Published in final edited form as:

Stat Biosci. 2013 May ; 5(1): 179–197. doi:10.1007/s12561-013-9086-9.

Screening for SNPs with Allele-Specific Methylation based on Next-Generation Sequencing Data

Bo Hu¹, Yuan Ji², Yaomin Xu¹, and Angela H Ting³

Yuan Ji: yji@northshore.org; Angela H Ting: tinga@ccf.org

¹Department of Quantitative Health Sciences, Cleveland Clinic

²Center for Clinical and Research Informatics, NorthShore University HealthSystem

³Genomic Medicine Institute, Cleveland Clinic

Abstract

Allele-specific methylation (ASM) has long been studied but mainly documented in the context of genomic imprinting and X chromosome inactivation. Taking advantage of the next-generation sequencing technology, we conduct a high-throughput sequencing experiment with four prostate cell lines to survey the whole genome and identify single nucleotide polymorphisms (SNPs) with ASM. A Bayesian approach is proposed to model the counts of short reads for each SNP conditional on its genotypes of multiple subjects, leading to a posterior probability of ASM. We flag SNPs with high posterior probabilities of ASM by accounting for multiple comparisons based on posterior false discovery rates. Applying the Bayesian approach to the in-house prostate cell line data, we identify 269 SNPs as candidates of ASM. A simulation study is carried out to demonstrate the quantitative performance of the proposed approach.

Keywords

DNA Methylation; ASM; Next-generation Sequencing; SNP

1 Introduction

DNA methylation is the most common epigenetic mechanism that plays a key role in transcription regulation [1], and it has been linked with many complex diseases [2, 3]. Since DNA methylation is reversible, it is a promising target for therapeutic interventions. Allele specific DNA methylation (ASM) is a form of allelic asymmetries where methylation occurs on one specific allele but not on the alternative allele. Locating genomic regions with ASM has been traditionally studied in the field of genome imprinting [4]. If a gene is imprinted, the inactive allele will be significantly more methylated than the actively expressed allele, which is also associated with allele specific expression. Another major type of ASM is X chromosome inactivation for females, where DNA methylation coordinates the random silencing of the X chromosome. X chromosome inactivation in any cell is maintained once established so that the inactivated allele is transcriptionally silenced. To date, only a handful of ASM regions have been identified [5].

The recent development of next-generation sequencing (NGS) technologies has the potential to greatly improve the ability to detect and understand ASM [6] since NGS applications are able to achieve high resolution at the level of single nucleotide. Methylation typically occurs

at the C5 position of cytosines within CpG sites (a cytosine next to a guanine). Bisulfite sequencing, a DNA sequencing method in combination with bisulfite conversion, is the current gold standard for methylation analysis, which converts unmethylated cytosines to uracils while leaving methylated cytosines intact. If bisulfite-based sequencing data are available, investigators can precisely determine the methylation state of each CpG site on the two alleles. Standard statistical tests such as the Fisher's exact test can then be used to compare the proportion of methylated CpG sites on the two alleles and test the hypothesis whether methylation is significantly associated with one specific allele [7]. However, bisulfite sequencing is costly and most studies using bisulfite sequencing to detect ASM have been restricted to targeted genomic regions. For example, Zhang et al. (2009) [8] analyzed chromosome 21, the shortest autosome in human, and only identified four distinct amplicons with ASM.

Due to the high cost and limitations of bisulfite sequencing, enrichment-based NGS technologies have been developed for whole-genome DNA methylation studies [9, 10], which are more affordable in a general laboratory setting. Popular enrichment-based NGS methods include MeDIP-seq [11] and MBD-seq [12]. A comprehensive comparison of these two sequencing methods and the bisulfite sequencing methods in profiling DNA methylation can be found in Harris et al. (2010) [13]. Unlike the bisulfite-based methods, enrichment-based sequencing methods typically only allow for binary calls of methylation. For general NGS data obtained from enrichment-based methods without bisulfite conversion, there is a lack of formal statistical tools for detecting ASM. In this paper, we propose a novel Bayesian approach to screen single nucleotide polymorphisms (SNPs) with ASM by incorporating SNP genotyping information with the sequencing data. Our approach separately models the data from biological samples with different genotypes. For each SNP, a Poisson-Binomial model or a Negative-Binomial-Binomial model with genotype-dependent parameters is applied to the short reads overlapping that SNP locus, leading to a posterior probability of methylation. The overall posterior probability of ASM for that SNP is obtained by combining the methylation probabilities across different genotypes. Source R code is available at <http://stat.cwru.edu/~bxh123>.

The remainder of the paper is organized as follows. In Section 2, we first introduce the motivating prostate cell line study. Then we define ASM as a probability event and propose the Bayesian approach for posterior inference. In particular, we introduce two hierarchical models, Poisson-Binomial and Negative-Binomial-Binomial, to model the short-read counts. We also describe how to screen SNPs with ASM by controlling for Bayesian false discovery rates. In Section 3, the proposed Bayesian approach is applied to the prostate cell line study, and statistical inference of ASM is conducted based on the in-house MBD-Seq data. A simulation study is also carried out to demonstrate the quantitative performance of our method. Conclusions and discussions are provided in Section 4.

2 Methods

2.1 A Prostate Cell Line Study

We investigated ASM based on DNA-seq data obtained from four prostate cell lines, including PrEC, RWPE, LNCaP, and Du145. These cell lines were processed using the methyl-CpG binding domain protein sequencing (MBD-seq) method, which produced about 14 millions short DNA reads. These short reads were mapped to the human genome Hg18 (NCBI Build 36.1, UCSC) using the BOWTIE algorithm [14].

For SNP calling, each cell line was processed twice on an Illumina array (Illumina Genome-wide DNA analysis BeadChip Human660w-QUAD). SNPs with inconsistent calls between these replicates were removed from further analysis. We also excluded SNPs with a single

homozygous genotype in all four cell lines since only one allele was represented for these SNPs. Moreover, the method proposed by Wang et al. (2007) [15] was applied to identify genomic regions with copy number variation (CNV). To limit false discovery due to CNV, SNPs that lie in the regions identified with CNV were also excluded. After these, we surveyed a total of 56,440 SNPs and retrieved the number of uniquely mapped short reads overlapping with each SNP. The SNP coordinates were obtained from the UCSC Hg18 assembly.

In a NGS methylation study, the count of short reads mapped to a genomic region typically represents the strength of the underlying methylation signal. In general, a zero count means no methylation for that region. Among the 56,440 SNPs, 44,594 had no reads mapped in all four cell lines. This left a total of 11,846 SNPs for ASM analysis. On the other hand, a genomic region can be considered as being methylated if the tag count is large. However, there is no established threshold of the read count for calling methylation in the literature. Laurent et al. (2010) [16] assumes that three mapped reads would be sufficient for calling methylation while others prefer larger thresholds [17]. When a SNP exhibits ASM, the number of reads mapped to its two alleles should differ greatly. Generally speaking, there should be a large number of reads mapped to one allele but few reads mapped to the alternative allele. Therefore an accurate quantification of the number of reads mapped to two alleles is crucial for ASM analysis.

Degner et al. [18] found that there could be a bias towards higher mapping rate of the allele in the reference genome, which was referred to as allele-specific mapping bias (ASMB). To correct such a bias, we applied the procedures proposed in [18] to the data in our prostate cell line study. With the ASMB-correction procedures, 11,693 SNPs were found to have reads mapped in at least one cell lines and were included in further ASM analysis, of which 11,654 (99.7%) were also found when not using the correction procedures. For our prostate study, applying ASMB-correction results in minor changes. In general, we recommend ASMB-correction as a necessary pre-processing step. More details about the MBD sequencing procedure, read mapping and data pre-processing are provided in the Supplementary Materials.

2.2 Definition of ASM

Suppose that a SNP carries alleles A and B. Then three different genotypes could be observed at this SNP locus, including AA, BB and AB. If this SNP exhibits ASM, either allele A or allele B is methylated while the alternative allele is not. Such a methylation asymmetry of the two alleles would imply two consistent patterns across the three possible genotypes. More specifically, in the case of A-allele specific methylation, this SNP will be methylated in samples with genotypes AA and AB but not methylated in sample with genotype BB. Similarly, this SNP will be only methylated in samples with genotypes BB and AB in the case of B-allele specific methylation.

To define ASM as a probability event, we introduce a set of latent variables s_g that represent all possible methylation states (allele-specific or not) of the SNP in samples with genotype g , with g taking a value in $\{0,1,2\}$, corresponding to “AA”, “AB” or “BB”, respectively. Table 1 summarizes the values of s_g and the corresponding methylation states. For the homozygous genotypes (AA and BB), we only need to consider two methylation states $s_0 = s_2 = \{0, 1\}$, in which 0 for no methylation and 1 for methylation, since the SNP only carries one particular allele at the SNP locus. For the heterozygous genotype (AB), methylation may occur on neither, one, or both alleles. Therefore we should consider four methylation states $s_1 = \{0, 1, 2, 3\}$, indicating no methylation, methylation on allele A only, methylation on allele B only and methylation on both alleles, respectively.

By the definition of s_g , the event of A-allele specific methylation (A-ASM) is equivalent to the cross-genotype event,

$$\text{A-ASM} = \{s_0=1, s_1=1 \text{ and } s_2=0\}.$$

Similarly, the event of B-allele specific methylation (B-ASM) corresponds to the event $\{s_0 = 0, s_1 = 2 \text{ and } s_2 = 1\}$. Therefore inference on ASM can be made by estimating the latent triplets (s_0, s_1, s_2) when multiple samples of different genotypes are available. It is worth emphasizing that the definition of ASM in this paper is restricted to the cases where one allele is methylated but the alternative allele is not. A broader definition of ASM could include the cases where both alleles are methylated at different levels. However, this requires a precise quantification of methylation level as a continuous variable instead of a binary call of methylation for the two alleles, which can only be achieved using the bisulfite-based sequencing technologies up to date. In the following section, we describe a Bayesian approach that computes the posterior probabilities of A- and B-ASM given the sequencing and genotyping data.

2.3 A Bayesian Approach

Suppose that sequencing data are obtained for m samples. For our data $m = 4$ representing the four cell lines. For sample $j, j = 1, \dots, m$, let G_j be the genotype of the SNP, n_j be the total number of short reads overlapping with the SNP locus, and a_j and b_j be the numbers of short reads on the A allele and the B allele, respectively. Since a short read will be mapped to either allele A or allele B, we should observe $n_j = a_j + b_j$ in general. However, due to a variety of reasons such as sequencing error [19] and the existence of multiple alleles [20], the sum $a_j + b_j$ could be smaller than n_j . In these rare cases, we may restrict the analysis to the reads mapped to the two alleles only and set $n_j = a_j + b_j$.

Modeling the read counts requires specifying a joint distribution for the pair (n_j, a_j) . Since a_j is the number of reads mapped to allele A among all n_j reads, it is natural to assume a Binomial distribution for a_j given n_j . The remaining piece is to model the marginal count n_j . In the literature, Poisson and Negative Binomial distributions are the two common distributions for modeling read counts in NGS data [21]. We first discuss our approach in the context of a Poisson distribution for n_j . More specifically, conditional on the observed genotype $G_j = g$ and the underlying methylation state $s_g = s$, the joint distribution of the count pair (n_j, a_j) is given as

$$\begin{aligned} P(a_j, n_j | s_g = s) &= \underbrace{P(a_j | n_j, s_g = s)}_{\text{Binomial}} \underbrace{P(n_j | s_g = s)}_{\text{Poisson}} \\ &= \binom{n_j}{a_j} \alpha_{gs}^{a_j} (1 - \alpha_{gs})^{n_j - a_j} \times \frac{\lambda_{gs}^{n_j}}{n_j!} \exp(-\lambda_{gs}), \end{aligned} \quad (1)$$

where $\lambda_{gs} \geq 0$ is the rate parameter of the Poisson distribution and $\alpha_{gs} \in [0, 1]$ is the percentage parameter of the Binomial distribution. The likelihood function for a single sample is given by

$$L_j = \prod_{s \in \mathcal{S}_g} [P(a_g, n_g | s_g = s) q_{gs}]^{\mathbf{I}(s_g = s)},$$

where $\mathbf{I}(\cdot)$ is the indicator function, and $q_{gs} = P(s_g = s)$ represents the probability that the methylation state s_g is s . With m independent samples, the overall likelihood function is then given by

$$L = \prod_{j=1}^m L_j = \prod_{g \in \{0,1,2\}} \prod_{j: G_j = g} \prod_{s \in \mathcal{S}_g} [P(a_g, n_g | s_g = s) q_{gs}]^{\mathbf{I}(s_g = s)}. \quad (2)$$

Denote $m_g = \sum_{j=1}^m \mathbf{I}(G_j = g)$, $a_g = \sum_{j: G_j = g} a_j$ and $n_g = \sum_{j: G_j = g} n_j$ for a fixed g , the joint distribution of the read counts $\mathbf{a} = (a_0, a_1, a_2)$, $\mathbf{n} = (n_0, n_1, n_2)$ and the methylation states $\mathbf{S} = (s_0, s_1, s_2)$ can be re-written as

$$P(\mathbf{a}, \mathbf{n}, \mathbf{S}) = \prod_{g \in \{0,1,2\}} \prod_{s \in \mathcal{S}_g} [P(a_g, n_g | s_g = s) q_{gs}]^{\mathbf{I}(s_g = s)}, \quad (3)$$

with $P(a_g, n_g | s_g = s) \propto \alpha_{gs}^{a_g} (1 - \alpha_{gs})^{n_g - a_g} \lambda_{gs}^{n_g} \exp(-m_g \lambda_{gs})$.

The goal is the posterior probability $P_{gs} = P(s_g = s | n_g, a_g)$. A full Bayesian implementation requires elicitation and calibration of prior distributions for the model parameters, including q_{gs} , α_{gs} and λ_{gs} . Computation cost will increase due to the need of Markov chain Monte Carlo simulations. Since our goal is to perform a fast and preliminary screen for ASM, we delay investigation of the full Bayesian approach for future research. Considering the large data size from NGS and the informative nature of the methylation biology, we decide to simplify the computation by eliciting the values of most parameters based on biological reasoning. Firstly, we assume a Dirichlet prior distribution for the probability vector $\bar{q}_g = (q_{gs}; s \in \mathcal{S}_g)^T$, that is, $\pi(\bar{q}_g) = \frac{1}{B(d_g)} \prod_{s \in \mathcal{S}_g} q_{gs}^{d_{gs}-1}$ with $B(d_g) = \Pi_s \Gamma(d_{gs}) / \Gamma(\sum_{s \in \mathcal{S}_g} d_{gs})$. The Dirichlet parameters $d_g = (d_{gs}; s \in \mathcal{S}_g)^T$ could be chosen to yield a small effect size by imposing the constraint $\sum_{s \in \mathcal{S}_g} d_{gs} = e$ for $e \in (0, 1)$ [22]. For example, we set equal d_{gs} with $\sum_{s \in \mathcal{S}_g} d_{gs} = 0.5$ in our numerical analyses. Similar numerical results (not shown) were obtained by randomly generating d_{gs} in the simulations.

Secondly, we specify the value of α_{gs} as given in Table 2, categorized by genotypes and methylation states. The justification of these values is summarized as follows. When ASM occurs at a SNP, the methylated allele will dominate in the sense that a high percentage of short reads will be mapped to the methylated allele. Since α_{gs} reflects the proportion of the short reads mapped to the A allele, we set α_{gs} at 0.9 when either the genotype is AA (i.e., there is no allele B) or only the A allele is methylated. In contrast, α_{gs} takes a small value of 0.1 when either the genotype is BB or only allele B is methylated. When the genotype is AB and there is no methylation, we assume a uniform distribution $U[0, 1]$ for α_{10} . When the genotype is AB and both alleles are methylated, we set $\alpha_{13} = 0.5$. A sensitivity analysis is later carried out to examine the specification of α_{gs} on screening SNPs with ASM.

Finally, to determine the rate parameter λ_{gs} that characterizes the strength of the methylation signal when the methylation state is s , we suggest an empirical approach based on the observed data. Recall that the state indicator $s = 0$ implies that the SNP is not methylated, and a positive value of s implies methylation. Thus λ_{g0} represents the mean read count when the SNP in the samples showing genotype g is not methylated. Its value should be small since if the SNP is not methylated, few reads will be mapped to its genomic location. In contrast, when s is positive, the SNP is methylated and λ_{gs} should be large. In summary, the mean read count could be assumed to follow a mixture of two distributions,

one centered at a small value around 0 and the other at a large positive value. Therefore we fit a two-component mixture Poisson model to the read counts by pooling n_j for all SNPs considered in the study. The density of the mixture Poisson model can be written as

$$f(n_j) = p\lambda_N^{n_j} \exp(-\lambda_N) + (1-p)\lambda_M^{n_j} \exp(-\lambda_M), \quad (4)$$

where p is the mixing weight; λ_N is the mean count for the unmethylated SNPs and $\lambda_M > \lambda_N$ is the mean count for the methylated SNPs. In other words, we assume that n_j follows a Poisson distribution with mean λ_N if the SNP is not methylated, and follows a Poisson distribution with mean λ_M when it is methylated. Fitting the mixture Poisson models has been implemented in most standard packages. In this paper, we used the R package FlexMix [23]. After obtaining the maximum likelihood estimates for λ_N and λ_M in (4), we set $\lambda_{g0} = \hat{\lambda}_N$ and $\lambda_{gs} = \hat{\lambda}_M$ for any $s > 0$. Here we apply the two-component mixture Poisson model to make binary calls of methylation based on the count data, which is widely used for general enrichment-based NGS data [13]. For bisulfite-based NGS data where the read counts accurately represent the underlying methylation level, another strategy could be to assume three Poisson distributions for the data, where the smallest λ denotes the mean read count for no methylation, the highest for two-allele methylation, and the middle value for one-allele methylation. These λ parameters can be then estimated by fitting a three-component mixture Poisson model.

Given these prior specifications, a closed form of the posterior probability $P_{gs} = P(s_g = s | n_g, a_g)$ exists as

$$P_{gs} = \frac{\alpha_{gs}^{a_g} (1 - \alpha_{gs})^{n_g - a_g} \lambda_{gs}^{n_g} \exp(-m_g \lambda_{gs}) d_{gs}}{\sum_{t \in S_g} \alpha_{gt}^{a_g} (1 - \alpha_{gt})^{n_g - a_g} \lambda_{gt}^{n_g} \exp(-m_g \lambda_{gt}) d_{gt}}. \quad (5)$$

The derivation of (5) are provided in the Supplementary Materials. The values of λ_{gs} are as discussed above. α_{gs} should be substituted, according to the values in Table 2. The calculation of P_{10} in (5) integrates out α_{10} over its uniform prior distribution $U(0, 1)$, leading to

$$P_{10} = \frac{B(a_1, n_1 - a_1) \lambda_{10}^{n_1} \exp(-m_1 \lambda_{10}) d_{10}}{B(a_1, n_1 - a_1) \lambda_{10}^{n_1} e^{-m_1 \lambda_{10}} d_{10} + \sum_{t \neq 0} \alpha_{1t}^{a_1} (1 - \alpha_{1t})^{n_1 - a_1} \lambda_{1t}^{n_1} e^{-m_1 \lambda_{1t}} d_{1t}},$$

where $B(\cdot, \cdot)$ is the Beta function. By the definition of A- and B-ASM, the triplets (P_{01}, P_{11}, P_{20}) and (P_{00}, P_{12}, P_{21}) are the posterior probabilities of interests.

Negative Binomial distribution is a popular alternative to the Poisson distribution for modeling NGS count data. The Negative Binomial distribution was proved to fit better to chip-sequencing (chip-seq) data, especially when there is over-dispersion in the count data [24]. Regarding DNA-seq methylation data, Yang et al. [21] showed that the Negative-Binomial distribution yielded more conservative results. Under the probability framework of our Bayesian approach, we can replace the Poisson density in (1) with a Negative-Binomial density, leading to

$$P(a_j, n_j | s_g = s) = \underbrace{P(a_j | n_j, s_g = s)}_{\text{Binomial}} \times \underbrace{P(n_j | s_g = s)}_{\text{Negative Binomial}} \quad (6)$$

$$= \binom{n_j}{a_j} \alpha_{gs}^{a_j} (1 - \alpha_{gs})^{n_j - a_j} \times \frac{\Gamma(\varphi_{gs} + n_j)}{\Gamma(\varphi_{gs}) n_j!} (1 - \theta_{gs})^{\varphi_{gs}} \theta_{gs}^{n_j},$$

where $\theta_{gs} = \frac{\mu_{gs}}{\varphi_{gs} + \mu_{gs}}$; μ_{gs} and $\varphi_{gs} > 0$ are the mean and dispersion parameters of the Negative-Binomial distribution, respectively. The variance of the Negative-Binomial distribution is $\mu_{gs} + \frac{\mu_{gs}^2}{\varphi_{gs}}$. This model is called a Negative Binomial-Binomial (NB-Binomial) model throughout this paper.

Following the derivation of (5), the posterior probability P_{gs} for the NB-Binomial model is

$$P_{gs} = \frac{\prod_{j:G_j=g} P(a_j, n_j | s_g = s) d_{gs}}{\sum_{t \in S_g} \prod_{j:G_j=g} P(a_j, n_j | s_g = t) d_{gt}}, \quad (7)$$

where the product $\prod_{j:G_j=g} P(a_j, n_j | s_g = t)$ is proportional to

$$\alpha_{gt}^{a_g} (1 - \alpha_{gt})^{n_g - a_g} (1 - \theta_{gt})^{m_g \varphi_{gt}} \theta_{gt}^{n_g} \frac{\prod_{j:G_j=g} \Gamma(\varphi_{gt} + n_j)}{\Gamma(\varphi_{gt})^{m_g}}.$$

Similar to the mixture Poisson model (4), μ_{gs} and φ_{gs} in (7) can be determined by fitting a two-component mixture Negative-Binomial model to the pooled counts n_j 's, where the density $f(n_j)$ is

$$p \frac{\Gamma(\varphi_N + n_j)}{\Gamma(\varphi_N) n_j!} (1 - \theta_N)^{\varphi_N} \theta_N^{n_j} + (1 - p) \frac{\Gamma(\varphi_M + n_j)}{\Gamma(\varphi_M) n_j!} (1 - \theta_M)^{\varphi_M} \theta_M^{n_j} \quad (8)$$

with $\theta_N = \frac{\mu_N}{\varphi_N + \mu_N}$ and $\theta_M = \frac{\mu_M}{\varphi_M + \mu_M}$, where p is the mixing weight; μ_M represent the mean read counts for the unmethylated and methylated SNPs, respectively; φ_N and φ_M are the two dispersion parameters. Since an unmethylated SNP should have a zero count in theory, one may assume a small μ_N and a large φ_N in fitting the mixture model (8). The posterior probability P_{gs} in (7) is then calculated by substituting the estimates for the corresponding parameters.

2.4 Probability of ASM and False Discovery Rate Control

The probability P_{gs} in (5) or (7) assesses the likelihood of the occurrence of the methylation state $s_g = s$ for genotype g . Since A-allele specific methylation is represented by ($s_0 = 1, s_1 = 1, s_2 = 0$), statistical inference for A-ASM should be based on (P_{01}, P_{11}, P_{20}). Similarly, the triplets (P_{00}, P_{12}, P_{21}) shall be inspected for the B-allele specific methylation.

Combining probabilities across the three genotypes is analogous to multi-dimension decision making [25]. Each genotype can be considered as a dimension in our setting. Since the probability P_{gs} is obtained from samples with genotype g and samples with different genotypes are independent of each other, the posterior probabilities of A-ASM and B-ASM are, respectively,

$$P^A = P_{01}P_{11}P_{20} \text{ and } P^B = P_{00}P_{12}P_{21}. \quad (9)$$

Combining the three probabilities multiplicatively is conservative in the sense that P^A and P^B become small as long as the data from one genotype show evidence against ASM. If data are missing for one or two genotypes, that is, $m_g = 0$ for some g , the calculation of P^A and P^B shall be based on the probabilities of the representing genotypes.

Since the event of ASM could be either the A-allele or the B-allele specific methylation and these two events are mutually exclusive, the overall probability of ASM is

$$P = P(\text{A-allele or B-allele specific Methylation}) = P^A + P^B. \quad (10)$$

For SNP i , $i = 1, \dots, n$, let P_i denote the probability of ASM. When screening a large number of SNPs, we need to determine an appropriate cut-off value for selecting significant SNPs with ASM, we adopt the method proposed by Newton et al. (2004) [26], which is a Bayesian analogy of FDR [27] that minimizes the false non-discovery rate subject to maintaining a fixed FDR.

For a targeted FDR of ρ , we compute the expected number of false discoveries as

$FD(k) = \sum_{i=1}^n (1 - P_i) \mathbf{I}(1 - P_i \leq k)$ for any given cut-off of k . The optimal cut-off k_ρ is the largest value among all k such that the observed FDR is smaller than ρ , that is,

$$k_\rho = \arg \max k, \text{ such that } \frac{FD(k)}{\sum_{i=1}^n (1 - P_i)} \leq \rho. \quad (11)$$

After determining the value of k_ρ in (11), we may select those SNPs with $\mathbf{I}(P_i - k_\rho) = 1$ as candidate ASM sites for further biological study. This Bayesian FDR control method has been shown to be optimal by Müller et al. (2004) [28] and Müller et al. (2007) [29] in the sense that k_ρ in (11) and the associated decision rule of $\mathbf{I}(P_i - k_\rho) = 1$ minimizes the false negative rate among all decisions with FDR less than ρ .

3 Numerical Results

3.1 Application to the Prostate Cell Line Study

A total of 11,693 SNPs were analyzed using the Bayesian screening approach described in Sections 2.2 and 2.3. Among them, 11,105 (95%) SNPs have heterozygous genotypes represented, and 3,513 (30%) have all three possible genotypes represented in these cell lines. For the SNPs with heterozygous genotypes, Figure 1 shows a series of box plots of the average read count per cell line stratified by the proportions of the reads mapped to the A alleles. It appears that those SNPs with large read counts and percentages either higher than 80% or less than 20% are potential candidates with ASM since a large count means methylation and a proportion close to one or zero indicates that methylation is associated with one specific allele.

We first applied the Poisson-Binomial model to the data. The probabilities of ASM in (10) were shown in Figure 2a. At a Bayesian FDR of 0.05, a total of 269 (2.3%) SNPs were selected as candidates with ASM. Note that without the ASMB correction procedure, the Poisson-Binomial model selected a total of 272 SNPs (96% overlap). The behavior of the Bayesian FDR procedure was shown in Figure 2b. For the selected SNPs, the percentages of the short reads on the methylated alleles are at least 75% (Figure 3). The average read counts

per cell line show a negative correlation with the percentages of the reads on the methylated alleles. If more than 90% of the short reads are mapped to the methylated allele, a median of 4 short reads per sample is required for calling ASM. The median count of short reads per sample increases to about 7 if the percentage lies in [80%, 90%], and to about 10 if the percentage is even lower than 80%.

The histogram in Figure 4a shows the distribution of the pooled counts of short reads mapped to all SNPs, where 53% of the counts are zero. For all the positive counts, the mean is 3.3, and the variance is 30.9. The much larger variance indicates that there might be over-dispersion in the data. We then applied the NB-Binomial model to the data. At a Bayesian FDR of 0.05, a total of 257 SNPs were detected, and 235 (91%) of them also appear in the set of SNPs identified using the Poisson-Binomial model. 22 SNPs were only detected by the NB-Binomial model while 34 SNPs were only detected by the Poisson-Binomial model. Figure 4b compares the estimated probabilities of ASM from the Poisson-Binomial and NB-Binomial models for the SNPs selected by either or both models. The majority of the probabilities from the two models are highly correlated, with the NB-Binomial model yielding slightly less calls of ASMs than the Poisson-Binomial model.

Figure 5 illustrates the sequencing and genotyping data of two SNPs selected by both models. For SNP rs3765696 located on chromosome 1, the genotype is heterozygous (CT) in the PrEC cell line and is homozygous (TT) in other three cell lines. In the PrEC cell line, 10 short reads overlap with this SNP, of which 9 (90%) are on the C allele and only 1 (10%) is on the T allele. This suggests that allele C is more prone to methylation than allele T. The data in other three cell lines are consistent, where only 2 reads are mapped at the SNP locus. The SNP rs2236998 is on chromosome 4 and is represented by all three genotypes in the four cell lines. In the PrEC cell line (genotype AG), 8 (80%) reads are on the G allele and 2 (20%) are on the A allele at the SNP locus, indicating G-allele specific methylation. In the LNCaP cell line with genotype GG, 7 reads are all on the G allele. In the RWPE and Du145 cell lines with genotype AA, only 1 read is mapped. Thus the data are consistent across these cell lines with the final call of G-ASM based on posterior inference. We examined more SNPs and found similar patterns, suggesting a reasonable fit of our proposed model.

3.2 A Simulation Study

In this simulation study we considered 10,000 SNPs, each replicated in 4 samples. The first stage of the simulation was to simulate the genotypes across multiple samples and the true methylation pattern across the three possible genotypes for each SNP. Due to the small sample size of 4, some SNPs may not have all three genotypes represented as in the prostate cell line study. Therefore, for genotype AA, we considered three cases: (1) genotype missing; (2) allele A methylated and (3) allele A not methylated. Similarly, we considered three cases for genotype BB and five cases for genotype AB. The total number of cross-genotype methylation pattern is then $3 \times 3 \times 5 = 45$. The proportion corresponding to each pattern was determined to approximate the prostate cell line data, and only 28 patterns were given positive proportions (see Table S2 in the Supplementary Materials). For each SNP, its methylation pattern across different genotypes was then randomly chosen from the list of 28 patterns according to the listed proportions.

Given a simulated methylation pattern, the second stage was to simulate the read counts. The total number of short reads overlapping with each SNP was generated from a Poisson or a Negative Binomial distribution, where the distribution parameters depend on the simulated genotype and methylation state. In the case of Poisson distribution, the total read count was generated from Poisson(0.5) regardless of the genotypes if a sample was assumed a non-methylation state. Otherwise, the total read count was generated from Poisson(λ_M). We varied the value of λ_M from 3 to 9 to examine the quantitative performance of our approach

in calling ASM. Given the total read count, a Binomial distribution with a percentage parameter 0.9 was assumed to generate the count of reads on the corresponding allele when the genotype is homozygous. For a heterozygous genotype, the binomial percentage parameter was 0.5, 0.9, 0.1 and 0.5, respectively, for the four methylation states considered in Table 1.

In the simulation study, the sensitivity and specificity of calling ASM were calculated at the Bayesian FDR of 0.05. We simulated the data 200 times for each λ_M . As the results, the sensitivity (average across 200 simulations) gradually increases with λ_M , the Poisson rate of the total read count for methylated samples (Figure 6a). The sensitivity is only 0.32 when λ_M is 3, and it becomes greater than 0.8 when λ_M is at least 7, which suggests that a read count of 7 is required to satisfactorily separate methylated SNPs from unmethylated ones. Meanwhile the specificity is very close to one in all simulation scenarios (Figure 6b), which confirms that combining probabilities across different genotypes multiplicatively is conservative in the sense that it reduces the false discovery rate. Table 3 summarizes the false positive and false negative rates in the simulations. The false positive rates are all smaller than the nominal level of 0.05 as expected, while the false negative rates decrease with the value of λ_M .

In the data-generation process, the Binomial percentage parameter α_{gs} was consistent with the suggested values in Table 2. To examine the effect of possible misspecification of the operating parameter α_{gs} on SNP screening, a further sensitivity analysis was performed. Recall that the value of α_{gs} was set at 0.9 to reflect the proportion of short reads on the methylated allele when ASM occurs. In the sensitivity analysis, we varied this value from 0.70 to 0.95. High sensitivity of calling ASM is achieved when $\alpha_{gs} = 0.8, 0.85$ and 0.90 (Figure 7a). The specificity is very close to one in all scenarios, with best results at 0.95 and slightly unsatisfactory results at 0.75, 0.80 and 0.85. In summary, we recommend a value of 0.9 for α_{gs} .

To examine the NB-Binomial model, the total read count was generated from NB(0.5, 100) regardless of the genotype if a sample had a non-methylation state. For the methylated samples, the total read count was generated from NB($\mu_M, 3$). Similar to the case of Poisson distribution, we examined the model performance by varying the value of μ_M from 3 to 9. The dispersion parameter was fixed at 3 so that the variance of the read count for the methylated SNPs ranged from 6 to 30. The sensitivity ranges from 0.25 to 0.82 (Figure 8a) as μ_M goes from 3 to 9, while the specificity is still close to one in all scenarios (Figure 8b). The sensitivity of detecting ASM is generally lower than the case of Poisson distribution, which suggests a higher coverage is needed for ASM detection when the count data present over-dispersion.

4 Discussion

We have proposed a Bayesian approach for screening SNPs with allele specific methylation based on enrichment-based NGS data and genotyping information from multiple samples. Probabilities of methylation are first derived for each genotype, and the overall probability of ASM is then derived by combining the probabilities of methylation across three possible genotypes multiplicatively. Our screening approach is based on a class of hierarchical models for the numbers of reads mapped to each SNP and its two alleles. A Poisson or Negative Binomial model is used to model the total read counts for a binary call of methylation, and a Binomial model is used to determine whether methylation is associated with one specific allele. We applied our approach to the data from a prostate cell line study, leading to 269 SNPs as candidates of ASM. The quantitative performance of the proposed models was also examined in a simulation study. If the count data can be well approximated

by a Poisson distribution, satisfactory performance of calling ASM can be achieved when the total read counts for methylated SNPs are at least 7, that is, sequencing depth greater than 7. If there is strong evidence of over-dispersion in the data and the Negative Binomial distribution has to be used, satisfactory sensitivity of calling ASM requires a larger sequencing depth (at least 9).

Our method has the merit of easy and fast computation. After obtaining the read counts from the sequencing data, the probabilities of ASM can be readily computed for multiple SNPs in a parallel fashion since all parameters in the Bayesian models are either pre-specified based on biology background or estimated based on count data before calculating the probabilities of ASM. It took about 15 minutes to run the R script for the prostate cell data on a Mac computer with a 2.4 GHz CPU and 4GB memory.

Our method defines genotype-dependent methylation states, and considers ASM as a cross-genotype event. Inference for ASM combines the posterior probabilities of methylation for different genotypes multiplicatively. An alternative is to set up a hierarchical model and define four overall methylation states for each SNP: no methylation, A-ASM, B-ASM and allele-free methylation. Such a model, when properly constructed coherently, is conceptually appealing. However, for practical data with high noise and often conflicting observations across genotypes, the alternative model could be vulnerable and sensitive to modeling assumptions, and deserves further research. Another extension would be to simultaneously estimate the model parameters and compute the posterior probabilities of ASM. Skelly et al (2011) [30] proposed a Bayesian hierarchical model to profile allele specific gene expression (ASE) based on RNA-seq data. ASE is also a form of allele asymmetry, where the two alleles have differential expression levels. RNA-seq data are similar to bisulfite-based DNA methylation sequencing data. The general hypothesis of ASE is that the proportion of the reads mapped to one allele differs from 0.5. Therefore the model in [30] can be adapted to ASM detection when bisulfite-based methylation data are available. On the other hand, enrichment-based methylation data do not quantify methylation level as RNA-data do for gene expression, since they often require a binary call of methylation, which is represented by the latent variable s_g in our model. To develop a full Bayesian approach, we need to consider non-degenerated prior distributions for the parameters, including α_{gs} and λ_{gs} for the Poisson-Binomial model (or μ_{gs} and ϕ_{gs} for the NB-Binomial model) for $g = 0, 1, 2$ and $s_g = s$. Such a full Bayesian approach would require careful elicitation of the hyper-parameters in the prior distributions, and the computation cost will increase due to the need of Markov chain Monte Carlo simulations for inference. Since our goal is to perform a fast screen for ASM, we delay the investigation of a full Bayesian approach for future research.

DNA methylation is known to occur mainly in CpG islands [31] or island shores [32]. There are 225,659 known SNPs on the CpG sites. For each SNP identified with ASM, bisulfite sequencing may be applied to further validate the methylation state of each CpG site within a local genomic region around that SNP. Zhang et al (2009) [33] also observed that ASM may occur in genomic regions that do not contain any SNPs based on bisulfite sequencing data. To identify regions unmarked by SNPs based on enrichment-based NGS data, it requires knowing the count of short reads overlapping with each CpG site on the two alleles. However, the sequencing reads currently available are generally not long enough so that for some short reads that overlap with CpG sites, whether they are mapped to A allele or B allele is unknown. For example, for a genomic region marked by a SNP, our model is not able to make use of the short reads that are just one base-pair way from this SNP. The rapid development of the NGS technologies is expected to solve these issues and our future research will include extending the current method to determine arbitrary genomic regions with allele specific methylation. Another future direction is to account for the effect of copy

number variation (CNV). An underlying assumption for our model is that read counts are mainly driven by methylation signal but not CNV. In the prostate cell line study, the method proposed by Wang et al. (2007) [15] was applied to identify genomic regions with possible CNV. To limit false discovery due to CNV, SNPs that lie in the regions identified with CNV were not considered in the analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank the reviewers and the Associated Editor for their helpful comments, which significantly improve the quality of this manuscript. The work is partially funded by National Cancer Institute (CA154356).

References

1. Jaenisch R, Bird A. Epigenetic regulation of gene expression: How the genome integrates intrinsic and environmental signals. *Nature Genetics*. 2003; 33:245–54. [PubMed: 12610534]
2. Robertson KD. DNA methylation and human disease. *Nat Rev Genet*. 2005; 6:597–610. [PubMed: 16136652]
3. Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. *Nature*. 2007; 447:433–40. [PubMed: 17522677]
4. Bird A. DNA methylation patterns and epigenetic memory. *Genes & development*. 2002; 16(1):6–21. [PubMed: 11782440]
5. Kerkel K, Spadola A, Yuan E, Kosek J, Jiang L, Hod E, Li K, Murty VV, Schupf N, Vilain E, Morris M, Haghighi F, Tycko B. Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nature genetics*. 2008; 40:904–8. [PubMed: 18568024]
6. Tycko B. Allele-specific DNA methylation: beyond imprinting. *Human molecular genetics*. 2010; 19:R210–20. [PubMed: 20855472]
7. Shoemaker R, Deng J, Wang W, Zhang K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome research*. 2010; 20:883–9. [PubMed: 20418490]
8. Zhang Y, Rohde C, Tierling S, Jurkowski TP, Bock C, Santacruz D, Ragozin S, Reinhardt R, Groth M, Walter J, Jeltsch A. DNA methylation analysis of chromosome 21 gene promoters at single base pair and single allele resolution. *PLoS genetics*. 2009; 5:e1000438. [PubMed: 19325872]
9. Hirst M, Marra MA. Next generation sequencing based approaches to epigenomics. *Briefings in functional genomics*. 2010; 9(5–6):455–65. [PubMed: 21266347]
10. Down TA, Rakyen VK, Turner DJ, Flicek P, Li H, Kulesha E, Gärf S, Johnson N, Herrero J, Tomazou EM, Thorne NP, Bäckdahl L, Herberth M, Howe KL, Jackson DK, Miretti MM, Marioni JC, Birney E, Hubbard TJ, Durbin R, Tavare S, Beck S. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nature biotechnology*. 2008; 26(7):779–85.
11. Jacinto FV, Ballestar E, Esteller M. Methyl-DNA immunoprecipitation (MeDIP): hunting down the DNA methylome. *BioTechniques*. 2008; 44(1):35–43. [PubMed: 18254377]
12. Serre D, Lee BH, Ting AH. MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic acids research*. 2010; 38:391–9. [PubMed: 19906696]
13. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao Y, Olshen A, Ballinger T, Zhou X, Forsberg KJ, Gu J, Echipare L, O'Geen H, Lister R, Pelizzola M, Xi Y, Epstein CB, Bernstein BE, Hawkins RD, Ren B, Chung WY, Gu H, Bock C, Gnirke A, Zhang MQ, Haussler D, Ecker JR, Li W, Farnham PJ, Waterland RA, Meissner A, Marra MA, Hirs M, Milosavljevic A, Costello JF. Comparison of sequencing-based methods to

profile DNA methylation and identification of monoallelic epigenetic modifications. *Nature biotechnology*. 2010; 28(10):1097–105.

14. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*. 2009; 10:R25. [PubMed: 19261174]
15. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research*. 2007; 17(11):1665–74. [PubMed: 17921354]
16. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J, Wei CL. Dynamic changes in the human methylome during differentiation. *Genome research*. 2010; 20:320–31. [PubMed: 20133333]
17. Zhang Y, Jeltsch A. The Application of Next Generation Sequencing in DNA Methylation Analysis. *Genes*. 2010; 1:85–101.
18. Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*. 2009; 25(24):3207–12. [PubMed: 19808877]
19. Zaranek AW, Levanon EY, Zecharia T, Clegg T, Church GM. A survey of genomic traces reveals a common sequencing error, RNA editing, and DNA editing. *PLoS genetics*. 2010; 6:e1000954. [PubMed: 20531933]
20. LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research*. 2009; 37:4181–93. [PubMed: 19570852]
21. Yang Y, Wang W, Li Y, Tu J, Bai Y, Xiao P, Zhang D, Lu Z. Identification of methylated regions with peak search based on Poisson model from massively parallel methylated DNA immunoprecipitation-sequencing data. *Electrophoresis*. 2010; 31(21):3537–44. [PubMed: 20925052]
22. Zhu M, Lu AY. The counter-intuitive noninformative prior for the Bernoulli family. *Journal of Statistical Education*. 2004; 12:1–10.
23. Leisch F. FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*. 2004; 11(8)
24. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature biotechnology*. 2008; 26(11):1293–300.
25. Triantaphyllou, E. Multi-Criteria Decision Making: A Comparative Study. Dordrecht, The Netherlands: Kluwer Academic Publishers (Springer); 2000.
26. Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*. 2004; 5:155–76. [PubMed: 15054023]
27. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*. 1995; 57:289–300.
28. Müller P, Parmigiani G, Robert C, Rousseau J. Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*. 2004; 99:990–1001.
29. Müller, P.; Parmigiani, G.; Rice, K. FDR and Bayesian multiple comparisons rules. In: Bernardo, J., et al., editors. *Bayesian Statistics*. Vol. 8. Oxford: Oxford University Press; 2007.
30. Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome research*. 2011; 21(10):1728–37. [PubMed: 21873452]
31. Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences of the United States of America*. 2002; 99(6): 3740–5. [PubMed: 11891299]
32. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, Cui H, Gabo K, Rongione M, Webster M, Ji H, Potash JB, Sabunciyan S, Feinberg AP. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shore. *Nature genetics*. 2009; 41(2):178–86. [PubMed: 19151715]

33. Zhang Y, Rohde C, Reinhardt R, Voelcker-Rehage C, Jeltsch A. Non-imprinted allele-specific DNA methylation on human autosomes. *Genome biology*. 2009; 10(12):R138. [PubMed: 19958531]

Figure 1.

Box plots of the average number of short reads per cell line stratified by the proportions of reads on the A alleles for the 11,105 SNPs with heterozygous genotypes

Figure 2.
 (a) Fitted probabilities of ASM for all SNPs using the Poisson-Binomial model (b) Bayesian FDR performance

Figure 3. Read counts and percentages of the reads on the methylated alleles for the selected SNPs using Poisson-Binomial model

Figure 4.

(a) Histogram of total read counts (n_s) mapped to all SNPs from 4 cell lines; (b) Fitted probabilities of ASM for SNPs selected using Poisson-Binomial and NB-Binomial model

Figure 5. MBD sequencing and genotyping data of two selected SNPs, rs3765696 and rs2236998. (PrEC, RWPE, NCaP, DU145) are the four prostate cell lines in our MBD-Seq experiment

Figure 6. Simulation results (Poisson-Binomial model): (a) Sensitivity of calling ASM with respect to λ_M ; (b) Specificity of calling ASM with respect to λ_M

Figure 7. Sensitivity (a) and specificity (b) of calling ASM with respect to the operating parameter α_{gs} in the Poisson-Binomial model

Figure 8. Simulation results (NB-Binomial model): (a) Sensitivity of calling ASM with respect to μ_M ; (b) Specificity of calling ASM with respect to μ_M

Table 1Definition of the latent variable s_g by genotypes

Methyl. State	Genotype		
	AA ($g = 0$)	AB ($g = 1$)	BB ($g = 2$)
	s_0	s_1	s_2
No Methyl.	0	0	0
A Methyl. B Not	1	1	0
B Methyl. A Not	0	2	1
A and B both Methyl.	1	3	1

Table 2

Values of α_{gs} in the Poisson-Binomial and NB-Binomial models by genotypes and methylation states

Genotype	Methylation state	
AA ($g = 0$)	No Methyl. ($s_0 = 0$)	$\alpha_{00} = 0.9$
	Methyl. ($s_0 = 1$)	$\alpha_{01} = 0.9$
BB ($g = 2$)	No Methyl. ($s_2 = 0$)	$\alpha_{20} = 0.1$
	Methyl. ($s_2 = 1$)	$\alpha_{20} = 0.1$
AB ($g = 1$)	No Methyl. ($s_1 = 0$)	$\alpha_{10} \sim U(0, 1)$
	A Methyl. ($s_1 = 1$)	$\alpha_{11} = 0.9$
	B Methyl. ($s_1 = 2$)	$\alpha_{12} = 0.1$
	A and B Methyl. ($s_1 = 3$)	$\alpha_{13} = 0.5$

Table 3

False positive and false negative rates (means (standard deviations)) of calling ASM with respect to λ_M (Poisson-Binomial model) in the simulations

λ_M	False positive rate	False negative rate
3	0.041 (0.008)	0.111 (0.002)
4	0.040 (0.006)	0.079 (0.002)
5	0.036 (0.005)	0.050 (0.002)
6	0.031 (0.004)	0.031 (0.002)
7	0.029 (0.004)	0.018 (0.002)
8	0.035 (0.004)	0.011 (0.001)
9	0.033 (0.004)	0.006 (0.001)