BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 2 2014, pages 172–179
doi:10.1093/bioinformatics/btt674

Genome analysis

Advance Access publication November 21, 2013

A classification approach for DNA methylation profiling with bisulfite next-generation sequencing data

Longjie Cheng* and Yu Zhu

Department of Statistics, Purdue University, West Lafayette, IN 47906, USA

Associate Editor: Michael Brudno

ABSTRACT

Motivation: With the advent of high-throughput sequencing technology, bisulfite–sequencing-based DNA methylation profiling methods have emerged as the most promising approaches due to their single-base resolution and genome-wide coverage. However, statistical analysis methods for analyzing this type of methylation data are not well developed. Although the most widely used proportion-based estimation method is simple and intuitive, it is not statistically adequate in dealing with the various sources of noise in bisulfite-sequencing data. Furthermore, it is not biologically satisfactory in applications that require binary methylation status calls.

Results: In this article, we use a mixture of binomial model to characterize bisulfite-sequencing data, and based on the model, we propose to use a classification-based procedure, called the methylation status calling (MSC) procedure, to make binary methylation status calls. The MSC procedure is optimal in terms of maximizing the overall correct allocation rate, and the false discovery rate (FDR) and false non-discovery rate (FNDR) of MSC can be estimated. To control FDR at any given level, we further develop an FDR-controlled MSC procedure, which combines a local FDR-based adaptive procedure with the MSC procedure. Both simulation study and real data application are carried out to examine the performance of the proposed procedures. It is shown in our simulation study that the estimates of FDR and FNDR of the MSC procedure are appropriate. Simulation study also demonstrates that the FDR-controlled MSC procedure is valid in controlling FDR at a prespecified level and is more powerful than the individual binomial testing procedure. In the real data application, the MSC procedure exhibits an estimated FDR of 0.1426 and an estimated FNDR of 0.0067. The overall correct allocation rate is >0.97. These results suggest the effectiveness of our proposed procedures.

Availability and implementation: The proposed procedures are implemented in R and are available at http://www.stat.purdue.edu/~cheng70/code.html.

Contact: cheng70@purdue.edu

Supplementary information: Supplementary data are is available at *Bioinformatics* online.

Received on May 20, 2013; revised on October 17, 2013; accepted on November 17, 2013

1 INTRODUCTION

DNA methylation is an epigenetic modification to the genomic DNAs by adding the methyl group to some C-5 positions of DNA sequences. It plays a crucial role in a variety of biological processes including cell development, imprinting and X-chromosome inactivation. It is prevalent at CpG positions with 60–90% of all CpGs being methylated in mammals, whereas it is much less frequent at non-CpG sites with only <3% of non-CpGs being found to be methylated. Unmethylated CpGs tend to cluster in small regions of DNA sequences called CpG islands, most of which coincide with promoter regions of many genes. The link between abnormal DNA methylation pattern and cancer is 2-fold (Robertson, 2005). First, a global hypomethylation is associated with genomic instability and is a common characteristic of cancer cells. Second, hypermethylation of CpG islands located at gene promoters results in suppression of gene expression and is conventionally observed in cancer cells. Therefore, it is desirable to reveal both genome-wide and promoter-specific DNA methylation patterns of a cell.

Various methods for genome-wide DNA methylation detection have been developed in the past 20 years. They can be classified into three categories, which are methylation-sensitive enzyme-based methods, enrichment-based methods and bisulfite conversion-based methods (Laird, 2010). In the past, array-based techniques, such as microarray technology, were the leading platforms to be combined with methods from all those three categories to survey methylation status. Although the application of these platforms enables comprehensive DNA methylation profiling at economical cost, they can only interrogate C sites of given regions at moderate resolution. Recently, next-generation sequencing (NGS) technology has been incorporated into all three categories of methods for genome-wide methylation profiling. Despite the fact that NGS-equipped methods have relative advantages over array-based methods, those methods from the first two categories are still subject to the same weaknesses they had when coupled with array-based techniques. More specifically, methylation-sensitive enzyme-based methods equipped with NGS technology remain restricted to the recognition sites of the particular enzymes used, and enrichment-based methods equipped with NGS technology do not overcome the disadvantage of moderate resolution. On the other hand, bisulfite conversion-based methods coupled with NGS technology, designated as bisulfite-sequencing methods, have emerged as the most promising methods, as they generate whole-genome DNA methylation profiles at single-base resolution. Among all bisulfite-sequencing methods, MethylC-Seq and reduced representation bisulfite sequencing (RRBS) are the two most popularly used methods.

In MethylC-Seq, genomic DNAs are first sonicated into smaller fragments. After going through end-repair and adapter ligation, these fragments are treated with bisulfite. The bisulfite treatment converts unmethylated cytosines into uracils and

*To whom correspondence should be addressed.

leaves methylated cytosines unchanged. Subsequent PCR amplification process further replaces uracils with thymines. These PCR-amplified fragments are then subject to standard sequencing technology to produce short sequencing reads, which are mapped back to the reference genome. Thus the unmethylated cytosines are distinguishable from methylated cytosines by examining sequencing reads (The workflow for MethylC-Seq experiment is given in Supplementary Material Section 1.) RRBS uses the same mechanism as MethylC-Seq. The major difference between RRBS and MethylC-Seq occurs in the first step, i.e. the way genomic DNAs are fragmented. In RRBS, genomic DNAs are digested with MspI, an enzyme which cuts all CCGG sites. These two methods have their relative advantages and disadvantages, which make them suitable for different research purposes. By the way genomic DNAs are digested in RRBS, CpG regions are substantially enriched in DNA fragments after size selection. Thus RRBS is more preferable than MethylC-Seq when the research is targeting regions with high density of CpG sites, such as CpG islands. On the other hand, because of its theoretical capacity of capturing methylation information from each C position in the whole genome, MethylC-Seq has become the golden standard for genome-wide DNA methylation analysis. As reported in Harris *et al*. (2010), when these two methods are applied to biological replicates of human embryonic stem cells, MethylC-Seq covers 95% of all CpGs, whereas RRBS shows a genome-wide CpG coverage of only 12%.

In the data generated by MethylC-Seq or RRBS, ideally there are only C reads or T reads for each covered C position of interest, depending on the methylation status. In other words, if a C position is methylated, then there should be only C reads at that site in the data; whereas if a C position is unmethylated, then there should be only T reads. However, owing to various sources of noise, in the real data generated by these two methods, there are both C reads and T reads for most of the target C sites. For instance, the process of bisulfite conversion needs to be carried out under specific experimental conditions (Smith *et al*., 2009). Failure to meet any of those conditions would lead to incomplete conversion, which further results in C reads at unmethylated C positions. Moreover, as a typical and inevitable result of applying NGS technology, there will be sequencing errors in the data, which means a small proportion of C reads will be miscalled to be T reads and vice versa. Because there are both C reads and T reads in the data, it is not straightforward to infer the true methylation status. The aim is then to make methylation call for each target C position based on the number of C reads and the number of T reads it receives, which becomes an interesting statistical problem.

In some studies concerning DNA methylation analysis, researchers use the ratio of C count to the total number of reads received at a site to quantify the methylation level at that site (Bock *et al*., 2010; Gu *et al*., 2010; Harris *et al*., 2010). The total number of reads received at a site is referred to as coverage or sequencing depth. Although this quantification approach has the virtue of being simple and straightforward, it does not use proper inference to deal with the noise in the data, and thus it is not statistically satisfactory. In other studies, researchers aim to make binary methylation calls for C positions of interest. There exist a few such approaches in the literature. In Harris *et al*. (2010), the proportion of C count at each CpG site was calculated and binary methylation call was made for each site with various choices of cutoff for the proportion. However, those choices were not statistically justified. A more sophisticated method was used in Lister *et al*. (2011), which applied a multiple testing procedure to identify methylated cytosines. In the MethylC-Seq experiment conducted by Lister *et al*. (2011), an unmethylated lambda DNA was spiked with the target genomic DNAs before sonication and was used to estimate the error rate at which a C count occurs at an unmethylated C position. As will be shown in Section 3 and Supplementary Material Section 5, the procedure used by Lister *et al*. (2011) is conservative in detecting unmethylated cytosines due to the underestimation of error rate and their choice of null hypothesis.

In this article, we propose to use a classification procedure based on a mixture of binomial model to make binary calls for methylation status. As a by-product, the memberships generated in this procedure can be used for quantifying the methylation levels. The performance of the proposed procedure can be assessed by correct allocation rates as well as false discovery rate (FDR) and false non-discovery rate (FNDR). Motivated by the concern of controlling FDR at any given level, we view our classification problem from a multiple testing perspective. Then a component based on local false discovery rate (*Lfdr*) is incorporated into the proposed classification procedure to provide an approach that is capable of controlling FDR at any prespecified level. On the basis of the Bayes rule, the proposed classification procedure is optimal in terms of maximizing the overall correct allocation rate. Simulation results demonstrate that the proposed classification procedure outperforms the conventional individual binomial testing procedure and is more accurate in detecting the true methylation status. Simulation results also show the validity of the proposed FDR-controlled procedure with various choices of FDR level. Owing to the fact that most non-CpG sites are unmethylated, when evaluating our method with real data, we focus exclusively on CpG sites. However, it is worth pointing out that our model can be applied to C positions of any context. The real dataset we use is from Lister *et al*. (2011), which was generated from a MethylC-Seq experiment. However, our method is not restricted to MethylC-Seq data, and it can also be applied to other bisulfite-sequencing data.

The rest of the article is organized as follows. Section 2 first gives the details of the mixture of binomial model and describes the proposed classification procedure, and then it presents several performance assessment methods of the proposed procedure. In the last part of Section 2, the FDR-controlled procedure is introduced. Section 3 reports the simulation results as well as the results from the real data application.

2 METHODS

2.1 Mixture of binomial model

As discussed in Section 1, MethylC-Seq experiment can roughly cover 95% of all CpGs. Those sites that do not receive any C read and T read are referred to as uncovered sites and will be excluded from methylation calling analysis. Suppose we consider M covered sites. These M sites can be the collection of all covered sites from a specific DNA segment of interest, a whole chromosome, or even the whole genome. For site i among these M sites, let X_i denote the total number of reads including both C and T reads and Y_i denote the number of C reads alone. Note

$Y_i \le X_i$. Let S_i be the indicator of the unobserved methylation status of site i, with $S_i = 0$ indicating site i is methylated and $S_i = 1$ indicating site i is unmethylated. If there is no error in the experiment, then $X_i = Y_i$ when site i is methylated, and $Y_i = 0$ when site i is unmethylated. In other words, there are only C reads for methylated sites and no C read for unmethylated sites. However, MethylC-Seq experiments are subject to both experimental errors and systematic errors. Thus, there are both C reads and T reads for most sites, or equivalently, $Y_i < X_i$ for most methylated sites and $Y_i > 0$ for most unmethylated sites.

There exist three main causes for experimental errors in MethylC-Seq experiments. First, incomplete conversion of unmethylated cytosine to uracil during bisulfite treatment results in C reads at unmethylated sites. In other words, the failure to convert unmethylated cytosine to uracil causes $Y_i > 0$ for unmethylated sites. We assume this non-conversion rate is e_{ic}, i.e. the probability that an unmethylated cytosine fails to convert to thymine. Second, overtreatment with bisulfite can lead to conversion of methylated cytosine to thymine (Laird, 2010). Suppose the miscoversion rate, or equivalently, the probability that a methylated cytosine converts to thymine is e_{mc}. Third, sequencing errors can potentially impact both methylated sites and unmethylated sites. For methylated sites, cytosines can be miscalled to be thymines and thus $Y_i < X_i$, and for unmethylated sites, bisulfite-converted thymines can be mistakenly read out as cytosines and thus $Y_i > 0$. Suppose the probability that a T read is miscalled to be a C read is e_{tc}, and the probability that a C read is miscalled to be a T read is e_{ct}. Experimental errors are unavoidable due to the random nature of sequencing technology and have to be incorporated in the model. On the other hand, systematic errors in bisulfite data can be identified and thus eliminated by carefully conducted data processing procedures. For MethylC-Seq experiment, deamination of methylated cytosine to thymine during cell development and those single nucleotide polymorphisms that a cytosine in the reference genome varies to a thymine in the sample DNA lead to systematic errors. Nevertheless, they can be detected by examining the nucleotide on the opposite strand of the C sites and thus can be eliminated from MethylC-Seq data (Laird, 2010). When they are not removed from the data, let e_{sys} denote the systematic error rate. For a more detailed review of potential sources of noise in bisulfite-sequencing data, see Krueger *et al.* (2012).

Let p_1 stand for the overall error rate for obtaining C reads at unmethylated sites caused by incomplete conversion, sequencing error and systematic errors. Similarly, let $1 - p_0$ denote the overall error rate for obtaining T reads at methylated sites caused by misconversion, sequencing error and systematic errors. It is clear that p_1 depends on e_{ic}, e_{tc} and e_{sys}, and p_0 depends on e_{mc}, e_{ct} and e_{sys}. The dependence of p_0 and p_1 on the various types of individual errors can be greatly simplified if the following three assumptions are imposed. First, there are no systematic errors in the data, i.e. $e_{sys} = 0$. Second, the two types of sequencing errors occur equally likely, which implies $e_{tc} = e_{ct}$. Third, the sample is not overtreated with bisulfite, or equivalently, $e_{mc} = 0$. Under these three assumptions, we postulate the relationship between the overall error rates and individual ones to be $p_0 = 1 - e_{tc} = 1 - e_{ct}$ and $p_1 = e_{ic} + e_{tc}$. Under the postulated relationship, if we can identify the overall error rates $1 - p_0$ and p_1, the individual error rates e_{tc}, e_{ct} and e_{ic} can also be identified. When any of the three aforementioned assumptions fail to satisfy, further information is needed to identify the various types of individual errors. Nevertheless, the overall error rates $1 - p_0$ and p_1 can still be estimated and methylation calling can be made by the procedure we will describe next. Owing to this reason, we shall use p_0 and p_1 in the rest of the article.

Based on the aforementioned discussion, we propose the following Binomial models as the conditional distribution of the C count at site i given the coverage X_i and methylation status S_i:

$$Y_i|(X_i = x, S_i = 0) \sim Bin(x, p_0);$$

$$Y_i|(X_i = x, S_i = 1) \sim Bin(x, p_1).$$

Here one important premise is that all the M sites of interest share the same error rates $1 - p_0$ and p_1. This assumption is commonly used in the literature on methylation analysis (Lister *et al.*, 2011 and Wu *et al.*, 2011).

Furthermore, suppose the proportion of methylated sites among these M sites is π, i.e. $P(S_i = 0) = \pi$ for any randomly selected site i. Then conditional on the sequencing depth at one site, the corresponding C count follows a mixture of two binomial distributions:

$$Y_i|(X_i = x) \sim \pi Bin(x, p_0) + (1 - \pi)Bin(x, p_1). \quad (1)$$

Even though MethylC-Seq data contain diverse types of errors, they are still assumed to carry information regarding the underlying methylation status in the sense that most methylated sites are dominated by C reads and most unmethylated sites are dominated by T reads. Therefore, it is reasonable to assume that p_1 and p_0 should satisfy $p_1 \ll p_0$. This assumption assures the identifiability of p_1 and p_0 and guarantees the validity of our procedure.

Suppose the coverages at the sites, i.e. X_i's, are independent and identically distributed with the same probability mass function (pmf) $f(x)$. For convenience, denote the pmf of the conditional distribution of C count of site i given $X_i = x$ defined in (1) as $g(y|x)$. For fixed i, the pmf of the joint distribution of (X_i, Y_i), denoted as $h(x, y)$, is given by $h(x, y) = g(y|x)f(x)$. Let $\phi = (p_0, p_1, \pi)$. Noticing that $f(x)$ does not involve ϕ, therefore we only need to use $g(y|x)$ for estimating ϕ.

Let $\mathbf{y} = (y_1, y_2, ..., y_M)$ be the observed C counts and $\mathbf{x} = (x_1, x_2, ..., x_M)$ the observed coverages. Then under the assumption that y_i is from a mixture of two binomial distributions given x_i, the log-likelihood function of ϕ can be written as follows.

$$l(\phi|\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{M} ln\{g(y_i|x_i)\} = \sum_{i=1}^{M} ln\{\pi g_{i0} + (1 - \pi)g_{i1}\},$$

where g_{i0} and g_{i1} are the pmfs of $Bin(x_i, p_0)$ and $Bin(x_i, p_1)$ for each i, respectively. The maximum likelihood estimate (MLE) of ϕ can be obtained by applying the well-established Expectation-maximization (EM) algorithm. However, our goal here is beyond estimating ϕ. What we want to achieve is to classify each site i to be either methylated or unmethylated on the basis of an adequate estimate of ϕ. Recall that for each i, S_i is an indicator of the true methylation status of site i with values equal to 0 or 1. Therefore, our goal is essentially to identify the value of S_i for each i.

Let θ_{i0} and θ_{i1} denote the posterior probabilities that $S_i = 0$ and $S_i = 1$, respectively, given $\mathbf{x}$, $\mathbf{y}$ and ϕ. The expressions of θ_{i0} and θ_{i1} are

$$\begin{aligned} \theta_{i0} &= P(S_i = 0|y_i, x_i, \phi) = \frac{\pi g_{i0}(y_i)}{\pi g_{i0}(y_i) + (1 - \pi)g_{i1}(y_i)}; \\ \theta_{i1} &= P(S_i = 1|y_i, x_i, \phi) = 1 - \theta_{i0}. \end{aligned} \quad (2)$$

Here θ_{i0} and θ_{i1} indicate how likely site i is methylated ($S_i = 0$) and unmethylated ($S_i = 1$), respectively, given the observed data and ϕ. Note that $\theta_{ir} (r = 0, 1)$ will play a role in the EM algorithm for computing the MLE of ϕ. In addition to facilitating the estimation of ϕ, θ_{ir} also plays a key role in the methylation status calling (MSC) procedure we will develop later.

Then the EM algorithm for computing the MLE of ϕ can be developed as follows. We start off with an initial estimate of ϕ and then compute the initial values of θ_{ir} given the initial values of ϕ. After the initial step, ϕ and θ_{ir} are iteratively updated. Conditional on the current values of θ_{ir}, we update ϕ by

$$\hat{p}_0 = \frac{\sum_{i=1}^{M} \hat{\theta}_{i0} y_i}{\sum_{i=1}^{M} \hat{\theta}_{i0} x_i}; \quad \hat{p}_1 = \frac{\sum_{i=1}^{M} \hat{\theta}_{i1} y_i}{\sum_{i=1}^{M} \hat{\theta}_{i1} x_i}; \quad \hat{\pi} = \frac{\sum_{i=1}^{M} \hat{\theta}_{i0}}{M}. \quad (3)$$

This new estimate of ϕ is then substituted back into (2) to yield new values of θ_{ir}. These two steps are repeated until certain convergence criterion is met. In our simulation study and real data application, the

convergence criterion is that the change of the log-likelihood function between two consecutive steps is below some prespecified value. A discussion of the convergence properties of EM algorithms can be found in Wu (1983). The derivation of (3) is given in Supplementary Material Section 3. Let $\hat{\phi}$ be the MLE of ϕ obtained from the EM algorithm and $\hat{\theta}_{ir}$ the estimate of θ_{ir} by plugging $\hat{\phi}$ into (2). We shall call $\hat{\theta}_{ir}$ memberships hereafter.

2.2 Classification-based MSC procedure

After the EM algorithm in the last subsection converges, we obtain the estimates $\hat{\phi}$ as well as the memberships $\hat{\theta}_{ir}$. The memberships can be further used to determine the methylation status of each site. We propose to use the following rule to make MSC. *For $i = 1, 2, ..., M$, site i is called to be methylated if $\hat{\theta}_{i0} > \hat{\theta}_{i1}$; otherwise, it is called to be unmethylated.* We shall refer to this classification procedure as the MSC procedure. Based on the Bayes rule, the MSC procedure is optimal in terms of maximizing overall correct allocation rate (McLachlan and Peel, 2000).

As mentioned in Section 1, sometimes researchers are interested in quantifying the methylation levels due to the heterogeneity of cell types or contamination during cell preparation. When the experiment is conducted on a mixture of different types of cells, it is valuable to directly use the membership $\hat{\theta}_{i0}$ to quantify the methylation level of each C position. In this case, the interpretation of the overall error rates $1 - p_0$ and p_1 is slightly different. More specifically, not only do they stand for the various types of noises caused by the bisulfite-sequencing experiment, they also represent the extent of cell type contamination. Because we are using a MethylC-Seq data of H9 human embryonic stem cells in our real data application, we will focus on the binary MSC in our article.

2.3 Performance assessment of the MSC procedure

We use individual and overall correct allocation rates to assess the performance of our proposed MSC procedure. Let Group 0 and Group 1 consist of all methylated sites and all unmethylated sites, respectively. Let M_0 and M_1 be the total number of methylated and unmethylated sites in the sample, respectively. Let M_{ij} be the number of sites that are from Group i and allocated to Group j by the MSC procedure, for $i = 0, 1$, and $j = 0, 1$. Let the total number of sites that are classified to Group 0 be U and let the total number of sites that are classified to Group 1 be V. The four possible outcomes from the proposed MSC procedure are listed in Table 1 with their corresponding frequencies.

The correct allocation rate for methylated sites (i.e. Group 0), denoted as P_0, is defined as the proportion of sites that are methylated and correctly allocated to Group 0 among methylated sites; similarly, the correct allocation rate for unmethylated sites (i.e. Group 1), denoted as P_1, is defined as the proportion of sites that are unmethylated and correctly allocated to Group 1 among unmethylated sites. The overall correct allocation rate, denoted as P, is defined as the proportion of correctly classified sites for both groups. Given Table 1, the correct allocation rates can be computed by $P_0 = \frac{M_{00}}{M_0}$, $P_1 = \frac{M_{11}}{M_1}$ and $P = \frac{M_{00}+M_{11}}{M_0+M_1}$. The quantities on the right side of these equations are unknown. Following Basford and McLachlan (1985), they can be estimated by $\hat{M}_0 = M\hat{\pi}$, $\hat{M}_1 = M(1-\hat{\pi})$, $\hat{M}_{00} = \hat{\theta}_{k0} I(\hat{\theta}_{k0} > \hat{\theta}_{k1})$ and $\hat{M}_{11} = \hat{\theta}_{k1} I(\hat{\theta}_{k0} \le \hat{\theta}_{k1})$, where $I(A)$ is an indicator of event A, such that $I(A)$ equals 1 if A is true and equals 0 otherwise. Thus P_0, P_1 and P can be estimated as follows.

$$\hat{P}_0 = \frac{1}{M\hat{\pi}} \sum_{k=1}^{M} \left\{ \hat{\theta}_{k0} I(\hat{\theta}_{k0} > \hat{\theta}_{k1}) \right\};$$

$$\hat{P}_1 = \frac{1}{M(1-\hat{\pi})} \sum_{k=1}^{M} \left\{ \hat{\theta}_{k1} I(\hat{\theta}_{k0} \le \hat{\theta}_{k1}) \right\};$$

$$\hat{P} = \frac{1}{M} \sum_{k=1}^{M} \left\{ \hat{\theta}_{k0} I(\hat{\theta}_{k0} > \hat{\theta}_{k1}) + \hat{\theta}_{k1} I(\hat{\theta}_{k0} \le \hat{\theta}_{k1}) \right\}.$$

As stated in Basford and McLachlan (1985), $\hat{P}_0 - P_0$, $\hat{P}_1 - P_1$ and $\hat{P} - P$ converge to 0 in probability when M goes to infinity. Therefore, $\hat{P}$, $\hat{P}_0$ and $\hat{P}_1$ can be used to assess the performance of the MSC procedure. Basford and McLachlan (1985) also proposed two versions of bootstrap-based methods to reduce the bias in estimating these correct allocation rates with $\hat{P}$, $\hat{P}_0$ and $\hat{P}_1$. However, we will not elaborate on the bias correction methods here. The reason is that, based on the simulation results reported in Supplementary Material Section S4, the bias of the estimated correct allocation rates for our model is hardly noticeable; see Supplementary Material Section S4 for more detail.

As mentioned in Section 1, Lister *et al.* (2011) used a multiple testing procedure to make binary methylation status calls. In their experiment, an unmethylated lambda DNA was spiked with the sample DNAs before bisulfite treatment. Because lambda DNA is known to be unmethylated, the proportion of resulting C reads at those C sites located within lambda DNA can be used to estimate non-conversion rate plus the sequencing error that T reads are miscalled to be C reads (p_1 in our case). We denote the resulting estimate as $\hat{p}_1^{Lis}$. Then the following hypothesis was tested for each C site in the whole genome simultaneously to detect methylated sites with FDR level 0.01.

$$H_{i0} : p = \hat{p}_1^{Lis} \quad vs \quad H_{ia} : p > \hat{p}_1^{Lis}.$$

Unlike the procedure used by Lister *et al.* (2011), our procedure does not need to borrow information from the unmethylated lambda DNA, instead, it can directly estimate p_1 as well as p_0 from the data.

The classification of two groups can also be viewed as a multiple testing problem once one of the groups is specified as the null (Storey, 2003). For our proposed MSC procedure, if we designate one group (e.g. methylated group) to be the null, then the FDR and FNDR can also be defined. Although the MSC procedure is optimal based on the Bayes rule, it is not ascertained that it has control over FDR, which is the most widely used criterion in multiple testing context. In the next subsection, we will view our classification approach from a multiple testing perspective. We will first show how to estimate the resulting FDR and FNDR for the MSC procedure. Then motivated by the concern that an FDR level other than the estimated FDR may be needed, we will develop an FDR-controlled MSC procedure.

Table 1. Possible outcomes from the MSC procedure and the FMSC procedure

Group	Classified as methylated (fails to reject H_0)	Classified as unmethylated (reject H_0)	Total
Group 0	M_{00}	M_{01}	M_0
Group 1	M_{10}	M_{11}	M_1
Total	U	V	M

2.4 MSC procedure with FDR control

We consider the following multiple testing problem after obtaining the estimated parameter $\hat{\phi}$ from Section 2.1:

$$H_{i0} : p = \hat{p}_0 \quad vs \quad H_{ia} : p = \hat{p}_1,$$

where $i = 1, 2, ..., M$. Because $\hat{\theta}_{i1} = 1 - \hat{\theta}_{i0}$ for any i, only $\hat{\theta}_{i0}$ are used as the test statistic and they are referred to as null memberships hereafter. It is clear that the proposed classification rule is equivalent to the testing rule that rejects H_{i0} if $\hat{\theta}_{i0} \le 0.5$. The four possible outcomes from the

MSC procedure given in Table 1 can be viewed as the four possible outcomes from the multiple testing perspective. And the frequencies for the outcomes from the aforementioned multiple testing rule are exactly the same as those for the outcomes from the MSC procedure.

By the definitions of FDR and FNDR, we have $FDR = E[\frac{M_{01}}{V}]$ and $FNDR = E[\frac{M_{10}}{U}]$. For the MSC procedure, $U = \#\{\hat{\theta}_{k0} > \hat{\theta}_{k1}\}$ and $V = \#\{\hat{\theta}_{k0} \le \hat{\theta}_{k1}\}$. Furthermore, based on the discussion in Section 2.3, we have $\hat{M}_{01} = \hat{M}_0 - \hat{M}_{00} = \sum_{k=1}^{M} \hat{\theta}_{k0} I(\hat{\theta}_{k0} \le \hat{\theta}_{k1})$ and $\hat{M}_{10} = \hat{M}_1 - \hat{M}_{11} = \sum_{k=1}^{M} \hat{\theta}_{k1} I(\hat{\theta}_{k0} > \hat{\theta}_{k1})$. Therefore, FDR and FNDR for the MSC procedure can be estimated as follows.

$$\widehat{FDR} = \frac{\hat{M}_{01}}{V} = \frac{\sum_{k=1}^{M} \hat{\theta}_{k0} I(\hat{\theta}_{k0} \le \hat{\theta}_{k1})}{\sum_{k=1}^{M} I(\hat{\theta}_{k0} \le \hat{\theta}_{k1})}; \quad (4)$$

$$\widehat{FNDR} = \frac{\hat{M}_{10}}{U} = \frac{\sum_{k=1}^{M} \hat{\theta}_{k1} I(\hat{\theta}_{k0} > \hat{\theta}_{k1})}{\sum_{k=1}^{M} I(\hat{\theta}_{k0} > \hat{\theta}_{k1})}. \quad (5)$$

Although FDR and FNDR can be estimated for the MSC procedure, this procedure cannot control FDR at an arbitrary level. In practice, it can be a concern, especially when the estimated FDR exceeds an acceptable level. Therefore, it is desirable to incorporate a FDR-controlling component into the MSC procedure. We shall investigate such a method next.

For the MSC procedure, the cutoff in the decision rule for rejecting the null hypothesis is 0.5. One way to control FDR is to adjust this cutoff according to the desirable FDR level. Suppose the prespecified FDR level is α. Then the goal here is to find a suitable cutoff c for null memberships such that the decision rule that rejects H_0 if

$$\hat{\theta}_{i0} \le c, \quad i = 1, 2, ..., M. \quad (6)$$

will have an FDR below α.

We follow an adaptive procedure developed by Sun and Cai (2007) to achieve the goal. In their original paper, Sun and Cai aimed to find a multiple testing procedure that is more efficient than the conventional *P*-value-based procedures. They first developed an *Lfdr*-based procedure for marginal FDR control and showed it is optimal in the sense that it controls marginal FDR at level α with the smallest marginal FNDR. Then they proposed a data-dependent adaptive procedure based on estimated *Lfdr* and proved that it asymptotically attains the performance of the optimal procedure. It was also demonstrated with numerical results that their adaptive procedure outperforms the conventional *P*-value-based procedures when marginal FDR is controlled at the same level. For our problem, recall that for site i, g_{i0} and g_{i1} are the pmfs of $Bin(x_i, p_0)$ and $Bin(x_i, p_1)$, respectively. The *Lfdr* of site i is given by

$$Lfdr_i = P(S_i = 0 | y_i, x_i, \phi) = \frac{\pi g_{i0}(y_i)}{\pi g_{i0}(y_i) + (1-\pi) g_{i1}(y_i)}.$$

Therefore, the null membership $\hat{\theta}_{i0}$ of site i is also an estimate of *Lfdr*. With this estimated *Lfdr* of each site, the adaptive procedure proposed by Sun and Cai (2007) can be incorporated into the MSC procedure.

Because $Fdr(z)$ is the average of $Lfdr(Z)$ for $Z \le z$ (Efron, 2007), the FDR of the decision rule (6) can be estimated by

$$\widehat{FDR}(c) = \frac{\sum_{i=1}^{M} \hat{\theta}_{i0} I(\hat{\theta}_{i0} \le c)}{\sum_{i=1}^{M} I(\hat{\theta}_{i0} \le c)}; \quad (7)$$

$$\widehat{FNDR}(c) = \frac{\sum_{i=1}^{M} (1 - \hat{\theta}_{i0}) I(\hat{\theta}_{i0} > c)}{\sum_{i=1}^{M} I(\hat{\theta}_{i0} > c)}. \quad (8)$$

When $c = 0.5$, which is the cutoff used by the MSC procedure, the resulting FDR and FNDR can be estimated by $\widehat{FDR}(0.5) = \sum_{i=1}^{M} \hat{\theta}_{i0} I(\hat{\theta}_{i0} \le 0.5) \Big/ \sum_{i=1}^{M} I(\hat{\theta}_{i0} \le 0.5)$ and $\widehat{FNDR}(0.5) = \sum_{i=1}^{M} (1 - \hat{\theta}_{i0}) I(\hat{\theta}_{i0} > 0.5) \Big/ \sum_{i=1}^{M} I(\hat{\theta}_{i0} > 0.5)$. These two estimates are exactly the same as $\widehat{FDR}$ and $\widehat{FNDR}$ given in (4) and (5) because $\hat{\theta}_{i1} + \hat{\theta}_{i0} = 1$. Therefore, $\widehat{FDR}(c)$ and $\widehat{FNDR}(c)$ given in (7) and (8) are extensions of $\widehat{FDR}$ and $\widehat{FNDR}$ to the general decision rule (6). Simulation results given in Supplementary Material Section 6 provide compelling evidence that the estimators in (7) and (8) are accurate in estimating the true FDR and FNDR.

Suppose the desirable FDR level is α. We apply the method developed by Sun and Cai (2007) to choose the cutoff c so that the resulting classification procedure will have its FDR controlled at α. The procedure is described as follows.

(1) Sort the null memberships in ascending order as $\hat{\theta}_{i_1 0}, \hat{\theta}_{i_2 0}, \ldots, \hat{\theta}_{i_M 0}$.
(2) Find $l = max\{j : \sum_{k=1}^{j} \hat{\theta}_{i_k 0}/j \le \alpha\}$.
(3) Then let $c = \hat{\theta}_{i_l 0}$ and all $H_{i_j 0}$ with $j \le l$ are rejected.
(4) Site i_j is called to be methylated if $j \le l$; otherwise, it is called to be unmethylated.

We shall refer to this procedure as the FDR-controlled MSC procedure at level α, or in short, the FDR-controlled MSC (FMSC) procedure at level α. Based on (7), the resulting FDR for the FMSC procedure can be estimated by $\widehat{FDR} = \sum_{k=1}^{l} \hat{\theta}_{i_k 0}/l$.

As mentioned in Section 2.3, Lister *et al.* (2011) used $H_{i0} : p = \hat{p}_1$ as the null hypothesis, i.e. the null hypothesis assumes that site i is unmethylated. In contrast, the null hypothesis we use here is $H_{i0} : p = \hat{p}_0$, or equivalently, it assumes that site i is methylated. Considering the fact that methylation is more prevalent in the sense that >60% of all CpG sites are expected to be methylated, it is more appropriate to assume the site is methylated in the null hypothesis instead of the other way around. Assuming the site is unmethylated in the null hypothesis leads to the consequence that a significantly higher proportion of the claimed unmethylated sites are methylated. Therefore, in terms of detecting unmethylated sites, our choice of null hypothesis produces more accurate results than the choice by Lister *et al.* (2011). See Supplementary Material Section S5 for more detail.

Sun and Cai (2007) showed that under several assumptions, the *Lfdr*-based adaptive method asymptotically attains the performance of the optimal method that controls marginal FDR at level α with the smallest marginal FNDR. Despite the discreteness and heterogeneity of the tests used for MSC, our simulation study in Section 3.1 shows the incorporation of this adaptive procedure into the MSC procedure leads to satisfactory results. Therefore, we believe the FMSC procedure is adequate in making methylation status calls when controlling FDR at a given level is of interest. When the interest is to control FNDR at a given level, an adaptive procedure similar to FMSC can be developed.

3 RESULTS

3.1 Simulation results

In this subsection, simulation results illustrating the behavior of our proposed procedures are presented. To carry out simulation study, we first use MethylC-Seq data of all CpG sites on Chromosome 1 of H9 human embryonic stem cells from Lister *et al.* (2011) to fit a coverage distribution $\hat{f}$ (Supplementary Material Section 2). Then we apply the mixture of binomial model to the same data to obtain $\hat{\phi} = (\hat{p}_0, \hat{p}_1, \hat{\pi})$ (Supplementary Material Section 7). The total number of CpG

sites in the simulation study is $M = 1000$. The general scheme of our simulation study is described as follows. Step 1: draw a random sample of M observations from $\hat{f}$ and use them as the coverage for M CpG sites. Let the simulated coverage of these M sites be $Z = (z_1, z_2, ..., z_M)$. For each of the M sites, generate its methylation status independently from $Bernoulli(\hat{\pi})$. Simulate C count for each site according to its methylation status and coverage. If the status for site i is methylated, the corresponding C count is generated from $\text{Bin}(z_i, \hat{p}_0)$; otherwise, it is generated from $\text{Bin}(z_i, \hat{p}_1)$. Denote the generated C counts as $R = (r_1, r_2, ..., r_M)$. Step 2: apply the mixture of binomial model to R and Z, obtain $\tilde{\phi} = (\tilde{p}_0, \tilde{p}_1, \tilde{\pi})$, compute the memberships and make methylation status call for each site using the MSC procedure. Step 3: for $i = 1, 2, ..., M$, compute the P-value, denoted as q_i, for testing $H_{i0} : p = \tilde{p}_0 \quad vs \quad H_{ia} : p = \tilde{p}_1$ using exact binomial test, which is, $q_i = \sum_{k=0}^{r_i} \binom{z_i}{k} (\tilde{p}_0)^k (1 - \tilde{p}_0)^{z_i - k}$. After obtaining the P-values, we apply the FDR-controlling procedure proposed by Benjamini and Hochberg (1995) at level $\alpha = 0.1$ to make methylation status calls. We shall refer this procedure to as the individual binomial testing (IBT) procedure. Step 4: use the FMSC procedure described in Section 2.4 to control FDR at three levels $\alpha = (0.1, 0.05, 0.01)$ separately. In the simulation study, for each site, five methylation status calls are made based on three different methods, which are the MSC procedure, the IBT procedure at level 0.1 and the FMSC procedure with three different choices of FDR level. By comparing these calls to the true methylation status, performances of these three procedures can be compared in terms of FDR and FNDR.

The comparison results based on $n = 100$ repeated simulations are displayed in Figure 1. Several observations can be made from the two plots in Figure 1. First, the median FDRs for MSC and FMSC at level 0.1 is around 0.1, and the corresponding median FNDRs are around 0.018 and 0.020, respectively. It shows that the MSC procedure produces similar FDR and FNDR results as the FMSC procedure at level 0.1. Second, the FDRs for the FMSC procedures at all three levels are well controlled. Third, the FNDR for the FMSC procedure at level 0.1 is notably smaller than the FNDR for the IBT procedure at level 0.1. It is caused by the fact that the IBT procedure at level 0.1 overcontrols FDR in the sense that the median FDR is only ~0.05. As a result, the FNDR for IBT is compromised. It suggests that the FMSC procedure is more powerful than the IBT procedure when their FDRs are controlled at the same level.

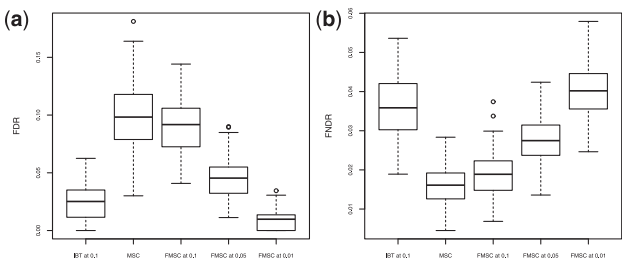

Fig. 1. (**a**) The box plots display FDRs for IBT at level 0.1, the MSC procedure and the FMSC procedure with FDR level 0.1, 0.05 and 0.01 from left to right. (**b**) The box plots display FNDRs for these methods in the same order

In the simulation study, we also applied other FDR-controlling procedures to q_is. They include the q-value method (Storey and Tibshirani, 2003) and procedures proposed by Storey (2002), Gilbert (2005) and Heyse (2011). The results are insensitive to the type of procedure used. Hence only the results from the FDR-controlling procedure proposed by Benjamini and Hochberg (1995) are shown here. Owing to the concern that the simulation results rely on $\hat{f}$ and $\hat{\phi}$, we also apply the proposed methods to a real MethylC-Seq data of H1 human embryonic stem cells from Lister *et al.* (2009) and use the estimates from this dataset to perform the simulation study. The relevant results are given in Supplementary Material Section 11.

3.2 Real data application

The MSC and FMSC procedures are applied to a real MethylC-Seq data of H9 human embryonic stem cells from Lister *et al.* (2011). Three FDR levels, 0.1, 0.05 and 0.01, are considered for FMSC. We first apply the procedures genome wide. The resulting estimate for $\phi = (p_0, p_1, \pi)$ is $\hat{\phi} = (0.9102, 0.1088, 0.8920)$. The MSC procedure is also applied to the same MethylC-Seq data chromosome wise. The results are given in Supplementary Material Section 7. In the genome-wide evaluation with MSC, 42 987 496 of 48 795 269 CpG sites are called to be methylated. For the chromosome-wise evaluation, a total of 43 097 321 CpG sites are declared to be methylated. The difference is ~109 thousands, which account for <0.3% of all covered CpG sites. The detailed comparison results are given in Supplementary Material Section 8. This high concordance suggests the consistency of the MSC procedure.

Correct allocation rates, estimated FDR and estimated FNDR for genome-wide analysis by MSC and FMSC at three FDR levels are also calculated. The results are given in Supplementary Table S5. For MSC, the correct allocation rates for the overall population and the methylated group are 0.9771 and 0.9810, respectively, whereas the rate for the unmethylated group is 0.9450. As for FDR and FNDR, the estimates for MSC are 0.1426 and 0.0067, respectively. For FMSC, as the FDR level decreases, the correct allocation rate for the overall population decreases slightly and the rate for the methylated group increases slightly, whereas the correct allocation rate for the unmethylated group is influenced more dramatically. It decreases from 0.9450 to 0.6395 as the FDR level decreases from 0.1 to 0.01. For FMSC at any of the three FDR levels, the resulting FDR is well controlled. And as expected, the estimated FNDR increases as the FDR level decreases. Based on these results, the performances of MSC and FMSC are acceptable. As mentioned in Section 3.1, the MSC and FMSC procedures are also applied to a real MethylC-Seq data of H1 human embryonic stem cells from Lister *et al.* (2009). The results are given in Supplementary Material Sections S12–S15.

Next, the whole-genome results from the MSC procedure are compared with those from the procedure used by Lister *et al.* (2011). The comparison results are shown in Table 2. Table 2 shows that these two procedures agree with each other on the methylation status calls of 47 619 954 CpG sites, which account for >97% of all covered CpG sites. For the sites that these two procedures make different methylation

Table 2. Comparison of whole-genome results from the MSC procedure and those from the procedure used by Lister *et al.* (2011) for all covered CpG sites

Lister	Our method		
	Methylated sites	Unmethylated sites	Total
Methylated sites	42 987 456	11 75 275	44 162 731
Unmethylated sites	40	46 32 498	46 32 538
Total	429 87 496	58 07 773	48 795 269

status calls, they disagree in two directions. There are only 40 CpG sites that our MSC procedure declares to be methylated but the procedure used by Lister *et al.* (2011) declares to be unmethylated, and we refer to this type of disagreement as the first direction. There are roughly 1.17 million CpG sites that are called to be unmethylated by the MSC procedure but called to be methylated by the procedure used by Lister *et al.* (2011), and we refer to this type of disagreement as the second direction.

Because there are only 40 CpG sites in the first direction but 1.17 million sites in the second direction, we will focus on the second direction in the subsequent analysis. A typical example in the second direction is that for a site with coverage 60 and C count 6, MSC declares it to be unmethylated, whereas the procedure used by Lister *et al.* (2011) declares it to be methylated. Several other typical cases are shown in Supplementary Table S6. As mentioned in the last paragraph, the null hypothesis for Lister *et al.* (2011) is that the site is unmethylated; therefore, *P*-value is computed as $p-value=\sum_{k=y_i}^{x_i}\binom{x_i}{k}(\hat{p}_1^{Lis})^k(1-\hat{p}_1^{Lis})^{x_i-k}$. Because Lister *et al.* (2011) used an extremely small $\hat{p}_1^{Lis}$, which is <0.01, the resulting *P*-value relies heavily on the C count in the sense that it decays to zero exponentially with increasing C count, regardless of coverage x_i and $\hat{\pi}$. Therefore, the C count threshold for declaring one site to be methylated based on the multiple testing procedure used by Lister *et al.* (2011) is generally low, even for sites with high coverage. However, for the MSC procedure, the null membership primarily depends on the proportion of C count at one site instead of C count alone. The cutoff for the proportion is around a half for all sites, which is intuitively more reasonable. Thus, the difference in the cutoff values for these two procedures becomes more evident when coverage increases. This difference is essentially caused by the underestimation of p_1 in the procedure used by Lister *et al.* (2011), and it demonstrates that the procedure used by Lister *et al.* (2011) lacks power in terms of detecting unmethylated sites, especially for sites with moderate to high coverage. Therefore, we believe the MSC procedure makes more accurate methylation status calls for this type of disagreement.

As a final evaluation, the methylation calls for those sites that MSC and the procedure used by Lister *et al.* (2011) disagree on are compared with the results obtained from Infinium Human Methylation 450K BeadChip. The Human Methylation 450K data used here are first analyzed by Merling *et al.* (2013). For those roughly 1.17 million sites that MSC and the procedure used by Lister *et al.* (2011) disagree on, 27 637 sites are covered by Human Methylation 450K BeadChip. We use 0.5 as the cutoff value to dichotomize the beta values in Human Methylation 450K BeadChip data to make binary methylation calls, and compare the calls to those obtained from MSC and the procedure used by Lister *et al.* (2011). The comparison result is given in Table 3. Table 3 shows that for nearly two-thirds of the 27 637 target sites, the methylation calls made by MSC are consistent with the calls made by Human Methylation 450K BeadChip. This suggests the calls made by the MSC procedure are more likely to be correct than those obtained by the procedure used by Lister *et al.* (2011).

Table 3. Third platform validation of the methylation calls for those sites that MSC and the procedure used by Lister *et al.* (2011) disagree on

Procedure	Number of sites that agree with the third platform	Number of sites that disagree with the third platform
MSC	18 090	9547
Lister's	9547	18 090

4 CONCLUSION

In this article, we develop a classification-based procedure and an adaptive FDR-controlled procedure to make binary DNA methylation status calls for bisulfite-based data. Based on our simulation study and real data evaluation, we believe the proposed classification procedure is the procedure of choice if an FDR level between 0.05 and 0.15 is satisfactory. Thus for analyzing bisulfite-based data, it is recommended to apply the classification procedure and estimate the resulting FDR first. If an FDR level other than the estimated FDR is of concern, the adaptive procedure at the desired FDR level can be used. Both the classification-based procedure and the adaptive FDR-controlled procedure are implemented in R and are available free online at http://www.stat.purdue.edu/~cheng70/code.html.

This work also points to several future research directions. First, in this work, the proposed procedures are applied either genome wide or chromosome wise. A more sophisticated choice of regions needs to be further explored, as it is assumed that all the CpG sites in the same region share the same error rates. Second, for adjacent CpG sites, their methylation statuses can be correlated. Hidden Markov models can be used to accommodate this type of correlation and may potentially lead to more accurate methylation status calls (Choi *et al.*, 2009; Qin *et al.*, 2010). Third, Bowtie was used to align the read sequences by Lister *et al.* (2011). This may lead to a bias toward the reference allele. How to correct this bias is worth exploring. Possible solutions include the methods proposed by Wu *et al.* (2010) and Yuan *et al.* (2013). Finally, the ultimate goal of methylation analysis is to detect differentially methylated sites or regions. How to accomplish this goal with our proposed model will be investigated in the future.

Funding: NSF-DMS-1000443.

Conflict of Interest: none declared

REFERENCES

Basford,K.E. and McLachlan,G.J. (1985) Estimation of allocation rates in a cluster analysis context. *J. Am. Stat. Assoc.*, **80**, 286–293.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, **57**, 289–300.

Bock,C. *et al.* (2010) Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nat. Biotechnol.*, **28**, 1106–1114.

Choi,H. *et al.* (2009) Hierarchical Hidden Markov Model with application to joint analysis of ChIP-chip and ChIP-seq data. *Bioinformatics*, **25**, 1715–1721.

Efron,B. (2007) Size, power and false discovery rates. *Ann. Stat.*, **35**, 1351–1377.

Gilbert,P. (2005) A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Appl. Stat.*, **54**, 143–158.

Gu,H. *et al.* (2010) Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nat. Methods*, **7**, 133–136.

Harris,R.A. *et al.* (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol.*, **28**, 1097–1105.

Heyse,J. (2011) A false discovery rate procedure for categorical data. In: *Recent Advances in Biostatistics: False Discovery Rates, Survival Analysis, and Related Topics*. World Scientific Publishing Company, New Jersey, pp. 43–58.

Krueger,F. *et al.* (2012) DNA methylome analysis using short bisulfite sequencing data. *Nat. Methods*, **9**, 145–151.

Laird,P.W. (2010) Principles and challenges of genomewide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.

Lister,R. *et al.* (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

Lister,R. *et al.* (2011) Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature*, **471**, 68–73.

McLachlan,G.J. and Peel,D. (2000) *Finite Mixture Models*. Wiley-Interscience, New York.

Merling,R.K. *et al.* (2013) Transgene-free iPSCs generated from small volume peripheral blood nonmobilized CD34+ cells. *Blood*, **121**, 98–107.

Qin,Z.S. *et al.* (2010) HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics*, **11**, 369.

Robertson,K.D. (2005) DNA methylation and human disease. *Nat. Rev. Genet.*, **6**, 597–610.

Smith,Z.D. *et al.* (2009) High-throughput bisulfite sequencing in mammalian genomes. *Methods*, **48**, 226–232.

Storey,J.D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc. Series B Stat. Methodol.*, **64**, 479–498.

Storey,J.D. (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.*, **31**, 2013–2035.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genome-wide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Sun,W. and Cai,T.T. (2007) Oracle and adaptive compound decision rules for false discovery rate control. *J. Am. Stat. Assoc.*, **102**, 901–912.

Wu,C.F.J. (1983) On the convergence properties of the EM algorithm. *Ann. Stat.*, **11**, 95–103.

Wu,G. *et al.* (2011) Statistical quantification of methylation levels by next-generation sequencing. *PLoS One*, **6**, e21034.

Wu,T.D. *et al.* (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.

Yuan,S. *et al.* (2013) Read-mapping using personalized diploid reference genome for RNA sequencing data reduced bias for detecting allele-specific expression. In: *BIBMW* 2012 *Workshop on Data-Mining of Next Generation Sequencing*. Philadelphia, PA.