Published in final edited form as:

Biometrics. 2016 June ; 72(2): 354–363. doi:10.1111/biom.12422.

Estimating DNA methylation levels by joint modeling of multiple methylation profiles from microarray data

Tao Wang¹, Mengjie Chen², and Hongyu Zhao^{1,*}

¹Department of Biostatistics, Yale University, New Haven, Connecticut, U.S.A.

²Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, U.S.A.

Summary

DNA methylation studies have been revolutionized by the recent development of high throughput array-based platforms. Most of the existing methods analyze microarray methylation data on a probe-by-probe basis, ignoring probe-specific effects and correlations among methylation levels at neighboring genomic locations. These methods can potentially miss functionally relevant findings associated with genomic regions. In this paper, we propose a statistical model that allows us to pool information on the same probe across multiple samples to estimate the probe affinity effect, and to borrow strength from the neighboring probe sites to better estimate the methylation values. Using a simulation study we demonstrate that our method can provide accurate model-based estimates. We further use the proposed method to develop a new procedure for detecting differentially methylated regions, and compare it with a state-of-the-art approach via a data application.

Keywords

DNA methylation index; Group fused lasso; Model-based analysis; Probe effect

1. Introduction

Playing a central role in normal human development and disease (Jones, 2012), DNA methylation is the most-studied epigenetic mechanism. Throughout the human genome, 70% of CpG sites are methylated (Lister et al., 2009). Short regions of high CpG-density, called CpG islands, are usually unmethylated and found in approximately 60% of gene promoters (Saxonov et al., 2006). The methylation level at a given CpG site is often measured as the percentage of methylation at that site, known as the β value.

Several high-throughput technologies, including microarray and next-generation sequencing, are currently available for genome-wide DNA methylation profiling (Laird, 2010). Although whole-genome bisulfite sequencing can map the methylation at single-base resolution, it

* hongyu.zhao@yale.edu.

Supplementary Materials

Web Appendices A, B, C, and D referenced in Sections 3 and 4 are available with this paper at the Biometrics website on Wiley Online Library. We have developed a Bioconductor package **MBAmethyl** that is available with the paper.

requires a high level of specialization and remains expensive and time-consuming. DNA methylation microarrays, on the other hand, have proved to be very useful for addressing various biological and medical questions. This, along with the current trend of decreasing cost and increasing coverage of these arrays, has brought about an explosion in their usage.

There is increasing evidence that DNA methylation is sensitive to environmental exposures and may be an interface through which environmental factors affect gene expression and health outcome (Tammen et al., 2013). Statistical analysis of DNA methylation profiles from microarray data is thus critically important for the discovery of novel DNA methylation-based biomarkers, and we are now experiencing a growth in the number of statistical tools for processing large-scale DNA methylation data (Bock, 2012). However, the analysis of such data is nontrivial because of the unique characteristics of DNA methylation (Ordway and Curran, 2002). It is further complicated by the large data size as well as many levels of variations introduced at different stages of the experiment (Siegmund, 2011). Like many other microarray-based measurements, DNA methylation values are also subject to platform-dependent probe effects, namely, CpG sites with the same true level of methylation may appear to have very different β values simply because some probes are more sensitive than others (Jaffe et al., 2012).

The probe affinity effect is well-known in the analysis of high-density oligonucleotide array probe-level data, where a measure of expression needs to be defined by summarizing intensities from a probe set corresponding to a specific gene. To account for probe-specific effects, several model-based approaches have been proposed and shown to perform better than simple measures such as the average difference and its variants. For example, Li and Wong (2001) developed a model-based expression index using a multiplicative model, while Irizarry et al. (2003) proposed an alternative measure motivated by a log scale linear additive model.

For oligonucleotide arrays these low-level analysis tools have worked well in practice. However, they are not appropriate for handling DNA methylation data, because here we are interested in the methylation status of all the probes in a probe set instead of a summary statistic. Nevertheless, when the data resolution is high, as is the case for the Infinium 450K data (Bibikova et al., 2011), the basic idea of averaging intensities across probes and removing probe-specific biases can be adapted to develop DNA methylation indexes for subsets of neighboring probes with similar true values of methylation. Current literature suggests that methylation levels are strongly correlated along the genome, and functional relationships between DNA methylation and gene expression are generally found in genomic regions rather than isolated CpG sites (Bock, 2012; Hansen et al., 2012). Furthermore, from a statistical perspective, analysis of DNA methylation data is ultimately concerned with finding spatial intervals exhibiting either a difference between biological groups of samples or an association with an environmental risk factor. Therefore, we should take advantage of the spatial distribution of DNA methylation values along the genome and make use of data from all samples in a study to infer methylation levels while adjusting for probe effects.

In the present paper we assume that for a cohort of biological samples with similar demographic characteristics, there are a common set of genomic regions in which true DNA

methylation levels remain unchanged. By adopting a flexible regression framework we propose a low-level model-based method that is able to draw strength from these similarities, and at the same time account for the probe effect. In particular, our model encompasses that of Li and Wong (2001) as a special case, and we recover these genomic regions using a penalized estimation procedure.

The rest of the paper is organized as follows. In Section 2, we propose a statistical model for describing the raw (observed) DNA methylation values, and then a method for estimating the true (unobserved) values. In Section 3, we conduct a simulation study to evaluate the performance of the model fitting procedure. In Section 4, we use the proposed method to develop a new procedure for detecting differentially methylated regions, and illustrate through a data example how to apply it to solve real problems. Finally, Section 5 gives a brief discussion of the methods and results. We have developed a Bioconductor package **MBAmethyl** that is available at <http://www.bioconductor.org/packages/release/bioc/html/MBAmethyl.html>. In Supplementary Materials we also provide some scripts used in the analysis.

2. Methods

The β value is defined as the ratio of intensities between methylated and unmethylated probes. Another popular metric for quantifying methylation levels is the M value, which is a logit transformation of the β value. Although studies have been conducted to compare different metrics, how to determine the best one usually depends on the nature of the problem under inquiry. Since the β value and M value have their own strengths and limitations, the investigators are generally recommended to use both of them (Du et al., 2010). As this section is concerned with the problem of estimating the true methylation levels in the presence of probe effects, the general methodology developed here puts no restriction on the choice of metrics, except that a continuous methylation measure is required.

2.1 Proposed model

For a gene or a probe set, let Y be the $p \times n$ matrix of raw DNA methylation measurements, for p probes and n samples. The (i, j) th entry in this matrix, Y_{ij} , denotes the observed methylation value of the i th probe of the j th sample. We assume

$$Y_{ij} = \phi_i \Theta_{ij} + \epsilon_{ij}, \quad (1)$$

where Θ_{ij} is the true (unobserved) methylation value, ϕ_i denotes the i th probe effect, and ϵ_{ij} is the random error with zero mean and finite variance. We further assume that $(\Theta_{1j}, \dots, \Theta_{pj})^\top$ is a piecewise constant vector for each sample j , and that the change-point locations are shared across samples, that is, there exist a positive integer K and a sequence of positions, $0 = t_0 < t_1 < \dots < t_K = p$, such that $\Theta_{ij} = \Theta_j^{(k)}$ for $t_{k-1} < i \leq t_k$, where $k = 1, \dots, K$. For identifiability of the parameters, we constrain the probe affinities such that

$$\sum_{t_{k-1} < i \leq t_k} \phi_i^2 = t_k - t_{k-1}.$$

We use the term *change-point* because, when $\phi_1 = \dots = \phi_p = 1$, model (1) becomes a multi-sequence change-point model that is often used for detecting DNA copy number variants in multiple array CGH profiles (Zhang et al., 2010; Picard et al., 2011). On the other hand, if $K = 1$, then model (1) reduces to the multiplicative model from the gene expression literature (Li and Wong, 2001). Note that the set of change-points divide the probe set into K genomic regions or blocks. Like gene expression indexes, we call $\Theta_j^{(k)}$ the methylation index for the k th genomic region in the j th sample.

Let $\phi = (\phi_1, \dots, \phi_p)^\top$ and $\Theta = (\Theta_{ij})$. Throughout this paper, for a matrix A , we use A_i to denote its i th row. To fit model (1), we estimate the parameters by minimizing the following criterion

$$Q_1(\phi, \Theta) = \sum_{i=1}^p \sum_{j=1}^n (Y_{ij} - \phi_i \Theta_{ij})^2 + \lambda \sum_{i=1}^{p-1} \frac{\|\Theta_{i+1\cdot} - \Theta_{i\cdot}\|_2}{w_i}, \quad (2)$$

where λ is a regularization parameter, $w_1, \dots, w_{p-1}$ are some pre-specified positive weights, and $\|\cdot\|_2$ stands for the l_2 norm.

The first term in (2) is the usual sum of squared errors, and the second term is the group fused lasso penalty (Bleakley and Vert, 2011) that enforces common change-points across samples. In this paper, we use the probe index-dependent weighting scheme

$$w_i = \sqrt{\frac{p}{i(p-1)}}, \quad i = 1, \dots, p-1.$$

When there is only one change-point, this weighting scheme was suggested and studied theoretically by Bleakley and Vert (2011) in the absence of probe effects. Note, however, that this scheme may not be appropriate in general and needs to be further investigated. We choose to use it for simplicity, and our limited experience shows that it works well empirically.

2.2 An alternating minimization algorithm

To solve the above optimization problem, we can use a simple alternating algorithm: that is, we first fix ϕ and estimate Θ , then we fix Θ and estimate ϕ , and we iterate between these two steps until the algorithm converges. Since criterion (2) is decreased at each step, convergence is guaranteed. In general, this algorithm converges to a local minimizer, since the penalty is not convex.

Two special cases are $\lambda = 0$ and $\lambda = \infty$, where $\lambda = 0$ imposes no penalty, and any solution will give an interpolating fit, while $\lambda = \infty$ forces $\Theta_{i+1\cdot} - \Theta_{i\cdot}$ to be zero for all $i = 1, \dots, p-1$,

and no change-point will exist. In the latter case, we alternate between fixing φ and solving for $\Theta_{1\cdot}$, and vice versa (Li and Wong, 2001). Specifically, for fixed φ , the solution for $\Theta_{1\cdot}$ is given by

$$\Theta_{1j} = \frac{\sum_{i=1}^p \phi_i Y_{ij}}{\sum_{i=1}^p \phi_i^2}, \quad j = 1, \dots, n.$$

For fixed $\Theta_{1\cdot}$, the solution for φ is given by

$$\phi_i = \frac{\sum_{j=1}^n \Theta_{1j} Y_{ij}}{\sum_{j=1}^n \Theta_{1j}^2}, \quad i = 1, \dots, p.$$

For the probe effects to be useful, we further normalize φ such that $\sum_{i=1}^p \phi_i^2 = p$.

We now consider the general case $\lambda \in (0, \infty)$. When φ is fixed, directly minimizing (2) with respect to Θ is challenging. We thus reformulate (2) in a more convenient form. Define the $(p-1) \times n$ matrix Δ by

$$\Delta_{i\cdot} = \frac{\Theta_{i+1\cdot} - \Theta_{i\cdot}}{w_i}, \quad i = 1, \dots, p-1.$$

Then, we have

$$\Theta_{i\cdot} = \Theta_{1\cdot} + \sum_{s=0}^{i-1} w_s \Delta_{s\cdot}, \quad i = 1, \dots, p.$$

Here, we assume that $w_0 = 0$ and $\mathbf{0}_0$ is a vector of zeros. The criterion now becomes

$$Q_2(\phi, \Theta_{1\cdot}, \Delta) = \sum_{i=1}^p \|Y_{i\cdot} - \phi_i \Theta_{1\cdot} - \phi_i \sum_{s=0}^{i-1} w_s \Delta_{s\cdot}\|_2^2 + \lambda \sum_{i=1}^{p-1} \|\Delta_{i\cdot}\|_2. \quad (3)$$

To further simplify the computation, we note that for any φ and Δ , the solution for $\Theta_{1\cdot}$ is given by

$$\Theta_{1\cdot} = \frac{\sum_{i=1}^p \phi_i Y_{i\cdot} - \sum_{i=1}^p \phi_i^2 \sum_{s=0}^{i-1} w_s \Delta_{s\cdot}}{\sum_{i=1}^p \phi_i^2} = \frac{\phi^\top Y - \phi^\top T \Delta}{\phi^\top \phi},$$

where

$$T = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ \phi_2 w_1 & 0 & 0 & \cdots & 0 \\ \phi_3 w_1 & \phi_3 w_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_{p-1} w_1 & \phi_{p-1} w_2 & \phi_{p-1} w_3 & \cdots & 0 \\ \phi_1 w_1 & \phi_p w_2 & \phi_p w_3 & \cdots & \phi_p w_{p-1} \end{pmatrix}$$

is a $p \times (p-1)$ matrix.

Plugging this into (3), we obtain

$$\begin{aligned} Q_3(\phi, \Delta) &= \sum_{i=1}^p \|Y_{i\cdot} - \phi_i \left(\frac{\phi^\top Y}{\phi^\top \phi} - \frac{\phi^\top T \Delta}{\phi^\top \phi} \right) - \phi_i \sum_{s=0}^{i-1} w_s \Delta_s\|_2^2 + \lambda \sum_{i=1}^{p-1} \|\Delta_{i\cdot}\|_2 \\ &= \|Q_\phi Y - Q_\phi T \Delta\|_F^2 + \lambda \sum_{i=1}^{p-1} \|\Delta_{i\cdot}\|_2, \end{aligned} \quad (4)$$

where $Q_\phi = I - (\phi^\top \phi)^{-1} \phi \phi^\top$, I is the $p \times p$ identity matrix, and $\|\cdot\|_F$ denotes the Frobenius norm. It is easy to see that Q_ϕ is a projection matrix.

Consequently, when ϕ is fixed, minimizing (4) with respect to Δ becomes a group lasso problem (Yuan and Lin, 2006). Hence, the solution can be efficiently computed using blockwise descent algorithms (e.g., Breheny and Huang (2015)). However, the computational burden increases dramatically as p and n increase. Instead, in this paper we use a fast, approximate algorithm based on the group lars (Yuan and Lin, 2006; Bleakley and Vert, 2011). The group lars adds change-points one after another, and by construction its solution path is piecewise linear as a function of the number of change-points. Accordingly, in the following we will index the solution by the number k of change-points rather than by the regularization parameter λ .

For fixed ϕ , the solution for Θ is then given by

$$\Theta_{1\cdot} = \frac{\sum_{i=1}^p \phi_i Y_{i\cdot} - \sum_{i=1}^p \phi_i^2 \sum_{s=0}^{i-1} w_s \Delta_s}{\sum_{i=1}^p \phi_i^2}$$

and

$$\Theta_{i\cdot} = \Theta_{1\cdot} + \sum_{s=0}^{i-1} w_s \Delta_s, \quad i = 2, \dots, p.$$

Furthermore, the probe index $i > 1$ is regarded as a change-point location when $\Theta_{i-1\cdot}$ is nonzero. The set of change-point locations then generates a linearly ordered partition $C_1, \dots, C_{k+1}$ of the index set $\{1, \dots, p\}$.

When Θ is fixed, minimizing (2) with respect to φ is a least squares problem. For fixed Θ , the solution for φ is given by

$$\phi_i = \frac{\sum_{j=1}^n \Theta_{ij} Y_{ij}}{\sum_{j=1}^n \Theta_{ij}^2}, \quad i=1, \dots, p.$$

We then normalize $\{\varphi_i, i \in C_l\}$ such that $\sum_{i \in C_l} \phi_i^2 = |C_l|, 1 \leq l \leq k+1$, where for a set A , we use $|A|$ to denote its size.

In summary, for a fixed number $k+1$ of change-points, our proposed algorithm iteratively updates Θ and φ , and proceeds as follows.

Step 0. Initialize φ . For example, $\varphi = (1, \dots, 1)^\top$.

Step 1. Update Θ .

Step 1.1. Update φ . Let φ denote the group lars solution to criterion (4) with k change-points. Identify the set of change-points and then generate a linearly ordered partition $C_1, \dots, C_{k+1}$ of the index set $\{1, \dots, p\}$.

Step 1.2. Update $\Theta_{1\cdot}$. Let

$$\Theta_{1\cdot} = \frac{\sum_{i=1}^p \phi_i Y_{i\cdot} - \sum_{i=1}^p \phi_i^2 \sum_{s=0}^{i-1} w_s \Delta_s}{\sum_{i=1}^p \phi_i^2}.$$

Step 1.3. Let

$$\Theta_{i\cdot} = \Theta_{1\cdot} + \sum_{s=0}^{i-1} w_s \Delta_s, \quad i = 2, \dots, p.$$

Step 2. Update φ . Let

$$\phi_i = \frac{\sum_{j=1}^n \Theta_{ij} Y_{ij}}{\sum_{j=1}^n \Theta_{ij}^2}, \quad i = 1, \dots, p,$$

then for $l = 1, \dots, k+1$, normalize $\{\varphi_i, i \in C_l\}$ so that $\sum_{i \in C_l} \phi_i^2 = |C_l|$.

Step 3. Iterate Steps 1 through 2 until convergence or until a maximum number of iterations is reached.

It is easy to see that the special case of $k = 0$ (i.e., $\lambda = \infty$) is also included within the above framework by simply setting $\lambda = (0)$ and $C_1 = \{1, \dots, p\}$ in Step 1.1.

We now use the above algorithm to efficiently compute a sequence of solutions indexed by k . Specifically, we start with $k = 0$ for which there is no change-point, and compute the solution for φ and $\Theta = (\Theta_1, \dots, \Theta_1)^\top$. Then, we increment k by one and compute the solution for φ and Θ . Then k is incremented by one again and the process is repeated. In practice, we adopt the following two rules to reduce the computational burden. First, the previous solution for φ can be used as a “warm start” for the next problem. Second, early stopping can be used to ease the computational burden and to avoid over-fitting. For example, suppose that $p = 200$ and $n = 100$; if we do not want more than $k = 30$ change-points in the final model, we may stop the algorithm after $k + 1 = 31$ steps (including $k = 0$). We call this procedure Group Fused lars (GFlars).

2.3 Tuning

The choice of k is of great importance under our framework. Because the estimates are obtained in a penalized regression context, methods that have been developed for selecting the tuning parameter for penalized regression can be applied. In this section we propose to use a criterion-based approach for selecting k . For a given k , denote the estimates of φ and Θ by $\varphi(k)$ and $\Theta(k)$, respectively. We define

$$IC(k) = \log \{RSS(k)\} + P(k), \quad (5)$$

where $RSS(k) = \sum_{i=1}^p \sum_{j=1}^n \{Y_{ij} - \phi_i(k) \Theta_{ij}(k)\}^2$ and $P(k)$ denotes the penalty term for the complexity of the fitted model.

Several criteria have been proposed to determine the number of change-points in single- and multi-sequence change-point models; see, for example, Zhang and Siegmund (2007; 2012) and Kurozumi and Tuvaandorj (2011). According to these works, the penalty terms of these criteria are different from those of the corresponding classical criteria because of the irregularity in the change-points. Following Kurozumi and Tuvaandorj (2011), we take

$$P(k) = \frac{\log(p)}{pn} \{(k+1) \times n + (p - k - 1) + 2 \times k\},$$

which corresponds to a modified Bayesian information criterion (BIC; Schwarz, 1978). We also consider a modified Akaike information criterion (AIC; Akaike, 1974; Kurozumi and Tuvaandorj, 2011) by taking

$$P(k) = \frac{2}{pn} \{(k+1) \times n + (p - k - 1) + 3 \times k\}.$$

Finally, we select k by minimizing $IC(k)$.

3. Simulations

In this section, we use simulated data to evaluate the performance of the proposed method. Throughout we apply GFlars in Section 2.2 to compute the estimates $\hat{\phi}(k)$ and $\hat{\Theta}(k)$, and then select k using criterion (5). Denote by $\hat{k}$ the estimated number of change-points.

We generate data according to model (1) with $p \in \{40, 60, 80, 100, 200, 300\}$ and $n \in \{20, 40, 80, 160\}$. Four cases are considered in the simulation study. In each case, the true

signals, $\Theta_{ij} = \sum_{k=1}^K \Theta_j^{(k)} I(t_{k-1} < i \leq t_k)$, are not fixed in advance, but are drawn randomly as follows.

In Case 1, $K = 1$, that is, there is no change-point. The $\Theta_j^{(0)}$ values are generated uniformly on the interval $[0.1, 0.9]$.

In Case 2, $K = 2$. We set $t_1 = p/2$. The $\Theta_j^{(0)}$ values are generated uniformly on $[0.1, 0.6]$ and the $\Theta_j^{(1)}$ values are generated uniformly on $[0.4, 0.9]$.

In Case 3, $K = 3$. Set $t_1 = \lfloor p/3 \rfloor$ and $t_2 = 2 \lfloor p/3 \rfloor$. The $\Theta_j^{(0)}$ values are generated uniformly on $[0.1, 0.6]$, the $\Theta_j^{(1)}$ values are generated uniformly on $[0.3, 0.7]$, and the $\Theta_j^{(2)}$ values are generated uniformly on $[0.4, 0.9]$.

In Case 4, $K = 5$. We set $t_k = kp/5$, $k = 1, \dots, 4$. The $\Theta_j^{(0)}$ values are generated uniformly on $[0.1, 0.5]$, the $\Theta_j^{(1)}$ values are generated uniformly on $[0.3, 0.7]$, the $\Theta_j^{(2)}, \Theta_j^{(4)}$ values are generated uniformly on $[0.5, 0.9]$, and the $\Theta_j^{(3)}$ values are generated uniformly on $[0.1, 0.9]$.

In all cases, the probe effects ϕ_i are generated by creating a random vector uniformly distributed in the hypercube $[0.5, 2]^p$ and normalizing it to satisfy

$\sum_{t_{k-1} < i \leq t_k} \phi_i^2 = t_k - t_{k-1}$, and the random errors ϵ_{ij} are drawn i.i.d. and uniformly from $[-0.2, 0.2]$.

We note that the simulated Y values loosely mimic the observed β values. Alternatively, we can generate Y values that mimic the observed M values. Since our methodology puts no restriction on the choice of metrics, its performance will be similar in the latter case.

We simulated 200 data sets for each configuration in each case. To assess the performance, we use (1) true positive rate, which is defined as the ratio of the number of correctly detected change-points to the number of true change-points; (2) false positive rate, which is defined as the ratio of the number of falsely detected change-points to the total number of non-change-point locations; and (3) mean squared errors of estimates of the parameters,

$\|\hat{\phi}(\hat{k}) - \phi\|_2^2/p$ and $\|\hat{\Theta}(\hat{k}) - \Theta\|_F^2/(np)$. These three criteria are commonly used in the statistics and biostatistics literature. Note that the true positive rate is not defined in Case 1.

The discussion below concentrates on the modified BIC. The corresponding results for the modified AIC are qualitatively similar and are put in Web Appendix A.

From Figure A1 (Web Appendix A), we see that the true positive rate is almost 100% in Case 2 in which there is one change-point. It drops as K increases. This is reasonable as K reflects the difficulty of the estimation problem. As can be seen from Figure A2 (Web Appendix A), the false positive rate is zero in Case 1 and is close to zero in other cases, indicating that our method has very high specificity. In Case 4, the true positive rate increases and the false positive rate decreases, as the number of probes increases (Figures A1 and A2). Since our method partitions a probe set into blocks, and then averages methylation values within each block while correcting for probe effects, the number of probes within each probe should have a positive effect on the true positive rate. These observations suggest that our method is able to adaptively discriminate between the case of no change-points and the cases with one or more change-points, and for moderately large p , it can reliably detect the locations of true change-points if they exist, while keeping the false positive rate relatively low. The new procedure also performs well in terms of parameter estimation. Both the average mean squared error of φ (see Figure A3 in Web Appendix A) and that of Θ are very small (Figure 1). The estimation accuracy deteriorates as K increases, and improves as p increases especially in Case 4. Therefore, for moderately large p , our method has good performance. Finally, we note that in general the performance improves as n increases. This is due to the interplay between the estimation of probe effects, which utilizes all samples, and the estimation of the true methylation values, which benefits from the former.

Note that some simulated Y values in the study above may be outside the range of 0 to 1. As suggested by one referee, we can truncate them before running the GFlars algorithm. We perform a simulation by considering Case 3 with $n = 80$, in which 5-6% of the simulated Y values are either less than 0 or larger than 1. From Figure A8 in Web Appendix A, we see that the performance on the truncated data is comparable to that on the raw data.

4. Detecting differentially methylated regions

4.1 The detection procedure

Differentially methylated regions (DMRs), as genomic regions with different methylation status under different conditions (e.g., tumor subtypes), are regarded as possible functional regions involved in gene transcriptional regulation (Song et al., 2005). Based on GFlars, we develop a procedure for detecting DMRs, without pre-specifying the genomic regions. For simplicity, we consider the detection problem under two conditions.

For a gene or a probe set, the procedure is described as follows.

Step 1. Run GFlars on all samples to identify K genomic regions adaptively. Let $\Theta_j^{(k)}$ denote the estimated methylation index for the k th region in the j th sample.

Step 2. For each region $k \in \{1, \dots, K\}$, separate $\{\Theta_j^{(k)}, j = 1, \dots, n\}$ according to the two conditions, and perform a two-sample t -test.

Step 3. Use an FDR controlling procedure to correct for multiple comparisons. Report a list of DMRs.

Here, we assume that samples under each condition share the same set of change points. However, the two change-point sets may differ between the two conditions. Therefore, the above procedure, which uses all the samples in a study to estimate probe effects, identifies the union of these two sets of change-points. It thus provides one possible but clearly not the best solution for detecting DMRs.

4.2 Real data application

We use a breast cancer dataset from The Cancer Genome Atlas project (TCGA, 2012) as an example to illustrate the usage of our procedure. This dataset consists of DNA methylation measurements from Infinium HumanMethylation450 BeadChip for breast cancer patients with different tumor subtypes. We focus on two subtypes: Her2 and Basal-like. The Her2 subtype is characterized by the expression of Her2. It tends to be more aggressive than other subtypes, and is relatively well understood with effective clinical treatments available. On the other hand, the Basal-like subtype (70% as triple-negative breast cancers) is poorly understood, with chemotherapy as the main therapeutic option (TCGA, 2012). As a result, characterization of the Basal-like subtype at the molecular level has important implications in both diagnosis and treatment. With data from 12 Her2 samples and 31 Basal-like samples, we want to detect DMRs between these two subtypes.

We downloaded level 3 methylation data (pre-normalized β values using the **methyllumi** package) from the TCGA data portal. We do not pursue further preprocessing since no significant batch effects were reported from previous studies. To facilitate the detection of DMRs, we divide the entire human genome into 200-probe non-overlapping windows (i.e., each window contains 200 probes). For each window, we apply our detection procedure (under the same name GFlars) to the observed methylation values. To control the False Discovery Rate (FDR), we use the Benjamini-Hochberg procedure. In the first part of the analysis, all regions with corrected p -values less than 0.01 are identified as DMRs. In total, we are able to identify 506 DMRs, including 132 with single CpG site and 374 with at least two probes. Among these, 116 DMRs have more than 10 probes.

We then compare GFlars with a popular DMR detection method called Bumphunter (Jaffe et al., 2012). Bumphunter collapses probes with a certain distance into small intervals. Within each interval, it uses the **loess** function in R to smooth the signals locally. Then, a permutation based procedure is applied to evaluate the statistical significance. We use the default parameters (except 250 KB windows for the smoothing procedure for the detection of large DMRs). Using the FDR cut-off of 0.01, we are able to identify 695 DMRs, including 258 with single CpG site and 437 with at least two probes. Among these, 156 DMRs share at least one probe with 153 DMRs from GFlars (hereafter called *shared DMRs*). Furthermore, 40 DMRs have more than 10 probes, and 16 of them are shared DMRs. Generally, DMRs from Bumphunter have smaller sizes than those from GFlars do.

In total, 4,477 probes are covered by GFlars DMRs and 2,337 are covered by Bumphunter DMRs. We further check the annotations of these probes using **Minfi** (Aryee et al., 2014).

The two probe sets have almost identical distributions with respect to their relation to CpG islands. More specifically, 27.5% of these probes are distributed on CpG islands, 7.5% on CpG islands Shelf, 21% on CpG islands Shore, and 44% on Open Sea. This is somewhat different from the distribution of all probes on the 450K array, where the corresponding proportions on CpG islands, Shelf, Shore, and Open Sea are 31%, 10%, 23%, and 36%, respectively. More differentially methylated probes are located in the Open Sea regions, which is consistent with previous observations (Timp et al., 2014). For the annotation with respect to functional genomic elements, no difference is observed between GFlars and Bumphunter, nor between any of those and the probe design. The above results are summarized in Figure 2.

To further compare GFlars and Bumphunter, we consider three cases, including: 1) use of two additional FDR cut-offs, 0.025 and 0.05, 2) selection of the top 1000, 500, and 200 regions ranked by p -values, and 3) use of two FDR cut-offs, 0.025 and 0.05, for Bumphunter, with a fixed FDR cut-off of 0.01 for GFlars. The results are shown in Table 1. Using the FDR cut-off of 0.01, 27% of the DMRs detected by GFlars are shared DMRs. Increasing the cut-off for Bumphunter up to 0.05, this number increases to 45%. For the top 1000 DMRs, the GFlars set shares 26% with the Bumphunter set. For the top 200 DMRs, the percentage of shared DMRs reduces to 9%.

For each shared DMR, we check the difference between the numbers of CpG sites covered by the two methods. Figure 3 shows the distributions of the difference for shared DMRs generated using different cut-off values. As we can see, these distributions are almost identical across cut-offs. Of the shared DMRs, around 30% have zero difference, over 60% have GFlars covering more CpG sites than Bumphunter, and about 40% have the difference being larger than 10.

For the shared DMRs, GFlars tends to cover more CpG sites and seems to provide more accurate boundaries, as evidenced by four typical examples shown in Figure 4. This is because GFlars can borrow strength from the neighboring sites to smooth the methylation values, and at the same time pool information on the same probe across samples to estimate the probe effect. The corresponding probe dependent effects are shown in Web Appendix B.

Finally, we investigate 10 large GFlars DMRs (at least 10 probes) with the smallest p -values (Table 2). All of them are shared DMRs, however with fewer probes covered by Bumphunter. We also characterize each region with its genomic location, overlapped gene list, and whether observed higher methylation in Basal samples. Interestingly, we find that three trefoil family genes, TFF1, TFF2, and TFF3, are differentially methylated between Basal and Her2. It has been reported that TFF1 stimulates migration of human breast cancer cells; TFF2 inhibits apoptosis in breast cell lines; and all three are characterized by a hormonally regulated transcriptional program and response to androgen (Prest et al., 2002; Siu et al., 2004; Doane et al., 2006). In summary, our results suggest that the methylation levels of these genes may be informative to discern breast cancer subtypes.

To test how robust our findings are to the choice of 200-probe non-overlapping windows, we re-apply our method on sliding windows with different numbers of probes shifted. The

results reported in Web Appendix C show that our method is very robust with the detected regions in Figure 4 highly repeatable. Finally, we consider using the M value and compare the results with those using the β value. Using an FDR cut-off of 0.05, 1,594 and 1,436 DMRs are identified respectively from M values and from β values, nearly half of which are shared DMRs. However, we observe that the M -value-based DMRs tend to be over-smoothed. From Figure D1 in Web Appendix D, we see that about 60% of the shared DMRs for the M values are at least 10 probes longer than those for the β value.

5. Discussion

We have proposed a group fused lasso method to construct DNA methylation indexes for subsets of neighboring probes with similar true values in methylation. Our method adaptively partitions a probe set (e.g., gene) into subsets, and within each subset it estimates the methylation level while correcting for probe-specific effects. As a model-based method, it is an extension to microarray methylation data of the well-known method for gene expression indexes developed by Li and Wong (2001). While the probe affinity effect is well-studied in the analysis of high-density oligonucleotide array probe-level data, it receives little attention in the methylation literature. To the best of our knowledge, the proposed method is the first for estimating DNA methylation levels by joint modeling of multiple methylation profiles from microarray data. Simulations have shown that it performs well in identifying subsets of probes and removing probe-specific effects.

Using the estimated methylation values, we can further detect differentially methylated regions. We have proposed a procedure by collapsing adjacent probes with the same estimated methylation level into regions, applying t -test for each region, and correcting the p -values for multiple comparisons. We have applied this procedure to a breast cancer dataset from TCGA, and have found that it performs better in detecting large differentially methylated regions than a state-of-the-art approach, Bumhunter. Even when the window size was set to be 250 KB, most regions detected by Bumhunter contained only one or two probes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

- Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974; 19:716–723.
- Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium dna methylation microarrays. *Bioinformatics*. 2014; 30:1363–1369. [PubMed: 24478339]
- Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, et al. High density dna methylation array with single cpg site resolution. *Genomics*. 2011; 98:288–295. [PubMed: 21839163]
- Bleakley, K.; Vert, J-P. The group fused lasso for multiple change-point detection. 2011. arXiv preprint arXiv:1106.4199

- Bock C. Analysing and interpreting dna methylation data. *Nature Reviews Genetics*. 2012; 13:705–719.
- Breheny P, Huang J. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*. 2015; 25:173–187. [PubMed: 25750488]
- Doane A, Danso M, Lal P, Donaton M, Zhang L, Hudis C, Gerald W. An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. *Oncogene*. 2006; 25:3994–4008. [PubMed: 16491124]
- Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*. 2010; 11:587. [PubMed: 21118553]
- Hansen KD, Langmead B, Irizarry RA, et al. Bsmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*. 2012; 13:R83. [PubMed: 23034175]
- Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003; 4:249–264. [PubMed: 12925520]
- Jaffe AE, Murakami P, Lee H, Leek JT, Fallin MD, Feinberg AP, Irizarry RA. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*. 2012; 41:200–209. [PubMed: 22422453]
- Jones PA. Functions of dna methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*. 2012; 13:484–492.
- Kurozumi E, Tuvaandorj P. Model selection criteria in multivariate models with multiple structural changes. *Journal of Econometrics*. 2011; 164:218–238.
- Laird PW. Principles and challenges of genome-wide dna methylation analysis. *Nature Reviews Genetics*. 2010; 11:191–203.
- Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*. 2001; 98:31–36.
- Lister R, Pelizzola M, Downen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, et al. Human dna methylomes at base resolution show widespread epigenomic differences. *Nature*. 2009; 462:315–322. [PubMed: 19829295]
- Ordway JM, Curran T. Methylation matters: modeling a manageable genome. *Cell Growth and Differentiation*. 2002; 13:149–162. [PubMed: 11971815]
- Picard F, Lebarbier E, Hoebeke M, Rigai G, Thiam B, Robin S. Joint segmentation, calling, and normalization of multiple cgh profiles. *Biostatistics*. 2011; 12:413–428. [PubMed: 21209153]
- Prest SJ, May FE, Westley BR. The estrogen-regulated protein, tff1, stimulates migration of human breast cancer cells. *The FASEB Journal*. 2002; 16:592–594. [PubMed: 11919164]
- Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of cpg dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*. 2006; 103:1412–1417.
- Schwarz G. Estimating the dimension of a model. *The Annals of Statistics*. 1978; 6:461–464.
- Siegmund KD. Statistical approaches for the analysis of dna methylation microarray data. *Human Genetics*. 2011; 129:585–595. [PubMed: 21519831]
- Siu L-S, Romanska H, Abel PD, Baus-Loncar M, Kayademir T, Stamp GW, Lalani E-N. Tff2 (trefoil family factor2) inhibits apoptosis in breast and colorectal cancer cell lines. *Peptides*. 2004; 25:855–863. [PubMed: 15177882]
- Song F, Smith JF, Kimura MT, Morrow AD, Matsuyama T, Nagase H, Held WA. Association of tissue-specific differentially methylated regions (tdms) with differential gene expression. *Proceedings of the National Academy of Sciences*. 2005; 102:3336–3341.
- Tammen SA, Friso S, Choi S-W. Epigenetics: the link between nature and nurture. *Molecular Aspects of Medicine*. 2013; 34:753–764. [PubMed: 22906839]
- TCGA. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012; 490:61–70. [PubMed: 23000897]

- Timp W, Bravo HC, McDonald OG, Goggins M, Umbricht C, Zeiger M, Feinberg AP, Irizarry RA. Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. *Genome Medicine*. 2014; 6:61. [PubMed: 25191524]
- Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*. 2006; 68:49–67.
- Zhang NR, Siegmund DO. A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*. 2007; 63:22–32. [PubMed: 17447926]
- Zhang NR, Siegmund DO. Model selection for high-dimensional, multi-sequence change-point problems. *Statistica Sinica*. 2012; 22:1507–1538.
- Zhang NR, Siegmund DO, Ji H, Li JZ. Detecting simultaneous changepoints in multiple sequences. *Biometrika*. 2010; 97:631–645. [PubMed: 22822250]

Figure 1.

Plots of the mean squared error (MSE) of $\Theta(\hat{k})$ versus the number p of probes for the modified BIC. Each panel shows four curves, corresponding to the sample sizes $n \in \{20, 40, 80, 160\}$, and each point on each curve represents the average of 200 trials.

Figure 2.
A. Venn diagram for Differentially Methylated Regions detected by GFlars and Bumhunter.
B. Boxplots of the distribution of the number of probes in the set of DMRs detected by GFlars uniquely, GFlars intersected with Bumhunter, Bumhunter uniquely, and Bumhunter intersected with GFlars. **C.** The annotation of GFlars detected probes, Bumhunter detected probes, and all probes on the 450K array with respect to functional genomic elements. **D.** The annotation of GFlars detected probes, Bumhunter detected probes, and all probes on the 450K array with respect to their relation to CpG islands.

Figure 3.
The distributions of the difference between the numbers of CpG sites for shared DMRs under different cut-offs.

Figure 4.

Four examples demonstrate the advantage of GFlars over Bumphunter. Each dot in each panel represents the β value from a TCGA breast cancer sample for a specific probe, with Basal samples shown in blue and Her2 shown in orange. The median values across the same subtype are shown in solid curves. The x-axis represents the genomic coordinates. GFlars and Bumphunter detected DMRs are shown in the dashed red box and solid black box, respectively.

Table 1

Comparisons of DMRs detected by GFlars and Bumphunter with different cut-offs.

GFlars Cut-off	Bumphunter Cut-off	GFlars Unique	Bumphunter Unique	GFlars Intersected	Bumphunter Intersected	GFlars Total Probes	Bumphunter Total Probes
FDR 0.01	FDR 0.01	369	539	137	156	4477	2337
FDR 0.01	FDR 0.025	303	1015	203	242	4477	3903
FDR 0.01	FDR 0.05	276	1374	230	283	4477	4968
FDR 0.025	FDR 0.025	618	914	286	343	8439	3903
FDR 0.05	FDR 0.05	969	1057	467	600	14835	4968
Top 1000	Top 1000	741	698	259	302	9157	3177
Top 500	Top 500	407	394	93	106	4017	1671
Top 200	Top 200	182	181	18	19	1465	610

Table 2

Ten large DMRs (at least 10 probes) with the smallest p -values identified by GFlars.

Chromosome	Location	Gene	Level in Basal	Number of probes	
				GFlars	Bumphunter
chr14	24802150:24804342	ADCY4	Low	12	10
chr7	101762511:101829412	CUX1	Low	22	6
chr6	158375684:158439902	SYNJ2	Low	27	2
chr3	112324301:112694062	CCDC80;CD200R1L;CD200R1	High	35	12
chr22	38337780:38434179	MICALL1;C22orf23;POLR2F;SOX10	Low	34	4
chr7	44058453:44061603	POLR2J4	High	13	1
chr10	102895043:102900491	TLX1	Low	17	7
chr22	45704902:45706726	FAM118A	Low	15	7
chr10	20230737:21456039	PLXDC2;NEBL;C10orf113	High	36	2
chr21	43729657:43792734	TFF3;TFF2;TFF1;TMPRSS3	High	27	12