# Comparing five statistical methods of differential methylation identification using bisulfite sequencing data

Xiaoqing Yu[1] and Shuying Sun[2*]

[1]Department of Biostatistics, Yale University, New Haven, CT 06511, USA.

[2]Department of Mathematics, Texas State University, San Marcos, TX 78666, USA.

[*]Corresponding author's email: yuxq1120@gmail.com

## Abstract

In this paper, we provide a comprehensive comparison analysis of five differential identification (DM) methods: methylKit, BSmooth, BiSeq, HMM-DM, and HMM-Fisher, which are developed for bisulfite sequencing (BS) data. We summarize the features of these methods from several analytical aspects, and compare their performances using both simulated and real BS datasets. Our comparison results are summarized below. First, parameter settings may largely affect the accuracy of DM identification. Different from default settings, modified parameter settings yield higher sensitivities and/or lower false positive rates. Second, all five methods show higher accuracies when identifying simulated DM regions that are long and have small within-group variation, but they have low concordance, probably due to the different approaches they have used when addressing the issues in DM identification. Third, HMM-DM and HMM-Fisher yield relatively higher sensitivities and lower false positive rates than others, especially in DM regions with large variation. Finally, we find that among the three methods (methylKit, BSmooth, and BiSeq) that involve methylation estimation, BiSeq can best present raw methylation signals. Therefore, based on these results, we suggest that users select DM identification methods based on the characteristics of their data and the advantages of each method.

**Key words:** differential methylation, bisulfite sequencing, HMM-DM, HMM-Fisher

**Supplementary information:** Available online.

## 1. INTRODUCTION

DNA methylation is an important epigenetic modification that plays a key role in regulating gene expression (Baylin and Bestor, 2002; Gopalakrishnan et al., 2008; LAW and Jacobsen, 2010; Suzuki and Bird, 2008). Differential methylation patterns are usually observed between diseased and normal samples, tissues and specimens, and individuals from a population. A wide range of methylation studies have shown that some genomic regions are differentially methylated between normal and disease specimens, as well as between different diseased conditions (Eckhardt et al., 2006; Hansen et al., 2011; Irizarry et al., 2009). Therefore, differentially methylated regions (DMRs) have been used as novel biomarkers for early detection and drug target identification of complex diseases such as cancers (Guzman et al., 2012; Strathdee and Brown, 2002; Wei et al., 2003).

Identifying differential methylation between two groups requires us to obtain methylation signals at each CG site (a cytosine-guanine pair) for the whole genome or some genomic regions. At each CG site, bisulfite treatment can convert unmethylated cytosine to uracil (later read as thymine) while leaving the methylated cytosine unchanged (Krueger et al., 2012). Therefore, bisulfite treatment combined with next-generation sequencing is a preferred method to measure methylation at single-base resolution. There are two main types of bisulfite sequencing technologies, whole-genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS) (Meissner et al., 2008). The former technique is comprehensive, yet costly (Laird, 2010), while the latter is cost-effective, but it only sequences the regions of genome with high GC contents. These technologies have been widely used to investigate DNA methylation patterns in human genomes (Hansen et al., 2011; Lister et al., 2009b; Lister et al., 2011; Sun et al., 2011). Because bisulfite-sequencing technologies can generate a tremendous amount of data with complex biological features, great efforts have been made to process and analyze such large datasets. For example, to deal with the asymmetric mapping issues in bisulfite-treated reads, several alignment tools have been developed, including BSMAP (Xi and Li, 2009), BRAT

(Harris et al., 2010), BS Seeker (Chen et al., 2010), BISMA (Rohde et al., 2010), SAAP-RRBS (Sun et al., 2012), Bismark (Krueger and Andrews, 2011), PASS-bis (Campagna et al., 2013), and RRBSMAP (Xi et al., 2012). Moreover, there are a few packages developed for the quality assessment of bisulfite sequencing data (Akalin et al., 2012; Lin et al., 2013; Sun et al., 2013; Sun et al., 2012).

Differential methylation (DM) identification has been an important research topic in epigenomic studies. During the last several years, a number of statistical and computational methods have been developed for DM identification. Most of these methods are included in the recent review paper by Robinson et al. (Robinson et al., 2014). For these methods, some are motivated by and/or designed for Illumina methylation array data, such as IMA (Wang et al., 2012), A-cluster (Sofer et al., 2013), Minfi (Aryee et al., 2014), MethyAnalysis (Du and Bourgon, 2014), DMRCate (Peters et al., 2015), Probe Lasso (Butcher and Beck, 2015), and BumpHunter (Jaffe et al., 2012). These methods may be applied directly or modified for bisulfite-sequencing data, but their performances on bisulfite sequencing data are not yet known. The others are motivated by and/or developed for bisulfite-treated methylation sequencing. For these bisulfite sequencing-based methods, according to their features, we summarize them into three types as shown below:

I. Methods are motivated by and/or developed only for identifying DM in one group of multiple samples, for example, QDMR (Zhang et al., 2011) and CpG-MPs (Su et al., 2013). It is obvious that these methods are not suitable for identifying DM between two groups or conditions.

II. Methods are motivated by and/or developed for only a pair of samples, for example, Fisher's exact test (Becker et al., 2011; Challen et al., 2011; Li et al., 2010; Lister et al., 2009a), BEAT (Akman et al., 2014), Bisulfighter (Saito et al., 2014), CpG-MPs, Methy-Pipe (Jiang et

al., 2014), and MethPipe (Song et al., 2013). These methods do not account for variation of methylation levels between replicates within one group.

III. Methods are motivated by and/or developed for two groups or conditions, and each group has multiple samples, for example, BSmooth (Hansen et al., 2012), methylKit (Akalin et al., 2012), BiSeq (Hebestreit et al., 2013), DMAP (Jayanth and Puranik, 2011), eDMR (Li et al., 2013), adjusted chi-square test (Xu et al., 2013), RADmeth (Dolzhenko and Smith, 2014), MethylSig (Park et al., 2014), DSS (Feng et al., 2014), and MOABS (Sun et al., 2014). For these methods, we summarize them based on their key features as shown below.

[a] Beta-binomial distribution based methods, e.g., RADmeth, MethylSig, DSS, and MOABS.

[b] Smoothing based methods, e.g., BSmooh and BiSeq.

[c] Regression-based methods, e.g., RADmeth, BiSeq, methylKit, and eDMR (extended methylKit).

[d] Statistical-test based methods (e.g., Fisher's exact test, chi-square test), e.g., DMAP, adjusted chi-square test, and BSmooth.

[e] Hidden Markov model (HMM) based methods: HMM-DM (Yu and Sun, 2015a, 2015b), HMM-Fisher (Sun and Yu, 2015a, 2015b), Bisulfighter, and MethPipe. Please note that Bisulfighter and MethPipe are developed for comparing two samples, not for two groups of samples. Because they used HMM, we include them here. HMM-DM and HMM-Fisher are two HMM-based methods recently developed by our group. The software packages associated with HMM-DM and HMM-Fisher can be found at https://github.com/xxy39/HMM-DM and https://github.com/xxy39/HMM-Fisher, respectively. The manuscripts and user manuals can be found at the above web links as well.

The above methods all have great ideas and try to incorporate some features of methylation data from different angles; however, the tradeoff made by each method is apparent. It is impossible to consider or incorporate everything in one statistical method (Robinson et al., 2014). Even though these methods have been developed, there is still not a consensus on statistical and computational approaches for analyzing this type of data. In addition, our own experience of comparing a few existing methods shows that different methods may not have a large percentage of agreement. This finding agrees with the comparison results shown in other papers (e.g., Figure 4 of DSS, Figure 1 of RADmeth, and Figure 2 of DMAP). In order to have a deeper understanding of the research topic of DM identification, we conduct systematical review and comparison for the two HMM-based methods recently developed by our group and the three commonly cited methods: BSmooth, methylKit, and BiSeq. Our goal is not to do a comparison of ALL available methods and report which one is the best, instead we mainly choose a few to compare in order to share some new perspective on DM identification.

Because BSmooth, methylKit, and BiSeq, HMM-DM, and HMM-Fisher are the methods we will review and compare, we will briefly review them in this section. As an R package for the analysis of BS data, methylKit (Akalin et al., 2012) models the differential methylation between groups using a logistic regression. A sliding linear model (Wang et al., 2011) is applied for converting p-values to q-values to correct for multiple testing. Another method, BSmooth (Hansen et al., 2012), is a pipeline to analyze WGBS data. It first smooths the methylation level via local likelihood estimation for each sample, and then tests for group differences using a modified *t*-test. In addition, BiSeq (Hebestreit et al., 2013) identifies DMR in targeted BS data only, so that it constrains the analysis to CG clusters. A hierarchical testing procedure is then applied to test for DMRs within clusters and control the given false discovery rate (FDR). Moreover, two hidden Markov model (HMM)-based methods, HMM-DM (Yu and Sun, 2015a) and HMM-Fisher (Sun and Yu, 2015a), have been recently developed by our group. The former estimates differential methylation status between two groups directly; while the latter estimates

the methylation states as F (Fully methylated), P (Partly methylated), and N (Not methylated) for each sample with an HMM, and then tests for group differences using Fisher's exact test.

This article is organized as follows. First, we review all five methods based on the six analysis aspects described in Figure 1. It is important to know that these six analysis aspects are not necessary to occur in the particular order listed in Figure 1, instead they can be performed in different steps and in different ways. Second, we use simulated data to evaluate the performance of these five methods. Finally, we apply all five methods to real bisulfite-treated methylation sequencing data and report the results of each method.

## 2. METHODS

This section includes two parts. In Part I, we review the five DM identification methods from six aspects: aligned bisulfite sequencing data format, quality control, smoothing, modeling, testing and defining DM regions, and further analysis. We also summarize and compare their features. In Part II, we introduce the simulated and real datasets that we will use to examine all five methods. We will then describe the workflow of our comparison analysis.

### 2.1 Part I: Overview of DM identification methods

Table 1 summarizes the main algorithms and basic functions used in each of the six analysis aspects for all five methods. In this section, we summarize these steps one by one.

#### *2.1.1 Aligned bisulfite sequencing data*

Bisulfite sequencing provides high throughput methylation data at the single base level. There are two main types of bisulfite sequencing protocols: whole-genome bisulfite sequencing (WGBS) and targeted bisulfite sequencing. WGBS measures methylation levels for an entire genome. Targeted bisulfite sequencing (e.g., reduced representative bisulfite sequencing, RRBS (Gu et al., 2010; Gu et al., 2011)) reduces the complexity of the genome by sequencing the CG enriched regions using restriction enzymes and DNA fragment size selection. Among the five statistical

methods developed for DM identification, BSmooth is designed for methylation data from the WGBS protocol only and BiSeq is for target BS only, while methylKit, HMM-DM, and HMM-Fisher are designed for data generated from any specific protocol. For all these methods, bisulfite sequencing data need to be preprocessed by alignment tools to determine methylation signals. As for the format of input data, methylKit takes the total number of reads and the percent of methylated reads at each CG site, while the other four methods take the total number of reads and number of methylated reads at each CG site.

### *2.1.2 Quality control*

Systematic sequencing errors and base-calling errors can affect the DM identification and downstream analysis. Therefore, it is critical to perform quality control on raw methylation ratio data. As an important indicator of methylation data quality, coverage is commonly considered in the quality control step of most of the five methods. Quality control does not need to be done only at the beginning of the analysis**.** In every step of DM identification, quality control has been conducted in various formats and to different degrees, which are summarized below for each method.

1) Before differential methylation detection, methylKit recommends that users filter out CG sites with relatively high coverage to remove potential PCR bias. In addition, to avoid bias introduced by a systematically more sequenced sample, methylKit can normalize sequencing coverage among samples.
2) In BSmooth, the quality control step is performed in the modeling part. CG sites with low coverage or no coverage are removed from modeling and testing. The threshold for low coverage can be defined by users based on their own data.
3) BiSeq is a DM identification method designed for targeted BS data. Therefore, it constrains the analysis to CG sites within CG clusters, which are regions with higher coverage and higher density of CG sites. These clusters are detected using a three-step strategy. First, CG sites that are covered in the majority (e.g., at least 75%) of samples are defined as frequently covered

CG sites. Second, it detects clusters within which the frequently covered CG sites are close to each other (e.g., at most 100 bp apart). Third, it retains only regions with a minimum number (e.g., 20) of frequently covered CG sites within the clusters.

4) For HMM-DM and HMM-Fisher, CG sites that are covered in only a minority of samples are removed; CG sites with very low coverage are also removed.

### *2.1.3 Smoothing*

Methylation levels of adjacent CG sites in a chromosome region tend to be similar (Eckhardt et al., 2006). Therefore, a smoothing algorithm that borrows information from neighbors is appropriate in this context (Jaffe et al., 2012). It not only reduces the required coverage, but also estimates methylation levels for the CG sites that are not covered by sequencing reads to avoid missing values. In addition, the falsely sequenced CG sites usually have low coverage and their methylation levels are dramatically different from their nearby sites. Smoothing the methylation level can correct these sequencing errors to some extent, but it may introduce some bias.

To account for the spatial correlation of methylation levels, both BiSeq and BSmooth smooth the raw methylation data before detecting differential methylation. In particular, the raw methylation level is smoothed via a local likelihood function weighted on coverage and distance. For each CG site, let the methylation level $y = m/n$, where $m$ is the number of methylated reads and $n$ is the total number of reads, and $m$ is modeled with a binomial distribution $B(n, y)$. Within a window of size $h$ around CG site $l$, CG sites are weighted by kernel functions, such that CG sites close to the CG site $l$ and with a high coverage are given high weight on the estimation of the methylation level at $l$. Despite BSmooth and BiSeq using a similar algorithm to smooth methylation data, they are different in the following ways:

1) Because higher coverage gets higher weight, unusually high coverage can introduce bias into smoothing. Therefore, before the CG clusters are set up for smoothing, BiSeq limits the coverage, e.g., to the 90% quantile of all CG sites.

2) BSmooth performs smoothing on the entire chromosome for each sample, while BiSeq estimates each pre-defined CG cluster separately.
3) In BSmooth, the smoothing window size (*h*) defined by users is the minimum size and the actual bandwidth is enlarged until at least *h* CG sites are included within the window. Therefore, the smoothing degree in BSmooth can be different for each sample and for each region. On the other hand, in BiSeq, *h* defined by users is a fixed window size, such that the intensity of smoothing is uniform for each sample.
4) BSmooth employs a local logistic regression for smoothing, which can lead to the problem of extrapolation. For example, when there is a long region (L bp) without reads covered, the methylation level of a CG site can be predicted by covered CG sites that are L bp away, resulting in an over-estimated methylation level of 0 or 1.

It is important to note that the size of the smoothing window has a large impact on the smoothing step. On one hand, large bandwidth may lead to over-smoothing issues, such that real signals are smoothed away while false signals are introduced. On the other hand, small bandwidth may not do much smoothing at all. Therefore, proper smoothing window size is critical, and the window length should be determined specifically for each dataset and each genomic region.

Other than smoothing, different methods are used to consider spatial correlation of methylation data. For example, HMM-Fisher and HMM-DM employ first-order Markov models. For a given CG site, these two models borrow methylation information from the previous CG site (see *Modeling*).

#### *2.1.4 Modeling*

Depending on the type of information that is modeled, the five methods can be grouped into three categories: modeling methylation levels, modeling methylation categories, and modeling differential methylation states directly.

*Modeling methylation levels*

In order to identify differential methylation between groups, methylKit, BSmooth, and BiSeq first model the methylation level for each CG site and then test for group differences. Detailed models and features of these methods are summarized below:

1) In methylKit, methylation level $y_i$ for sample $i$=1,..,n is modeled by a logistic regression: $\log\left(\frac{y_i}{1-y_i}\right) = \beta_0 + \beta_1 * x_i$, where $x_i = 1$ for the test group and $x_i = 0$ for the control group; $\beta_0$ is the log odds of the control group, and $\beta_1$ is the log odds ratio between the test and control groups. This logistic regression framework can be generalized to more than two groups or data types, with covariates incorporated into the model.
2) In BSmooth, the methylation level $y_{ij}$ for sample $i$ at location $j$ is assumed to be a sample-specific smooth function of genome location $l_j$, $f_i(l_j)$. Then a modified $t$-statistics $t(l_j)$ is formed, with the location-dependent standard deviation floored to the 75$^{th}$ percentile and smoothed using a running mean with a window size of 101. This $t(l_j)$ statistic is later used to test for methylation difference between groups.
3) Instead of estimating methylation levels for all CG sites, BiSeq constrains the analysis to CG sites within clusters defined in the quality control step. Within each cluster, the methylation level at each CG site, $y$, is modeled by a beta distribution with $E(y) = \mu$ and $Var(y) = \mu(1-\mu)/(1+\varphi)$, where $\varphi$ is a precision parameter. The mean of methylation $y_j$ at position $j$ is modeled with a beta regression. The group difference is then tested using the Wald test.

*Modeling methylation categories*

Instead of modeling methylation levels as continuous values, HMM-Fisher models methylation levels as categorical data for each sample separately. In the HMM step, hidden states $h$ are estimated as N (Not-methylated), P (Partly-methylated), and F (Fully-methylated). At each CG site $j$, the transition between current state $h_j$ and the next state $h_{j+1}$ is allowed, and staying in the same state has a higher probability, while the transition between N and F has a lower probability. The emission probabilities (the probabilities of observing methylation levels $O_j$ given the hidden

state of this CG site $h_j$) are modeled by the truncated normal distributions. Based on the biological meanings of the hidden states, the means of truncated normal distributions for N, P, and F states are set as 0, 0.5, and 1 respectively. Therefore, for each sample, the methylation state at each CG site is estimated as N, P, or F.

*Modeling differential methylation states*

All of the four methods described above choose to model methylation levels first, and detect methylation differences afterward. Different from these methods, HMM-DM directly models the differential methylation between groups. Therefore, the hidden states are defined as Hyper (hypermethylated in the test group), Hypo (hypomethylated in the test group), and EM (equally methylated in both groups). The transition probabilities are estimated from the data using dirichlet distributions. As for the emission probabilities, HMM-DM uses beta distributions. For Hyper and Hypo states, at a given CG site *i*, two beta distributions with different means are used to model methylation levels of the control and test groups separately to ensure differential methylation between the two groups. For the EM state, all samples from both control and test groups are modeled using the same distribution that assumes no differences between groups. Beta distributions are used for each CG site separately and all parameters are estimated from the data. Thus, the result of HMM-DM is the differential methylation status for each CG site – hyermethylated in the test group, hypomethylated in the test group, or no differences between groups. Hypermethylated and hypomethylated CG sites with relatively large mean differences are defined as differentially methylated CG sites. Setting a relatively large mean difference is to ensure the biological significance of the identified CG sites.

### *2.1.5 Testing and Defining DMRs*

Depending on the testing strategies employed, the five methods can be grouped into three categories: controlling FDR, not controlling FDR, and not test-based.

*Controlling FDR*

For the analysis of either WGBS or RRBS data, the number of CG sites can go up to millions. Therefore, it is important to deal with the multiple testing issue (Storey, 2002; Storey and Tibshirani, 2003) in detecting differential methylation (Bock, 2012). Two methods, methylKit and BiSeq, incorporate multiple testing corrections during their analysis. For methylKit, a sliding linear model (Wang et al., 2011) is used to correct the p-values obtained from the logistic regression model to q-values. CG sites with associated q-values below a certain threshold and having large mean differences are defined as differentially methylated sites. As for BiSeq, a much more complex algorithm, two-step hierarchical testing (Benjamini and Heller, 2007) is used to correct for multiple testing. The first step is to detect CG clusters containing at least one differentially methylated location and to control a size-weighted FDR (WFDR) on clusters. To control the WFDR, the weighted Benjamini–Hochberg method (Benjamini and Hochberg, 1997) is applied on p-values of clusters, which are calculated from the p-values (Wald test in the Modeling step) for CG sites within the clusters. Clusters with small p-values are selected and considered as candidates for DMRs. In the second step, the equally methylated CG sites within the selected CG clusters are removed and a location-wise FDR is controlled (Benjamini and Hochberg, 1997; Benjamini et al., 2006). Therefore, the result of this hierarchical testing is a list of differentially methylated CG sites within clusters. The adjacent differentially methylated CG sites that locate within the same cluster and have the same direction of methylation differences are defined as one DMR. Thus, CG sites within a DMR are all hypermethylated or all hypomethylated.

*Not controlling FDR*

The other two test-based methods, BSmooth and HMM-Fisher, do not control FDR in their analysis. In BSmooth, after getting the statistics $t(l_j)$, DMRs are defined as groups of consecutive CG sites for which all $|t(l_j)| > c$, where $c$ is a positive cutoff selected based on the marginal empirical distribution of $t(l_j)$. In addition, CG sites with a large distance (e.g., 300 bp)

are not allowed to be in the same DMR. In HMM-Fisher, the categorical data obtained from the HMM step is used to test for a group difference by Fisher's exact test. To better incorporate the information in neighboring CG sites and thus reduce the impact of small sample size and sequencing error, consecutive CG sites are combined if their distance is short (e.g., < 100 bp) and the sample size is very small. CG sites with large mean differences and small p-values are then identified as differentially methylated CG sites. Finally, these identified CG sites are pooled into DMRs based on their p-value and physical distance.

*Not test-based*

HMM-DM is not a test-based method. The output results of hidden Markov models are the estimated differential methylation status of CG sites – hyper, EM, and hypo. To identify DMRs, the differentially methylated CG sites (hyper and hypo) detected from hidden Markov model are formed into regions based on their DM status, physical distance, and posterior probability.

### *2.1.6 Further analysis*

To understand the biological impact of differential methylation, the identified CG sites or regions should be put into genomic context for further analysis. All of the five methods provide tools for differential methylation visualizations and/or annotations.

Visual representation of data can be very useful for the interpretation of DMRs. Visualization tools for differential methylation can be divided into three types: 1) plots of the methylation levels for all samples with identified DMRs (BSmooth, BiSeq, HMM-Fisher, and HMM-DM); 2) summary statistics for DMRs, e.g., number of hyper- and hypo-methylation events per chromosome (methylKit); and 3) web-based genome browsers (UCSC Genome Browser and Integrated Genome Viewer) that allow users to view methylation data along with their genetic annotation (methylKit and BiSeq).

Annotating differential methylation regions can help to predict their functional impact and to find potential disease-related events for further analysis. Most methods can annotate identified DMRs with CpG islands, CpG island shores, genes, and promoter regions (methylKit, BiSeq,

HMM-DM, and HMM-Fisher). In addition, users may be interested in specific genomic regions that are related to certain diseases. Therefore, genetic annotation can also be performed for user-specified CpG sites in methylKit, HMM-Fisher, and HMM-DM.

#### *2.1.7 Summarizing key features*

Table 2 summarizes the key features of the five methods we have reviewed above. BSmooth and BiSeq are designed for specific BS protocol only, while the other three can be applied to both WGBS and target sequencing data. All five methods have corresponding R packages or pipelines available online. Coverage is considered to be an important indicator for sequencing quality and is used as a common criterion in the quality control step. Spatial correlation, the key characteristic of DNA methylation, is considered in most methods by borrowing information from neighboring CG sites. Two test-based methods, methylKit and BiSeq, intend to correct for multiple testing issues in DMR identification. DMR visualizations and genomic annotation tools are available in all five methods.

### 2.2 Part II: Datasets and Comparison Analysis

#### *2.2.1 Real methylation sequencing data*

To compare the five methods, we use publicly available DNA methylation sequencing data (*GSE27003*) (Sun et al., 2011) generated using the Reduced Representation Bisulfite Sequencing (RRBS) protocol (Gu et al., 2010; Gu et al., 2011) from eight breast cancer cell lines, including four estrogen receptor positive (ER+) and four negative (ER-) samples. We then use the software package BRAT (Harris et al., 2010) to trim off bases with low quality from both ends of the reads and to align reads afterwards. Methylation levels are obtained for all CG sites in eight samples using the BRAT *acgt-count* function. After removing CG sites with extremely low methylation coverage, 77,822 CG sites from chromosome 1 are used for further analysis.

#### *2.2.2 Simulation data*

To mimic the complex DNA methylation patterns, all DMRs are simulated based on methylation levels and the variation status of the "control group" of a real dataset. In particular, we take the first 10,000 CG sites of the four ER+ samples from the data described earlier as a control group, and the same 10,000 CG sites of the four ER- samples as a test group. For the test group, the methylation levels are simulated using the control group as a background. Specifically, DMRs in the test group are obtained by adding differential methylation signals with various lengths and intensities to the background. Simulated DMRs are generated this way to preserve the natural changes in methylation patterns across CG sites and the variation patterns among samples. The specific simulation procedure is explained as follows:

First, CG sites are categorized into five methylation classes based on their methylation level and variation status in the control group (see Supplemental Figure 1):

1) H (high methylation), where the methylation levels of all four control samples are $\geq 0.6$, such that the within-group variation is relatively small.
2) L (low methylation), where the methylation levels of all four control samples are $\leq 0.4$, such that the within-group variation is relatively small.
3) M (median methylation), where the mean of four control samples is within the range of (0.4, 0.6).
4) M-H (median-high methylation), where the mean is $\geq 0.6$ but the variation across the four samples is relatively large compared to class H.
5) M-L (median-low methylation), where the mean is $\leq 0.4$ but the variation across the four samples is relatively large compared to class L.

Second, based on the methylation classes, consecutive CG sites of the same class are grouped together, generating four types of regions: two types with small variations, H region and L region; and two types with large variations, M-H region and M-L region. The defined regions are further fine-tuned such that M class CG sites are allowed in M-H and M-L regions with low frequencies. This step generates 2459 methylation regions.

Third, from the regions generated above, we randomly choose 80 DMRs with various methylation statuses and sizes (1 - 76 CG sites) to create methylation differences (see Supplemental Table 1). These DMRs cover 929 differentially methylated CG sites. Then, methylation levels for the test group in these DMRs are sampled from uniform distributions (Table 3). Since the region types are defined based on the control group, we simulate test samples with lower methylation levels for H and M-H DMRs and with higher methylation levels for L and M-L DMRs to create a contrast. In addition, to ensure a true difference in DMRs with larger variation and/or smaller size, we use more stringent uniform distributions for H-M and H-L DMRs and DMRs with ≤ 3 CG sites.

***2.2.3 Comparison analysis***

All five methods are compared by exploring the effect of parameter settings on the DMR identification results. We first use the default settings for each method, and then modify the settings based on the features of each method and the characteristics of the dataset. In order to compare the performance of the five methods, we analyze their results using both simulated and real data.

For the simulated data, sensitivities and false positive rates are calculated for different cutoffs of statistics in each method and the ROC curves are plotted accordingly. Moreover, the simulated DMRs are separated into classes based on their lengths and within-group variation. In particular, the 80 simulated DMRs are separated into three classes based on their sizes: long DMRs with >20 CG sites, median DMRs with 3 – 20 CG sites, and short DMRs with ≤ 2 CG sites. As for the variation levels, the 80 DMRs are grouped into two categories based on their within-group variation: small-variation DMRs (H and L regions) and large-variation DMRs (M-H and M-L regions). The sensitivities for each class of DMRs are then calculated and compared between methods.

For the real data, we compare the differentially methylated CG sites identified by each method and draw Venn diagrams to visualize the results. In addition, for the three methods that

involve the estimation of methylation levels, we evaluate the effect of their estimation by plotting the mean differences between groups for identified DM CG sites. Finally, we plot their estimated mean differences vs. their raw mean differences for CG sites with different coverage cutoffs to investigate the effect of coverage in estimation for the three methods.

## 3. RESULTS

### 3.1 Simulated Dataset

*Default and Modified Settings*

We first apply all methods to the simulated dataset with their default parameter settings (column 2 of Table 4) and cutoffs of statistics (column 2 of Table 5).

1) For methylKit, the coverage of sequencing reads is normalized between samples to avoid bias introduced by systematically more sequenced sample; CG sites with q-statistics $\leq 0.01$ are considered to be differentially methylated sites.
2) For BSmooth, the minimum number of methylation loci in a smoothing window is set to be 70; the minimum length of a smoothing window is set to be 5; and the maximum gap between two methylation loci (i.e., before the smoothing is broken across the gap) is set to be $10^8$ bp. In the modified *t*-test step, the variance is estimated for the control group. Any CG site with a statistics beyond 2 is identified as a differentially methylated CG.
3) For BiSeq, the analysis is first constrained to CG clusters with at least 20 CG sites, where the distance between any two CG sites within a cluster is $\leq 100$ bp, then methylation levels of the CG sites within these clusters are smoothed with a window of 80 bp. To define a differentially methylated region, the cluster-wise FPR is set at 0.1, and the CG-wise FPR is set at 0.05.
4) For HMM-DM, the differentially methylated CG sites are defined as the DM CG sites with posterior probabilities $> 0.4$.
5) For HMM-Fisher, the cutoff of p-value is set at 0.05 to identify DM CG sites.

Table 6A shows the number of identified DM CG sites, number of true positive (sensitivity), and number of false positive (false positive rate) in the five methods with default settings. HMM-DM, HMM-Fisher, and methylKit all yield high sensitivities, while BSmooth and BiSeq show much lower sensitivities and low false positive rates. These differences are due to the cluster pattern of simulation data and the different degrees of spatial correlation each method incorporated. The simulation data are generated based on the real breast cancer dataset, where the CG sites form into relatively small clusters. In methylKit, the methylation level is estimated for each CG site separately. For HMM-DM and HMM-Fisher, the state of each CG site only depends on one previous CG site. Therefore, the estimated methylation levels or estimated DM patterns obtained from these three methods are not heavily influenced by neighboring CG sites. However, in BSmooth and BiSeq, the smoothing windows are much larger (at least 70 CG sites for BSmooth and 80 bp for BiSeq). In BSmooth, the smoothing is only broken when the two consecutive CG sites are more than $10^8$ bp away. Therefore, the short differentially methylated regions can be easily underestimated. In addition, BiSeq constrains the analysis to CG clusters with relatively long length and high CG content so that smaller clusters are left out from the testing for differential methylation.

For the purpose of a fair comparison, we modify the settings of BSmooth and BiSeq to be similar to the other methods (column 3 of Table 4). In particular, all clusters are used for analysis in BiSeq, and the smoothing window size is set to be much smaller (at least 5 CG sites in BSmooth and 25 bp in BiSeq). The parameter settings for methylKit, HMM-DM and HMM-Fisher stay the same as the default. For each method, sensitivities and false positive rates are calculated for different cutoffs of statistics. We then choose the cutoffs that yield relatively higher sensitivity and relatively lower false positive rate (column 3 in Table 5) and show their results in Table 6B. With the modified settings and the chosen cutoffs, all five methods identify a similar number of DM CG sites. In particular, although the sensitivity of methylKit only drops by 10% compared to the default settings (Table 6A), the number of false positives significantly decreases

from 914 to 387. In BSmooth, both the sensitivity and false positive rate are increased with the modified settings. Moreover, the number of DM CG sites called by BiSeq significantly increases from 491 (Table 6A) to 1,234 (Table 6B), yielding a much higher sensitivity. In summary, the five methods perform better with the modified settings as shown in Table 6B. Therefore, we use the modified settings in all further analysis.

*Method Comparison*

To compare the performance of the five approaches, we also show their ROC (Receiver Operating Characteristics) curves with modified settings in Figure 2. Because two FPR levels have to be chosen for BiSeq, we plot three ROC curves each with a fixed cluster-wise FPR (q) and different CG-wise FPRs ($q_2$). In general, HMM-DM and HMM-Fisher achieve higher sensitivities than the others for false positive rates lower than 5%. Out of the three q values chosen for BiSeq, q = 0.9 yields the highest sensitivity by sacrificing the false positive rate. MethylKit can achieve a sensitivity as high as 95% but with a false positive rate of almost 10%. Among all approaches, BSmooth shows the lowest sensitivity and the highest false positive rate. This is because BSmooth is more sensitive to the length than to the intensity of the differential methylation signal. Therefore, long regions that are only slightly different between the two groups (e.g., mean difference $\leq 0.05$) are ranked much higher than smaller regions with strong differential methylation signals.

For each approach we choose the "optimal" cutoff that shows relatively higher sensitivity and relatively lower false positive rate than other cutoffs in the ROC curve analysis (Figure 2, circle on each curve). We then compare the results of these "optimal" cutoffs in detail. The overall sensitivity and false positive rate of "optimal" cutoffs are shown in Table 6B. Among all methods, HMM-DM and HMM-Fisher achieve the highest sensitivity with the lowest FPR, while BSmooth yields the lowest sensitivity of 66.15% and the highest FPR of 5.13%. Table 7 depicts their sensitivity in DMRs with different lengths and variation levels. HMM-DM shows high sensitivity in all five classes of DMRs, especially in DMRs with large variation, while BSmooth has the

lowest sensitivity among all methods. Compared with the other DMRs, the DMRs with ≤ 2 CG sites have much lower sensitivities in all five approaches. This can be explained by the fact that almost all approaches incorporate spatial correlation when identifying DM CG sites and regions, therefore small regions with one or two CG sites are more likely to be weighted out by their neighboring background CG sites. In particular, HMM-Fisher shows a relatively lower sensitivity (66.67%) for small DMRs. This is probably because HMM-Fisher combines the neighboring CG sites in the Fisher's exact test step, the signal of a single DM CG is very likely to be balanced out by the neighboring background CG sites. As for the variation types, all methods work well in regions with small within-group variations in both groups, which is a relatively easy case to identify DMRs. However, for the regions with large within-group variation, BSmooth shows a much lower sensitivity of 10.35% compared to other methods and other cases.

### 3.2 Breast Cancer Dataset

We also compare the five approaches using the real breast cancer dataset mentioned in the Methods section. In chromosome 1, a total of 77,822 CG sites are considered. To ensure that the identified DM CG sites have biological significance rather than statistical significance alone, only CG sites with mean differences ≥ 0.3 are identified as DM. For methylKit, BSmooth, and BiSeq, the mean difference is calculated using the estimated methylation levels. For HMM-DM and HMM-Fisher that do not estimate methylation levels directly, the mean difference is calculated using the raw methylation levels. In addition, DM CG sites in which ER- has higher methylation level than ER+ are defined as hypermethylated, and DM CG sites in which ER- has lower methylation level are defined as hypomethylated. With the default settings (column 2 of Table 1) and default cutoff of statistics (column 2 of Table 2), the five approaches show dramatically different results (Supplemental Table 2). Then we use the modified settings (column 3 of Table 1) for further analysis. Posterior probability > 0.4 in HMM-DM and $p \leq 0.05$ in HMM-Fisher are used to define DM CG sites. The cutoff in BSmooth ($-1.8 \leq q \leq 1.8$) is selected based on the plot of q-statistics following the instruction of the BiSeq user manual. The cutoffs in methylKit ($q <$

$10^{-14}$) and BiSeq ($q = 0.5$, $q_2 = 0.99$) are selected so that these two methods can get a similar number of DM CG sites as the others. Table 8 shows the number of DM CG sites (hyper- and hypomethylated) by each method, where the majority of DM CG sites have higher methylation in the ER- than the ER+ group. All methods identify around 2000 DM CG sites, except that BiSeq only identifies 766 DM CG sites.

Figure 3 shows the Venn diagrams comparing all approaches. Because BiSeq identifies significantly fewer DM CG sites than the others, we first compare the other four methods without BiSeq (Figure 3A). In total, 4752 DM CG sites are detected, with 12.96% detected by all four, 15.63% by any three, 19.97% by any two, and 51.44% by only one method. We then add BiSeq to the comparison (Figure 3B). The number of DM CG sites shared by all methods decreases from 616 to 387, while the percentages of DM CG sites identified by any two or only one method stay similar. In both the four-method and five-method comparisons, different methods show low concordance. This is probably because the five methods address differential methylation identification from different angles and employ different algorithms.

BSmooth, methylKit, and BiSeq all use the estimated methylation levels to test for differential methylation, while HMM-DM and HMM-Fisher do not estimate the methylation level for each CG site. To investigate the effect of estimation, we plot the absolute value of the raw mean differences for the DM CG sites identified by each method in Figure 4. In HMM-DM and HMM-Fisher, all DM CG sites show mean difference $\geq 0.3$ since DM CG sites are defined based on the magnitude of the raw mean differences. For the other three methods, DM CG sites are required to have an estimated mean difference $\geq 0.3$. Therefore, this plot examines the agreement between estimated and raw mean differences for the identified DM CG sites. Both BSmooth and BiSeq smooth the methylation levels using local likelihood estimation incorporating the information of distance, coverage, and neighboring CG sites. BiSeq shows a similar pattern as HMM-DM and HMM-Fisher. Only 4.18% of identified DM CG sites have mean differences less than 0.3, suggesting a strong agreement between estimated and raw mean differences. However,

17.77% of the DM CG sites in BSmooth have raw mean differences < 0.3 while their estimated mean differences are actually ≥ 0.3. The difference between BSmooth and BiSeq may be because BSmooth has a larger smoothing effect, even though the smoothing window size is comparable in these two methods. In BiSeq the smoothing window size is fixed at 25 bp, while in BSmooth the given window size is a minimum size and can be enlarged to any number as long as the consecutive CG sites are within 100 bp (column 3 of Table 1). Among all the methods, the estimated mean differences of methylKit are most different from the raw mean differences. While all DM CG sites by methylKit have estimated mean differences ≥ 0.3, 32.79% of them show raw mean differences < 0.3. This is probably because methylKit estimates the methylation level for each CG separately, with only the coverage incorporated. In addition, this finding also suggests that, even though the DM CG sites in methylKit are identified on the basis of statistical significance, a large percentage of them may not be real differential methylation signals.

Coverage is a factor that all three methods methylKit, BSmooth, and BiSeq consider when they estimate or smooth the methylation levels. CG sites with higher coverage are usually given higher weight in the estimation. To check the effect of coverage in estimation for these three methods, we plot their estimated mean differences vs. their raw mean differences for CG sites with different coverage in Figure 5. As we have expected, CG sites with higher coverage (≥ 30 X) show a better agreement between estimated and raw values in all three methods. When comparing the three methods, the estimation of BiSeq has the best agreement with the raw data, while methylKit shows the lowest concordance between estimated and raw mean difference. This observation is consistent with our previous finding obtained based on how well the estimates can represent the raw data: BiSeq > BSmooth > methylKit.

## 4. DISCUSSION

For the breast cancer data, BiSeq identifies fewer DM CG sites than the other methods. In fact, the majority of CG sites fail the cluster-wise FDR control. Even with a large cluster-wise FDR of

0.9, only 2,596 out of the 77,822 CG sites are available for further analysis. This is probably because that FDR control can lead to a low sensitivity or a high false negative rate under certain circumstances (Pawitan et al., 2005). There are at least two factors determining the FDR characteristics of a DMR detection study: (1) the proportion of truly differentially methylated CG sites and (2) the sample size. To guarantee a small FDR and a high sensitivity, there needs to be a large percentage of CG sites that are truly differentially methylated, as well as a large sample size. However, in the breast cancer data, less than 5% CG sites are identified as DM by the other methods, suggesting that only a small proportion is truly differentially methylated. Moreover, the sample size of the data is relatively small, with four samples in each group. Therefore, for a dataset with a higher percentage of true DM or a larger sample size, BiSeq may yield a high sensitivity when a small FDR is controlled, as in the simulated data and dataset used in the BiSeq paper (Hebestreit et al., 2013).

To explore the effect of parameter settings in HMM-DM and HMM-Fisher, we also modify their parameters as we did with the other methods. In HMM-DM, key parameters such as the priors for transition and emission probabilities are estimated from the data directly. There are only two parameters that might need to be changed: (1) the number of CG sites to break the Markov chain and (2) the direchlet prior for transition probabilities. Similarly, in HMM-Fisher, there are only two parameters that might need to be modified: (1) the standard deviation of the truncated normal distribution of emission probabilities for the three states and (2) the dirichlet prior for the transition probabilities. Different settings of these parameters are applied to the two methods, and we get similar results as the default settings (Supplemental Tables 3 and 4). Therefore, we only report the results of the default settings in this paper.

When comparing the five methods using simulated data, ROC curves are plotted with the y-axis ranging from 0.5 to 1. This is because all the five methods have sensitivities much higher than 0.5. Therefore, although the traditional ROC curves usually have a y-axis of 0 to 1, we use a smaller range to zoom in for a better illustration.

As for the real breast cancer data analysis, the five methods show low concordance in the identified DM sites. This is probably due to several reasons. First, methylation sequencing is still a relatively new research area. Different resources from both biological and technological aspects may contribute to its complexity. Second, to identify differential methylation, each method approaches the question from different angles and has its own advantages and disadvantages. Therefore, we suggest that users select a DM identification method based on the characteristics of the data and the advantages of each method. In addition, for the purpose of validation and further analysis, users may first select the identified DMRs that are relatively long and have small within-group variation to guarantee a high accuracy, and then move on to shorter regions and DM sites with larger within-group variation.

## 5. CONCLUSION

In this paper, we have provided a comprehensive comparison analysis of methods available for the identification of differential methylation in bisulfite sequencing data. First, it is important to explore the effect of parameter settings on the accuracy and efficiency of DMR identification. The simulation data analysis shows that the modified parameter settings can yield higher sensitivities and/or lower false positive rates, especially for methylKit, BSmooth, and BiSeq. Second, to compare the five methods, we have evaluated their performances in simulated DMRs with different lengths and within-group variation. All five methods can better identify DMRs that are relatively long and have small within-group variation. Among all methods, HMM-DM and HMM-Fisher exhibit relatively high sensitivities and low false positive rates, especially in DMRs with large within-group variation. Third, we have compared the five methods using a real breast cancer dataset; however, a low concordance is observed. We have also investigated the effect of methylation estimation. Our results show that among the three methods that involve methylation estimation, BiSeq can best present the raw methylation signals. Therefore, in view of the above findings, we recommend that users select DMR identification methods based on the

characteristics of the data and the different advantages that each method has. We also recommend that, when validating and further analyzing the identified DMRs, users choose long DMRs that have small within-group variation as a priority.

**ACKNOWLEDGEMENTS**

Xiaoqing Yu's stipend was provided by the Case Comprehensive Cancer Center when she was a graduate student at Case Western Reserve University. This work was also supported by Dr. Shuying Sun's start-up funds and the Research Enhancement Program provided by Texas State University.

**REFERENCES**


- Akalin, A., M. Kormaksson, S. Li, F. E. Garrett-Bakelman, M. E. Figueroa, A. Melnick and C. E. Mason (2012): "methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles," Genome Biology, 13, R87.
- Akman, K., T. Haaf, S. Gravina, J. Vijg and A. Tresch (2014): "Genome-wide quantitative analysis of DNA methylation from bisulfite sequencing data," Bioinformatics, 30, 1933-1934.
- Aryee, M. J., A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen and R. A. Irizarry (2014): "Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays," Bioinformatics, 30, 1363-1369.
- Baylin, S. and T. H. Bestor (2002): "Altered methylation patterns in cancer cell genomes: Cause or consequence?," Cancer Cell, 1, 299-305.
- Becker, C., J. Hagmann, J. Muller, D. Koenig, O. Stegle, K. Borgwardt and D. Weigel (2011): "Spontaneous epigenetic variation in the Arabidopsis thaliana methylome," Nature, 480, 245-249.
- Benjamini, Y. and R. Heller (2007): "False Discovery Rates for Spatial Signals," Journal of the American Statistical Association, 102, 1272-1281.
- Benjamini, Y. and Y. Hochberg (1997): "Multiple Hypotheses Testing with Weights," Scandinavian Journal of Statistics, 24, 407-418.
- Benjamini, Y., A. M. Krieger and D. Yekutieli (2006): "Adaptive linear step-up procedures that control the false discovery rate," Biometrika, 93, 491-507.
- Bock, C. (2012): "Analysing and interpreting DNA methylation data," Anglais, 13, 705-719.
- Butcher, L. M. and S. Beck (2015): "Probe Lasso: A novel method to rope in differentially methylated regions with 450K DNA methylation data," Methods (San Diego, Calif.), 72, 21-28.
- Campagna, D., A. Telatin, C. Forcato, N. Vitulo and G. Valle (2013): "PASS-bis: a bisulfite aligner suitable for whole methylome analysis of Illumina and SOLiD reads," Bioinformatics, 29, 268-270.
- Challen, G. A., D. Sun, M. Jeong, M. Luo, J. Jelinek, J. S. Berg, C. Bock, A. Vasanthakumar, H. Gu, Y. Xi, S. Liang, Y. Lu, G. J. Darlington, A. Meissner, J.-P. J. Issa, L. A. Godley, W. Li and M. A. Goodell (2011): "Dnmt3a is essential for hematopoietic stem cell differentiation," Nature genetics, 44, 23-31.
- Chen, P. Y., S. J. Cokus and M. Pellegrini (2010): "BS Seeker: precise mapping for bisulfite sequencing," BMC Bioinformatics, 11, 203.
- Dolzhenko, E. and A. D. Smith (2014): "Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments," BMC Bioinformatics, 15, 215-215.
- Du, P. and R. Bourgon (2014): "methyAnalysis: DNA methylation data analysis and visualization," R package version 1.10.0.
- Eckhardt, F., J. Lewin, R. Cortese, V. K. Rakyan, J. Attwood, M. Burger, J. Burton, T. V. Cox, R. Davies, T. A. Down, C. Haefliger, R. Horton, K. Howe, D. K. Jackson, J. Kunde, C. Koenig, J. Liddle, D. Niblett, T. Otto, R. Pettett, S. Seemann, C. Thompson, T. West, J. Rogers, A. Olek, K. Berlin and S. Beck (2006): "DNA methylation profiling of human chromosomes 6, 20 and 22," Nature Genetics, 38, 1378-1385.
- Feng, H., K. N. Conneely and H. Wu (2014): "A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data," Nucleic Acids Research, 42, e69-e69.
- Gopalakrishnan, S., B. O. Van Emburgh and K. D. Robertson (2008): "DNA methylation in development and human disease," Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 647, 30-38.

- Gu, H., C. Bock, T. S. Mikkelsen, N. Jager, Z. D. Smith, E. Tomazou, A. Gnirke, E. S. Lander and A. Meissner (2010): "Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution," Nature Methods, 7, 133-136.
- Gu, H., Z. D. Smith, C. Bock, P. Boyle, A. Gnirke and A. Meissner (2011): "Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling," Nature Protocols, 6, 468-481.
- Guzman, L., M. Depix, A. Salinas, R. Roldan, F. Aguayo, A. Silva and R. Vinet (2012): "Analysis of aberrant methylation on promoter sequences of tumor suppressor genes and total DNA in sputum samples: a promising tool for early detection of COPD and lung cancer in smokers," Diagnostic Pathology, 7, 87.
- Hansen, K., B. Langmead and R. Irizarry (2012): "BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions," Genome Biology, 13, R83.
- Hansen, K. D., W. Timp, H. C. Bravo, S. Sabunciyan, B. Langmead, O. G. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, E. Briem, K. Zhang, R. A. Irizarry and A. P. Feinberg (2011): "Increased methylation variation in epigenetic domains across cancer types," Nature Genetics, 43, 768-775.
- Harris, E. Y., N. Ponts, A. Levchuk, K. L. Roch and S. Lonardi (2010): "BRAT: bisulfite-treated reads analysis tool," Bioinformatics, 26, 572-573.
- Hebestreit, K., M. Dugas and H. U. Klein (2013): "Detection of significantly differentially methylated regions in targeted bisulfite sequencing data," Bioinformatics, 1647–1653.
- Irizarry, R. A., C. Ladd-Acosta, B. Wen, Z. Wu, C. Montano, P. Onyango, H. Cui, K. Gabo, M. Rongione, M. Webster, H. Ji, J. B. Potash, S. Sabunciyan and A. P. Feinberg (2009): "The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores," Nature Genetics, 41, 178-186.
- Jaffe, A. E., P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg and R. A. Irizarry (2012): "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies," International Journal of Epidemiology, 41, 200-209.
- Jayanth, N. and M. Puranik (2011): "Methylation Stabilizes the Imino Tautomer of dAMP and Amino Tautomer of dCMP in Solution," The Journal of Physical Chemistry B, 115, 6234-6242.
- Jiang, P., K. Sun, F. M. F. Lun, A. M. Guo, H. Wang, K. C. A. Chan, R. W. K. Chiu, Y. M. D. Lo and H. Sun (2014): "Methy-Pipe: An Integrated Bioinformatics Pipeline for Whole Genome Bisulfite Sequencing Data Analysis," PLoS ONE, 9, e100360.
- Krueger, F. and S. Andrews (2011): "Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications," Bioinformatics, 27, 1571-1572.
- Krueger, F., B. Kreck, A. Franke and S. Andrews (2012): "DNA methylome analysis using short bisulfite sequencing data," Nature Methods, 9, 145 - 151.
- Laird, P. W. (2010): "Principles and challenges of genome-wide DNA methylation analysis," Nat Rev Genet, 11, 191-203.
- LAW, J. A. and S. E. Jacobsen (2010): "Establishing, maintaining and modifying DNA methylation patterns in plants and animals," Anglais, 11, 204-220.
- Li, S., F. Garrett-Bakelman, A. Akalin, P. Zumbo, R. Levine, B. To, I. Lewis, A. Brown, R. D'Andrea, A. Melnick and C. Mason (2013): "An optimized algorithm for detecting and annotating regional differential methylation," BMC Bioinformatics, 14, S10.
- Li, Y., J. Zhu, G. Tian, N. Li, Q. Li, M. Ye, H. Zheng, J. Yu, H. Wu, J. Sun, H. Zhang, Q. Chen, R. Luo, M. Chen, Y. He, X. Jin, Q. Zhang, C. Yu, G. Zhou, J. Sun, Y. Huang, H. Zheng, H. Cao, X. Zhou, S. Guo, X. Hu, X. Li, K. Kristiansen, L. Bolund, J. Xu, W. Wang, H. Yang, J. Wang, R. Li, S. Beck, J. Wang and X. Zhang (2010): "The DNA Methylome of Human Peripheral Blood Mononuclear Cells," PLoS Biology, 8, e1000533.
- Lin, X., D. Sun, B. Rodriguez, Q. Zhao, H. Sun, Y. Zhang and W. Li (2013): "BSeQC: quality control of bisulfite sequencing experiments," Bioinformatics, 29, 3227-3229.

- Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren and J. R. Ecker (2009a): "Human DNA methylomes at base resolution show widespread epigenomic differences," Nature, 462, 315-322.
- Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren and J. R. Ecker (2009b): "Human DNA methylomes at base resolution show widespread epigenomic differences," Nature, 462, 315-322.
- Lister, R., M. Pelizzola, Y. S. Kida, R. D. Hawkins, J. R. Nery, G. Hon, J. Antosiewicz-Bourget, R. O'Malley, R. Castanon, S. Klugman, M. Downes, R. Yu, R. Stewart, B. Ren, J. A. Thomson, R. M. Evans and J. R. Ecker (2011): "Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells," Nature, 471, 68-73.
- Meissner, A., T. S. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. E. Bernstein, C. Nusbaum, D. B. Jaffe, A. Gnirke, R. Jaenisch and E. S. Lander (2008): "Genome-scale DNA methylation maps of pluripotent and differentiated cells," Nature, 454, 766-770.
- Park, Y., M. E. Figueroa, L. S. Rozek and M. A. Sartor (2014): "MethylSig: a whole genome DNA methylation analysis pipeline," Bioinformatics, 30, 2414-2422.
- Pawitan, Y., S. Michiels, S. Koscielny, A. Gusnanto and A. Ploner (2005): "False discovery rate, sensitivity and sample size for microarray studies," Bioinformatics, 21, 3017-3024.
- Peters, T. J., M. J. Buckley, A. L. Statham, R. Pidsley, K. Samaras, R. V Lord, S. J. Clark and P. L. Molloy (2015): "De novo identification of differentially methylated regions in the human genome," Epigenetics & Chromatin, 8, 6.
- Robinson, M. D., A. Kahraman, C. W. Law, H. Lindsay, M. Nowicka, L. M. Weber and X. Zhou (2014): "Statistical methods for detecting differentially methylated loci and regions," Frontiers in Genetics, 5, 324.
- Rohde, C., Y. Zhang, R. Reinhardt and A. Jeltsch (2010): "BISMA - Fast and accurate bisulfite sequencing data analysis of individual clones from unique and repetitive sequences," BMC Bioinformatics, 11, 230.
- Saito, Y., J. Tsuji and T. Mituyama (2014): "Bisulfighter: accurate detection of methylated cytosines and differentially methylated regions," Nucleic Acids Research, 42, e45.
- Sofer, T., E. D. Schifano, J. A. Hoppin, L. Hou and A. A. Baccarelli (2013): "A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure," Bioinformatics, 29, 2884-2891.
- Song, Q., B. Decato, E. E. Hong, M. Zhou, F. Fang, J. Qu, T. Garvin, M. Kessler, J. Zhou and A. D. Smith (2013): "A Reference Methylome Database and Analysis Pipeline to Facilitate Integrative and Comparative Epigenomics," PLoS ONE, 8, e81148.
- Storey, J. D. (2002): "A direct approach to false discovery rates," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64, 479-498.
- Storey, J. D. and R. Tibshirani (2003): "Statistical significance for genomewide studies," Proceedings of the National Academy of Sciences, 100, 9440-9445.
- Strathdee, G. and R. Brown (2002): "Aberrant DNA methylation in cancer: potential clinical interventions," Expert Reviews in Molecular Medicine, 4, 1-17.
- Su, J., H. Yan, Y. Wei, H. Liu, H. Liu, F. Wang, J. Lv, Q. Wu and Y. Zhang (2013): "CpG_MPs: identification of CpG methylation patterns of genomic regions from high-throughput bisulfite sequencing data," Nucleic Acids Research, 41, e4-e4.
- Sun, D., Y. Xi, B. Rodriguez, H. Park, P. Tong, M. Meong, M. Goodell and W. Li (2014): "MOABS: model based analysis of bisulfite sequencing data," Genome Biology, 15, R38.
- Sun, S., A. Noviski and X. Yu (2013): "MethyQA: a pipeline for bisulfite-treated methylation sequencing quality assessment," BMC Bioinformatics, 14, 259.

- Sun, S. and X. Yu (2015a): "HMM-Fisher: identifying differential methylation using a hidden Markov model and Fisher's exact test," Manuscript submitted for publication.
- Sun, S. and X. Yu (2015b): "HMM-Fisher," GitHub repository, https://github.com/xxy39/HMM-Fisher.
- Sun, Z., Y. W. Asmann, K. R. Kalari, B. Bot, J. E. Eckel-Passow, T. R. Baker, J. M. Carr, I. Khrebtukova, S. Luo, L. Zhang, G. P. Schroth, E. A. Perez and E. A. Thompson (2011): "Integrated Analysis of Gene Expression, CpG Island Methylation, and Gene Copy Number in Breast Cancer Cells by Deep Sequencing," PLoS ONE, 6, e17490.
- Sun, Z., S. Baheti, S. Middha, R. Kanwar, Y. Zhang, X. Li, A. S. Beutler, E. Klee, Y. W. Asmann, E. A. Thompson and J.-P. A. Kocher (2012): "SAAP-RRBS: streamlined analysis and annotation pipeline for reduced representation bisulfite sequencing," Bioinformatics, 28, 2180-2181.
- Suzuki, M. and A. Bird (2008): "DNA methylation landscapes: provocative insights from epigenomics," Anglais, 9, 465 - 476.
- Wang, D., L. Yan, Q. Hu, L. E. Sucheston, M. J. Higgins, C. B. Ambrosone, C. S. Johnson, D. J. Smiraglia and S. Liu (2012): "IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data," Bioinformatics, 28, 729-730.
- Wang, H., L. Tuominen and C. Tsai (2011): "SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures," Bioinformatics, 27, 225-231.
- Wei, S., R. Brown and T. Huang (2003): "Aberrant DNA methylation in ovarian cancer: is there an epigenetic predisposition to drug response?," Annals of the New York Academy of Sciences, 983, 243-250.
- Xi, Y., C. Bock, F. Müller, D. Sun, A. Meissner and W. Li (2012): "RRBSMAP: A Fast, Accurate and User-friendly Alignment Tool for Reduced Representation Bisulfite Sequencing," Bioinformatics, 28, 430-432.
- Xi, Y. and W. Li (2009): "BSMAP: whole genome bisulfite sequence MAPping program," BMC Bioinformatics, 10, 232.
- Xu, H., R. H. Podolsky, D. Ryu, X. Wang, S. Su, H. Shi and V. George (2013): "A Method to Detect Differentially Methylated Loci With Next-Generation Sequencing," Genetic Epidemiology, 37, 377-382.
- Yu, X. and S. Sun (2015a): "HMM-DM: identifying differentially methylated regions using a hidden Markov model," Manuscript submitted for publication.
- Yu, X. and S. Sun (2015b): "HMM-DM," GitHub repository, https://github.com/xxy39/HMM-DM.
- Zhang, Y., H. Liu, J. Lv, X. Xiao, J. Zhu, X. Liu, J. Su, X. Li, Q. Wu, F. Wang and Y. Cui (2011): "QDMR: a quantitative method for identification of differentially methylated regions by entropy," Nucleic Acids Research, 39, e58-e58.

## Tables

**Table 1**. Algorithms and functions in each analysis aspect for five methods.

| | **MethylKit** | **BSmooth** | **BiSeq** | **HMM-Fisher** | **HMM-DM** |
|---|---|---|---|---|---|
| **Quality control and preprocessing** | Coverage normalization | Removing low coverage | Constraining on CpG cluster | Removing low coverage | Removing low coverage |
| **Smoothing** | No smoothing involved | Locally weighted logistic regression | 1) Limiting the high coverage<br>2) Weighted local likelihood | One first order hidden Markov model for each sample | First order hidden Markov model for both groups |
| **Modeling** | Modeling methylation level using logistic regression | 1) Quality control<br>2) Modeling methylation level using linear regression | Modeling methylation level using beta regression | Modeling methylation category using HMM | Modeling DM status using HMM |
| **Testing** | Sliding linear model to correct p-value | Modified *t*-test | 1) Cluster-wise testing<br>2) CG-wise testing | Fisher's exact test | No testing involved |
| **Defining DMRs** | Single-CG level | Region-level | Region-level | 1) Single-CG level<br>2) Summarize into regions | 1) Single-CG level<br>2) Summarize into regions |
| **Further analysis** | Annotation and visualization | Visualization | Annotation and visualization | Annotation and visualization | Annotation and visualization |

**Table 2**. Key features of five DM identification methods.

| | MethylKit | BSmooth | BiSeq | HMM-Fisher | HMM-DM |
|---|---|---|---|---|---|
| Data type | WGBS<br>Targeted BS | WGBS | Targeted BS | WGBS<br>Targeted BS | WGBS<br>Targeted BS |
| R package/code | package methylKit | Bioconductor package bsseq | Bioconductor package biseq | Pipeline HMM-Fisher | Pipeline HMM-DM |
| Limit high coverage | √ | × | √ | × | × |
| Remove low coverage | √ | √ | √ | √ | √ |
| Spatial correlation | × | √ | √ | √ | √ |
| Multiple testing correction | √ | × | √ | × | Not applicable |
| DMRs visualization | √ | √ | √ | √ | √ |
| Genomic annotation | √ | × | √ | √ | √ |

√: the method has a specific feature

×: the method does not have a specific feature

**Table 3**. Uniform distributions that are used to simulate the test samples in DMRs.

| | > 3 CG sites | ≤ 3 CG sites |
|---|---|---|
| H DMRs | Uniform (0, 0.4) | Uniform (0, 0.2) |
| L DMRs | Uniform (0.6, 1) | Uniform (0.8, 1) |
| M-H DMRs | Uniform (0, 0.3) | Uniform (0, 0.2) |
| M-L DMRs | Uniform (0.7, 1) | Uniform (0.8, 1) |

**Table 4**. The default and modified settings of all five methods.

| | Default settings | Modified settings |
|---|---|---|
| MethylKit | Normalizing read coverage | Same as the default |
| BSmooth | Smooth window ≥ 70 CG/1000 bp, distance ≤ $10^8$ bp | Smooth window ≥ 5 CG/25 bp, distance ≤ 100bp |
| BiSeq | Cluster ≥ 20 CG sites, distance ≤ 100 bp, smooth window = 80 bp | Cluster ≥ 1 CG, distance ≥ 100 bp, smooth window = 25 bp |
| HMM-DM | Partition = 200 CG, transition prior = dirichlet (10, 10, 10) | Same as the default |
| HMM-Fisher | Transition prior = dirichlet (1,1,1), the standard deviation of the emission distribution is 0.12, 0.15, and 0.13 for N, P, and F states respectively | Same as the default |

**Table 5**. The cutoff of statistics for both default and modified settings using simulated data.

| | Cutoff for default settings | Cutoff for modified settings |
|---|---|---|
| MethylKit | $q < 0.01$ | $q < 10^{-10}$ |
| BSmooth | $-2 \le q \le 2$ | $-4.6 \le q \le 4.6$ |
| BiSeq | q (cluster-wise FPR) =0.1<br>$q_2$ (CG-wise FPR) =0.05 | q = 0.9<br>$q_2$= 0.1 |
| HMM-DM | Posterior probability > 0.4 | Posterior probability > 0.8 |
| HMM-Fisher | $p \le 0.05$ | $p \le 0.03$ |

**Table 6**. Results of the five methods from the simulated dataset.

"Optimal" cutoff: the cutoff of statistics for each method that shows relatively higher sensitivity and relatively lower false positive rate than other cutoffs. The "optimal" cutoffs are: $q \leq 10^{-7}$ for methylKit, $-4.6 \leq t \leq 4.6$ for BSmooth, $q = 0.9$ and $q_2 = 0.1$ for BiSeq, posterior probability $\geq 0.8$ for HMM-DM, and $p \leq 0.03$ for HMM-Fisher.

**A.** Default parameter settings and default cutoff of output statistics

| | **Called DM** | **True positive (sensitivity)** | **False positive (FP rate)** |
|---|---|---|---|
| MethylKit | 1841 | 927 (99.89%) | 914 (10.08%) |
| Bsmooth | 500 | 460 (49.52%) | 40 (0.44%) |
| BiSeq | 491 | 435 (46.82%) | 56 (0.63%) |
| HMM-DM | 1220 | 922 (99.25%) | 298 (3.29%) |
| HMM-Fisher | 1174 | 903 (97.20%) | 271 (2.99%) |

**B.** Modified settings and "optimal" cutoff of output statistics

| | **Called DM** | **True positive (sensitivity)** | **False positive (FP rate)** |
|---|---|---|---|
| MethylKit | 1207 | 820 (88.27%) | 387 (4.27%) |
| Bsmooth | 1085 | 619 (66.63%) | 466 (5.13%) |
| BiSeq | 1234 | 894 (96.23%) | 340 (3.75%) |
| HMM-DM | 1206 | 908 (97.74%) | 298 (1.77%) |
| HMM-Fisher | 1124 | 903 (97.20%) | 221 (2.44%) |

**Table 7**. Sensitivities of the five approaches for DMRs with different lengths and variation levels. Shown are comparison results of five approaches with their "optimal" cutoff values. Sensitivities are calculated for DMRs with > 20 CG sites, DMRs with 3-20 CG sites, DMRs with ≤ 2 CG sites, as well as DMRs with small and large within group variation (see the first column). The number of DM CG sites within each region type is shown in parenthesis in column 1.

| | methylKit | BSmooth | BiSeq | HMM-DM | HMM-Fisher |
|---|---|---|---|---|---|
| DMRs >20 (414 CG) | 93.22% | 72.59% | 99.72% | 99.72% | 99.03% |
| DMRs 3-20 (488 CG) | 85.95% | 63.14% | 94.71% | 97.26% | 97.34% |
| DMRs ≤ 2 (27 CG) | 70.37% | 59.26% | 81.48% | 81.48% | 66.67% |
| Small variation DMRs (649 CG) | 90.79% | 86.89% | 99.38% | 99.38% | 99.23% |
| Large variation DMRs (280 CG) | 82.50% | 10.35% | 88.93% | 93.93% | 92.50% |

**Table 8**. The number of DM, hypermethylated, and hypomethylated CG sites identified by each method.

| | Called DM | Hypermethylated | Hypomethylated |
|---|---|---|---|
| MethylKit | 2507 | 1722 | 785 |
| Bsmooth | 2285 | 1612 | 673 |
| BiSeq | 766 | 633 | 133 |
| HMM-DM | 2326 | 1789 | 537 |
| HMM-Fisher | 1917 | 1513 | 404 |

## Figures

**Figure 1**. Six analysis aspects of DMR identification methods.

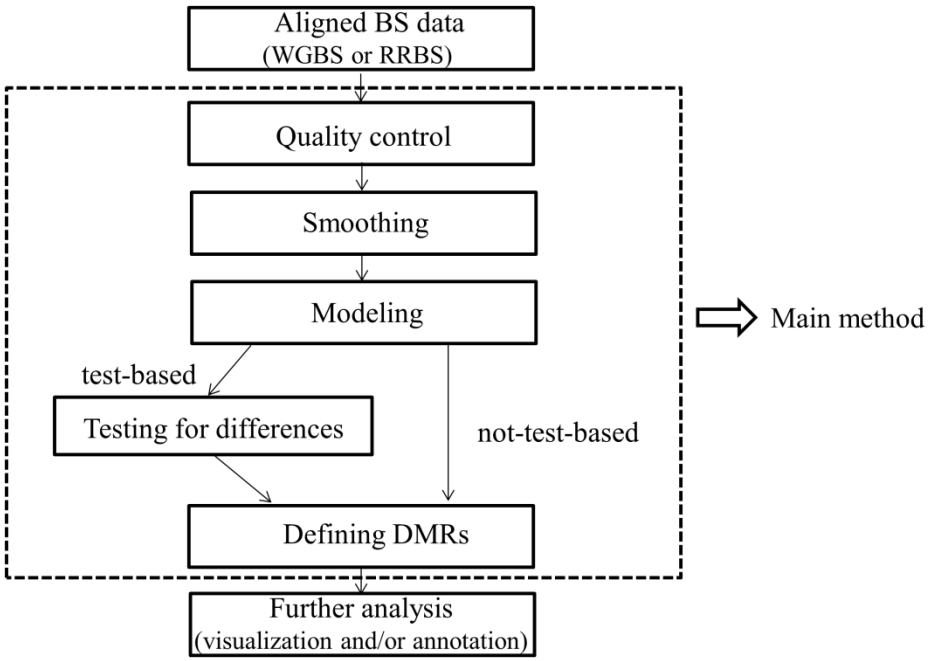



Figure 1

**Figure 2**. ROC curves for differentially methylated CG sites identified by the five methods. Shown are ROC curves for different q-value thresholds (methylKit; purple dash line), different *t*-statistics (BSmooth; orange dash line), different q (cluster-wise FPR) and $q_2$ (CG-wise FPR) values (BiSeq; colored solid line), different posterior probability cutoffs (HMM-DM; black solid line), and different p-value thresholds (HMM-Fisher; black dash line). Each ROC curve for BiSeq is generated from a chosen q value with different $q_2$ values. The circle on each curve shows the "optimal" cutoff that shows relatively higher sensitivity and relatively lower false positive rate than other cutoffs. For BiSeq, the "optimal" cutoff is found with q = 0.9.

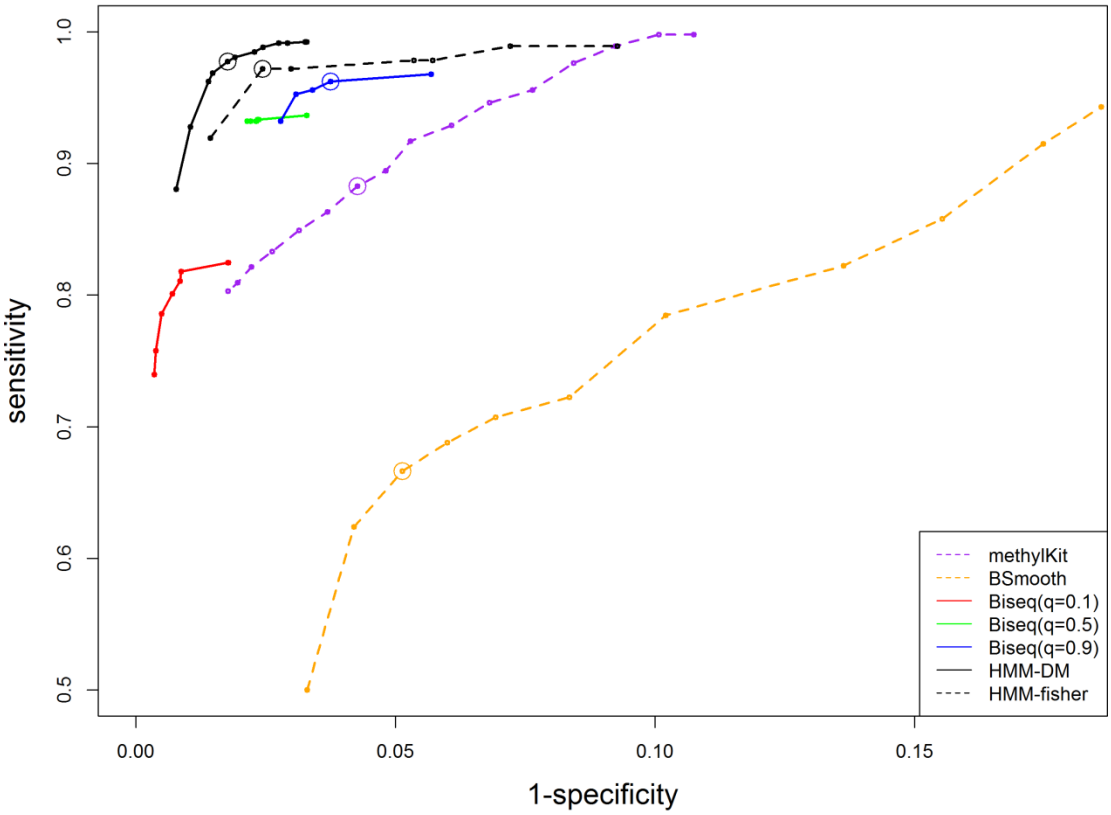
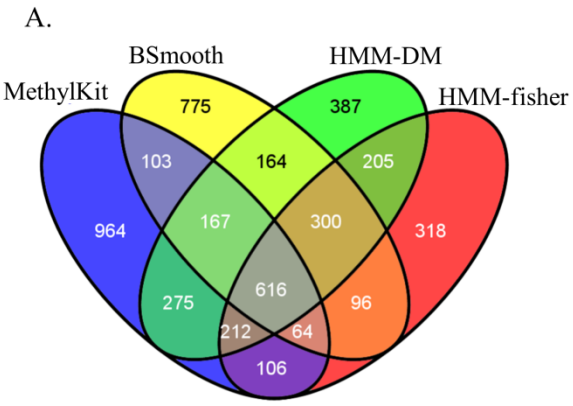


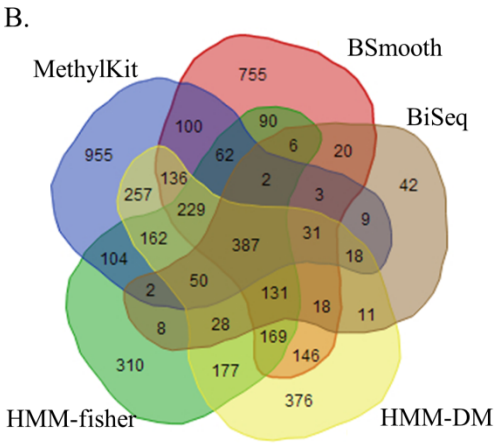

Figure 2

**Figure 3**. Compare the DM CG sites identified by all five approaches.

(A) Compare all methods except BiSeq. Shown are the Venn diagram of comparison results and number (percentage) of DM CG sites identified by all four, any three, any two, and only one method. (B) Compare all five methods. Shown are the Venn diagram of comparison results and number (percentage) of DM CG sites identified by all five, any four, any three, any two, and only one method.




| Total | 4752 |
|---|---|
| By all 4 | 616(12.96%) |
| By any 3 | 743 (15.63%) |
| By any 2 | 743(19.97%) |
| By only 1 | 2444(51.44%) |




| Total | 4794 |
|---|---|
| By all 5 | 387(8.07%) |
| By any 4 | 443 (9.24%) |
| By any 3 | 604 (12.60%) |
| By any 2 | 922 (19.23%) |
| By only 1 | 2438 (50.86%) |

Figure 3

**Figure 4**. Absolute values of raw mean differences for DM CG sites identified by all methods. For each CG, mean difference is calculated as the mean methylation level in ER+ group minus the mean methylation level in ER- group. The dashed line indicates raw mean difference of 0.3. The percentage of DM CG sites with raw mean difference < 0.3 is 32.79% for methylKit, 17.77% for BSmooth, 4.18% for BiSeq, and 0% for HMM-DM and HMM-Fisher.

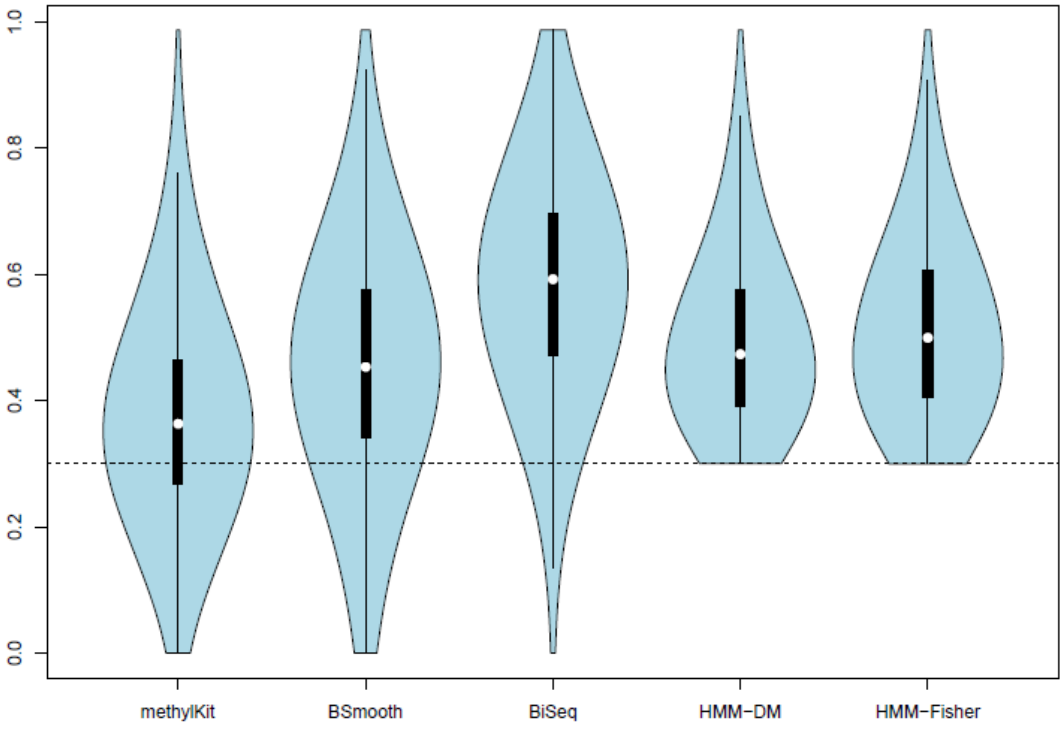



Figure 4

**Figure 5**. Plots of estimated mean differences vs. raw mean differences for CG sites with different coverage.

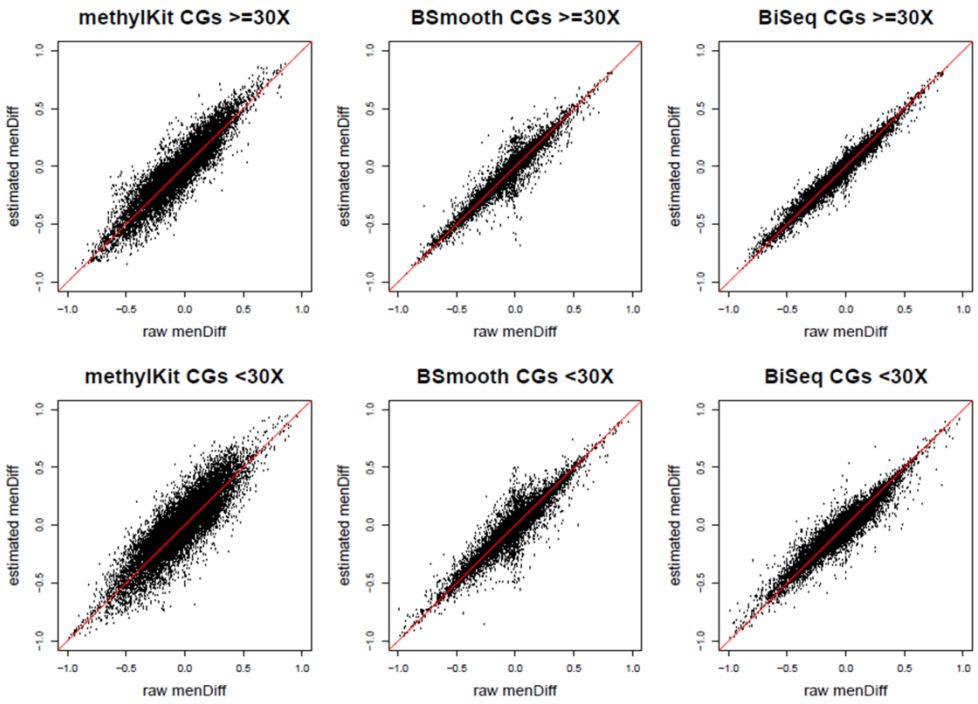



Figure 5