BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 22 2010, pages 2849–2855
doi:10.1093/bioinformatics/btq553

Gene expression

Advance Access publication September 29, 2010

A statistical framework for Illumina DNA methylation arrays

Pei Fen Kuan[1,2,*], Sijian Wang[3], Xin Zhou[1] and Haitao Chu[4]

[1]Department of Biostatistics, [2]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599, [3]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI 53792 and [4]Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The Illumina BeadArray is a popular platform for profiling DNA methylation, an important epigenetic event associated with gene silencing and chromosomal instability. However, current approaches rely on an arbitrary detection *P*-value cutoff for excluding probes and samples from subsequent analysis as a quality control step, which results in missing observations and information loss. It is desirable to have an approach that incorporates the whole data, but accounts for the different quality of individual observations.

Results: We first investigate and propose a statistical framework for removing the source of biases in Illumina Methylation BeadArray based on several positive control samples. We then introduce a weighted model-based clustering called LumiWCluster for Illumina BeadArray that weights each observation according to the detection *P*-values systematically and avoids discarding subsets of the data. LumiWCluster allows for discovery of distinct methylation patterns and automatic selection of informative CpG loci. We demonstrate the advantages of LumiWCluster on two publicly available Illumina GoldenGate Methylation datasets (ovarian cancer and hepatocellular carcinoma).

Availability: `R` package `LumiWCluster` can be downloaded from http://www.unc.edu/~pfkuan/LumiWCluster

Contact: pfkuan@bios.unc.edu

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

Received on June 18, 2010; revised on September 9, 2010; accepted on September 23, 2010

1 INTRODUCTION

DNA methylation is an important epigenetic modification and plays critical roles in transcriptional regulation, chromosomal stability, genomic imprinting and X-inactivation (Rakyan *et al.*, 2008). Numerous literatures have established the influence of DNA methylation in transcriptional aberrations in human diseases including various types of cancer (Esteller, 2007; Irizarry *et al.*, 2009; Koga *et al.*, 2009). DNA methylation occurs in cytosines of CpG dinucleotides in human in a non-random fashion across the genome. In particular, CpG-rich regions (known as CpG islands) are usually hypomethylated, whereas repetitive genomic sequences are hypermethylated in normal cells. Over the past decade, there has been vast research in studying alterations of DNA methylation in cancer. A general observed trend of perturbed DNA methylation includes hypomethylation of oncogenes and hypermethylation of tumor suppressor genes, leading to genomic instability and tumorigenesis (Esteller, 2007; Irizarry *et al.*, 2009).

Several platforms are available for DNA methylation profiling that includes high-throughput arrays and more recently, the next-generation sequencing instruments. The experimental approaches in high-throughput array-based methylation include bisulfite conversion-based methods, restriction enzyme-based methods and immunoprecipitation-based methods (Down *et al.*, 2008). A popular robust methylation profiling platform via bisulfite conversion is the Illumina GoldenGate and Infinium Methylation Assays based on the BeadArray technology. This technology utilizes 3 μm silica beads which are replicated ~30 times on the array and has emerged as an attractive platform for genotyping, expression and methylation analysis (Dunning *et al.*, 2008b; Lynch *et al.*, 2009; Xie *et al.*, 2009). This technology requires less sample input but produces high-quality data (Dunning *et al.*, 2008b; Xie *et al.*, 2009), thereby reducing the cost of array experiments. The increasing popularity of Illumina BeadArray technology is apparent given the numerous scientific publications since 2008. However, there is only a handful of statistical framework for analyzing Illumina BeadArray gene expression (Dunning *et al.*, 2008b; Wong *et al.*, 2008), and limited work is available for methylation array counterpart (Lynch *et al.*, 2009). Existing work for gene expression BeadArray falls into the categories of data preprocessing and differential expression detection. This includes background correction methods (Dunning *et al.*, 2008b; Xie *et al.*, 2009), variance-stabilizing techniques (Dunning *et al.*, 2008a) and modified test statistics for differential gene expression in Illumina BeadArray (Wong *et al.*, 2008).

Most of the framework for gene expression analysis is based on the assumption that the majority of the genes are not differentially expressed. In contrast, many sites are expected to be methylated (Irizarry *et al.*, 2008), and therefore the assumptions in gene expression are not applicable to methylation experiments. The goal of our article is to provide a statistical framework for array-based methylation profiling on Illumina BeadArray technology, by studying the source of biases and the data-generating mechanism of Illumina methylation assays. Specifically, we propose a model for correcting the source of biases in Illumina Methylation BeadArray and introduce a weighted model-based approach for clustering the methylation profiles. The framework of our weighted model-based clustering can also be directly applied to other Illumina BeadArray platforms, e.g. gene expression BeadChip, because it does not rely on the assumption that most beads are from the null distribution of no differential expression. In the next section, we describe the

*To whom correspondence should be addressed.

© The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

methylation data structure and introduce our proposed statistical framework.

2 MOTIVATION

Methylation levels in Illumina methylation assays are quantified by the *beta* value using the ratio of intensities between methylated (M) and unmethylated (U) alleles. Specifically,

$$beta = \frac{\max(M,0)}{\max(M,0)+\max(U,0)+100}$$

where M and U are the red and green dyes, respectively, for the GoldenGate and VeraCode Methylation assays, whereas for Infinium assay, M and U are signals A and B (produced by two different bead types and reported in the same color), respectively. The constant 100 is to regularize *beta* when both M and U values are small (Bibikova *et al.*, 2006). The *beta* values are continuous and range from 0 (unmethylated) to 1 (completely methylated). Each locus reports an average *beta* value obtained from the average of M's and U's across approximately 30 bead replicates, and individual bead-level measurements are not readily available (Dunning *et al.*, 2008b; Wong *et al.*, 2008). A standard summary output from BeadStudio (Illumina software to process raw intensities) includes four columns for each sample, i.e. (1) average *beta*, (2) average M, (3) average U and (4) detection P-values, for each locus. Therefore, our proposed framework will be based on the average *beta* values for convenience.

The detection P-value reported by BeadStudio can be used as a quality control measure of probe performance. The detection P-value is defined as $1-P$-value computed from the background model characterizing the chance that the signal was distinguishable from negative controls (Supplementary Materials). Standard protocol by Illumina recommends excluding probes that have a detection P-value greater than an *arbitrary* cutoff of 0.05. On the other hand, Marsit *et al.* (2009) excluded samples that consist of $\geq 25\%$ observations with detection P-values $\geq 1\times 10^{-5}$, as well as probes (CpG loci) with median detection P-values >0.05, whereas Hernandez-Vargas *et al.* (2010) excluded probes with detection P-values >0.01 in $>10\%$ of the samples. In Section 3.2, we will introduce a modeling framework that avoids arbitrary choice of detection P-value threshold.

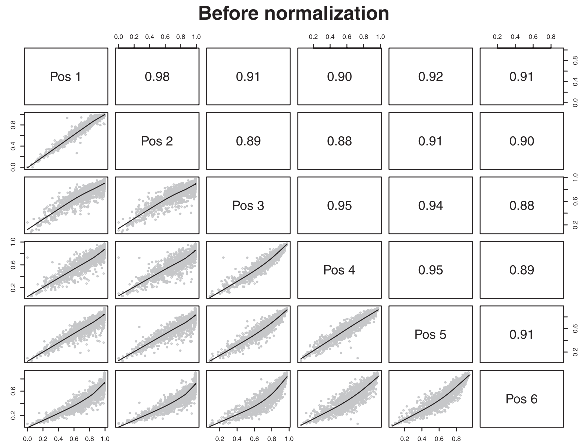

Fig. 1. Pair plots of the six positive controls before normalization. Upper panel prints the pairwise correlation coefficient.

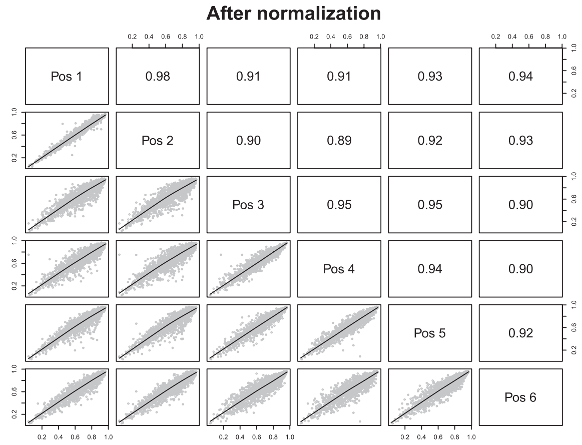

Fig. 2. Pair plots of the six positive controls after quantile normalization. Upper panel prints the pairwise correlation coefficient.

We will now explore and illustrate the source of biases present in Illumina Methylation BeadArray based on the data generated by the Thomas-Conway Lab at UNC-Chapel Hill. The methylation experiment is performed on the GoldenGate Cancer Panel I methylation panel, which interrogates 1505 CpG loci/probes associated with 803 cancer-related genes (tumor suppressor genes, oncogenes, genes involved in DNA repair, cell-cycle control, differentiation, apoptosis, X-linked and imprinted genes) where 28.6% contain one CpG site per gene, 57.3% contain two CpG sites and 14.1% have three or more sites (Illumina, 2006). The probe length varies between 41 bp and 59 bp with median 50 bp, whereas the number of CG dinucleotides for each probe varies between 1 and 7. A pair of allele-specific oligonucleotide (ASO) and locus-specific oligonucleotide (LSO) measures the methylation level of a specific CG dinucleotide for a probe under the assumption that flanking CG dinucleotides within the same probe exhibit similar methylation status. In other words, the observed methylation level for the measured CG dinucleotide should not be affected by the number of flanking CG dinucleotides. To investigate the potential source of biases in Illumina Methylation assays, we utilize six positive control samples from our dataset, in which all the cytosines in CG dinucleotides are expected to be methylated and any deviation from methylated status indicates the presence of technical biases. We apply quantile normalization to these six positive control samples. Figures 1 and 2 compare the pairwise correlations among these positive controls before and after normalization, respectively. Both plots show the high correlations among the positive controls, with improvement after normalization.

In addition, array-based technology are known to be affected by the sequence and thermodynamic properties, e.g. melting temperature and GC content in protein–DNA binding and gene expression experiments (Dunning *et al.*, 2008b; Wei *et al.*, 2008). In Figure 3, we plot the quantile normalized average *beta* values pooled from all positive controls against probe length, number of CG dinucleotides within each probe, melting temperature and GC content. Individual plots for each positive control sample are given in Supplementary Materials. Melting temperature is computed according to Wei *et al.* (2008), whereas GC content is the percentage of C and G nucleotides for a given probe. As evident from Figure 3,

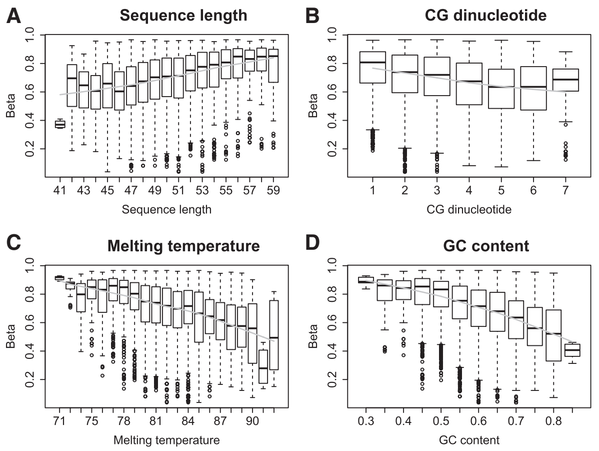

Fig. 3. The effects of (**A**) sequence length, (**B**) number of CG dinucleotides, (**C**) melting temperature and (**D**) GC content bias. Gray line is the lowess approximation.

the observed methylation level is influenced by the sequence and thermodynamics properties of the probes. Sequence length bias in methylation array is also observed in Lynch *et al.* (2009). Since melting temperature is a function of GC content and sequence length, and GC content is highly correlated with the number of CG dinucleotides, our modeling framework will incorporate sequence length and GC content. In gene expression and protein–DNA binding arrays, GC content exhibits increasing trend with intensities due to the three hydrogen bonds compared with two hydrogen bonds in AT pairs. However, the decreasing trend observed in methylation arrays can be attributed to the loss of efficiency in binding for a probe with more CG dinucleotides, because the CG dinucleotides within a probe are expected to have similar methylation status. Although CG dinucleotides yield a more straightforward interpretation, we choose GC content which has more distinct values for better function approximation.

3 METHODS

3.1 Estimating L_j and GC_j biases

For notational brevity, we denote the average *beta* value and detection *P*-value for each locus j, $j=1,\dots,p$ and sample i, $i=1,\dots,n$ as β_{ij} and p_{ij}, respectively. Let L_j and GC$_j$ denote the probe length and GC content for locus j, respectively. Since $\beta_{ij}\in(0,1)$, beta distribution arises as a natural distribution for modeling the observed β_{ij} (Houseman *et al.*, 2008; Siegmund *et al.*, 2004). However, maximum likelihood estimation of the unknown parameters (α,β) in a beta distribution does not have a closed form and relies on numerical methods (Ji *et al.*, 2005). In this article, we consider an alternative for modeling β_{ij} via a logit transformation,

$$y_{ij} = \log\left(\frac{\beta_{ij}}{1-\beta_{ij}}\right)\in\mathbb{R}$$

To avoid a logit tranformation of $\beta=0$, we add an $\epsilon=10^{-4}$ to β as

$$\beta = \frac{\max(\bar{M},\epsilon)}{\max(\bar{M},\epsilon)+\max(\bar{U},\epsilon)+100}$$

where $\bar{M}$ and $\bar{U}$ are the average M and U values across ∼30 replicates as mentioned in Section 2. We model

$$y_{ij} = s_{ij}+h_1(L_j)+h_2(GC_j)$$

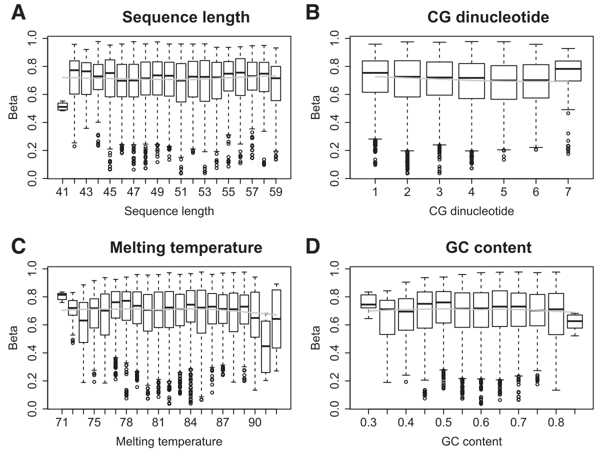

Fig. 4. The effects of (**A**) sequence length, (**B**) number of CG dinucleotides, (**C**) melting temperature and (**D**) GC content bias after correction. Gray line is the lowess approximation.

where s_{ij} is the true methylation level and h_1, h_2 characterize the bias arising from sequence length and GC content. We estimate h_1, h_2 from the positive control sample instead of treatment sample itself. This is to safeguard against removing actual methylation signals due to the potential confounding effect with GC content, i.e. hypomethylation in CpG islands (CG-rich regions) of normal cells (Esteller, 2007; Irizarry *et al.*, 2008). Let

$$\begin{aligned}h_1(L_j) &= \alpha_1 I(L_j<44)+\alpha_2 L_j I(44\le L_j<57)\\&\quad+\alpha_3 I(L_j\ge 57)\\h_2(GC_j) &= \gamma_1 I(GC_j<0.4)+\gamma_2 GC_j I(0.4\le GC_j<0.8)\\&\quad+\gamma_3 I(GC_j\ge 0.8)\end{aligned}$$

under the constraints (1) $\alpha_1=44\alpha_2$, (2) $\alpha_3=57\alpha_2$, (3) $\gamma_1=0.4\gamma_2$ and (4) $\gamma_3=0.8\gamma_2$ for continuity at the knots. That is, we model h_1 and h_2 as piecewise constant + linear + constant. The knots are chosen so that the number of observations in $\{L_j<44\}$ and $\{L_j\ge 57\}$ are comparable with the number of observations in $\{L_j=n\}$ for $n=44,\dots,56$ as well as to avoid over (under)-estimation for large (small) L_j, and vice versa for GC$_j$.

In Figure 4, we plot the corrected $\hat{\beta}_{ij}$ against the attributes, where

$$\hat{\beta}_{ij} = \frac{\exp(y_{ij}-\hat{h}_1(L_j)-\hat{h}_2(GC_j))}{1+\exp(y_{ij}-\hat{h}_1(L_j)-\hat{h}_2(GC_j))}$$

for the positive control. As evident from this figure, the bias of these four attributes are reduced substantially.

3.2 A weighted model-based approach

One of the most common applications of DNA methylation is in identifying subgroups with distinct methylation patterns (Christensen *et al.*, 2009; Houseman *et al.*, 2008; Marsit *et al.*, 2009; Shen *et al.*, 2009; Siegmund *et al.*, 2004) via unsupervised clustering techniques. Numerous clustering methods have been developed, including non-parametric (e.g. agglomerative hierarchical clustering) and model-based approaches. Model-based clustering assumes that the data is generated from a finite mixture model, in which each mixture component corresponds to a cluster. It has emerged as a popular technique and allows for statistical inference (e.g. selecting number of clusters and estimating membership probability) to be carried out (Fraley and Raftery, 2002; Siegmund *et al.*, 2004). We let $\boldsymbol{y}_i=(y_{i1},\dots,y_{ip})^T$ to be a vector of logit-transformed *beta* values for sample i. We assume that $\boldsymbol{y}_i$ is generated from a mixture of K multivariate normal distributions. Specifically,

$$\boldsymbol{y}_i \sim \sum_{k=1}^{K} \pi_k f_k, \ \sum_{k=1}^{K} \pi_k = 1,$$

$$f_k \sim \mathcal{N}(\mu_k + h_1(L) + h_2(GC), \Sigma_k), \ k \in \{1, \dots, K\}$$

where h_1, h_2 are pre-estimated from positive control samples as shown in Section 3.1. Define $\tilde{y}_{ij} = y_{ij} - \hat{h}_1(L_j) - \hat{h}_2(GC_j) \sim \sum_{k=1}^{K} \pi_k \tilde{f}_k = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$. Let $\boldsymbol{\theta}_k = (\pi_k, \mu_k, \Sigma_k)$ be the unknown parameters. The mixture model log-likelihood function for the whole data is given by

$$\mathcal{L} = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \tilde{f}_k(\tilde{\boldsymbol{y}}_i; \boldsymbol{\theta}_k)$$

As pointed out in Section 2, standard preprocessing steps for Illumina methylation assays include a quality control of probe measurements by excluding probes that have detection P-values (p_{ij}) larger than an arbitrary cutoff. This step results in missing observations and information loss by discarding a subset of probes. To avoid using a hard threshold for quality assessment, we would like to assign a weight for each sample which reflects its quality rather than discarding a sample completely from the analysis. We introduce an alternative model-based clustering on a weighted likelihood-based approach. The objective of a weighted likelihood function is to assign a different weight to each sample, in which samples with higher weights have more influence in estimating the mixture parameters for cluster structure inference. The weighted mixture model log -likelihood function is given by

$$\mathcal{L}_W = \sum_{i=1}^{n} w_i \log \sum_{k=1}^{K} \pi_k \tilde{f}_k(\tilde{\boldsymbol{y}}_i; \boldsymbol{\theta}_k)$$

where w_i is the weight of sample i. Without loss of generality, we assume $\sum_{i=1}^{n} w_i = 1$. The detection P-values arise as a natural weight function, since samples with large detection P-values are less reliable. A possible choice of weight function is $w_i = \text{median}_j(\log p_{ij}) / \sum_{i=1}^{n} \text{median}_i(\log p_{ij}) \in (0,1)$. Note that the criteria in Marsit *et al.* (2009) which excluded samples that consist of $\geq$25% observations with detection P-values $\geq 1 \times 10^{-5}$ is a special case by defining $w_i = I[Q3_j(p_{ij}) < 1 \times 10^{-5}] / \sum_{i=1}^{n} I[Q3_j(p_{ij}) < 1 \times 10^{-5}]$, where $Q3$ is the third quartile.

Weighted model-based clustering has been shown to outperform the non-weighted method in both simulations and real datasets from remote sensing images in geology (Richards *et al.*, 2009). In addition, Seo *et al.* (2004) showed that detection P-values weighting in computing Pearson's correlation coefficient improved the performance of expression profiling in Affymetrix microarrays. The mixture modeling framework can be recast in an expectation–maximization (EM) framework for estimating the unknown parameters $\boldsymbol{\theta}_k$. We introduce z_{ik} to be the unobserved indicator latent variable taking value 1 if sample i belongs to cluster k and 0 otherwise. The complete weighted log-likelihood function is

$$\mathcal{L}_{CW} = \sum_{i=1}^{n} \sum_{k=1}^{K} w_i z_{ik} [\log \pi_k + \log(\tilde{f}_k(\tilde{\boldsymbol{y}}_i; \boldsymbol{\theta}_k))]$$

In clustering DNA methylation profiles for identifying subgroups among the n samples, we are faced with the well-known 'large p, small n' problem. One strategy is to apply dimension reduction, e.g. principal component analysis (PCA) to the p loci, followed by clustering on the reduced space. However, treating dimension reduction and clustering as two separate steps may destroy the cluster structure in the data (Raftery, 2003; Wang and Zhu, 2008). Moreover, each PCA is a linear combination of all CpG loci, and does not allow for automatic selection of important CpG loci. Therefore, our goal is to incorporate a variable selection in model-based clustering approach, which identifies important CpG loci (variable selection) and subgroups among the n samples (clustering) simultaneously based on a penalized criterion (Pan and Shen, 2007; Wang and Zhu, 2008). We consider a penalized complete weighted log-likelihood to achieve the goal:

$$\mathcal{L}_{PCW} = \sum_{i=1}^{n} \sum_{k=1}^{K} w_i z_{ik} [\log \pi_k + \log(\tilde{f}_k(\tilde{\boldsymbol{y}}_i; \boldsymbol{\theta}_k))] - J(\boldsymbol{\Omega}) \qquad (1)$$

where $\boldsymbol{\Omega} = \{\mu_{kj}, k = 1, \dots, K; j = 1, \dots, p\}$ and $J(\boldsymbol{\Omega})$ is a penalty function. Several choices of penalty functions are available, e.g. Pan and Shen (2007) proposed an L_1-norm penalty function which takes the form $J(\boldsymbol{\Omega}) = \sum_{k=1}^{K} \sum_{j=1}^{p} |\mu_{kj}|$. As pointed out by Wang and Zhu (2008), however, there is a natural group structure among μ_{kj}'s, i.e. for each j, we can treat μ_{kj}, $k = 1, \dots, K$ as a group since they are associated with the same CpG locus. The L_1-norm penalty function ignores this group structure and treats μ_{kj} individually. As a result, it tends to keep many unimportant loci in the model.

To circumvent this problem, Wang and Zhu (2008) proposed a penalty function that incorporates group information and shrinks μ_{kj}'s more effectively. In addition, some loci have large detection P-values across all samples and are unreliable. Therefore, we would like to impose heavier penalty on these loci. We achieve this by introducing g_j to be weight of locus j, where larger g_j values indicate more reliable probes. One possible choice is $g_j = \text{median}_i(\log p_{ij}) / \sum_{j=1}^{p} \text{median}_i(\log p_{ij}) \in (0,1)$ (here we take median across samples, cf. w_i: median across loci). We generalize the proposed penalty function by Wang and Zhu (2008) by including the detection P-values as follows:

$$J(\boldsymbol{\Omega}) = \sum_{j=1}^{p} \frac{\gamma_j}{g_j \max_k(|\tilde{\mu}_{kj}|^{\alpha})} + \lambda \sum_{k=1}^{K} \sum_{j=1}^{p} \frac{|\theta_{kj}|}{|\tilde{\mu}_{kj}|^{\alpha}}$$

where $\mu_{kj} = \gamma_j \theta_{kj}$, $\tilde{\mu}_{kj}$'s are the unpenalized estimates of cluster means and α is a non-negative tuning parameter. Under this penalty function, loci with large detection P-values will be assigned a higher penalty, and are more likely to be excluded in the variable selection. λ is a tuning parameter that controls the sparsity, i.e. small (large) λ results in the selection of more (fewer) CpG loci. Additional details on the proposed penalty function are given in Supplementary Materials. We further assume that $\Sigma_k = \Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$ as in Wang and Zhu (2008). That is, the covariance matrices are the same across different clusters and are diagonal, a common assumption for a high dimension and small sample size problem. Further theoretical justifications for adopting the diagonal structure of the covariance matrix are provided by Bickel and Levina (2004).

At the t iteration of the EM algorithm, the E-step computes

$$\hat{z}_{ik}^{(t)} = \frac{\hat{\pi}_k^{(t-1)} \tilde{f}_k^{(t-1)}(\tilde{\boldsymbol{y}}_i; \hat{\boldsymbol{\theta}}_k^{(t-1)})}{\sum_{c=1}^{K} \hat{\pi}_c^{(t-1)} \tilde{f}_c^{(t-1)}(\tilde{\boldsymbol{y}}_i; \hat{\boldsymbol{\theta}}_k^{(t-1)})}$$

for $i = 1, \dots, n$ and $k = 1, \dots, K$.

The M-step involves maximizing Equation (1) with respect to (π_k, μ_{kj}, σ_j^2). Following the derivations in Wang and Zhu (2008) with modifications to incorporate the weights w_i, g_j and that $\sum_{i=1}^{n} w_i = 1$,

$$\hat{\pi}_k^{(t)} = \sum_{i=1}^{n} w_i \hat{z}_{ik}^{(t-1)},$$

$$\hat{\sigma}_j^{2(t)} = \sum_{k=1}^{K} \sum_{i=1}^{n} w_i \hat{z}_{ik}^{(t-1)} (\tilde{y}_{ij} - \hat{\mu}_{kj}^{(t)})^2,$$

The estimates for γ_j and θ_{kj} are not trivial, since the penalty function is singular at the origin point. However, similar to the derivation in Wang and Zhu (2008), we can update estimates of γ_j and θ_{kj} iteratively by the following explicit formula, which makes our method easy to implement in practice:

$$\hat{\gamma}_j = I_{(\exists k, \theta_{kj} \neq 0)} \left(\sum_{k=1}^{K} \frac{\xi_k \mu_{kj}^0}{\hat{\theta}_{kj} \sum_{k=1}^{K} \xi_k} - \frac{\hat{\sigma}_j^2}{g_j \max_k(|\tilde{\mu}_{kj}|^{\alpha}) \sum_{k=1}^{K} \xi_k} \right)_+$$

$$\hat{\theta}_{kj} = I_{(\gamma_j > 0)} \text{sign}(\mu_{kj}^0) \left(\frac{|\mu_{kj}^0|}{\gamma_j} - \frac{\lambda \hat{\sigma}_j^2}{|\tilde{\mu}_{kj}|^{\alpha} \gamma_j^2 \sum_{i=1}^{n} w_i \hat{z}_{ik}} \right)_+$$

where $\xi_k = \sum_{i=1}^{n} w_i \hat{z}_{ik} \hat{\theta}_{kj}^2$ and $\mu_{kj}^0 = \frac{\sum_{i=1}^{n} w_i \hat{z}_{ik} \tilde{\boldsymbol{y}}_i}{\sum_{i=1}^{n} w_i \hat{z}_{ik}}$. The E-step and M-step are iterated till convergence.

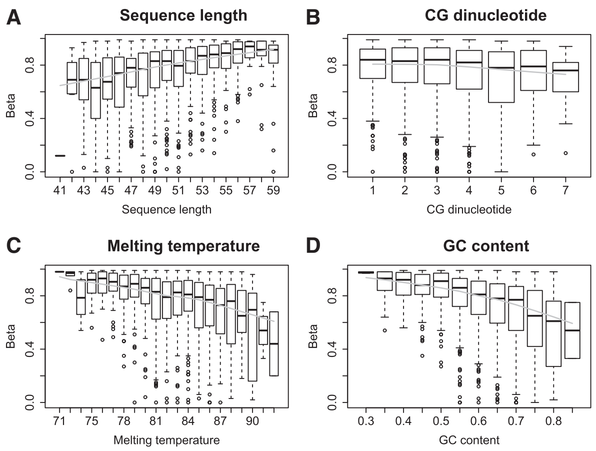

Fig. 5. The effects of (**A**) sequence length, (**B**) number of CG dinucleotides, (**C**) melting temperature and (**D**) GC content bias on positive control of Houshdaran *et al.* (2009). Gray line is the lowess approximation.

As in Pan and Shen (2007) and Wang and Zhu (2008), we choose the tuning parameter λ and the number of clusters K by minimizing the Bayesian information criterion (BIC). To account for the weights w_i in the likelihood functions in which $\sum_{i=1}^{n} w_i = 1$, we define a modified BIC as follows:

$$\mathrm{BIC} = -2n\sum_{i=1}^{n} w_i \log\left(\sum_{k=1}^{K}\hat{\pi}_k\tilde{f}_k(\tilde{\mathbf{y}}_i;\hat{\boldsymbol{\theta}}_k)\right)+P\log n$$

where P is the total number of non-zero estimates in $\hat{\mu}_{kj}$, $\hat{\sigma}_j^2$ and $\hat{\pi}_k$. The first term is the right-hand side is exactly the regular $-2\sum_{i=1}^{n}$loglik when $w_i = 1/n$, $\forall i$. We name our method LumiWCluster (Il*Lumi*na *W*eighted model-based *Cluster*ing).

4 RESULTS

4.1 DNA methylation studies in ovarian cancer

We illustrate our proposed method on the ovarian epithelial carcinoma tumors and cell lines methylation dataset from Houshdaran *et al.* (2009). In Figure 5, we show the presence of sequence bias on the positive control sample in this dataset. The observed pattern in consistent with our six positive control samples (Fig. 3). Using the estimated h_1 and h_2 from our six positive controls, we adjust for the observed bias in the ovarian methylation dataset. Although h_1 and h_2 are estimated from an independent source of data, we show that the effect of sequence bias is reduced significantly for the ovarian methylation dataset (Fig. 6).

This ovarian cancer dataset consists of 27 primary tumors (15 serous, 9 endometrioid and 3 clear cell) and 15 cell lines. By applying our proposed weighted clustering approach (LumiWCluster) to this dataset, the optimal number of clusters chosen is $K=4$. Details are provided in Supplementary Materials. We also compare the clustering results from Gaussian mixture model without penalty and weights, i.e. $J(\mathbf{\Omega})=0$, $w_i = 1/n\ \forall i$, $g_j = 1/p\ \forall j$. The optimal number of clusters chosen is $K=3$. We refer to this model as GMM-nopenalty. This model is a special case of Mclust by Fraley and Raftery (2002). Mclust is a non-penalized standard model-based clustering which allows for different functional forms of the covariance matrices, where all CpG loci are retained in the resulting clustering. We also include the clustering results from Mclust, *k*-means and PAM (partitioning around medoids, a more robust version of k-means) (Kaufman and Rousseeuw, 1990) which identify $K=4$, 2 and 2 as the optimal number of clusters, respectively. The optimal number of clusters chosen by *k*-means and PAM is based on the 'silhouette' criterion (Rousseeuwl, 1987). As shown in Table 1, *k*-means and PAM are unable to separate cell lines from primary tumors; whereas LumiWCluster, GMM-nopenalty and Mclust yield comparable cluster membership, where cell lines are separated from primary tumors, but there are some mixing among the three tumor subtypes. However, an advantage of LumiWCluster is that it automatically shrinks 554 CpG loci to 0, which implies that these loci do not contribute to the clustering. This refines the set of CpG loci which are important in the resulting cluster structure. These 554 CpG loci include the eight sites which have median detection $P > 0.05$ and are excluded using the filtering criterion in Marsit *et al.* (2009). We also demonstrate that the ability of LumiWCluster in selecting important CpGs yields tighter clusters in Supplementary Materials. In addition, LumiWCluster results in smaller BIC compared with GMM-nopenalty and Mclust, indicating a better model fit. We provide additional information on the advantages of incorporating the detection *P*-values in our proposed clustering approach in Supplementary Materials.

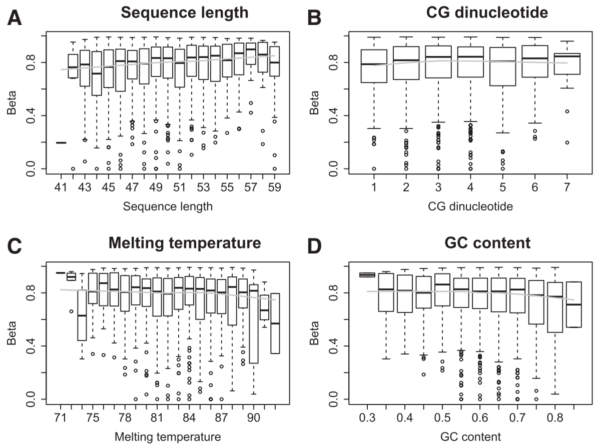

Fig. 6. The effects of (**A**) sequence length, (**B**) number of CG dinucleotides, (**C**) melting temperature and (**D**) GC content bias after correction on positive control of Houshdaran *et al.* (2009). Gray line is the lowess approximation.

Table 1. Clustering result for the ovarian cancer data

Cluster	LumiWCluster	GMM-nopenalty	Mclust
1	15CL	15CL	15CL
2	1CC, 2E, 9S	1CC, 2E, 9S	1CC, 2E, 9S
3	2CC, 5E, 3S	2CC, 7E, 6S	2CC, 5E, 3S
4	2E, 3S	–	2E, 3S
Cluster	*k*-means	PAM	
1	15CL, 1CC, 2E 9S	13CL, 1CC, 2E, 9S	
2	2CC, 7E, 6S	2CL, 2CC, 7E, 6S	

CC, clear cell; CL, cell lines; E, endometrioid; S, serous.

Table 2. Clustering result for the HCC-cirrhosis data

Cluster	LumiWCluster	GMM-nopenalty	Mclust
1	20C	17C	17C
2	20N	11N	11N
3	–	3C, 9N	3C, 9N
Cluster	k-means	PAM	GMM-nopenalty (with subset of CpGs selected by LumiWCluster)
1	17C	17C	20C
2	11N	11N	20N
3	3C, 9N	3C, 9N	–

C, HCC with cirrhosis samples; N, normal liver tissues.

4.2 DNA methylation studies in HCV-cirrhosis

Our next example is on a data set measuring DNA methylation of hepatocellular carcinoma (HCC) (Archer *et al.*, 2010) from Illumina GoldenGate Methylation BeadArray. This data set consists of 20 samples from HCC with cirrhosis and 20 normal liver tissues. Similar to Section 4.1, we compare the clustering results from LumiWCluster, GMM-nopenalty, Mclust, k-means and PAM on the normalized data (corrected for sequence length and GC content biases). LumiWCluster selects $K = 2$, whereas the rest of the methods select $K = 3$ as the optimal number of clusters. LumiWCluster automatically shrinks 639 CpG to zero and results in perfect separation between the C and N samples. However, the other methods contain misclassification of these samples (Table 2). In Supplementary Materials, we also provide the clustering results for the other methods by fixing $K = 2$. Unlike LumiWCluster, these methods still misclassify the two different types of samples.

Next, we run GMM-nopenalty on the subset of CpG loci selected by LumiWCluster [referred to as ‘GMM-nopenalty (with subset of CpGs selected by LumiWCluster)’ in Table 2]. Interestingly, the optimal number of clusters is chosen to be 2 and results in perfect classification. This highlights that LumiWCluster is able to select informative CpG loci that differentiate cirrhosis from normal tissues.

To further illustrate the advantage of LumiWCluster in selecting important CpG loci, we carry out the non-parametric Wilcoxon rank-sum test for comparing cirrhosis and normal group samples on each of these 1505 CpG loci. The P-values are adjusted using the false discovery rate (FDR) control (Benjamini and Hochberg, 1995). At FDR of 0.05, 597 loci are significant, of which 578 overlap with the 866 loci selected by LumiWCluster. This shows that LumiWCluster is able to retain statistically significant loci. As a comparison, 19 CpG loci have median detection $P > 0.05$ and will be excluded using the filtering criterion in Marsit *et al.* (2009). Among these 19 loci, 15 of them were shrunk to 0 by LumiWCluster. Figure 7A and B shows the distribution of the β values for two CpG loci that were not shrunk to 0, which appear to be informative in differentiating HCC with cirrhosis from normal liver tissues. We also include two CpG loci that have median detection $P > 0.05$ and were shrunk to 0 by LumiWCluster (Fig. 7C and D). This again demonstrates the ability of LumiWCluster in selecting informative CpGs which can differentiate the two groups despite being completely unsupervised.

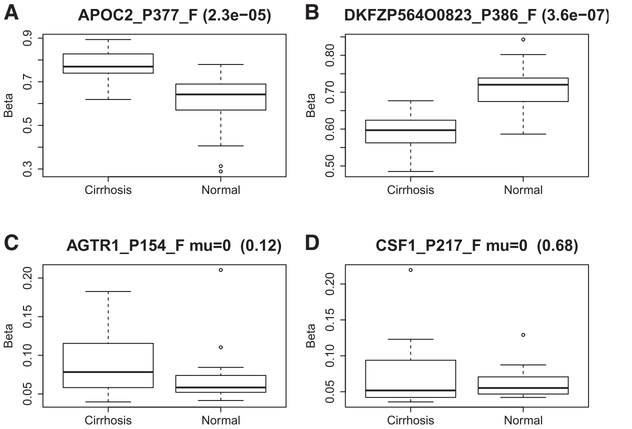

Fig. 7. (**A**, **B**) Distribution of β values for three CpGs which are omitted by the criterion in Marsit *et al.* (2009), but are not shrunk to zero in our weighted model-based clustering. (**C**, **D**) Example of CpGs which are omitted by the criterion in Marsit *et al.* (2009) and are shrunk to zero in our weighted model-based clustering. The numbers printed in the parenthesis are the adjusted P-values.

5 DISCUSSION

The delineation of DNA methylation patterns is important in understanding how these epigenetic changes might lead to aberrant expression patterns and disease (Laird, 2010). Advancements in biotechnology have enabled high-throughput profiling of DNA methylation, including the Illumina GoldenGate and Infinium BeadArray via bisulphite conversion. These platforms are robust, highly reproducible and require less starting materials. In the first part of this study, we illustrated the source of biases present in Illumina Methylation arrays and proposed a model for correcting these biases.

A common approach in analyzing Illumina Methylation data includes omitting CpG loci and samples that exhibit detection P-values larger than an arbitrary cutoff. This hard thresholding step often results in missing observations and information loss by discarding a subset of probes. We proposed a weighted model-based approach called LumiWCluster that weights each CpG locus/sample by its detection P-values for clustering DNA methylation profiles. In this article, we set the weights as the median detection P-values across samples (or CpG loci) which appears to perform well in the two case studies. Optimal selection of weight functions is beyond the scope of this article and will be an interesting future research direction.

ACKNOWLEDGEMENTS

We thank Drs Nancy Thomas, Kathleen Conway and Sharon Edmiston for providing the control samples and the reviewers for valuable comments.

Funding: NCI (grants 5-P30-CA16086-34 and 1R21CA134368-01A1 to P.K. and H.C., in parts); Susan G. Komen Foundation (grant KG081397 to P.K. and H.C., in parts).

Conflict of Interest: none declared.

REFERENCES

Archer,K. *et al.* (2010) High-throughput assessment of CpG site methylation for distinguishing between HCV-cirrhosis and HCV-associated hepatocellular carcinoma. *Mol. Genet. Genomics*, **283**, 341–349.

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

Bibikova,M. *et al.* (2006) High-throughput DNA methylation profiling using universal bead arrays. *Genome Res.*, **16**, 383–393.

Bickel,P. and Levina,E. (2004) Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. *Bernoulli*, **10**, 989–1010.

Christensen,B. *et al.* (2009) Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. *PLoS Genet.*, **5**, e1000602.

Down,T. *et al.* (2008) A Bayesian deconvolution strategy for immunoprecipitation based DNA methylome analysis. *Nat. Biotechnol.*, **26**.

Dunning,M. *et al.* (2008a) Spike-in validation of an Illumina-specific variance-stabilizing transformation. *BMC Res. Notes*, **1**, 18.

Dunning,M. *et al.* (2008b) Statistical issues in the analysis of Illumina data. *BMC Bioinformatics*, **9**, 85.

Esteller,M. (2007) Cancer epigenomics: DNA methylomes and histone-modifications maps. *Nat. Rev. Genet.*, **8**, 286–298.

Fraley,C. and Raftery,A. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.*, **97**, 611–631.

Hernandez-Vargas,H. *et al.* (2010) Hepatocellular carcinoma displays distinct DNA methylation signatures with potential as clinical predictors. *PLoS One*, **5**, e9749.

Houseman,A. *et al.* (2008) Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distribution. *BMC Bioinformatics*, **9**, 365.

Houshdaran,S. *et al.* (2009) DNA methylation profiles of ovarian epithelial carcinoma tumors and cell lines. *PLoS ONE*, **5**, e9359.

Illumina (2006) GoldenGate methylation cancer panel I. Available at http://www.illumina.com/technology/goldengate_methylation_assay.ilmn.

Irizarry,R. *et al.* (2008) Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Res.*, **18**, 780–790.

Irizarry,R. *et al.* (2009) The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nat. Genet.*, **41**, 178–186.

Ji,Y. *et al.* (2005) Applications of beta-mixture models in bioinformatics. *Bioinformatics*, **21**, 2118–2122.

Kaufman,L. and Rousseeuw,P. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis* . Wiley-Interscience.

Koga,Y. *et al.* (2009) Genome -wide screen of promoter methylation identifies novel markers in melanoma. *Genome Res.*, **19**, 1462–1470.

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

Lynch,A. *et al.* (2009) Considerations for processing and analysis of Goldengate-based two-colour illumina platforms. *Stat. Methods Med. Res.*, **18**, 437–452.

Marsit,C. *et al.* (2009) Epigenetic profiling reveals etiologically distinct patterns of DNA methylation in head and neck squamous cell carcinoma. *Carcinogenesis*, **30**, 416—422.

Pan,W. and Shen,X. (2007) Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.*, **80**, 1145–1164.

Raftery,A. (2003) Discussion of "Bayesian clustering with variable selection and transformation selection" by liu et al. *Bayesian Stat.*, **7**, 266–271.

Rakyan,V. *et al.* (2008) An integrated resource for genome-wide identification and analysis of human tissue-specific differential methylated regions (tDMRs). *Genome Res.*, **18**, 1518–1529.

Richards,J. *et al.* (2009) Weighted model-based clustering for remote sensing image analysis. *Comput. Geosci.*, **14**, 125–136.

Rousseeuwl,P. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Seo,J. *et al.* (2004) Interactively optimizing signal-to-noise ratios in expression profiling, project-specific algorithm selection and detection p-value weighting in Affymetrix microarrays. *Bioinfomatics*, **20**, 2534–2544.

Shen,R. *et al.* (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.

Siegmund,K. *et al.* (2004) A comparison of cluster analysis methods using DNA methylation data. *Bioinformatics*, **20**, 1896–1904.

Wang,S. and Zhu,J. (2008) Variable selection for model-based high dimensional clustering and its application to microarray data. *Biometrics*, **64**, 440–448.

Wei,H. *et al.* (2008) A study of the relationships between oligonucleotide properties and hybridization signal intensities from NimbleGen microarray datasets. *Nucleic Acids Res.*, **36**, 2926–2938.

Wong,W. *et al.* (2008) On the necessity of different statistical treatment for Illumina BeadChip and Affymetrix GeneChip data and its significance for biological interpretation. *Biol. Direct*, **3**.

Xie,Y. *et al.* (2009) Statistical methods of background correction for Illumina BeadArray data. *Bioinformatics*, **25**, 751–757.