A BAYESIAN FRAMEWORK TO DETECT DIFFERENTIALLY METHYLATED LOCI IN BOTH MEAN AND VARIABILITY WITH NEXT GENERATION SEQUENCING

By

Shuang Li

Submitted to the Faculty of the School of Graduate Studies

of the Georgia Regents University in partial fulfillment

of the Requirements of the Degree of

Doctor of Philosophy

July

2015

A BAYESIAN FRAMEWORK TO DETECT DIFFERENTIALLY METHYLATED LOCI IN BOTH MEAN AND VARIABILITY WITH NEXT GENERATION SEQUENCIN

This thesis is submitted by Shuang Li and has been examined and approved by an appointed committee of the faculty of the School of Graduate Studies of Georgia Regents University.

The signatures which appear below verify the fact that all required changes have been incorporated and that the thesis has received final approval with reference to content, form, and accuracy of presentation.

This dissertation is therefore in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____ _____

Date Major Advisor

_____

Department Chairperson

_____

Dean, School of Graduate Studies

Shuang Li A Bayesian framework to Detect Differentially Methylated Loci in Both Mean and Variability with Next Generation Sequencing (Under the direction of Dr. Hongyan Xu and Dr. Varghese George)

DNA methylation at CpG loci is the best known epigenetic process involved in many complex diseases including cancer. In recent years, next-generation sequencing (NGS) has been widely used to generate genome-wide DNA methylation data. Although substantial evidence indicates that difference in mean methylation proportion between normal and disease is meaningful, it has recently been proposed that it may be important to consider DNA methylation variability underlying common complex disease and cancer. We introduce a robust hierarchical Bayesian framework with a Latent Gaussian model which incorporates both mean and variance to detect differentially methylated loci for NGS data. To identify methylation loci which are associated with disease, we consider Bayesian statistical hypotheses testing for methylation mean and methylation variance using a two-dimensional highest posterior density region. To improve computational efficiency, we use Integrated Nested Laplace Approximation (INLA), which combines Laplace approximations and numerical integration in a very efficient manner for deriving marginal posterior distributions. We performed simulations to compare our proposed method to other alternative methods. The simulation results illustrate that our proposed approach is more powerful in that it detects less false positives and it has true positive rate comparable to the other methods.

INDEX WORDS: DNA methylation, Next Generation sequencingl, Hierarchical Bayesian framework, Latent Gaussian model, Integrated Nested Laplace Approximation, Laplace approximation

ACKNOWLEDGMENTS

First, I would like to thank my major advisor Dr. Hongyan Xu and my co-advisor Dr. Varghese George for agreeing to advise me and for introducing me to the exciting area of Epigenetics. Their guidance, enthusiasm and encourage have been instrumental in proposing, executing, and completing this work.

Second, I would like to acknowledge my committee members: Dr. Duchwan Ryu, Dr. Robert H. Podolsky, Dr. Xiaoling Wang and Dr. Huidong Shi, who graciously agreed to serve on my committee. They have provided helpful feedback and support in my academic research life.

I would also like to thank my family for their patience, support and love. They have made this trip a wonderful experience.

Finally I would like to thank the Graduate school and the department of biostatistics at GRU for providing me the educational and financial opportunity to study the Ph.D. in biostatistics.

Table Of Contents

List of Tables

List of Figures

1 Introduction

1.1 The Biological Problem

DNA methylation at CpG dinucleotides plays important roles in gene expression and cell differentiation. Modifications in DNA methylation induce many complex diseases, including cancer. Of all methylation alterations, hypermethylation, which represses transcription of the promoter regions of tumor suppressor genes, has been studied primarily as an important contributor to gene silencing. In addition, hypomethylation has been recognized as a cause of oncogenesis [1]. Biomedical studies of the connection between methylation modifications and disease will be invaluable for disorder prevention and disease treatment. Extracting information on methylation patterns from biological systems is critical for understanding the biological role of DNA methylation. Recently, various high-throughput platforms based on Next Generation Sequencing (NGS) have been developed and applied to genome-wide DNA methylation analysis. Large amounts of methylation data have been generated from NGS platforms in recent years thanks to the rapid development of sequencing technologies, and several statistical methods have been proposed for detecting differentially methylated loci when comparing the methylation proportion mean between the normal group and the disease group.

Although differences in mean methylation are meaningful criteria for identifying disease risk biomarkers, several biomedical researchers have recently proposed that differences in methylation variability might be important indicators for common complex diseases such as cancer. Feinberg and Irizarry [2] showed that variation in methylation level is associated with important genes involved in normal development of the organism and morphogenesis. Hansen et al. [3] found increased methylation variability in cancer tissues, and pointed out that hypomethylation gene blocks have extremely high methylation variability. Teschendorff and Widschwendter [4] showed that, in some cases, methylation variability is more reliable than mean methylation for identifying cancer risk biomarkers. Therefore, biomedical problems attributable to differential methylation variability should also be monitored.

In order to detect differences in mean and/or variance of methylation proportions between disease group and normal group, researchers may utilize statistical methods for mean and variance

separately, and then combine the results. For example, if the normality assumption is satisfied for the methylation proportion, one may use Student's t test to compare means and the F-test for homoscedasticity to compare variances. Two approaches have been proposed for identifying difference in either methylation mean or methylation variance using microarray data. Ahn and Wang [5] proposed a modified score test that incorporates changes in either mean or variance to detect methylation related markers of disease. Wang et al.[6] proposed a method based on the ratio of the methylation proportion density functions of the two groups to identify differences in either mean or variance. These two methods incorporate the information contained in both the methylation mean and the methylation variance; thus, significant p-values indicate significant differences in mean or variance or both. Although these two methods have proposed for microarray data, they can be applied to NGS data using the estimated methylation proportions.

Most of the currently available statistical methods for NGS data were developed for microarray data, so they rely on estimated methylation proportions instead of NGS counts when they are applied to NGS data. The variation in the estimates is often ignored. To overcome this drawback, we propose to develop a flexible Bayesian framework that includes both methylation mean and methylation variance in a hierarchical model using NGS counts. This method is flexible enough to model the NGS count explicitly, thus accounting for any inter-individual variation in coverage. The proposed method applies the Integrated Nested Laplace Approximation (INLA) to derive the posterior distributions with high accuracy; therefore, it does not suffer from the problems of Markov Chain Monte Carlo sampling (MCMC) in terms of large computational burden and potential lack of convergence. This proposed method will yield a two dimensional highest posterior density (HPD) region for the joint posterior density function of the Bayesian model. The application of a Bayesian approach based on HPD regions should yield high power for testing the difference of methylation levels in mean and/or in variance between groups and avoids the usual assumption between sample mean and variance.

1.2 DNA Methylation

DNA methylation is a biochemical process that involves the covalent addition of a methyl group to the cytosine nucleotides. This process typically occurs at the 5' of cytosine in CpG dinucleotides,

which are regions of DNA where a cytosine nucleotide occurs to a guanine nucleotide. DNA methyltransferases (DNMTs)[7, 8] and the methyl-CpG binding proteins (MBDs) [9–11] work together to maintain DNA methylation patterns. Normally, approximately 60% to 90% of all CpGs are methylated in mammals [12–14]. DNA methylation often occurs in repetitive genomic regions, including microsatellite sequences, which are known to lead to genetic disorders [15–18]. Unmethylated CpGs are often located in the regulatory regions of many genes, including gene promoters [19], the modification of which leads to inappropriate gene expression.

DNA methylation plays an important role in regulating gene expression through various cellular processes including X chromosome inactivation, genomic imprinting and chromosome stability [20–22]. DNA methylation modifications modulate gene expressions that are known to have a large impact in normal cell development. Several DNA methylation modifications involved in developing primordial germ cells and in fertilized oocytes are inheritable. Ultimately, these methylation modifications produce stable alterations in gene expression. DNA methylation is also associated with histone modification, and plays a crucial role in the basis of chromatin structure[12, 23, 24].

Aberrant DNA methylations, both hypermethylation (gain of methylation) and hypomethylation (loss of methylation), have been associated with a large number of human diseases, including cancer [25, 26]. Hypermethylation within the promoter regions is commonly known to silence certain tumor suppressor genes and to affect many cellular processes. This process represses DNA transcription by inhibiting the binding of transcription factors or MBDs [27–30]. In many cancers, hypermethylation occurs in genomic regions with a high frequency of CpG sites (CpG Islands) [31, 32]. In contrast, hypomethylation frequently occurs in the early stages of neoplasm development[25, 33] and is linked to chromosomal instability and loss of imprinting [34]. Furthermore, it has been shown to be prognostic for tumor progression, disease severity, and metastatic potential.

In recent years, the interest in epigenome-wide association studies (EWAS) has increased tremendously due to the rapid development of genotyping technologies [35–38], including several platforms for genome-wide DNA methylation studies. Next Generation Sequencing (NGS) has been commonly used platform for extracting methylation pattern information from biological system, providing huge sequencing capacity, cost-effectiveness, and broad application.

1.3 Next Generation Sequencing

NGS has been a primary technology for extracting methylation information from biological systems since its invention [39]. Its underlying principle is similar to capillary electrophoresis-based (CE) DNA sequencing, which utilizes the identification of a small DNA fragment sequentially from a DNA template strand. NGS extends this process by performing thousands of biochemical reactions in a massively parallel way, rather than being limited to only one or a few DNA fragments. With the rapid development of high throughput sequencing instruments and advancement of modern bioinformatics tools, the production rate for NGS platforms has dramatically increased and the cost has sharply decreased in the past few years [40].

The availability of sequencing library preparation kits obtained through different approaches provides NGS platforms with a broad range of applications in human genomic research in particular, examples include whole genome sequencing, targeted region sequencing and RNA-seq [41]. The ability to generate large volumes of sequencing data rapidly makes the NGS platforms especially powerful for whole-genome sequencing and methylation studies. So far, the Illumina Solexa sequencing platform has been the most widely used NGS platform for whole-genome DNA methylation sequencing analysis. The Illumina sequencing platform produces read length that ranges from 35bp to 75bp [42], and enables considerable genome coverage. For example, the Illumina Genome Analyzer achieves 86% coverage of the 43 million cytosines genomewide [43]. With the development of high throughput sequencing instruments and advancement of modern bioinformatics tools, NGS platforms can produce sequence reads of pooled samples simultaneously in a single run, enabling researchers to save time and money.

The application of these methods to targeted region sequencing focuses on a subset of genes or defined regions in a genome. The availability of the latest NGS library preparation kits allows researchers to perform sequencing on the regions of the genome of greatest interest, rather than on whole-genome sequencing. The ability to generate large amount of sequencing data at a considerable depth of coverage enables NGS platform to identify both common and rare variants. Hence, these applications are particularly useful for discovering genetic variants to fulfill various study objectives.

NGS has also been successfully applied to performing RNA sequencing in recent studies of RNAs. The availability of RNA sequencing workflow, from sample preparation kit to data analysis, enables

researchers to perform thorough investigations of the entire transcriptome, which is the set of all transcripts in a cell, including mRNAs, rRNAs, tRNAs and non-coding RNAs [44]. Understanding transcription is very important in investigating the functional elements of the genome and the molecular constituents of cells and tissues. The availability of faster and cheaper NGS platforms in recent years allows researchers to perform precise transcriptomic analysis, which can be used for monitoring expression of specific genes.

1.4 DNA methylation profiling with NGS

More relevant to the present study, various sequencing methods based on bisulfite conversion have been widely applied to determine genomic methylation patterns [45]. Treatment of DNA with bisulfite converts unmethylated cytosines to uracils, but leaves methylated cytosines intact. Therefore, after bisulfite conversion, methylated and unmethylated cytosines can be distinguished according to DNA sequence, with methylated sites appearing as cytosine and unmethylated sites appearing as thymine (Figure 1).

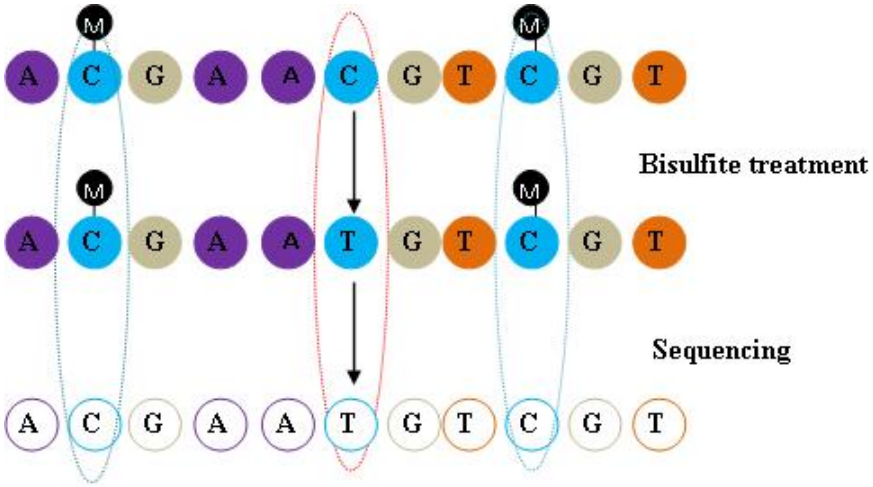

Figure 1: **Example of bisulfite conversion. Bisulfite converts unmethylated cytosine to thymine, but leaves methylated cytosines intact.**

Recently, several high-throughput approaches based on a bisulfite-treated DNA template combined with NGS have been developed to determine methylation status. These methods have the advantages of large production rate, high quality, substantial coverage and reduced cost. With the NGS approaches, information about the methylation status at each CpG site is collected, and the counts of methylated and unmethylated molecules at each CpG site are generated for each individual. Figure 2 shows an example of methylation profiling with NGS data for two samples.

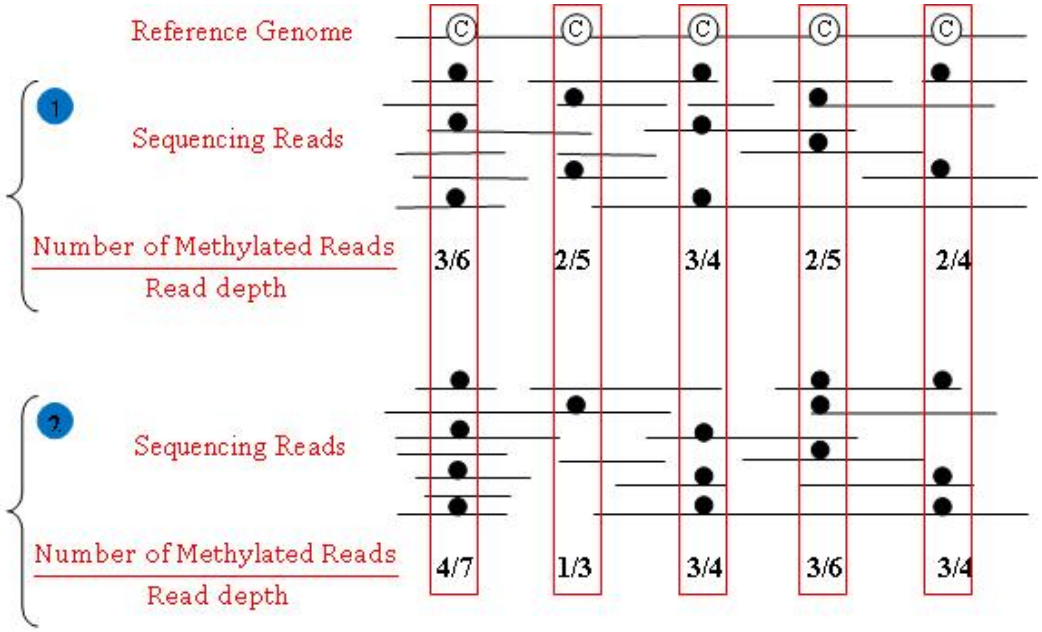

Figure 2: **Example of methylation data frame with next generation sequencing.**

With NGS platforms, DNA methylation measurements are represented by the counts of methylated molecules and the number of times a CpG site is read during the sequencing process, referred to as "read depth" at each CpG dinucleotide for each individual. Suppose we have a two-group design, disease and normal, with N_1 individuals in the disease group and N_2 individuals in the normal group. For the NGS genome-wide methylation data at one CpG site, let n_{ij} be the read depth and let y_{ij} be the count of methylated molecules for individual j in group i at this CpG site. For NGS data, methylation proportions can be estimated using the counts of methylated molecules and the read depths at one CpG site. With our example above, the methylation proportion $\hat{p}_{ij}$ for individual j in group i at the $k^{'}th$ CpG site can be estimated as $\hat{p}_{ij} = y_{ij}/n_{ij}$.

The methylation proportion p is a commonly used measurement of methylation level in genome-wide studies. Several statistical approaches use the estimated methylation proportion $\hat{p}$ as the input in test the methylation difference between the disease group and the normal group. In many cases, a given NGS dataset will not encompass the entire CpG region of a target genome because of the difficulty in sequencing or to mapping certain CpG regions. Therefore, even if the depth of coverage is good in most CpG regions, depth of coverage is not constant across a genome. The estimated p ignores depth of coverage, which loses information. Additionally, estimation of p could be easily affected by factors such as the sampling process, library preparation and batch effects.

1.5 Review of Statistical Methods

1.5.1 Statistical methods for detecting differences in methylation mean

Large amounts of methylation data have been generated from NGS platforms in recent years due to the rapid development of sequencing technologies. Several statistical methods have been proposed to detect differentially methylated loci by identifying difference in methylation mean between normal and disease groups using the high-throughput data. There include 1) Student's t-test; 2) a cluster-based approach that treats the count data as clustered within groups; 3) Bayesian framework approaches and 4) logistic regression-based Wald statistics. These methods will be discussed in detail in the following subsections.

1.5.1.1 Student's t-test

Student's t-test is commonly used to detect differences in mean methylation level using the estimated methylation proportions $\hat{p}$ [5, 46–48]. This approach converts methylation information at each CpG sites to an estimated proportion, thus removing differences associated with unequal read depths among individuals. There are several disadvantages with this approach. First, as the data are sample proportions that range between 0 and 1; thus the normality assumption is likely to be violated, especially with small sample sizes or outliers. Second, the methylation proportion is calculated as the ratio of the count of methylated molecules over the read depth. This approach is different from that used in microarray experiments, and could easily be affected by the experimental process. Student's t-test focus on the difference in methylation means, with

eithrt homogeneity or heterogeneity assumed for the population variances. As such, it offers little information regarding the difference in methylation variances.

1.5.1.2 Cluster-based Rao-Scott chi-square test Xu et al. [49] proposed to use the Rao-Scott chi-square test [50] for differential methylation analysis with NGS data. The key principle of this approach is to treat the NGS reads as clusters within each individual, and to adjust the estimated methylation levels for the effect of clustering. Under this approach, the overall methylation proportion $\hat{\beta}$ for each group is estimated as the ratio of the total methylation count over the total read depth within the group. The overall variance of the estimated proportion for each group is estimated as the mean of the squared difference between the methylation proportion and its expected value, based on the calculated overall methylation proportion for all the individuals within the group. Without clustering, the variances of the sample methylation proportions from a binomial distribution. Therefore the design effects because of clustering is given by the ratio of the variance without clustering and the variance with clustering. The clustering effect is adjusted by dividing the estimated proportion and its variance for each group by the design effect. The adjusted proportions and their adjusted variances are then used to derive a chi-square test statistic with 1 degree of freedom under the null hypothesis of no differential methylation. This approach takes the differences in depth coverage into account; thus decreasing the bias from in estimation of the methylation proportion. However, this method suffers from the fact that it does not account for the effects of covariates. Xu et al. focus on only the differences in methylation mean with their adjusted variance technique. No information provided about the difference in methylation variance.

1.5.1.3 Bayesian Approaches The Bayesian approach provides a flexible framework for modeling the NGS counts and the methylation proportions. This Bayesian framework captures the distribution of the NGS counts and incorporates prior information about the methylation proportions. More importantly, the Bayesian framework is applicable to different null hypotheses of interest, providing the flexibility to identify differences in methylation mean or methylation variance. The underlying principle of the Bayesian approach is Bayes' rule. Suppose that $\boldsymbol{y}^{'} = \{(y_1, \cdots, y_n)\}$ is a vector of n observations whose probability distribution $p(\boldsymbol{y}^{'}|\theta)$ depends on the parameter θ. Suppose also that

θ itself has a probability density function $p(\theta)$. The conditional distribution of θ is

$$p(\theta|\boldsymbol{y}^{'}) = \frac{p(\boldsymbol{y}^{'}|\theta)p(\theta)}{p(\boldsymbol{y}^{'})}$$

The probability density function $p(y|\theta)$ is called the likelihood of θ for given y, which is written as $l(\theta|y)$. According to Bayes' rule, the probability density for θ posterior to the data $\boldsymbol{y}$ is proportional to the product of the distribution of θ prior to the data times the likelihood for θ given y. That is,

$$\textit{posterior distribution} \propto \textit{likelihood} * \textit{prior distribution}$$

The Bayesian approach is widely applied in parameter estimation, model determination and hypothesis testing. In essence, the Bayesian approach treats parameters as random variables and inference is based on the posterior distribution of these random variables.

Markov Chain Monte Carlo (MCMC) commonly used to generate samples from the posterior distributions in Bayesian analysis. Its algorithms offer a sequence of samples, named a Markov chain, whose stationary distribution is the target posterior distribution[1]. Normally, the quality of the Markov chain improves as the number of iterations increases. However, it is difficult to determine how many iterations are needed for a Markov chain to converge to its stationary distribution. It is important to make sure that all random variables are converged in order to obtain valid MCMC posterior samples. In a hierarchical model, such as we are considering here, we have posterior parameter vectors of high dimension. It is not easy to solve these convergence problems even if we choose a very large number of "burn-in" iterations [51].

Bayesian models have been successfully applied in modeling sequencing data; for example, ChIP-Seq data [52] and RNA-Seq data [53]. The Bayesian hierarchical framework offers flexibility in modeling the complex process of generating sequencing counts. Spyrou et al. proposed a statistical algorithm, BayesPeak, with a hidden Markov model (HMM), and modeled the counts with a Gamma-Poisson mixture distribution for ChIP-Seq data. Markov chain Monte Carlo algorithms were used with the posterior distributions to detect enriched locations in the genome. Wu et al. [54]

[1] A Markov chain has stationary transition probabilities if the conditional distribution of X_{n+1} given X_n does not depend on n, where n is the state.

proposed two methods for NGS data, using methyl-seq approach and reduced representation bisulfite sequencing (RRBS) approach; one is based on maximum likelihood estimation and the other is based on Bayesian estimation with a Gamma-Poisson mixed model. They demonstrated that the maximum likelihood method yields biased estimates at extreme methylation levels, while the Bayesian hierarchical model can adjust this bias in a flexible manner. However, Wu et al. only provided a statistical approach for parameters estimation, and did not offer a formal statistical approach for Bayesian hypothesis testing in order to detect differential methylation.

The beta-binomial distribution is commonly used to model NGS counts. When considering the sample methylation proportion at a particular site, the NGS counts are binomial distributed, while the true methylation proportion is assumed to follow a beta prior under the Bayesian framework. McCallum et al. [55] provide a Bayesian framework with a Beta-Binomial hierarchical model to detect differentially methylated loci between the normal group and the disease group. The maximum likelihood estimated hyperparameters were chosen and the posterior distribution of the methylation proportion, defined as $\pi_{\beta|D}$was derived from the combination of the likelihood function and the estimated hypermeters using MCMC. These authors demonstrated that this Bayesian framework approach is more powerful than Fisher's exact test when there are samll samples sites at many sites across the genome and can control the false discovery rate (FDR) very well.

Hardcastle and Kelly [56] developed a Bayesian framework based on the Beta-Binomial distribution and proposed a Wald statistic for identifying differences in methylation means. Feng et al. [57] also proposed a Wald statistic based on the posterior distribution of a beta-binomial hierarchical model to identify difference in methylation means. Both methods made use of maximum likelihood estimation of the parameters of the posterior distribution in their Wald statistics; however, Feng recommeded that methylation variance estimation be based on the dispersion of a beta-binomial distribution. These two methods are based on maximum likelihood estimation of the parameters of the posterior distribution rather than Bayesian sampling; thus, they do not suffer from the disadvantages of MCMC in terms of computational burden and convergence issues. However, these two methods are not currently able to identify differences in methylation variance. Another consideration in the beta-binomial model is the dependent relationship between the mean and the variance of the methylation proportion if a binomial distribution is assumed. A beta-distributed methylation proportion

with a mean close to either 0 or 1 generally has a smaller variance than a methylation proportion distribution with a mean of 0.5. Therefore, if it is of interest to identify differences in methylation variance between the normal and disease groups, it is not available to model the variance using a beta-binomial model.

1.5.1.4 Logistic Regression Approach

Logistic regression is a commonly used approach for detecting differential methylation levels with NGS data [58–61]. This method models the log odds based on the methylation proportion p at each CpG site using a logit transformation with group as the predictor:

$$\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 * X_i,$$

where X_i denotes the treatment group indicator for individual i, $i = 1, 2, \cdot, n$. The differential methylation is determined according to whether the parameter β_1 for the group variable is significantly different from zero. If the null hypothesis is rejected, it implies that the methylation proportions are significantly different between the disease and the control groups. The standard inference procedure for testing this logistic regression coefficient is based on the empirical likelihood function, which eliminates any assumption concerning the distribution of methylation values. As usual, the Wald test statistic for β_1 is based on the maximum likelihood estimate divided by its estimated standard error. Akalin et al. [62] developed the R package methylKit for determining differential methylation across all regions using logistic regression. The use of logistic regression offers great flexibility in the models that can be fit and this method is able to identify any significant difference in methylation proportions. The logistic regression approach accounts for bias due to potential confounders by including them in the model as independent variables; thus, the statistical results are adjusted for the effect of covariates. However, it assumes that the methylation proportion is the same within groups, thus ignoring the within group variation. It also ignores the effect of variations in coverage among individuals.

1.6 Statistical methods for detecting differences in methylation mean or methylation variance

The approaches described in the previous subsection focus on detecting differences in methylation mean or proportion between disease and normal groups. Mostly recently, differences in methylation variance between the disease and the normal groups have been paid more attention. Most researchers choose to apply the methods for identifying differences in methylation mean and methylation variance separately, and then obtain the differentially methylated loci in mean and/or variance from their union or intersection loci. Ahn et al. [5] proposed a joint score-based statistical method for identifying differences in either methylation mean or methylation variance. Chen et al.[6] proposed a semiparametric method focused on testing differences in either methylation mean or methylation variance. Both methods consider the statistical hypotheses $H_0 : \mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$ vs $H_1 : \mu_1 \neq \mu_2$ or $\sigma_1^2 \neq \sigma_2^2$; here, μ_1 and μ_2 are the mean methylation levels for disease and normal, respectively, and σ_1^2 and σ_2^2 are the variances of the methylation level for disease and normal, respectively. Thus, these tests yield significant p-values when either the mean or the variance is significantly different.

1.6.0.5 Ahn's joint score statistics

Ahn and Wang [5] proposed a joint score statistic by combining two score tests for methylation mean and methylation variance respectively. The score statistic for methylation mean was based on the logistic regression model,

$$logit[P(Y_i = 1)] = \alpha + \beta X_i,$$

where Y_i denotes the trait value (1 for diseases and 0 for controls) and X_i denotes the methylation value. These score statistics utilized the information from the first and second derivatives of the likelihood function for above logistic regression model.The score statistic for variability was based on the logistic regression model,

$$logit[P(Y_i = 1)] = \alpha + \beta Z_i,$$

where Y_i denotes the trait value, $Z_i = (X_i - \bar{X})^2$, and X_i denotes the methylation value. Similarly, this score statistic utilized the information from the first and second derivatives of the likelihood function for above logistic regression model.

As opposed to other methods, Ahn's score test for the methylation mean takes into account the second derivative of the likelihood function, which is a measure of the variance, resulting in more power when testing heterogeneity in methylation variability. Similarly, Ahn's score test for methylation variance depends on the methylation mean, which provider more power for detecting difference in means. These score statistics were not used to test methylation mean or methylation variance separately in the simulations in this paper.

The joint score statistic combines the above two score statistics and takes the correlation between X and Z into account via the variance-covariance matrix. The simulation results in Ahn and Wang showed that this joint statistic increased the power for detecting true positively differentially methylated loci when compared to student's t test in the presence of heterogeneity of methylation variability between normal and disease groups. However, direct comparisons of power between the joint score statistics based on both mean and variance vs. Student's t test of the difference in mean are not possible because of their different null hypotheses.

1.6.0.6 Chen's semiparametric test Recently, Chen et al. [6] proposed a semiparametric test based on a generalized exponential model for identifying differentially methylated Loci in both methylation mean and methylation variance with a two-group design. This approach assumes different distributions of the methylation β values, $f(x)$ and $g(x)$, in the disease and normal groups at the one CpG site, respectively. The semiparametric model is assumed as following

$$\frac{g(x)}{f(x)} = \exp\{\alpha + \beta_1 h_1(x) + \beta_2 h_2(x)\}, \tag{1}$$

where $f(x)$ denotes the distribution of methylation β values in the normal group, and $g(x)$ denotes the distribution of methylation β values in the disease group, α denotes the intercept, β_1 and β_1 are the parameters which are introduced to capture the methylation mean and methylation variance, and $h_1(x)$ and $h_2(x)$ are functions specific to the model. For example, if $f(x)$ is the $Normal(\mu_1, \sigma_1^2)$

p.d.f and $g(x)$ is the $Normal(\mu_2, \sigma_2^2)$ p.d.f, the semiparametic model yields

$$\frac{g(x)}{f(x)} = \exp\{\frac{1}{2}\log\frac{\sigma_1^2}{\sigma_2^2} + \frac{1}{2}(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}) + (\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2})x + (\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2})x^2\} \quad (2)$$

In order to test for the equivalence of the methylation mean and methylation variance between disease and control groups, this method tests $H0 : \beta_1 = \beta_2 = 0$ in the model specified by equation (1). For example, under the normal distributional assumptions, the null hypothesis corresponding to equation (2) is $\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2} = \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} = 0$. The Wald test, score test, empirical likelihood ratio test and pseudo likelihood ratio test based on equation (2) were all performed for the appropriate null hypothesis in Chen et al. 's simulation study. There results showed that this method is robust in terms of power when compared to student's t test and regression-based tests in the presence of heterogeneity of methylation variability between normal and disease groups.

Ahn's joint score test and Chen's semiparametric method both incorporate the mean methylation level and the variance methylation level; thus, significant p-values from both methods indicate either significant differences in methylation mean or in methylation variance. Additional approaches should be applied in order to determine whether the detected differentially methylated loci have a difference in mean or in variance after applying these two methods. Ahn's joint score test accounts for the effect of covariates in the model. Both methods suffer from not accounting for the differences in coverage between samples and the variability of the methylation proportion within a group.

1.7 Proposal Overview

In this dissertation, we propose to develop a robust Bayesian framework that incorporates both methylation mean and methylation variance and is designed to detect differentially methylated loci with NGS data. To improve computational time, we will use the Integrated Nested Laplace Approximation (INLA) to compute the posterior densities with high accuracy, and then use credible region based the joint posteriors to perform Bayesian inference. Chapter 1 contains the introduction and a literature review of existing statistical methods. Chapter 2 describes the components of our proposed methodology. The first subsection of Chapter 2 gives a brief review of the INLA method,

the second subsection introduces our proposed Bayesian framework with NGS data, and the last subsection describes the two-dimensional HPD region test. Chapter 3 illustrates the application of our method to both simulated and real data. Finally, Chapter 4 contains the summary and the discussion.

2 Methodology

Our proposed method models NGS counts by using a Bayesian hierarchical framework to identify differences in both methylation mean and methylation variance. This method enables us to model the NGS counts explicitly by accounting for the inter-individual variation in coverage, and incorporates both methylation mean and methylation variance in the model. Thus our proposed method will identify the CpG sites with significant mean differences and/or significant variance differences. We increase computational efficiency by using the INLA to derive posterior distributions. Furthermore, we utilize a two dimensional HPD region for differences in mean and variance for decision making in the Bayesian context. The use of two- dimensional HPD regions enables our proposed approach to test mean and/or variance differences at the same time without requiring the independence assumption. In this chapter, we describe the components of our methodology.

2.1 Review for Markov chain Monte Carlo sampling

Markov chain Monte Carlo (MCMC) sampling is a common approach for obtaining a Bayesian posterior distribution. This method applies a stochastic simulation technique that utilizes Markov chain properties: all of the random variables at the current stage of the model are required to evaluate the next stage samples. Thus, the observed values of a random variable, as well as its related random variables have a strong association between one stage and the next. A model with a large number of random variables requires a more sampling extensive processes, thus it requires extensive computing time. Gibbs sampling [63] is a preferred MCMC method for the hierarchical Bayesian framework. Suppose we want to obtain samples of values of random variables $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$ from their joint probability distribution $p(\theta_1, \cdots, \theta_n)$. Denote the sample for the i_{th} stage as $\boldsymbol{\theta}^i = (\theta_1^i, \cdots, \theta_n^i)$. The sample values of the j_{th} random variable at the i_{th} stage, θ_j^i, can be sampled by applying the Gibbs algorithm to the conditional distribution

$$p(\theta_j^i | \theta_1^i, \cdots, \theta_{j-1}^i, \theta_{j+1}^{i-1}, \cdots, \theta_n^{i-1}). \tag{3}$$

The above process begins with some initial values $\boldsymbol{\theta}^0 = (\theta_1^0, \cdots, \theta_n^0)$. A good choice of initial values decreases the computational time required for the Markov chains to convergence. However,

selection of "good" initial values is difficult in practice. In order to reduce the possibility of bias caused by the choice of initial values, samples from the iterations within the initial phase, the "burn-in" iterations, are usually discarded. Due to the variability in the convergence computational time, this approach has another problem in implementation: determining the number of burn-in iterations. In some cases, the computational time required to reach convergence is hours or days. Additionally, it is necessary to perform multiple statistical analyses to determine a lack of convergence; for example, Gelman-Rubin diagnosis of convergence [64], quantile plots, and autocorrelation plots [65]. However, all of the diagnostics methods are able detect false convergence [66]. Researchers suggested that one use multiple diagnostic methods; thus, the samples that do not lack convergence according to these diagnostics will be reliable [67]. Since NGS includes millions of CpG sites, to check for convergence for each of them makes this method practically impossible. Thus, large computational burden and potential lack of convergence make the application of MCMC sampling with NGS counts inefficient.

2.2 Integrated Nested Laplace approximation (INLA)

To solve the potential issues with MCMC, we propose to use INLA to obtain posterior distributions. INLA, proposed by Rue et al. [68], is an approximation algorithm for deriving a Bayesian posterior distribution. This method models the expectation of a response variable that belongs to an exponential family with a structured additive regression model and introduces a latent Gaussian field that includes all potential Gaussian variables for this model. The posterior marginals for the latent Gaussian field are derived in a closed form using Laplace approximation [69]. Thus, this method does not suffer from the problems of MCMC sampling in terms of large computational time and convergence issues. The following subsections briefly summarize the INLA approach.

2.2.1 Latent Gaussian Models

Suppose the response variable Y belongs to an exponential family with density

$$f_Y(y;\theta,\psi) = exp\{\frac{y\theta - b(\theta)}{a(\psi)} + c(y,\psi)\}, \tag{4}$$

where θ is the "natural" parameter and ψ is the scale parameter. INLA models μ_j, the expectation of y_j, where j denotes the j_{th} individual, using a link funtion $g(\mu_j) = \eta_j$ with the structured additive regression model,

$$\eta_j = \beta_0 + \boldsymbol{z}_j^T \boldsymbol{\beta} + \sum_{m=1} f_m(u_{mj}) + \epsilon_j, \tag{5}$$

where β_0 is the intercept term, β is a p-dimensional vector which represents the linear regression coefficients for the fixed covariate vector $\boldsymbol{z}_j = (z_{1j}, z_{2j}, \cdots, z_{pj})^T$, $f(.)$ denotes unknown functions of the covariates $\boldsymbol{u}$, and the ϵ_j are unstructured error terms. In summary, the structured additive regression model includes three components, the fixed covariates part; the unknown functions part, and the unstructured error term part. A latent vector which includes all of the Gaussian variables, is denoted by $\boldsymbol{x} = \{\beta_0, \boldsymbol{\beta}, f(.), \epsilon_j\}^T$.

INLA is based on the assumption that the prior distribution for any element in the latent Gaussian vector $\boldsymbol{x}$ follows a Gaussian distribution. For the fixed covariates $\boldsymbol{\beta}^* = (\beta_0, \beta_1, \cdots, \beta_p)^T$, INLA assumes that the $\beta' s$ are independent and follow a multivariate Gaussian distribution with a fixed mean vector $\boldsymbol{\mu}_1 = (\mu_{\beta_0}, \mu_{\beta_1}, \cdots, \mu_{\beta_p})^T$ and covariance matrix $\boldsymbol{\Sigma}_1$, that is $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. The choices of the mean vector and the covariance matrix are based on prior knowledge. However, when there is limited prior information, a weakly informative multivariate Gaussian prior is an effective choice. For the unknown functions, $f(.)$, INLA incorporates a nonparametic model, such as a smoothing spline, and assumes their prior distributions are Gaussian distributed; that is, $f(.) \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. For the unstructured error terms $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^T$, INLA assumes that their prior distribution is a multivariate Gaussian distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_3$, that is $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_3)$. If $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ are independent and identically distributed (i.i.d.), then $\boldsymbol{\Sigma}_3 = \sigma^2 \boldsymbol{I}_n$. Let the vector $\boldsymbol{\vartheta}_1$ consist of all of the random variables in $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$. INLA defines these random variables to be hyperparameters that are not required to have Gaussian distributions. The prior distributions for these hyperparameters must be specified.

With the specification of the three components as descried above, the prior for the latent variable vector $\mathbf{x}$ follows a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T, \mathbf{0})^T$ and

covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & & \mathbf{0} \\ & \boldsymbol{\Sigma}_2 & \\ \mathbf{0} & & \boldsymbol{\Sigma}_3 \end{pmatrix}. \tag{6}$$

The covariate matrix $\boldsymbol{\Sigma}_2$ depends on the model chosen for the specific application, and will not be discussed further. When the component of $\boldsymbol{\epsilon}$ are i.i.d., $\boldsymbol{\Sigma}_3 = \sigma^2 \boldsymbol{I}_n$, and σ^2 is the only hyperparameter in $\boldsymbol{\Sigma}_3$. We can write the prior of the latent variable vector as $\boldsymbol{x}|\boldsymbol{\vartheta}_1 \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\vartheta}_1$ is the vector of the hyperparameters in $\boldsymbol{\Sigma}$.

Based on the above discussion, the response variable Y belongs to an exponential family with "natural" parameter θ and scale parameter ψ. INLA models μ, a function of the parameter θ, with a Latent Gaussian vector $\boldsymbol{x}$, which was discussed above. Based on the function between θ and the Latent Gaussian model $\boldsymbol{x}$ with the hyperparameters $\boldsymbol{\vartheta}_1$, the density of the response variable Y can be written as

$$f_Y(y; \theta, \psi) = f_Y(y; \boldsymbol{x}, \boldsymbol{\vartheta}_1, \psi).$$

Here, the hyperparameters $\boldsymbol{\vartheta}_1$ and ψ are independent. Let $\boldsymbol{\vartheta} = \{\boldsymbol{\vartheta}_1, \emptyset\}$; then, $p(\boldsymbol{x}|\boldsymbol{\vartheta}_1) = p(\boldsymbol{x}|\boldsymbol{\vartheta})$. The model specification is completed by assigning a prior density to the hyperparameters $\boldsymbol{\vartheta}$. The joint posterior can then be written as

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{\vartheta}|y) &\propto p(\boldsymbol{\vartheta}) p(\boldsymbol{x}|\boldsymbol{\vartheta}) \textstyle\prod_j p(y_j|\boldsymbol{x}, \boldsymbol{\vartheta}) \\ &\propto p(\boldsymbol{\vartheta}) p(\boldsymbol{x}|\boldsymbol{\vartheta}_1) \textstyle\prod_i p(y_j|\boldsymbol{x}, \boldsymbol{\vartheta}_1, \psi) \\ &\propto p(\boldsymbol{\vartheta}) |\boldsymbol{\Sigma}|^{-1/2} \exp[-1/2(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) + \textstyle\sum \log\{p(y_j|\boldsymbol{x}, \boldsymbol{\vartheta}_1, \psi)\}]. \end{aligned} \tag{7}$$

2.2.2 Integrated Nested Laplace Approximation

The posterior marginal distributions of interest can be written as

$$p(x_k|\boldsymbol{y}) = \int p(x_k|\boldsymbol{\vartheta}, \boldsymbol{y}) p(\boldsymbol{\vartheta}|\boldsymbol{y}) d\boldsymbol{\vartheta}, \tag{8}$$

and

$$p(\vartheta_q|\boldsymbol{y}) = \int p(\boldsymbol{\vartheta}|\boldsymbol{y})d\boldsymbol{\vartheta}_{-q}, \tag{9}$$

where k denotes the k_{th} element in vector $\boldsymbol{x}$ and q denotes the q_{th} elements in vector $\boldsymbol{\vartheta}$; ϑ_{-q} is the vector of ϑ which removes the q_{th} elements. We must now approximate the posterior distributions $p(\boldsymbol{\vartheta}|\boldsymbol{y})$, $p(x_k|\boldsymbol{y})$ and $(\vartheta_q|\boldsymbol{y})$. INLA provides an efficient and accurate way for deriving these posterior approximations in a closed form expression. Therefore, problems of convergence and computational burden that occur with MCMC sampling are non-existent in INLA. INLA combines Laplace approximations and integration in a very efficient manner. Here are the three main steps in using INLA to derive posterior approximations.

1. Approximation of $p(\boldsymbol{\vartheta}|\boldsymbol{y})$ using a Laplace approximation [69]:

$$\tilde{p}(\boldsymbol{\vartheta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\vartheta},\boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{\vartheta},\boldsymbol{y})p(\boldsymbol{x}|\boldsymbol{\vartheta},\boldsymbol{y})}{p(\boldsymbol{y})p(\boldsymbol{x}|\boldsymbol{\vartheta},\boldsymbol{y})} \propto \frac{p(\boldsymbol{x},\boldsymbol{\vartheta},\boldsymbol{y})}{p(\boldsymbol{x}|\boldsymbol{\vartheta},\boldsymbol{y})} \approx \frac{p(\boldsymbol{x},\boldsymbol{\vartheta},\boldsymbol{y})}{\tilde{p}_{LA}(\boldsymbol{x}|\boldsymbol{\vartheta},\boldsymbol{y})}|_{\boldsymbol{x}=x^*(\boldsymbol{\vartheta})} \tag{10}$$

where $\tilde{p}_{LA}(\boldsymbol{x}|\boldsymbol{\vartheta},\boldsymbol{y})$ is the Laplace approximation to $p(\boldsymbol{x}|\boldsymbol{\vartheta},\boldsymbol{y})$, and $\boldsymbol{x} = x^*(\boldsymbol{\vartheta})$ is its mode, which is the solution of $l^{'}(x) = 0$ when $l^{''}(x) < 0$.

2. Let $\boldsymbol{x} = (x_1, \cdots, x_k)^T$. The approximation of $p(\boldsymbol{x}_k|\boldsymbol{\vartheta}, y)$ based on the Laplace approach is,

$$\tilde{p}(x_k|\boldsymbol{\vartheta},\boldsymbol{y}) \propto \frac{p(\boldsymbol{y}|\boldsymbol{\vartheta},\boldsymbol{x})p(\boldsymbol{x}|\boldsymbol{\vartheta})p(\boldsymbol{\vartheta})}{\tilde{p}_G(\boldsymbol{x}_{-k}|x_k,\boldsymbol{\vartheta},\boldsymbol{y})}|\boldsymbol{x}_{-k} = \boldsymbol{x}^*_{-k}(\boldsymbol{\vartheta}), \tag{11}$$

where $\tilde{p}_G(\boldsymbol{x}_{-k}|x_k,\boldsymbol{\vartheta},\boldsymbol{y})$ is the Gaussian approximation to the function $p(\boldsymbol{x}_{-k}|x_k,\boldsymbol{\vartheta},\boldsymbol{y})$, and $\boldsymbol{x}^*_{-k}(\boldsymbol{\vartheta})$ is the mode of the distribution.

3. The last step is to obtain the marginal posteriors using nested approximations,

$$\tilde{p}(x_k|\boldsymbol{y}) = \int \tilde{p}(x_k|\boldsymbol{\vartheta},\boldsymbol{y})\tilde{p}(\boldsymbol{\vartheta}|\boldsymbol{y})d\boldsymbol{\vartheta}, \tag{12}$$

$$\tilde{p}(\vartheta_q|\boldsymbol{y}) = \int \tilde{p}(\boldsymbol{\vartheta}|\boldsymbol{y})d\boldsymbol{\vartheta}_{-q}. \tag{13}$$

INLA is an approximation method and yields a faster and more accurate posterior samples. For example, the distribution plots in Figure 3 show the comparison between the posterior marginals approximated by INLA in tne solid blue line and by MCMC in histograms for 10 samples of parameters in a binomial distribution. The approximations returned by INLA are basically instantaneous, about 0.29 seconds, while MCMC matchs INLA's accuracy after about 118 seconds of computational time using Just Another Gibbs Sampler (JAGS) in R, which is a program for the statistical analysis of Bayesian hierarchical models using MCMC [70]. Additionally, several minutes were spent for the diagnostics to check for convergence with MCMC.

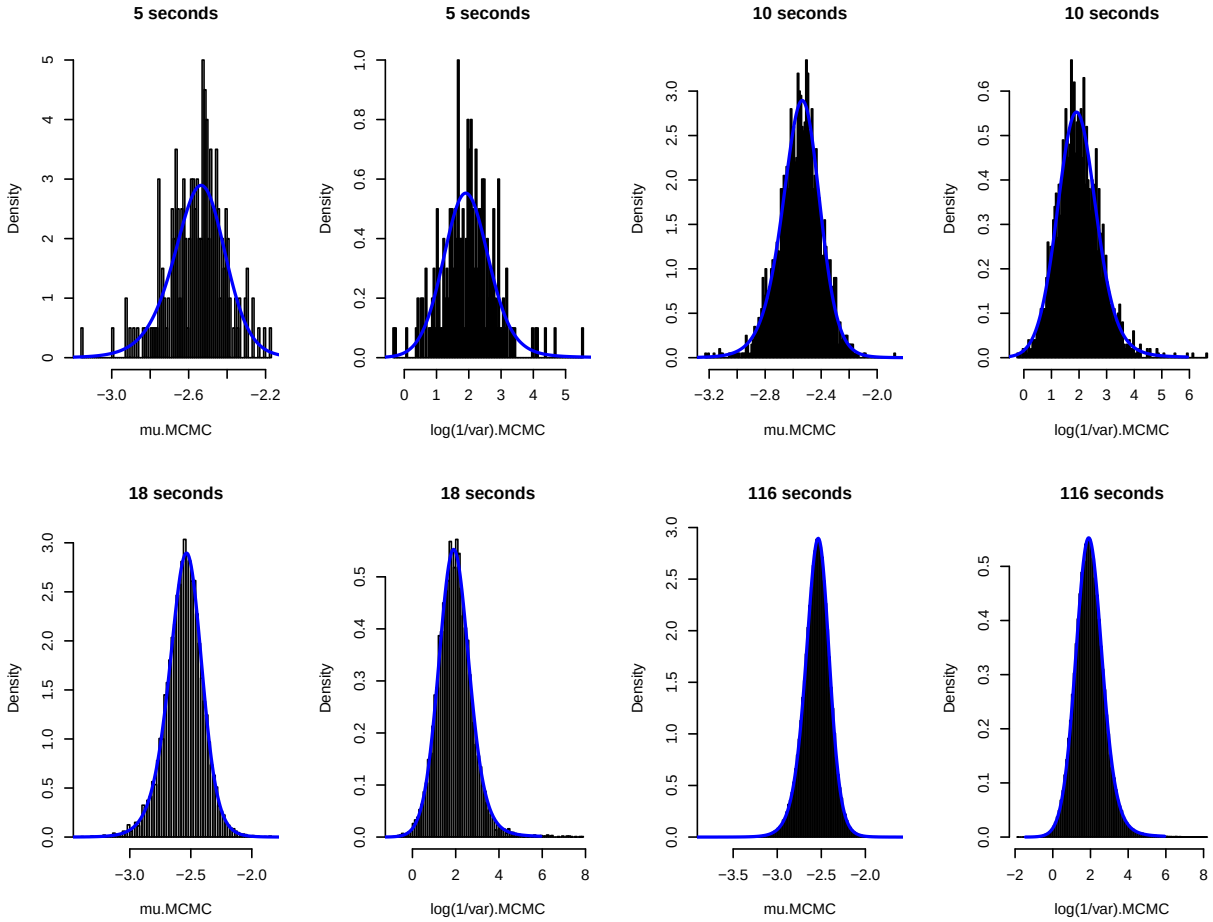

Figure 3: **Comparison of posterior marginals approximated by INLA (solid blue lines) and MCMC (histograms).**

In addition, for large datasets or complex hierarchical models that include several random variables, INLA yields samples from posterior marginals much faster than MCMC. Table 1 composes the run times of MCMC and INLA for posteriors of all of the random variables with different sample sizes. MCMC chains used for the count data have length 100,000 with 10,000 burn-in iterations and have generated using JAGS in R on a desktop PC. The running times range from about 2 minutes to about 20 minutes depending on the data sapmple sizes. For the INLA approach, the results were generated using the R inteface on the same computer in less than 12 seconds for all scenarios.

Table 1: Running time of MCMC sampling and INLA in seconds

	N=10	N=20	N=50	N=100
MCMC	118	302	729	1125
INLA	0.29	3.22	5.40	11.16

2.3 Proposed Bayesian framework with NGS

This section presents the proposed Bayesian framework based on INLA for inference on differential methylation. Suppose we have a two-group design, diseased and normal, with N_1 individuals in the diseased group and N_2 individuals in the normal group. For the NGS genome-wide methylation data at one CpG site, let n_{ij} be the read depth and let y_{ij} be the count of methylated molecules for individual j in group i at this CpG site. Let p_{ij} denote the true methylation proportion for individual j in group i for this CpG site. Using the NGS genome-wide methylation data, we can model y_{ij} with a binomial distribution,

$$y_{ij} \sim Bin(n_{ij}, p_{ij}), \tag{14}$$

where $i = 1, 2; j = 1, \cdots, N_i$.

To identify differential methylation associated with the disease, we consider tests for methylation mean and methylation variability. Based on the binomial distribution in equation (14), a structured

additive regression model with a logit link function $\eta_{ij} = \log(\frac{p_{ij}}{1-p_{ij}})$ can be constructed, that is

$$\eta_{ij} = \log(\frac{p_{ij}}{1-p_{ij}}) = \beta_{i0} + \epsilon_{ij}, \tag{15}$$

where β_{i0} is the overall effect and ϵ_{ij} is the random effect. When covariates such as age and sex are of interest, the model can be extended to:

$$\eta_{ij} = \log(\frac{p_{ij}}{1-p_{ij}}) = \beta_{i0} + \boldsymbol{z}_{ij}^T\boldsymbol{\beta}_i + \epsilon_{ij}, \tag{16}$$

where β_{i0} is the overall effect, $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \cdots, \beta_{ip})^T$ contains the covariate-specific effects, $\boldsymbol{z}_{ij} = (z_{i1j}, z_{i2j}, \cdots, z_{ipj})^T$ is a vector of covariates, and ϵ_{ij} is a Gaussian random effect. Therefore, the expected value of η_{ij} is linked to the linear predictor which accounts for the fixed effects of the covariates, and the variability of η_{ij} is linked to the Gaussian random effects.

2.4 Posteriors of proposed model by INLA

Based on the binomial model and the structured additive regression model without covariates, a latent variable vector that includes all of the Gaussian variables can be constructed as $\boldsymbol{x}_1 = (\beta_{10}, \epsilon_{11}, \cdots, \epsilon_{1n})^T$ for the disease group and $\boldsymbol{x}_2 = (\beta_{20}, \epsilon_{21}, \cdots, \epsilon_{2n})^T$ for the control group. We assume that $\boldsymbol{x}_1$ follows a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_1 = (\mu_{\beta_0}, \mathbf{0})^T$. In our analyses of the simulation data and real data, we choose an informative multivatiate Gaussian distribution with $\boldsymbol{\mu}_1 = (\mathbf{0})^T$ and covariance matrix

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_1^2 \end{bmatrix}$$

with the hyperparameter $\vartheta = \{\sigma_1^2\}$. We chose a logGamma(0.001,0.001) as the prior for σ_1^2. Then the posterior for the disease group, can be written as

$$\begin{aligned} p(\boldsymbol{x}_1, \sigma_1^2|\boldsymbol{y}_1) &\propto p(\sigma_1^2)p(\boldsymbol{x}_1|\sigma_1^2)\prod_j p(y_{1j}|x_1, \sigma_1^2)) \\ &\propto p(\sigma_1^2)|\boldsymbol{\Sigma}_1|^{-1/2}\exp[-1/2\boldsymbol{x}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{x}_1 + \sum \log\{p(y_{1j}|x_1, \sigma_1)^2\}]. \end{aligned} \tag{17}$$

Similarly, $\boldsymbol{x}_2$ is assumed to a multivariate normal distribution with $\boldsymbol{\mu}^2 = (\mathbf{0})^T$ and covariate matrix

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} \sigma_2^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_2^2 \end{bmatrix}$$

with hyperparameter $\vartheta = \{\sigma_2^2\}$. We chose a logGamma(0.001, 0.001) as the prior for σ_2^2. Then the posterior for the disease, can be written as

$$\begin{aligned} p(\boldsymbol{x}_2, \sigma_2^2|\boldsymbol{y}_2) &\propto p(\sigma_2^2)p(\boldsymbol{x}_2|\sigma_2^2)\prod_j p(y_{2j}|x_2, \sigma_2^2)) \\ &\propto p(\sigma_2^2)|\boldsymbol{\Sigma}_2|^{-1/2}\exp[-1/2\boldsymbol{x}_2^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{x}_2 + \sum \log\{p(y_{2j}|x_2, \sigma_2)^2\}]. \end{aligned} \tag{18}$$

We get very similar posteriors using MCMC and INLA under the same prior distributions when MCMC converges. For example, when prior of $\sigma^2 \sim logGamma(0.001, 0.001)$ and the two posteriors are almost identical Figure 4. The runing time of MCMC for this example is about 2 minutes, where the INLA approach required about 0.4 seconds.

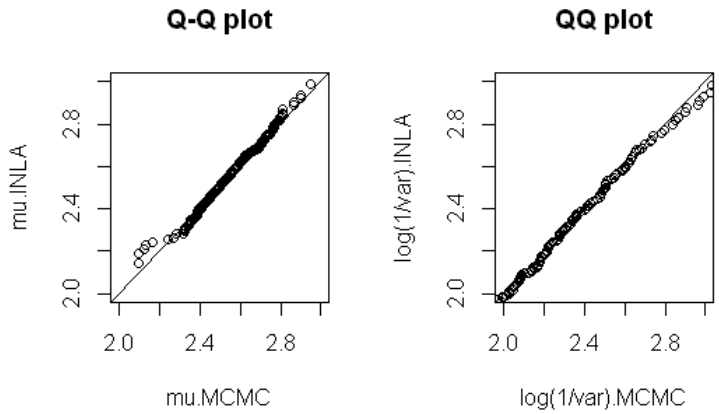

Figure 4: **The left Q-Q plot is for μ by INLA and MCMC; the right Q-Q plot is for $1/\log(\sigma^2)$ by INLA and MCMC.**

2.5 Bayesian decision making using the highest posterior density (HPD) region

2.5.1 Bayesian HPD region

Box and Tiao 2011 [71] introduced the concept of the highest posterior density (HPD) credible interval. In the case of a one dimensional parameter θ, a HPD interval is a credible interval for θ with $100(1-\alpha)\%$ Bayesian coverage, that is $p[\theta_L \leq \theta \leq \theta_U|D] = 1-\alpha$, where θ_L and θ_U are the lower and upper limits, respectively, and D indicates the data set. A HPD interval is similar to a confidence interval in that it specifies a range of possible values for the parameter. The values between θ_L and θ_U have the $100(1-\alpha)\%$ highest posterior density; this indicates that $100(1-\alpha)\%$ of the posterior parameter space will be in this range.

For a higher (≥ 2) dimensional parameter $\boldsymbol{\theta}$, Box and Tiao extended the idea of a HPD interval to a HPD region. A $100(1-\alpha)\%$ HPD region of $\boldsymbol{\theta}$ can be defined as

$$p\{\boldsymbol{\theta} \in Region|D\} = 1-\alpha. \tag{19}$$

Definition 1. *Let $p(\theta|D)$ be a posterior density function, and Θ be the parameter space. A* $100(1-$

α)% HPD region for θ is a subset $\Theta_c \in \Theta$ defined by

$$\Theta_c = \{\theta : p(\theta|D) \geq k\}, \tag{20}$$

where k is the largest number such that

$$\int_{\theta:p(\theta|D)\geq k} p(\theta|D)d\theta = 1 - \alpha. \tag{21}$$

The value k can be thought of as a horizontal plane over the posterior density whose intersections with the posterior define regions of probability $1 - \alpha$. Any value included in the HPD region has a larger probability than any value outside the HPD region. Some properties of the HPD region are provided in the following remarks.

Remark 1. *For a given probability $1 - \alpha$, the HPD region is the smallest possible area in the parameter space θ.*

Remark 2. *If Θ_c is a HPD region of size $(1 - \alpha)$, for any point $\theta_1 \in \Theta_c$ and any point $\theta_2 \notin \Theta_c$, $p(\theta_1|D) \geq p(\theta_2|D)$.*

Remark 3. *If $\emptyset = f(\theta)$ defines a one-to-one transformation from θ to $\emptyset$, any region of size $1 - \alpha$ in the space of θ transforms into a region of the same size in the space of $\emptyset$, but the HPD region for θ will not transform into an HPD region of θ, unless the transformation is linear.*

In 1973, Box and Tiao also introduced a Bayesian-decision making approach using the Bayesian HPD region; namely, the hypothesis $H_0 : \theta = \theta_0$ can be tested by checking whether θ_0 lies inside a HPD region of content $(1 - \alpha)$ or not. This depends only on the posterior distribution $p(\theta|D)$ regardless of which prior we chose. It follows that the null hypothesis parameter value θ_0 is covered by the $100(1 - \alpha)\%$ HPD interval if and only if

$$Pr\{p(\theta|D) > p(\theta_0|D)\} \leq 1 - \alpha. \tag{22}$$

From the properties of HPD regions, we accept $H_0 : \theta = \theta_0$ if θ_0 lies inside a HPD region. This

is equivalent to the event that $p(\theta_0|D) \geq k$, where k is a particular chosen positive constant from equation (21). We reject $H_0 : \theta = \theta_0$ if θ_0 lies outside the HPD region. This means that the hypothesized parameter value θ_0 has a smaller probability of being covered by the HPD region of size $1 - \alpha$, with $p(\theta_0|D) < k$.

2.6 Proposed Bayesian decision making approach based on Bayesian HPD region with NGS counts

Our proposed approach is a robust Bayesian framework that incorporates both mean and variance to detect differentially methylated loci using NGS counts. According to the hierarchical Bayesian model for NGS counts in equation (15), the random sampling of the parameter vector $\{\mu_1, \sigma_1^2, \mu_2, \sigma_2^2\}$, where μ_1 and σ_1^2 are the logit transformed methylation proportion mean and variance in the disease group, respectively, μ_2 and σ_2^2 are the logit transformed methylation proportion mean and variance in the control group, respectively, can be accomplished using the INLA approach. To detect differentially methylated loci with the two group design, we consider the difference of means $\Delta_\mu = \mu_1 - \mu_2$ and the difference of log(variances) $\Delta_{\sigma^2} = log(\sigma_1^2/\sigma_2^2) = log(\sigma_1^2) - log(\sigma_2^2)$ using the Bayesian approach. For the two-dimensional parameter space $\theta = (\Delta_\mu, \Delta_{\sigma^2})$, the $100(1-\alpha)\%$ HPD regions for $\theta = (\Delta_\mu, \Delta_{\sigma^2})$ can be defined by

$$\Theta_c = \{\theta : p(\theta|D) \geq k\},$$

where k is the largest value such that

$$\int_{(\Delta_\mu, \Delta_{\sigma^2}):p(\Delta_\mu, \Delta_{\sigma^2}|D)\geq k} p(\Delta_\mu, \Delta_{\sigma^2}|D)d(\Delta_\mu, \Delta_{\sigma^2}) = 1 - \alpha. \tag{23}$$

The $100(1-\alpha)\%$ highest posterior density region for $\theta = (\Delta_\mu, \Delta_{\sigma^2})$ is an area inside a contour line with connecting points where $p(\Delta_\mu, \Delta_{\sigma^2}|D)$ has the same value k as in equation (23). Since $p(\Delta_\mu, \Delta_{\sigma^2}|D)$ is a function of two variables, a plot of the posterior distribution involves a two-dimensional figure. The $100(1-\alpha)\%$ highest posterior density region can be obtained by plotting probability density contours in the $(\Delta_\mu, \Delta_{\sigma^2})$ plane based on equation (23).

The key principle behind using the joint posterior HPD region for testing the difference in means and variances is that the samples inside the contour line have larger probability than the samples outside the contour line. Using the joint posterior HPD area of $(\Delta_\mu, \Delta_{\sigma^2})$ to test the difference in means and the difference in variances offers four distinct conclusions described in Table 2. The first conclusion is that both the mean and variance are not significant when the line $\Delta_\mu = 0$ and the line $\Delta_{\sigma^2} = 1$ intersect with the HPD contour (Figure 5). The second conclusion is that the difference in means is not significant but the difference in variances is significant, when the line $\Delta_\mu = 0$ intersects with the HPD contour lines only (Figure 6). The third conclusion is that the difference in means is significant but the differences in variance in not-significant, when the line $\Delta_{\sigma^2} = 1$ intersects with the HPD contour lines only (Figure 7). The fourth conclusion is that both the mean and variance differences are significant; this occurs when neither the line $\Delta_\mu = 0$ nor the line $\Delta_{\sigma^2} = 1$ intersect with the HPD contour lines (Figure 8).

Table 2: Four distinct conclusions.

Both the difference in mean and variance are not significant.
The difference in variance is significant, but the difference in mean is not significant.
The difference in mean is significant, but the difference in variance is not significant.
The difference in both the mean and the variance are significant.

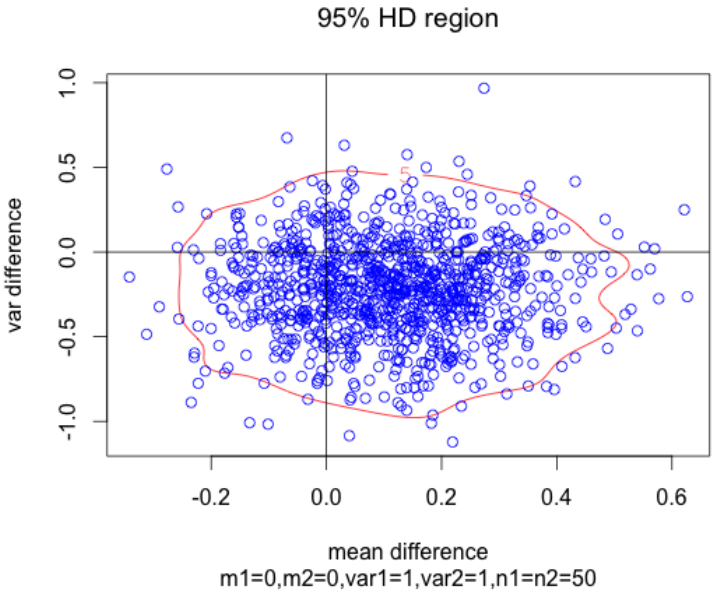

Figure 5: **Both the difference in mean and variance are not significant.**

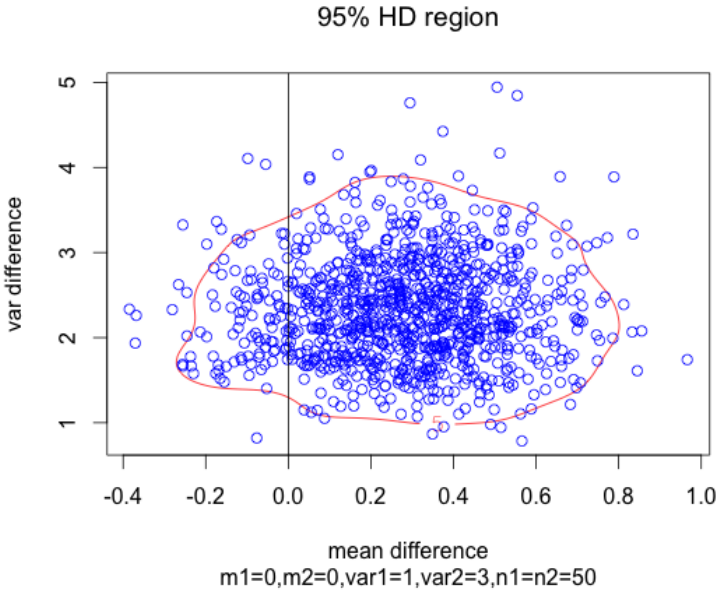

Figure 6: **The difference in variance is significant, but the difference in mean is not significant .**

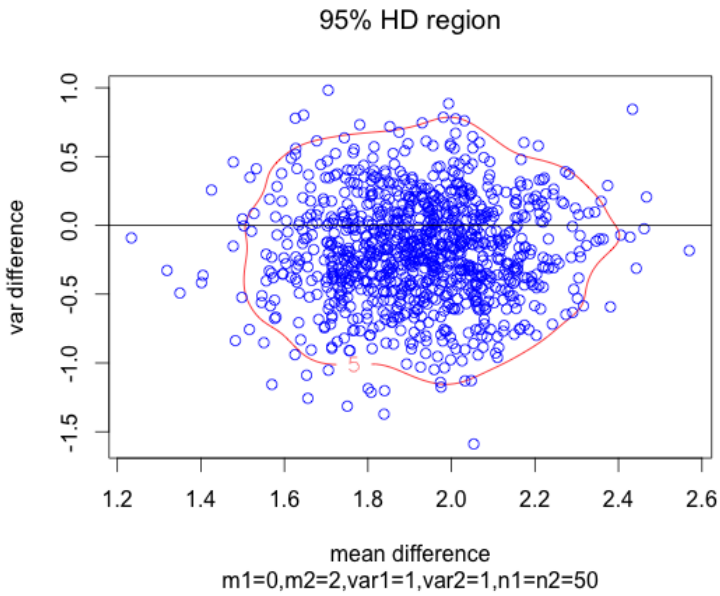

Figure 7: **The difference in mean is significant, but the difference in variance is not significant.**

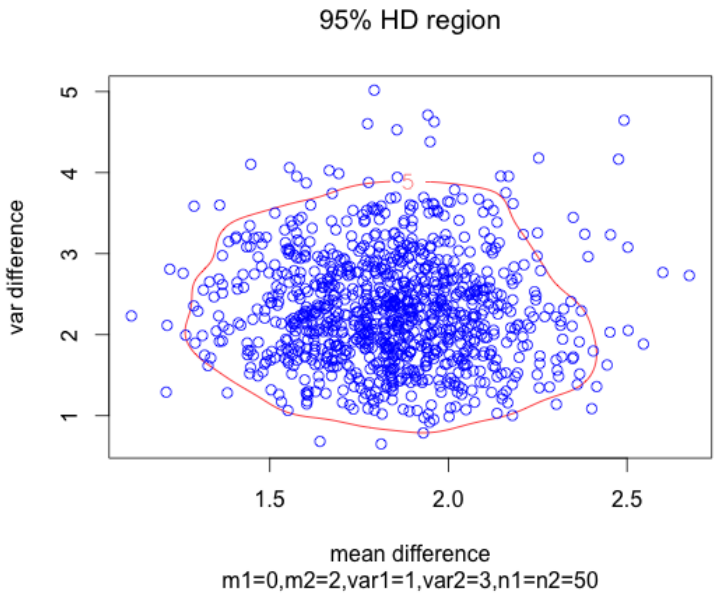

Figure 8: **The difference in both the mean and the variance are significant.**

3 Results

3.1 Simulation Studies

The simulation studies aim to compare the statistical properties of the methods in detecting differentially methylated loci in methylation mean and/or in methylation variance with our randomly generated NGS counts between the normal group and the disease group. The simulations made comparisons between our proposed approach with Rao-Scott chi-square test, student's t test and logistic regression based Wald statistics in identifying difference in mean methylation, and with Ahn's joint score statistics and Chen's semiparametric approach in identifying difference in either methylation mean or methylation variance. The comparisons have been made in terms of true positive rate and false positive rate for detecting differentially methylated loci by each method. Here the true positive rate is the probability of the truly differentially methylated loci, given the simulation data with significant difference in either methylation means or methylation variances, and the false positive rate is the probability of the falsely differentially methylated loci, given the simulation data with no significant difference in both methylation means and methylation variances. For the frequentist approaches, a gene is considered differentially expressed when its p-value is less than the nominated significant level (α). While for our proposed Bayesian approach, a gene is considered differentially expressed in mean when the line of $\Delta\mu = 0$ is excluded from the $(1-\alpha)\%$ HPD contour lines, and a gene is considered differentially expressed in variance when the line of $\Delta\sigma^2 = 1$ is excluded from the $(1-\alpha)\%$ HPD contour lines.

3.2 Data Generation

The structured additive regression model in our hierarchical Bayesian framework is based on a logit link function for the methylation proportion. Therefore, the first step of our data generation is to randomly generate the logit transformed methylation proportion, $logit(p)$ from a normal distribution $N(\mu, \sigma^2)$. The second step is to generate methylation proportion, p through inverse-logit transformation. The last step is to randomly generate the NGS counts from a binomial distribution. The simulation studies also considered the effects of various parameters in the data generation: (1) the effects of sample size; (2) mean of the logit transformed methylation proportion; and (3) variance

of the logit transformed methylation proportion. The read depth n_{ij} in both normal and disease were generated from a censored normal distribution with mean of 30 and variance of 13, and the minimum value is 5 [49]. The following subsections describe the simulation parameters.

3.2.1 Sample size

Sample size for the studies is defined as the number of individuals in the disease group or the normal group. Equal sample sizes were assigned for both groups in the simulations. To examine the effects of the sample size in the simulation studies, various sample sizes have been considered. The sample sizes used in the simulations were: 10, 20, 50, 100, 200 and 500.

3.2.2 Values of logit transformed methylation proportion mean

Methylation proportion is defined to be the ratio of the methylated molecules over the read depth at each CpG site. It follows that the value of methylation proportion varies between 0 and 1. We generated the logit transformed methylation proportions from a normal distribution with mean μ and variance σ^2 first, and then generated the methylation proportion through the inverse-logit transformation. Due to the nature of the logit function, the value of logit transformed methylation proportion ranges from $-\infty$ to $+\infty$. To simulate the value of μ from this infinite range is impossible. An efficient way to solve this problem is to choose μ's value within a small range, with [-4, +4] that includes a large proportion of its values. The values of logit transformed methylation mean μ used in the simulation studies are: -4, -2, -0.5, 0, 0.5, 2 and 4. When we choose the variance as 1, the expectation values of these inverse-logit transformed methylation proportion are 0.0281, 0.1555, 0.3982, 0.5, 0.6021, 0.8445 and 0.9719 respectively.

3.2.3 Values of logit transformed methylation proportion variance

Our simulation studies generated the logit transformed methylation proportion from a normal distribution with variance which has a theoretical range from 0 to $+\infty$. There is limited information on the variance value from literature. This wide range increases the difficulty of choosing the variance value efficiently. To solve this problem we selected the values of variance based on the selected means because the variance of the data is related to the mean of the data. The selection criterion

is to choose the values of variance which are not far away from the values of means. Based on the selected means in the simulations that are, 0, 2, 4, the values of logit transformed methylation variance that will be used in the simulation studies are: 1, 2, 3 and 4. When we choose the mean as 0, the variance values of these inverse-logit transformed methylation proportion are 0.0434, 0.0688, 0.0857 and 0.0986 respectively.

3.2.4 Histograms of the generated methylation proportion

We plot histograms to show the distributions of the simulated logit transformed methylation proportions as well as the distributions of the methylation proportions without logit transformation, which is obtained from inverse-logit transformation, that is $logit^{-1}(p^*) = \frac{exp(p^*)}{1+exp(p^*)}$, where p^* represents logit transformed methylation proportion. Figure 9 panels A-D represent histograms of simulation scenarios with logit transformed methylation proportion from normal distribution with mean zero and variance 1,2,3 or 4 respectively. In these simulation studies, the corresponding mean of methylation proportions $E(p) = 0.5$ and the variance is 0.0434, 0.0684, 0.0857 or 0.0986 respectively. The histograms of methylation proportion suggest that their distributions are symmetric when the mean of the logit transformed methylation proportions equals zero. The histograms further indicate that these distributions are different with different variance of logit transformed methylation proportions. When the variance equals to 1, the distribution is unimodal and looks similar to the $Beta$ density, however when the variance is large to 3 or 4, the distribution is bimodal.

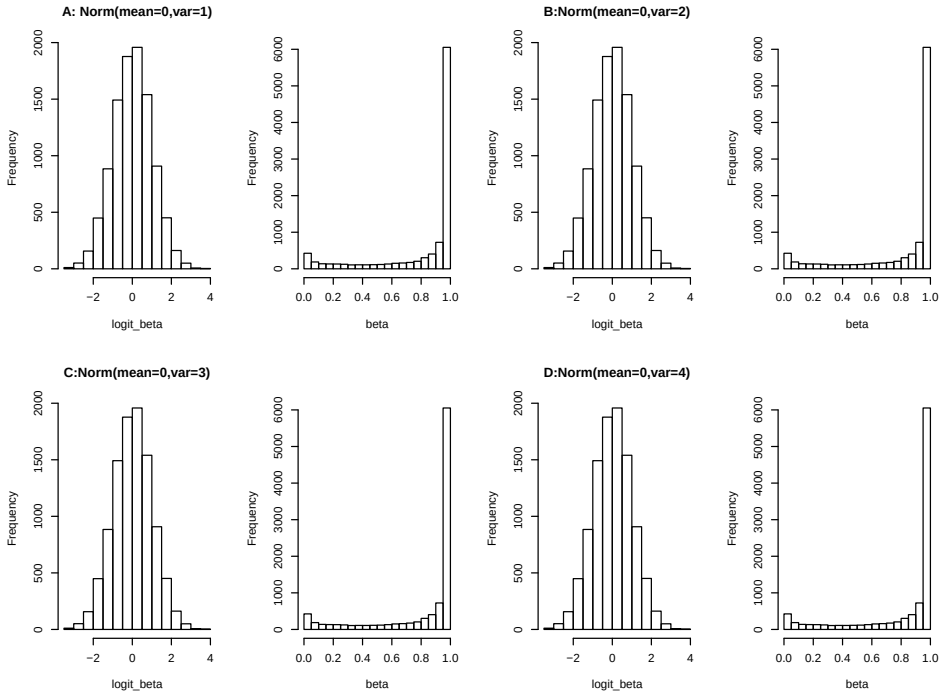

Figure 9: Histograms for the simulated logit transformed methylation proportions with normal distribution of mean zero, variance 1, 2, 3 or 4 and histograms of their inverse-logit methylation proportions.

Figure 10 A-D represent histograms of simulated logit transformed methylation proportions which follow a normal distribution with mean of 2 and variance of 1,2,3, or 4 respectively, as well as histograms of the original methylation proportion after inverse-logit transformation. These histograms show that the methylation proportions have right-skewed distributions. The mean of methylation proportion based on the generated logit methylation proportion is not constant, which decreases with the increasing variance in table 3 . With larger variance, the distribution is more likely to be skewed toward to one. In contrast, Figure 11 A-D represent histograms of simulated logit transformed methylation proportion as well as the histograms of their inverse-logit transformed methylation proportion under the simulation scenarios of mean -2 and variance 1,2,3 or 4 respectively. The histograms of methylation proportions show that their distribution are skewed to left with a bump near zero. In these simulation scenarios, the mean of methylation proportion increases with the increasing variance. With larger variance, the distribution is more likely to be skewed toward

zero. Table 3 shows the mean of methylation proportion under different simulation scenarios.

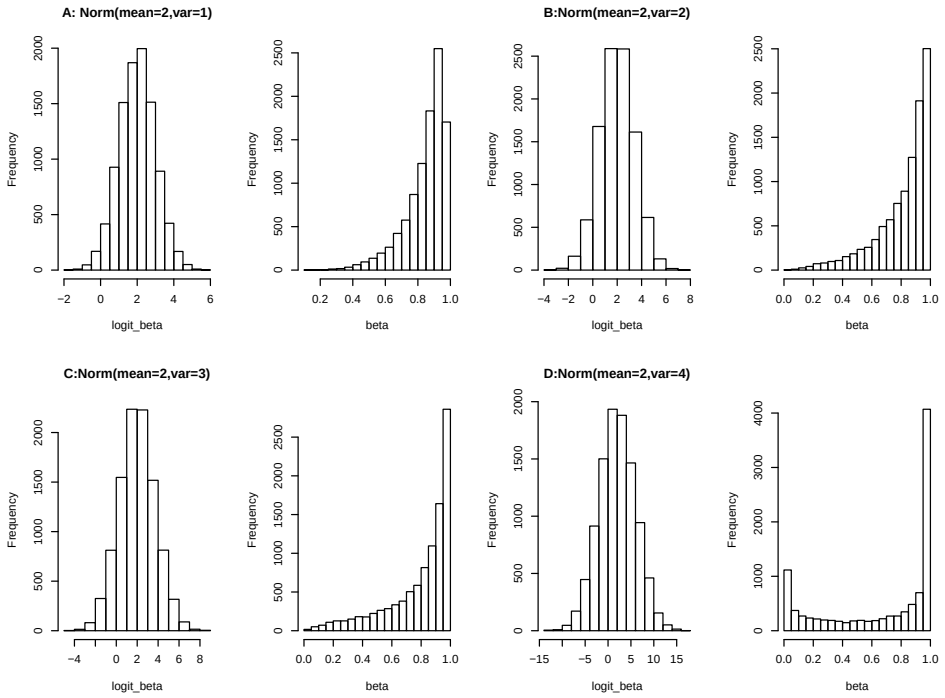

Figure 10: Histograms for the simulated logit transformed methylation proportions with normal distribution mean 2 and their inverse-logit methylation proportions.

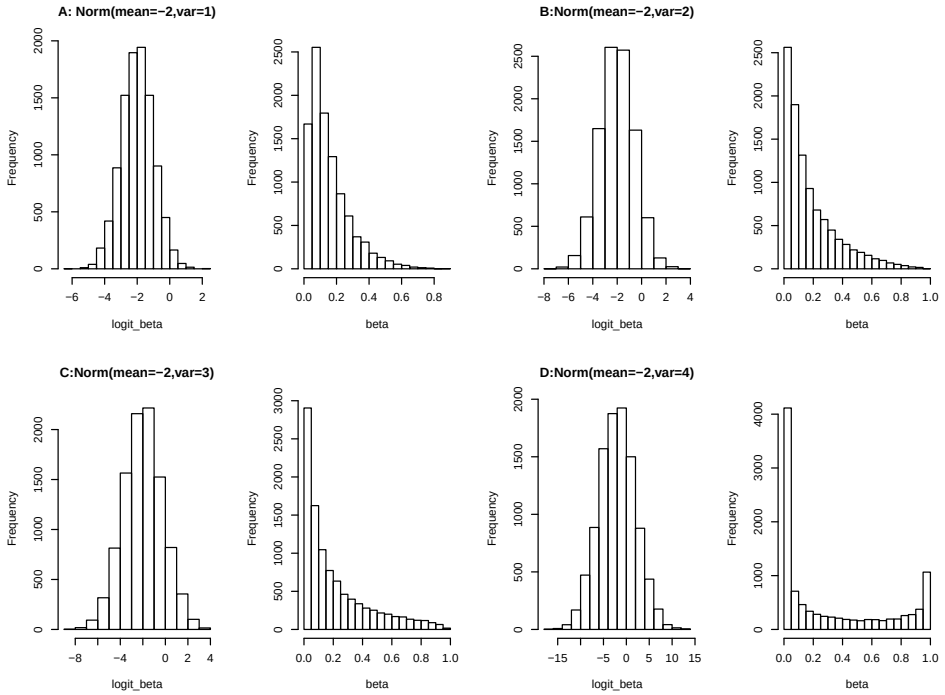

Figure 11: Histograms for the simulated logit transformed methylation proportions with normal distribution mean -2 and their inverse-logit methylation proportions.

Table 3: Mean and variance of inverse-logit transformed methylation proportions.

$logit(p)$	p		$logit(p)$	p	
	Mean	Variance		Mean	Variance
N(0.5, 1)	0.6020	0.0406	N(-0.5, 1)	0.3982	0.0406
N(0.5, 2)	0.5753	0.0959	N(-0.5, 2)	0.4247	0.0959
N(0.5, 3)	0.5574	0.1329	N(-0.5, 3)	0.4427	0.1329
N(0.5, 4)	0.5454	0.1574	N(-0.5, 4)	0.4546	0.1574
N(2, 1)	0.8447	0.0155	N(-2, 1)	0.1554	0.0155
N(2, 2)	0.8161	0.0330	N(-2, 2)	0.1838	0.0330
N(2, 3)	0.7938	0.0484	N(-2, 3)	0.2065	0.0484
N(2, 4)	0.7751	0.0619	N(-2, 4)	0.2246	0.0619
N(4, 1)	0.9719	0.0010	N(-4, 1)	0.0028	0.0010
N(4, 2)	0.9594	0.0042	N(-4, 2)	0.0407	0.0042
N(4, 3)	0.9459	0.0094	N(-4, 3)	0.0542	0.0094
N(4, 4)	0.9322	0.0158	N(-4, 4)	0.0675	0.0158
N(0, 1)	0.5	0.0434			
N(0, 2)	0.5	0.0688			
N(0, 3)	0.5	0.0857			
N(0, 4)	0.5	0.0986			

We further plotted histograms of simulated logit transformed methylation proportion as well as the histograms of methylation proportion with simulation scenarios of mean 4 in Figure 12 and mean -4 in 13. Figure 12 suggest that the distribution of methylation proportions are strongly skewed toward one. In contrast, Figure 13 suggest that the distribution of methylation proportions are strongly skewed toward zero. The mean and variance of the logit-transformed and original methylation proportions are represented in table 3.

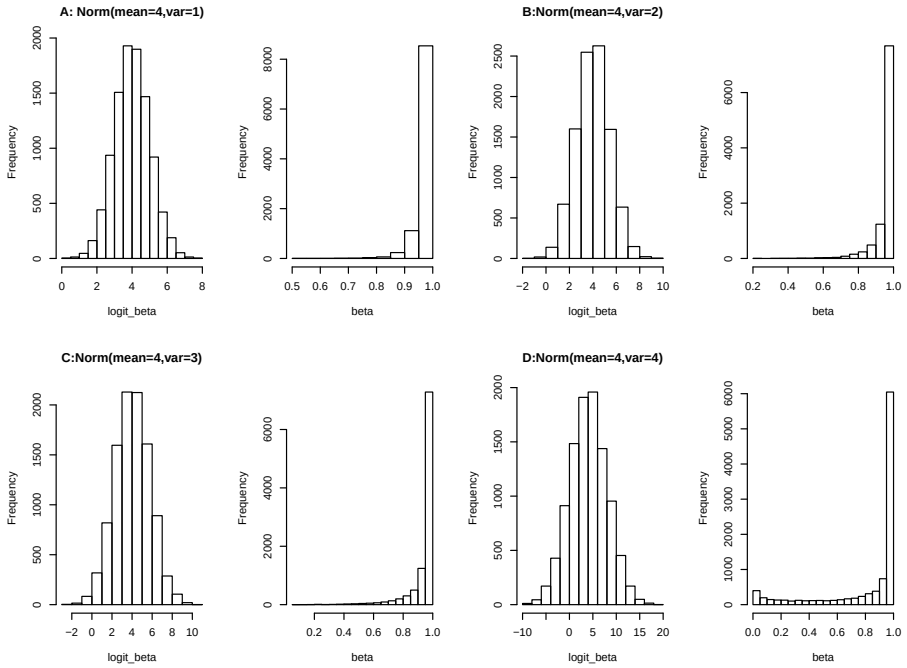

Figure 12: Histograms for the simulated logit transformed methylation proportions with normal distribution mean 4 and their inverse-logit methylation proportions.

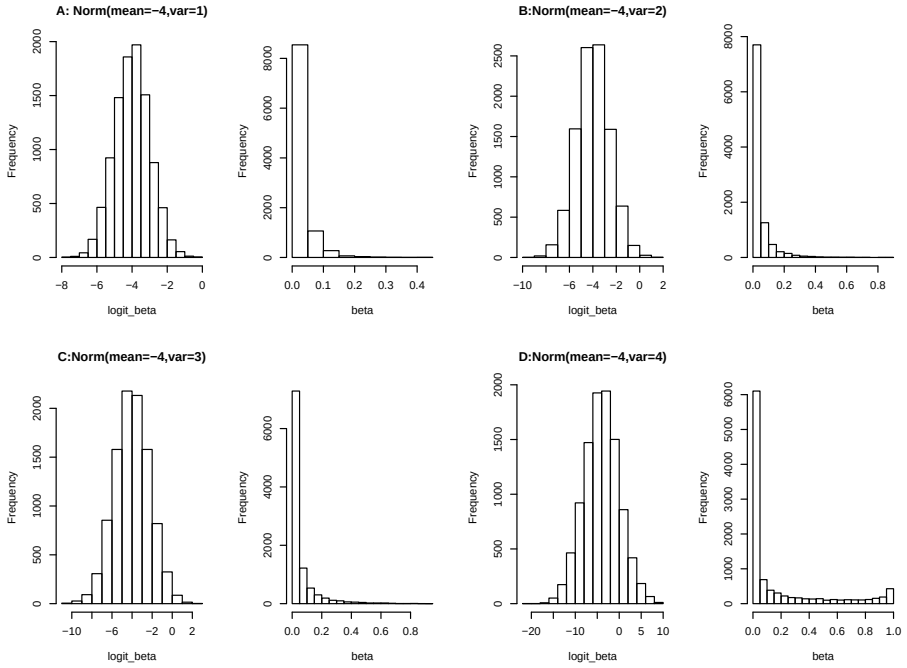

Figure 13: Histograms for the simulated logit transformed methylation proportions with normal distribution mean -4 and their inverse-logit methylation proportions.

The simulations use combinations of different values of parameters including the sample sizes, the mean and the variance of the logit transformed methylation proportion. No matter which combination of simulation parameters we chosen, the methylation proportions are always between 0 and 1. In order to ensure quality estimations, 5,000 replicates will be generated for each combination of simulation parameters.

3.3 Simulation Results

3.3.1 False positive rate

This section describes the results for identifying differentially methylated loci under the simulation scenarios of equal mean and equal variance of logit transformed methylation proportion. In these simulation scenarios, the simulated logit transformed methylation proportions have identical distributions with both mean and variance between diseases and controls. Therefore the mean and the variance of methylation proportions after inverse-logit transformation are also identical between two

groups. We performed 5,000 replicates for each sample size with significant level α equals to 5% or 1%. Among these 5,000 replicates, the number of replicates with significant differentially methylated loci were recorded. The comparison in these simulation scenarios is the false positive rate, which is the proportion of the number of significant replicates over the 5,000 replicates. The methods in the comparison include Student t test, Rao Scott statistics, Ahn's score statistics, logistic regression based Wald statistics and Chen's method.

Figure 14 shows false positive rates for the six approaches at 5% or 1% significance level when logit transformed methylation proportions were generated from a normal distribution with mean 0 and variance 1. The mean of the methylation proportion is 0.5 and its variance is 0.0433. The distribution plots of the logit transformed methylation proportion and its methylation proportion are showed in Figure panels 9. Figure 14 shows a very high false positive rate (around 0.5) for logistic regression based Wald statistics for all of the sample sizes. Actually, the high level false positive rate pattern appears in all of the simulation scenarios with the assumption of no difference in neither methylation mean nor methylation variance. This pattern suggests that logistic regression based Wald statistics detects large number of significant methylated CpG sites under no difference assumptions. This method will be excluded in the following figures so that the results of other methods can be shown more clearly.

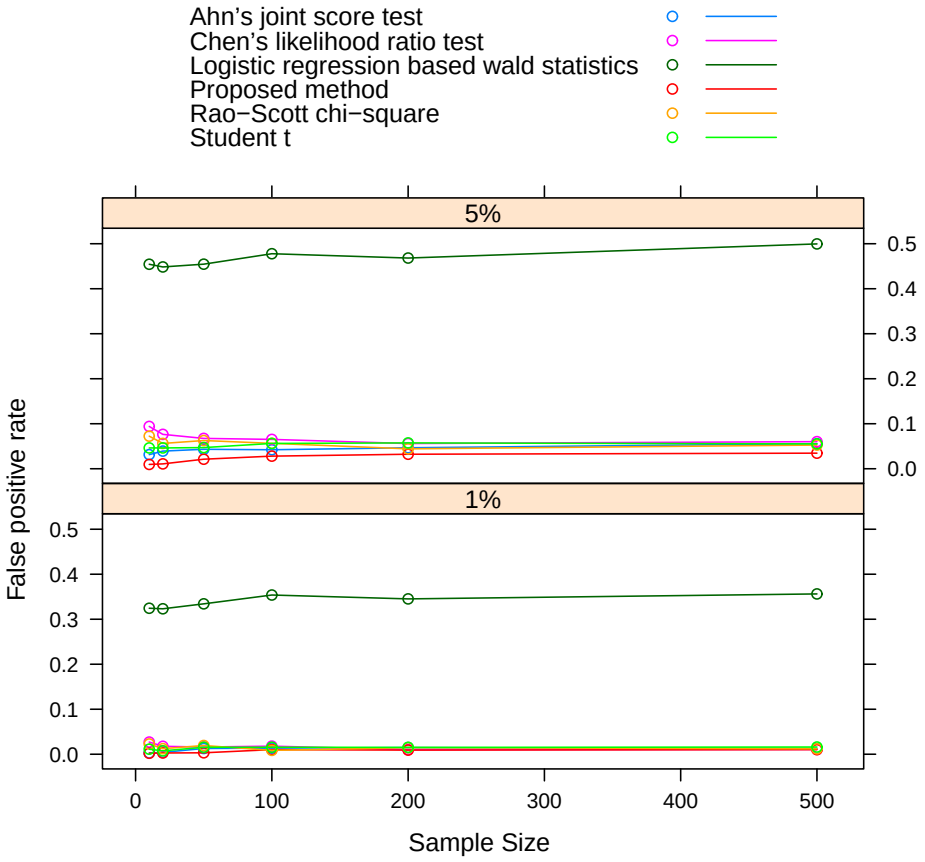

Figure 14: False positive rate for the six tests at 5% and 1% significance levels when logit transformed methylation proportion were generated from a normal distribution with $\mu_1 = \mu_2 = 0$ and $\sigma_1^2 = \sigma_2^2 = 1$.

Figure 15 shows false positive rates for the other five approaches at 5% or 1% significance level when the logit transformed methylation proportions were generated from a normal distribution with mean 0 and variance 3. The distribution plots of the logit transformed methylation proportion and its methylation proportion are showed in Figure panels 9. In this simulation, the mean of methylation proportion is 0.5 and the variance of methylation proportion is 0.0856. Figure 16 shows false positive rates under the simulation scenario with mean 2 and variance 1. The distribution plots of the logit transformed methylation proportion and its methylation proportion are showed in Figure panels 10. In this simulation, the mean of methylation proportion is 0.6971 and the variance is 0.0333. These two figures show a false positive rate patterns that is consistent under all

of the simulation scenarios. They show that our proposed approach has the lowest number of false positive rate among the methods. The false positive rates of our approach are less than the nominal α level over all of the simulated sample sizes, whereas the false positive rates are larger than the nominal α level for other four approaches. The false positive rate of our method increases with the increasing sample size. As can be seen from these figures, the Rao-Scott chi-square test and Chen's likelihood ratio test has the highest false positive rate in detecting differentially methylated genes when the sample size is small, while the value decreases as the sample size increasing. When sample size increases to 200, there is not much difference between the Rao-Scott chi-square test, Student's t-test, Ahn's joint score test and Chen's likelihood ratio test.

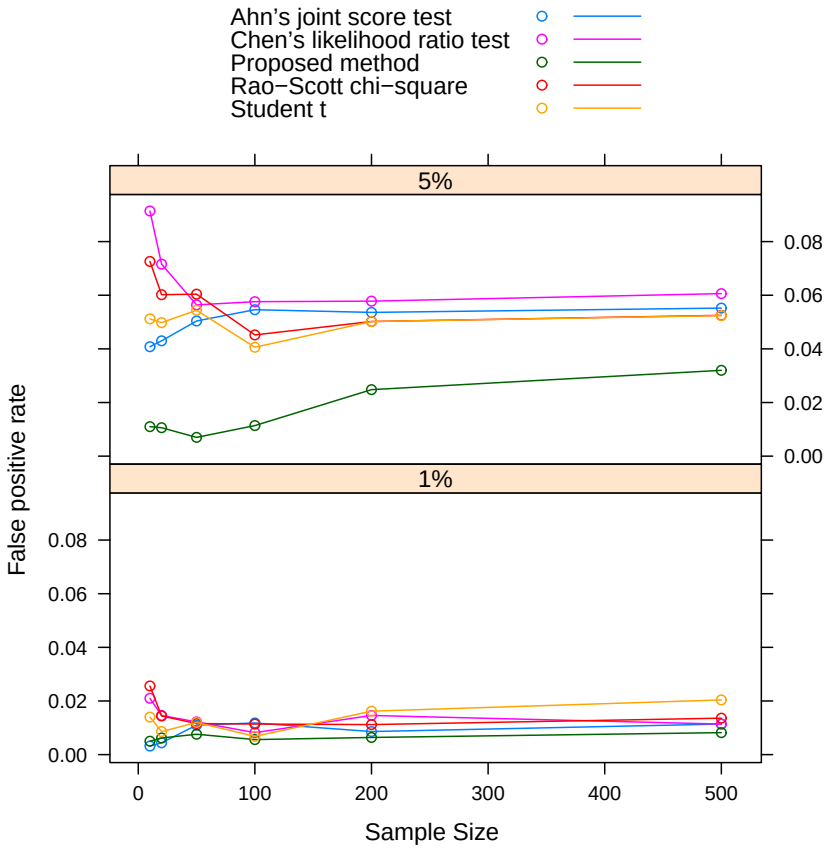

Figure 15: False positive rate for the five tests at 5% and 1% significance levels when logit transformed methylation proportion were generated from a normal distribution with $\mu_1 = \mu_2 = 0$ and $\sigma_1^2 = \sigma_2^2 = 3$.

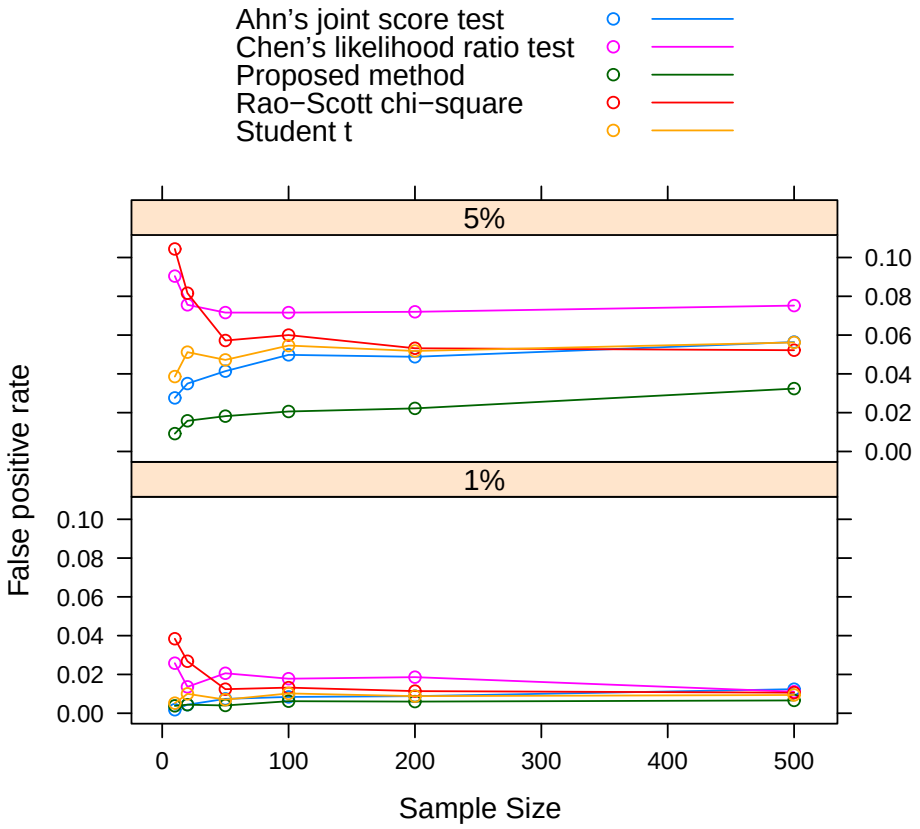

Figure 16: False positive rate for the five tests at 5% and 1% significance levels when logit transformed methylation proportion value were generated from a normal distribution with $\mu_1 = \mu_2 = 2$ and $\sigma_1^2 = \sigma_2^2 = 1$.

3.3.2 True positive rate for unequal mean

This section describes the results for simulation scenarios with unequal mean and equal variance. As described in table 3, the mean and the variance of the methylation proportion after inverse-logit transformation depend on the distribution of the logit-transformed methylation proportion. In these simulation scenarios, we simulated data with unequal means and equal variances for both simulated logit transformed methylation and the inverse-logit transformed methylation proportion. We performed 5,000 replicates for each sample size with α equals to 0.05 or 0.01. Among these 5,000 replicates, the numbers of replicates with significant differentially methylated loci were recorded. The comparison in these simulation scenarios is the true positive rate, which is the proportion of

the number of significant runs over the 5,000 replicates.

Figure 17 shows the true positive rate from the simulation scenario with mean 0.5 in diseases and -0.5 in controls. Both variances in diseases and controls equal to 2. In this simulation scenario, the means of methylation proportion are 0.5902 and 0.4101 for diseases and controls respectively; the variances are 0.0654 for both groups. Figure 17 shows a very high true positive rate (around 0.8) when sample size is 10 and the true positive rate is 1 when sample size is 50 with logistic regression based Wald statistics. In the previous subsection, the figures indicate that this method has the highest false positive rate in all of the simulation combinations with equal mean and equal variance. Additionally, this method detects the most when the means are unequal and the variances are equal. We conclude that this method is not a valid methods in terms of both false positives and true positives. Therefore, the following figures excluded logistic regression method.

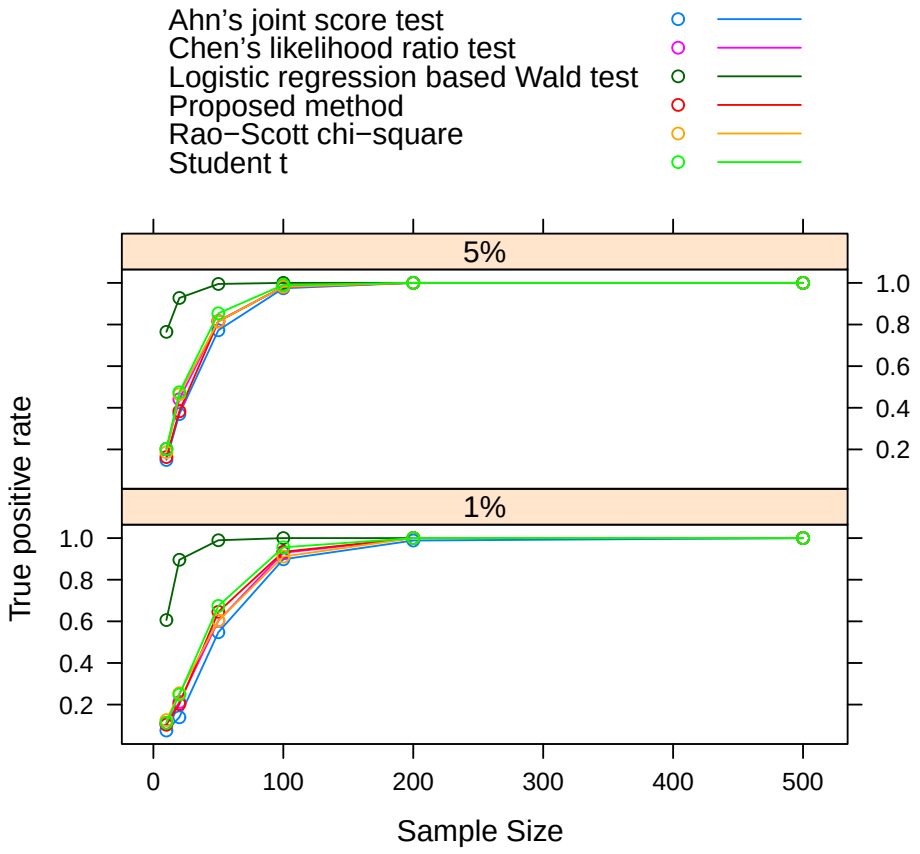

Figure 17: True positive rate for the six tests at 5% and 1% significance levels when logit transformed methylation proportion value were generated from a normal distribution with $\mu_1 = 0.5, \mu_2 = -0.5$ and $\sigma_1^2 = \sigma_2^2 = 2$.

Our proposed approach and Ahn's joint score statistics detect the least number of significantly differentially methylated CpG sites when sample size is 10 in all of the simulation scenarios. When sample size increases to 100, there is not much difference of true positive rate between these five methods. Chen's likelihood ratio test detects the highest number of significantly differentially methylated CpG site when the sample size is 10. Additionally, this figure introduces an increasing true positive rate trends with the increasing sample size in these 5 methods.

To illustrate the changes of true positive rate with the changes of variances with our proposed method, we create the plots with the same mean difference and different sample variances. Figure 18 represents the simulation scenario with the mean of logit transformed methylation proportions equals

to 1 in diseases and -1 in controls. This figure suggests that the true positive rate decreases as the simulated variance increases.

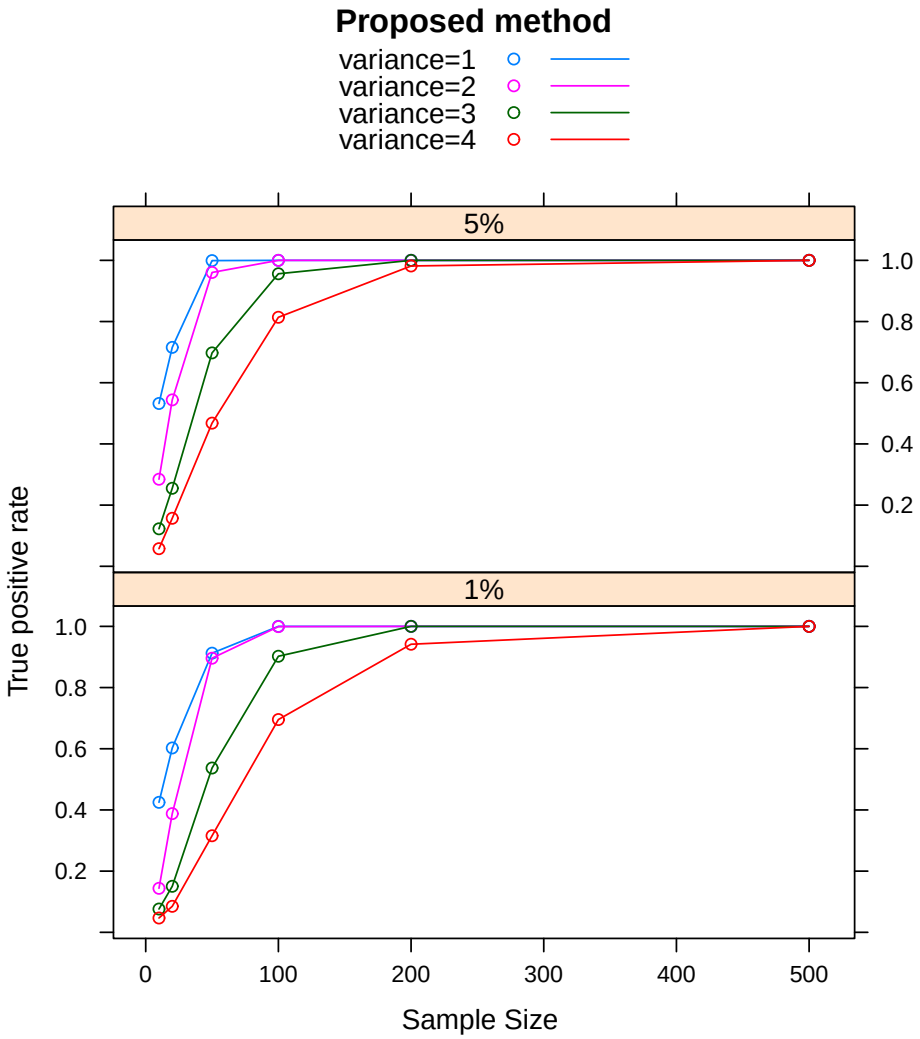

Figure 18: True positive rate for the proposed method at 5% and 1% significance levels when logit transformed methylation proportion value were generated from a normal distribution with $\mu_1 = 1, \mu_2 = -1$.

3.3.3 True positive rate for unequal variance

This section describes the results for simulation scenarios with equal mean but unequal variance. Under this assumption, the mean of the logit transformed methylation proportion are identical but

their variances are different between disease group and control group. The relationship between the simulated logit transformed methylation proportion and the methylation proportion after inverse-logit transformation suggest that the mean of methylation proportion always equals to 0.5 when the simulated logit transformed methylation proportion follows a normal distribution with mean zero. We performed 5,000 replicates for each sample size with α equals to 0.05 or 0.01 and the numbers of replicates with significantly differentially methylated loci were recorded. Student's t test and Rao-scott Chi square test focus on testing methylation mean with assumption of their unequal variance in this section. Our proposed method, Ahn's joint score statistics and Chen's likelihood ratio test detect differentially methylation CpG sites with their mean or variance.

Under the assumption of equal mean but unequal variance for disease and control, student's t test and Rao-Scott chi square have identified CpG sites with mean methylation difference. The detected CpG sites using these two methods are false positives. Figure 19 shows the false positive rates for these two methods with different variances. These figures suggest that both methods have identified CpG sites with difference in methylation mean but ignore their variance information.

Our proposed method, Ahn's joint score statistics and Chen's likelihood ratio test detect CpG site with either methylation mean or methylation variance. The detected significant CpG sites indicate that they are significantly different in methylation mean or methylation variance. Therefore under the assumption of difference in either mean or variance, the detected CpG sites using these three methods are true positives. Figure 20 shows the increasing true positive trends with increasing sample size increase using these three methods. When the sample is small ($n_1 = n_2 = 10$), our proposed method detects the least number of significantly methylated CpG site among these three methods. When the sample size increases to 50, the numbers of detected CpG site are not much different between these three methods. Figure 22 represents the true positive rates for our proposed method with different variances. This figures indicate that our proposed methods detect the least number of significant methylation CpG site when the variance ratio, which is the ratio of the variances in control group and disease group is more close to one. The variance ratio equals to one indicates that there is no significant difference between disease and controls with methylation variances.

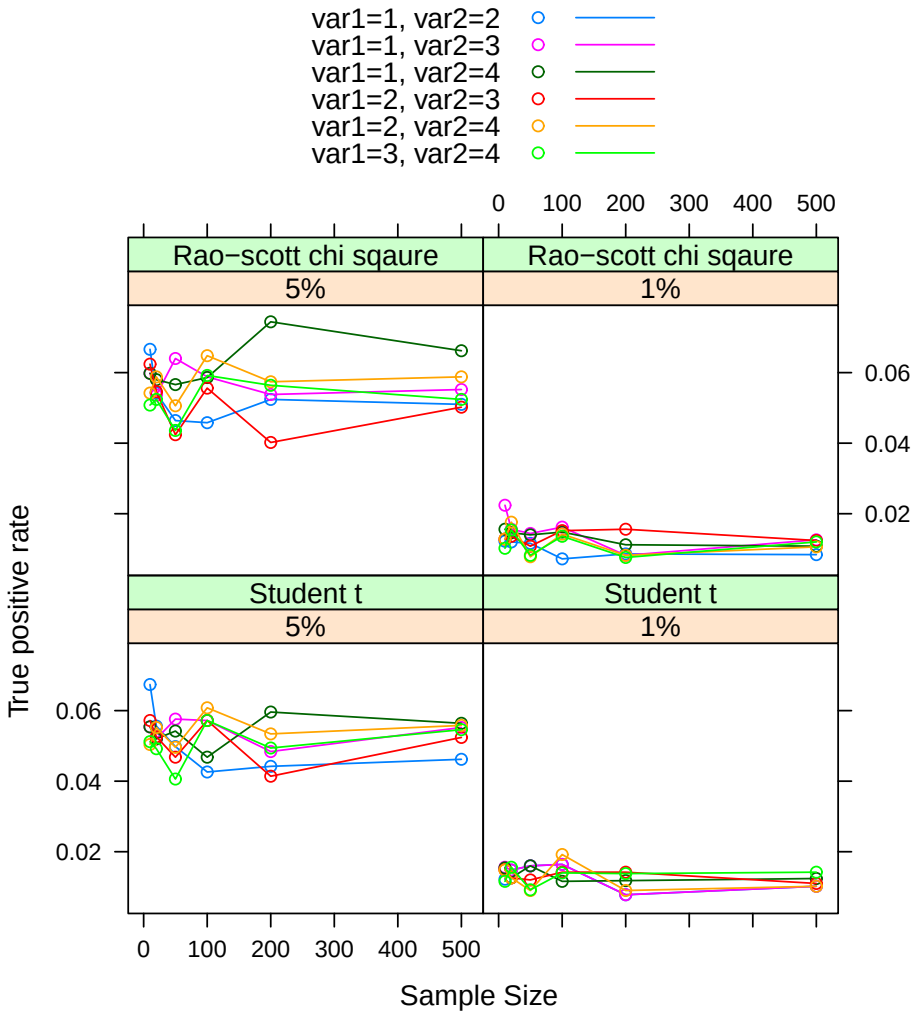

Figure 19: True positive rate for the methods at 5% and 1% significance levels when logit transformed methylation proportion value were generated from a normal distribution with $\mu_1 = \mu_2 = 0$, and $E(p) = 0.5$ for both disease group and control group.

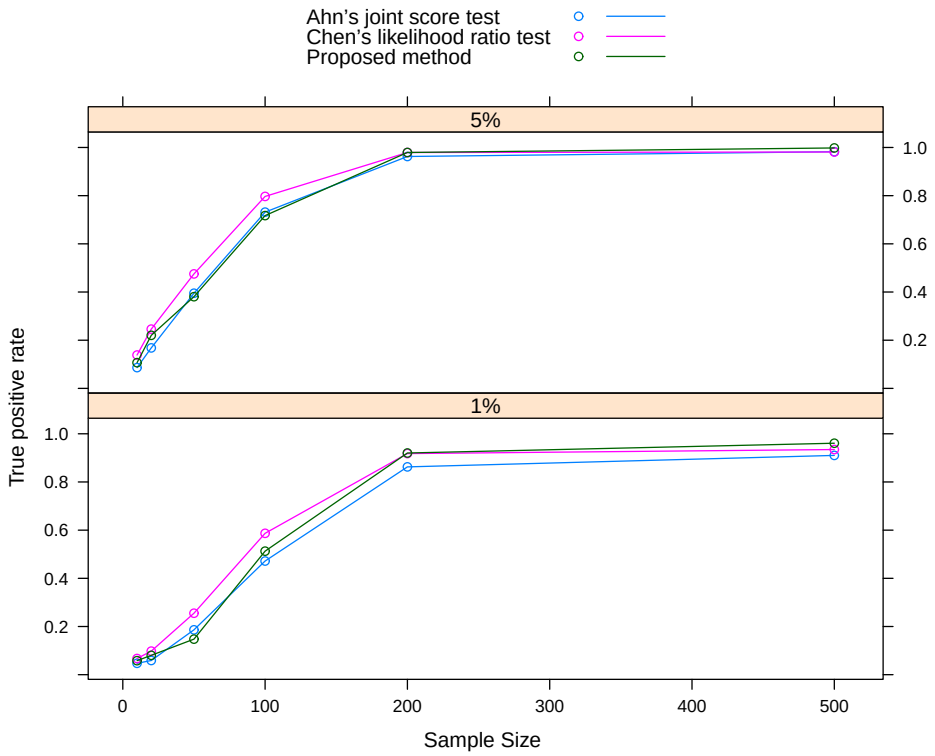

Figure 20: True positive rate for the methods at 5% and 1% significance levels when logit transformed methylation proportion value were generated from a normal distribution with $\mu_1 = \mu_2 = 0$, and $E(p) = 0.5$ for both disease group and control group.

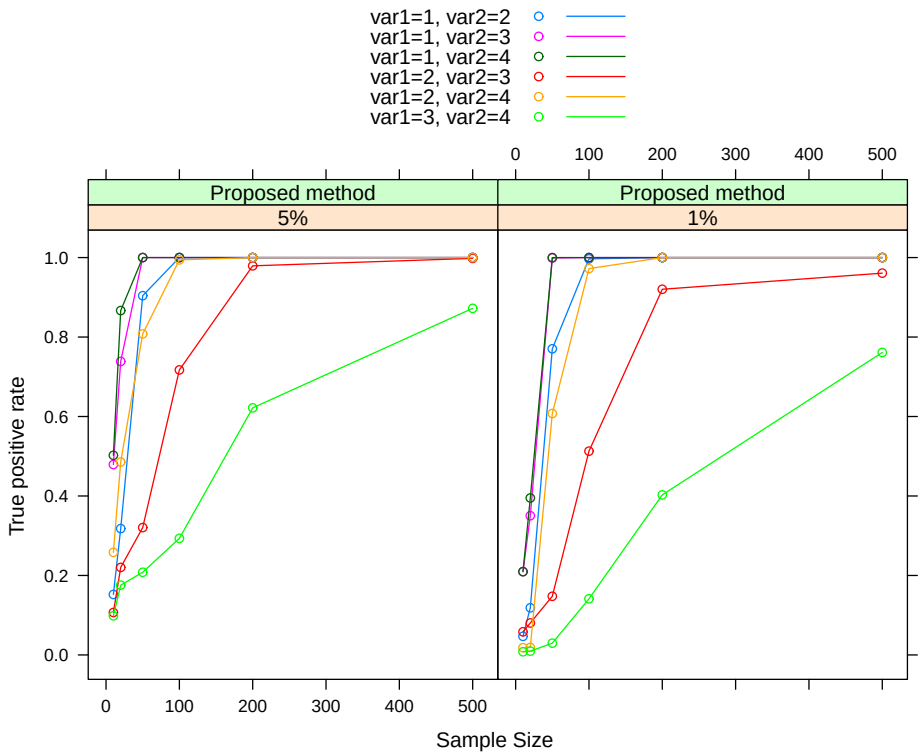

Figure 21: True positive rate for the proposed method at 5% and 1% significance levels when logit transformed methylation proportion value were generated from a normal distribution with $\mu_1 = \mu_2 = 0$, and $E(p) = 0.5$ for both disease group and control group.

3.3.4 True positive rates for unequal mean and unequal variance

This section describes the results for simulation scenarios with unequal mean and unequal variance. Under this assumption, both the mean and the variance of the logit transformed methylation proportion are different between disease and control. Therefore both the mean and the variance of the methylation proportion after logit-inverse transformation are different. We still performed 5,000 replicates for each sample size with α equals to 0.05 or 0.01 and recorded the number of replicates with significantly differentially methylated for each method. The comparison here is the true positive rate, which is defined by the proportion of replicates with significant differential methylation among the 5,000 replication. Table 4 is one simulation scenario example. This table shows that our proposed method detects less number of significantly methylated CpG site than Chen's likelihood

ratio test and Ahn's joint score statistics when the sample size equals 10. However when the sample size increases to 50, there is no substantial difference between these three methods. This table also shows that the logistic regression based Wald statistics detects the most number of significant methylation CpG site. This pattern is similar in all of the previous simulation scenarios for this approach.

Table 4: True positive rate for the six tests at 5% and 1% significance levels when logit transformed methylation proportion value were generated based on the normal distribution with $\mu_1 = 0$ and $\mu_2 = 1$ and $\sigma_1 = 1$ and $\sigma_2 = 2$.

$\mu_\Delta = (\mu_1 - \mu_2)/\sigma_1 = 1$ $\sigma_\Delta = \sigma_1/\sigma_2 = 1/2$	N	10	20	50	100	200	500
Proposed Bayesian method	5%	0.2464	0.4562	0.9916	1	1	1
Student t test		0.1798	0.3834	0.7592	0.9676	1	1
Rao-Scott chi-square test		0.1568	0.3582	0.7046	0.9494	1	1
Ahn's joint score statistics		0.3348	0.6702	0.9901	1	1	1
Logistic regression based wald test		0.5246	0.8574	0.9756	1	1	1
Chen's likelihood ratio test		0.4986	0.7328	0.9915	1	1	1
Proposed Bayesian method	1%	0.1026	0.2004	0.9176	0.9928	1	1
Student t test		0.0986	0.1812	0.5286	0.8848	0.9982	1
Rao-Scott chi-square test		0.0826	0.1634	0.4768	0.8402	0.9924	1
Ahn's joint score statistics		0.2004	0.3946	0.9002	0.9901	1	1
Logistic regression based wald test		0.4476	0.7894	0.9546	0.9980	1	1
Chen's likelihood ratio test		0.2680	0.4564	0.9210	1	1	1

One simulation scenarios that needs to be discussed here is under the assumption of equal non-zero mean and unequal variance. In the data simulation, the means of the logit transformed methylation proportions are equal to some non-zero value but the variances are different between diseases and controls. As described in the table 3, when the simulated logit transformed methylation proportion follows a normal distribution with an equal, non-zero mean and different variance, both the mean and the variance of the methylation proportion after logit-inverse transformation are different. For example, when the simulated logit transformed proportions follow a normal distribution of mean 2 and variances are 1 and 2 respectively, the mean of methylation proportions is 0.8447 or 0.8161 respectively. Figure 22 represents the detected significantly methylated CpG sites for the simulation scenarios with the mean is 2 in both groups, and the variance is 2 and 1 for disease group or control group respectively. The distribution plots of the logit transformated methylation proportion and

its methylation proportion are showed in Figure panels 10. Under this situation, the mean and the variance of the methylation proportion are different between groups. Figure 22 suggests that Student's t test and Rao-scott Chi square test detect less numbers of significantly methylated CpG sites under this simulation assumption.

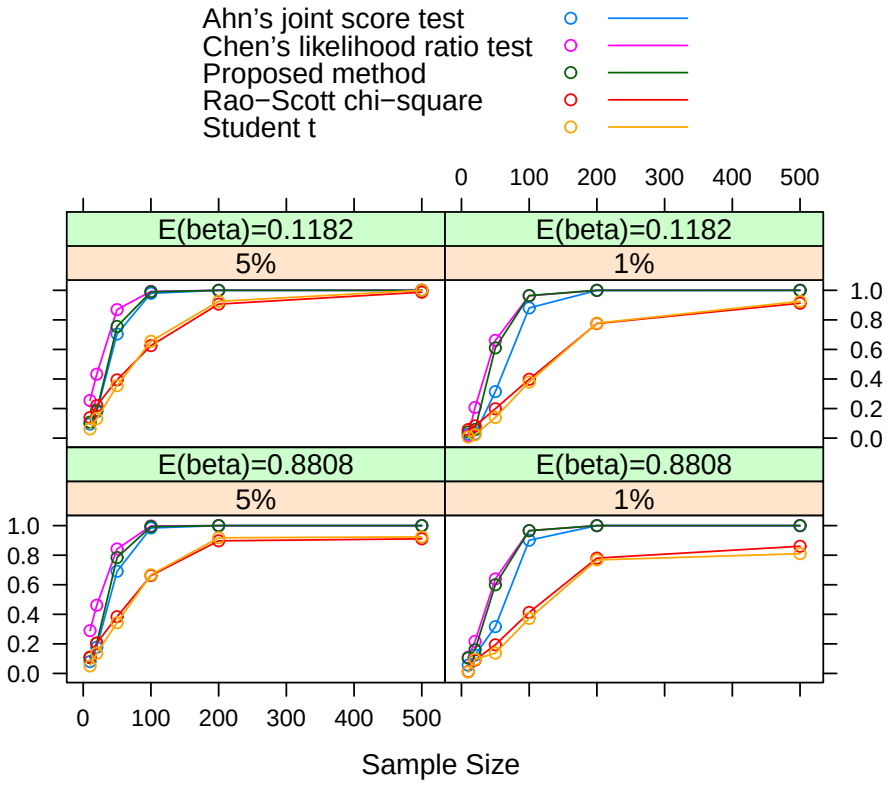

Figure 22: True positive rate for the methods at 5% and 1% significance levels when logit transformed methylation proportion value were generated from a normal distribution with $\mu_1 = \mu_2 = 2$.

In summary, our proposed method has the lowest false positive rate under all of the simulation combinatons. For the true positive rate, our method is comparable to other methods for simulations with unequal mean and/or unequal variance.

3.4 Real data analysis

In order to evaluate the performance of these methods on a more realistic NGS data, in addition to the simulation data above, we also consider the real NGS data on chronic lymphocytic leukemia (CLL). Several analyses indicate that the patients with high CD38 expression are associated with bad prognosis [72]. We divided the patients into two groups: 10 individuals with CD38 >= 30 and 30 individuals with CD38 <30. This methylation data were generated using reduced representation bisulfite sequencing (RRBS) approach. Using this approach, we obtained genome-wide methylation data with complete information at 231,643 CpG sites from CLL peripheral blood samples. We applied our proposed test for differential methylation on this data set, and we also applied Student's t-test, Rao-scott Chi Square statistics, Ahn's joint score statistics, Logistic regression based Wald statistics and Chen's likelihood ratio test to the same data set. Table 5 shows the numbers of differentially methylated loci from these tests with 5% or 1% significance level, respectively. This table shows that logistic regression based Wald statistics detects the most differently methylated CpG sites at both 5% and 1% level. According to the simulation results, this method has the highest false positive rate under the assumption of equal sample mean and equal sample variance. Therefore, the findings of this could be the mostly false positives with the real data analyses and we will further investigate this with the permutation data set. Rao-scott Chi square and Student's t test detect the second most significantly methylated CpG sites as shown in the Table 5. Our proposed method, Ahn's joint score statistics and Chen's likelihood ratio test detect the significantly methylated CpG sites with either methylation mean or methylation variance. Among these three methods, Chen's method detect the most loci and our proposal method detects the second most loci at 5%. However, our proposed method detects the highest number of significantly methylated CpG site at 1%.

Table 5: **Real data analyses** Number of detected CpG site for the six tests at 5% and 1% significance levels

	Proposed test	Rao-Scott	Student t	Ahn's	GLM	Chen's LRT
5%	7,729	21,755	20,854	7,330	63,222	16,164
1%	2,357	5,298	5,671	985	43,136	2,287

We further examined the results of our proposed method and other four alternative methods. Among the detected CpG sites by our proposed method, Table 6 presents the numbers of CpG sites

also detected by the other alternative methods. For example, among 7,278 detected CpG sites by the proposed method at 5% level, 3,539 (48.61%) sites are detected by Rao Scott statistics; among 2,357 detected CpG sites by the proposed method at 1% level, 979 (41.54%) sites are detected by Rao scott statistics.

Table 6: **Real data analyses** Number of detected CpG site with real data by both our proposed method and alternative methods at 5% and 1% significance levels.

	Rao-Scott	Student t	Ahn's	Chen's LRT
5%	3,539 (48.61%)	2,787 (38.29%)	1,310 (18.00%)	3,196 (43.91 %)
1%	979 (41.54%)	605 (25.67%)	201 (8.53%)	559 (23.73%)

Our proposed method reveals more details for the detected CpG sites. Among the 7,729 detected loci at 5% significant level, Table 7 shows that 1,565 CpG sites are significantly different with methylation mean only, 5,430 CpG sites are significantly different with methylation variance only and 234 CpG sites are significantly different with both methylation mean and methylation variance. Among the 2,357 detected loci at 1% level, 549 CpG sites are significant with methylation mean, 1,776 CpG sites are significant with methylation variance only and 32 CpG sites are significant with both mean and variance.

Table 7: **Real data analyses** Number of detected CpG site in mean and/or in variance for the proposed tests at 5% and 1% significance levels.

Proposed test	mean difference only	variance difference only	mean & variance difference
5%	1,565	5,430	234
1%	549	1,776	32

3.5 Permutation data analysis

We also randomly assigned these 40 individuals into two groups with sizes of 10 and 30 ignoring their CD38 status. This permutation data is generated under the assumption that there is no significant methylation difference between groups. We then applied our proposed test for differential methylation on this data set, and we also applied Student's t-test, Rao-scott Chi Square statistics, Ahn's joint score statistics, Logistic regression based Wald statistics and Chen's likelihood ratio test to the same data set. Table 8 shows the numbers of differentially methylated loci from these tests with 5% and 1% significance levels, respectively.

Table 8: **Permutation data analyses** Number of detected CpG site with permutation data for the six tests at 5% and 1% significance levels

	Proposed test	Rao-Scott	Student t	Ahn's	GLM	Chen's LRT
5%	6,459	16,717	21,614	9,206	66,547	15,370
1%	2,197	8,064	4,273	1,433	45,061	2,135

This table shows that logistic regression based Wald statistic has the highest number of differently methylated CpG sites with both 5% and 1%. According to the simulation results, this method has the highest false positive rate under the assumption of equal methylation mean and equal methylation variance. Here, this method have identified more significant loci with permutation data than the real data at both 5% and 1% significant levels. Therefore, this method might detect the mostly false positives with the permutation data analyses. Similar to the real data set, Rao-scott Chi square detects the second most number of significantly methylated CpG site as shown in the Table 8. Our proposed method detects the least number of significantly methylated CpG sites in this permutated data.

After further comparison between Table 5 and Table 8, we found that the analyses of one permutated data found a similar number of sites as the analyses of the real observed data for each method. The permutated data was generated under tha assumption that there was no significant methylation difference between groups. The expected results should be that the analyses of the permutated data found less number of sites. The slight difference between the observation data and the permutation data implied that none of the methods worked on the real data based on the comparison with the permutated data results.

In summary, we applied these six methods to a real observed data and one permutated data. Each method found their own significant CpG sites. However, there is slight difference of number of sites between the analyses of real observed data and one permutated data. We concluded that none of the methods worked on the real data based on the comparison with the permutated data results.

4 Summary and Discussion

The development and application of the proposed Bayesian methods are particularly valuable for biomedical researchers in that they allow for identifying differentially methylated loci with mean difference and / or variability difference. Both difference in methylation mean and methylation variability are important indicators for identifying disease risk biomarkers. Therefore the methods for identifying mean difference and/or variability will offer a high level of resolution for disorder prevention and disease treatment.

In the chapter one, we have given an introduction to the DNA methylation and the high-throughput platforms based on NGS. We have given the review about the existed statistical methods in identifying differentially methylated CpG sites. The methods for testing the difference in methylation mean only ignore information provided by the methylation variance. Therefore, they may detect less significantly methylated sites in the case of heterogeneity in methylation variances. Ahn's joint score test and Chen's semiparametric test incorporate the mean methylation information and the variance methylation information in the test. Their significant p-values indicate either significant difference in methylation mean or in methylation variance. In Chapter 2 we developed a flexible Bayesian framework that includes both methylation mean and methylation variance in a hierarchical model using NGS counts. We modeled the methylation proportion with a Latent Gaussian Model with covariates and applied INLA to derive the posterior distributions. After that we introduced a two dimensional highest posterior density region for the Bayesian hypotheses. INLA is derived by a Gaussian approximations which save much computation time for posterior sampling.

In chapter 3 we have presented the simulation results and the real data analysis results. Simulation results indicate that our proposed method have the smallest false positive rate under the equal mean and equal variance scenarios over all of the simulation sample size. Logistic regression based Wald statistics have the largest false positive rate for all of the simulation scenarios. When the simulated methylation proportion under the assumption of unequal mean and/or unequal variance, our proposed method detect less significantly differentially methylated CpG sites when sample size is small. When sample size increase to 100, there is not much difference of true positive rate between these five methods. In the real data analyses, logistic regression based Wald statistics have the largest number of significantly methylated loci under both real data set and its permutation

data set. Our proposed method detected least significantly methylated loci in the permutation data set than the real data set. However, the analyses of one permutation of the data found a similar number of sites as the analyses of the real observed data for each methods. The slight difference between the observation data and the permutation data implied that none of the method worked well for this observed data.

Most of the currently methods for NGS data in the review chapter rely on estimated methylation proportion as the input, rather than the raw counts with NGS. In many cases, the coverage for each CpG site is different across a genome. The variance of the estimated methylation proportion ignores the difference of the coverage and treats all the CpGs in the same way, thus it is subject to potential loss of power. Our robust Bayesian framework flexibly models the NGS counts and incorporates both methylation mean and methylation variance in a hierarchical model.

Most methods reviewed in this dissertation have distribution assumptions. For example, student t test has normality assumption; logistic regression model and Ahn's joint score statistics have logistic linear model assumption, Chen's semiparametric method has exponential distribution assumption and our proposal Bayesian model assumes latent Gaussian model. The value of methylation proportion varies between 0 and 1, but its distribution is unknown in most cases. Our proposed simulation covers the mean methylation levels in a large range (0.018, 0.982), and the variance methylation levels in a reasonable range based on the randomly selected samples from the real data. Therefore, our simulation results cover most potential situations with NGS data. The simulation results in terms of true positive rate and false positive rate are collected for comparisons. Methods with higher true positive rate and lower false positive rate are more powerful and more accurate to detect differentially methylated loci with NGS data.

As described in the table 3, when the simulated logit transformed methylation proportion follows a normal distribution with zero mean and difference variance between the normal and the disease groups, the means of methylation proportion after logit-inverse transformed equal to 0.5 in both groups. However when the simulated logit transformed methylation proportion follows a normal distribution with an equal non-zero mean and different variance, both the mean and the variance of the methylation proportion are different. Our simulated data considered this relationship between the methylation proportions and the logit transformed methylation proportions.

To improve computational efficiency, we use Integrated Nested Laplace Approximation (INLA), which combines Laplace approximations and numerical integration in a very efficient manner for deriving marginal posterior distributions. This method assumes Gaussian distribution and performs Laplace approximation. It can achieve more accurate results when the distribution is near to a Gaussian distribution. However, if the distribution is far away from a Gaussian distribution, it might produce bias.

In summary, we developed a flexible Bayesian approach to detect differentally methylation loci in methylation mean and/or methylation variance. It has lowest false positive rate among the methods compared. It also has true positve rate comparable to the other methods. It can account for co-variates and is computatially faster than MCMC methods.

Bibliography

1. Partha M Das and Rakesh Singal. Dna methylation and cancer. *Journal of Clinical Oncology*, 22(22):4632–4642, 2004.

2. Andrew P Feinberg and Rafael A Irizarry. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proceedings of the National Academy of Sciences*, 107(suppl 1):1757–1764, 2010.

3. Kasper Daniel Hansen, Winston Timp, Héctor Corrada Bravo, Sarven Sabunciyan, Benjamin Langmead, Oliver G McDonald, Bo Wen, Hao Wu, Yun Liu, Dinh Diep, et al. Increased methylation variation in epigenetic domains across cancer types. *Nature genetics*, 43(8):768–775, 2011.

4. Andrew E Teschendorff and Martin Widschwendter. Differential variability improves the identification of cancer risk markers in dna methylation studies profiling precursor cancer lesions. *Bioinformatics*, 28(11):1487–1494, 2012.

5. Surin Ahn and Tao Wang. A powerful statistical method for identifying differentially methylated markers in complex diseases. In *Pac Symp Biocomput. http://www. ncbi. nlm. nih. gov/pubmed/23424113*. World Scientific, 2013.

6. Yong Chen, Yang Ning, Chuan Hong, and Shuang Wang. Semiparametric tests for identifying differentially methylated loci with case–control designs using illumina arrays. *Genetic epidemiology*, 38(1):42–50, 2014.

7. Roland Lauster. Evolution of type ii dna methyltransferases: a gene duplication model. *Journal of molecular biology*, 206(2):313–321, 1989.

8. Albertas Timinskas, Viktoras Butkus, and Arvydas Janulaitis. Sequence motifs characteristic for dna [cytosine-n4] and dna [adenine-n6] methyltransferases. classification of all dna methyltransferases. *Gene*, 157(1):3–11, 1995.

9. Yi Zhang, Gary LeRoy, Hans-Peter Seelig, William S Lane, and Danny Reinberg. The

dermatomyositis-specific autoantigen mi2 is a component of a complex containing histone deacetylase and nucleosome remodeling activities. *Cell*, 95(2):279–289, 1998.

10. Matthew J Scanlan, Yao-Tseng Chen, Barbara Williamson, Ali O Gure, Elisabeth Stockert, John D Gordan, Oezlem Tuereci, Ugur Sahin, Michael Pfreundschuh, and Lloyd J Old. Characterization of human colon cancer antigens recognized by autologous antibodies. *International journal of cancer*, 76(5):652–658, 1998.

11. Mun-Kit Choy, Mehregan Movassagh, Hock-Guan Goh, Martin R Bennett, Thomas A Down, and Roger SY Foo. Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated. *BMC genomics*, 11(1):519, 2010.

12. Melanie Ehrlich, Miguel A Gama-Sosa, Lan-Hsiang Huang, Rose Marie Midgett, Kenneth C Kuo, Roy A McCune, and Charles Gehrke. Amount and distribution of 5-methylcytosine in human dna from different types of tissues or cells. *Nucleic acids research*, 10(8):2709–2721, 1982.

13. Kerry Lee Tucker. Methylated cytosine and the brain: a new base for neuroscience. *Neuron*, 30(3):649–652, 2001.

14. David N Cooper and Michael Krawczak. Cytosine methylation and the fate of cpg dinucleotides in vertebrate genomes. *Human genetics*, 83(2):181–188, 1989.

15. Tetsuya Kamikihara, Takahiro Arima, Kiyoko Kato, Takao Matsuda, Hidenori Kato, Tsutomu Douchi, Yukihiro Nagata, Mitsuyoshi Nakao, and Norio Wake. Epigenetic silencing of the imprinted gene zac by dna methylation is an early event in the progression of human ovarian cancer. *International journal of cancer*, 115(5):690–700, 2005.

16. Angela H Ting, Kornel E Schuebel, James G Herman, and Stephen B Baylin. Short double-stranded rna induces transcriptional gene silencing in human cancer cells in the absence of dna methylation. *Nature genetics*, 37(8):906–910, 2005.

17. R Keith Slotkin and Robert Martienssen. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, 8(4):272–285, 2007.

18. Jeffrey A Yoder, Colum P Walsh, and Timothy H Bestor. Cytosine methylation and the ecology of intragenomic parasites. *Trends in genetics*, 13(8):335–340, 1997.

19. Elias Daura-Oller, Maria Cabre, Miguel A Montero, Jose L Paternain, and Antoni Romeu. Specific gene hypomethylation and cancer: New insights into coding region feature trends. *Bioinformation*, 3(8):340, 2009.

20. Keith D Robertson. Dna methylation and human disease. *Nature Reviews Genetics*, 6(8):597–610, 2005.

21. Anne C Ferguson-Smith and M Azim Surani. Imprinting and the epigenetic asymmetry between parental genomes. *Science*, 293(5532):1086–1089, 2001.

22. Jeannie T Lee. Molecular links between x-inactivation and autosomal imprinting: X-inactivation as a driving force for the evolution of imprinting? *Current biology*, 13(6):R242–R254, 2003.

23. Rudolf Jaenisch and Adrian Bird. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, 33:245–254, 2003.

24. Manel Esteller and James G Herman. Cancer as an epigenetic disease: Dna methylation and chromatin alterations in human tumours. *The Journal of pathology*, 196(1):1–7, 2002.

25. Jeffrey Craig and Nicholas C Wong. *Epigenetics: a reference manual*. Horizon Scientific Press, 2011.

26. Andrew P Feinberg and Benjamin Tycko. The history of cancer epigenetics. *Nature Reviews Cancer*, 4(2):143–153, 2004.

27. Chih-Lin Hsieh. Dependence of transcriptional repression on cpg methylation density. *Molecular and cellular biology*, 14(8):5487–5494, 1994.

28. Pearlly S Yan, Huidong Shi, Farahnaz Rahmatpanah, Timothy HC Hsiau, Andrew HA Hsiau, Yu-Wei Leu, Joseph C Liu, and Tim Hui-Ming Huang. Differential distribution of dna methylation within the rassf1a cpg island in breast cancer. *Cancer research*, 63(19):6178–6186, 2003.

29. Elizabeth E Cameron, Stephen B Baylin, and James G Herman. p15ink4b cpg island methylation in primary acute leukemia is heterogeneous and suggests density as a critical factor for transcriptional silencing. *Blood*, 94(7):2445–2451, 1999.

30. Alessio Amatu, Andrea Sartore-Bianchi, Catia Moutinho, Alessandro Belotti, Katia Bencardino, Giuseppe Chirico, Andrea Cassingena, Francesca Rusconi, Anna Esposito, Michele Nichelatti, et al. Promoter cpg island hypermethylation of the dna repair enzyme mgmt predicts clinical response to dacarbazine in a phase ii study for metastatic colorectal cancer. *Clinical Cancer Research*, 19(8):2265–2272, 2013.

31. Jean-Pierre Issa. Cpg island methylator phenotype in cancer. *Nature Reviews Cancer*, 4(12):988–993, 2004.

32. Adrian P Bird. Cpg-rich islands and the function of dna methylation. *Nature*, 321(6067):209–213, 1985.

33. Naoko Nakamura and Keizo Takenaga. Hypomethylation of the metastasis-associated s100a4 gene correlates with gene activation in human colon adenocarcinoma cell lines. *Clinical & experimental metastasis*, 16(5):471–479, 1998.

34. Hannes M Müller, Michael Oberwalder, Heidi Fiegl, Maria Morandell, Georg Goebel, Matthias Zitt, Markus Mühlthaler, Dietmar Öfner, Raimund Margreiter, and Martin Widschwendter. Methylation changes in faecal dna: a marker for colorectal cancer screening? *The Lancet*, 363(9417):1283–1285, 2004.

35. Howard J Edenberg, Daniel L Koller, Xiaoling Xuei, Leah Wetherill, Jeanette N McClintick, Laura Almasy, Laura J Bierut, Kathleen K Bucholz, Alison Goate, Fazil Aliev, et al. Genome-wide association study of alcohol dependence implicates a region on chromosome 11. *Alcoholism: Clinical and Experimental Research*, 34(5):840–852, 2010.

36. Elias Zintzaras and Joseph Lau. Trends in meta-analysis of genetic association studies. *Journal of human genetics*, 53(1):1–9, 2008.

37. Daniëlle van Manen, Angélique B van't Wout, Hanneke Schuitemaker, et al. Genome-wide association studies on hiv susceptibility, pathogenesis and pharmacogenomics. *Retrovirology*, 9(70):1–8, 2012.

38. Mukesh Verma. Epigenome-wide association studies (ewas) in cancer. *Current genomics*, 13(4):308, 2012.

39. Elaine R Mardis. The impact of next-generation sequencing technology on genetics. *Trends in genetics*, 24(3):133–141, 2008.

40. Neil Hall. Advanced sequencing technologies and their wider impact in microbiology. *Journal of Experimental Biology*, 210(9):1518–1525, 2007.

41. Chandra Shekhar Pareek, Rafal Smoczynski, and Andrzej Tretyn. Sequencing technologies and genome sequencing. *Journal of applied genetics*, 52(4):413–435, 2011.

42. Michael L Metzker. Sequencing technologies—the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

43. Shawn J Cokus, Suhua Feng, Xiaoyu Zhang, Zugen Chen, Barry Merriman, Christian D Haudenschild, Sriharsa Pradhan, Stanley F Nelson, Matteo Pellegrini, and Steven E Jacobsen. Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature*, 452(7184):215–219, 2008.

44. Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

45. Daniel Zilberman and Steven Henikoff. Genome-wide analysis of dna methylation patterns. *Development*, 134(22):3959–3965, 2007.

46. Zhongxue Chen, Qingzhong Liu, and Saralees Nadarajah. A new statistical approach to detecting differentially methylated loci for case control illumina array methylation data. *Bioinformatics*, 28(8):1109–1113, 2012.

47. Janette Mareska Rumbajan, Toshiyuki Maeda, Ryota Souzaki, Kazumasa Mitsui, Ken Higashimoto, Kazuhiko Nakabayashi, Hitomi Yatsuki, Kenichi Nishioka, Ryoko Harada, Shigehisa Aoki, et al. Comprehensive analyses of imprinted differentially methylated regions reveal epigenetic and genetic characteristics in hepatoblastoma. *BMC cancer*, 13(1):608, 2013.

48. Ryan KC Yuen, Ruby Jiang, Maria S Peñaherrera, Deborah E McFadden, and Wendy P Robinson. Genome-wide mapping of imprinted differentially methylated regions by dna methylation profiling of human placentas from triploidies. *Epigenetics & chromatin*, 4(1):1–16, 2011.

49. Hongyan Xu, Robert H Podolsky, Duchwan Ryu, Xiaoling Wang, Shaoyong Su, Huidong Shi, and Varghese George. A method to detect differentially methylated loci with next-generation sequencing. *Genetic epidemiology*, 37(4):377–382, 2013.

50. JNK Rao and AJ Scott. A simple method for the analysis of clustered binary data. *Biometrics*, pages 577–585, 1992.

51. John K Kruschke. What to believe: Bayesian methods for data analysis. *Trends in cognitive sciences*, 14(7):293–300, 2010.

52. Christiana Spyrou, Rory Stark, Andy G Lynch, and Simon Tavaré. Bayespeak: Bayesian analysis of chip-seq data. *BMC bioinformatics*, 10(1):299, 2009.

53. Sika Zheng and Liang Chen. A hierarchical bayesian model for comparing transcriptomes at the individual transcript isoform level. *Nucleic acids research*, 37(10):e75–e75, 2009.

54. Guodong Wu, Nengjun Yi, Devin Absher, and Degui Zhi. Statistical quantification of methylation levels by next-generation sequencing. *PloS one*, 6(6):e21034, 2011.

55. Kenneth McCallum, Wenxin Jiang, and Ji-Ping Wang. An empirical bayes approach for methylation differentiation at the single nucleotide resolution. *Computer Science*, 5(2), 2010.

56. Thomas J Hardcastle and Krystyna A Kelly. Empirical bayesian analysis of paired high-throughput sequencing data with a beta-binomial distribution. *BMC bioinformatics*, 14(1):135, 2013.

57. Hao Feng, Karen N Conneely, and Hao Wu. A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic acids research*, 42(8):e69–e69, 2014.

58. Altuna Akalin, Francine E Garrett-Bakelman, Matthias Kormaksson, Jennifer Busuttil, Lu Zhang, Irina Khrebtukova, Thomas A Milne, Yongsheng Huang, Debabrata Biswas, Jay L Hess, et al. Base-pair resolution dna methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. *PLoS genetics*, 8(6):e1002781, 2012.

59. Naiara G Bediaga, Amelia Acha-Sagredo, Isabel Guerra, Amparo Viguri, Carmen Albaina, Irune Ruiz Diaz, Ricardo Rezola, María Jesus Alberdi, Joaquín Dopazo, David Montaner, et al. Dna methylation epigenotypes in breast cancer molecular subtypes. *Breast Cancer Res*, 12(5):R77, 2010.

60. Hong-Qiang Wang, Lindsey K Tuominen, and Chung-Jui Tsai. Slim: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics*, 27(2):225–231, 2011.

61. Youngik Yang, Kenneth Nephew, and Sun Kim. A novel k-mer mixture logistic regression for methylation susceptibility modeling of cpg dinucleotides in human gene promoters. *BMC bioinformatics*, 13(Suppl 3):S15, 2012.

62. Altuna Akalin, Matthias Kormaksson, Sheng Li, Francine E Garrett-Bakelman, Maria E Figueroa, Ari Melnick, Christopher E Mason, et al. methylkit: a comprehensive r package for the analysis of genome-wide dna methylation profiles. *Genome Biol*, 13(10):R87, 2012.

63. Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

64. Alan E Gelfand, Susan E Hills, Amy Racine-Poon, and Adrian FM Smith. Illustration of bayesian inference in normal data models using gibbs sampling. *Journal of the American Statistical Association*, 85(412):972–985, 1990.

65. Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

66. Mary Kathryn Cowles and Bradley P Carlin. Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

67. Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

68. Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

69. Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

70. Martyn Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March*, pages 20–22, 2003.

71. George EP Box and George C Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.

72. Silvia Deaglio, Kapil Mehta, and Fabio Malavasi. Human cd38: a (r) evolutionary story of enzymes and receptors. *Leukemia research*, 25(1):1–12, 2001.