 참조 서열조각들과 결합할 때, 이들 형광물질이 내는 빛의 세기를 측정하는 방법이 마이크로어레이 기반의 메틸화 정도 측정법에 사용되어 왔다. (MeDIP-chip) 또한 면역침강 반응과 차세대 염기서열 시퀀싱 (NGS) 기술이 결합하여, 면역침강반응이 일어난 DNA 조각들을 전체 유전체상에 대응시켜 그 조각들의 밀도를 측정하여 메틸화 된 정도를 추

정하는 방법도 개발되었다. (MeDIP-seq). 그러나 이 두 가지의 방법들은 미리 정해진 염기서열단위의 메틸화 정도만을 축정할 수 있거나, (MeDIP-chip) 최소 약 50개정도의 염기서열의 해상도 수준의 정보만을 얻을 수 있기 때문에 (MeDIP-chip, MeDIP-seq) 기술 자체의 한계를 내포하고 있다.

이러한 단점들을 극복하기 위해서, 차세대 시퀀싱 기술과 bisulfite 처리를 결합한 BS-seq이라는 방법이 개발되었다. 이는 메틸화가 안된 사이토신이 bisulfite 처리를 통해서 티민으로 바뀌는 원리를 이용한 것으로서, bisulfite 처리 이후의 DNA 조각들을 전체 유전체 상에 대응시켜, 각 사이토신 위치마다 몇 개씩의 티민과 사이토신을 포함한 DNA조각들이 붙어있는지의 정보를 가지고, 이들 각각이 얼마나 메틸화가 되어 있는지를 파악할 수 있게 되었다. 그리고 각 사이토신에 대한 메틸화의 정보를 얻게 됨으로써, 이를 다루기 위한 통계적 방법개발의 필요성이 대두되었다. 첫번째는 메틸화 판정법으로서, 각 염기서열마다 이 부위가 메틸화가 되어있는지 아닌지를 판정하는 방법이다. 두번째는 다르게 메틸화된 영역을 검출하기 위한 검정 방법 개발이다.

이러한 두가지 주제에 대해서, 우리는 새로운 통계적 검정 방법을 제안하였다. 첫번째로 이진형 메틸화 판정방법에 대해서, 우리는 베이즈 분류기와 주변정보를 동시에 이용한 Bis-Class 라는 방법을 제안하였다. 이 방법은 메틸화가 되어있다는 정보가, 공간적으로 상관관계가 있다는 생물학적으로 알려져 있는 사실을 이용하여 메틸화가 전체적으로 덜되어 있는 종에서 시퀀싱 커버리지가 낮은 영역들에 대해서 전반적인 검출력을 높인 방법이다. 우리는 시뮬레이션과 실제 데이터 분석을 통해서 기존에 제안된 이항분포 검정 및 FDR 을 이용한 방법보다 우수한 분류성능을 보임을 확인하였다. 또한 다르게 메틸화된 영역검출에 대한 통계적 방법으로 우리는 변형된 Cochran-Mantel-Haenzel (CMH) 검정을

제안하였다. CMH 검정은 본래 독립적으로 층화된 자료들에 대해서 조건부 독립을 검정하는 방법으로서, 우리는 여기에 인접한 유전자 염기서열간의 상관성을 포함시켰으며, 시뮬레이션 연구 및 꿀벌의 bisulfite sequencing 자료의 실제 분석을 통해 기존에 사용되었던, 피셔의 정확검정보다 제 1종오류를 좀더 잘 보정하고, 다양한 상황에 적용할 수 있는 검정을 확인하였다.

우리는 차세대 염기서열 시퀀싱 기술에 기반한 BS-seq 자료의 통계적 분석을 위해 새롭게 제안된 두 가지의 방법들이 메틸화와 각종 생물학적 현상들의 연관성 분석에 널리 사용될 것으로 기대하며 그로 인해 각종 생물학적 인과관계를 밝히는 데 도움이 될 것으로 기대한다.

주요어: DNA 메틸화, 차세대 염기서열 시퀀싱 (NGS), bisulfite 처리, 이진형 메틸화 판정법, 다르게 메틸화된 영역의 검정

학 번: 2008-20272