Bioinformatics, 32(5), 2016, 664–672
doi: 10.1093/bioinformatics/btv659
Advance Access Publication Date: 11 November 2015
Original Paper

OXFORD

Genome analysis

Identification of differentially methylated loci using wavelet-based functional mixed models

Wonyul Lee and Jeffrey S. Morris*

Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, Houston, TX, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on December 4, 2014; revised on November 4, 2015; accepted on November 5, 2015

Abstract

Motivation: DNA methylation is a key epigenetic modification that can modulate gene expression. Over the past decade, a lot of studies have focused on profiling DNA methylation and investigating its alterations in complex diseases such as cancer. While early studies were mostly restricted to CpG islands or promoter regions, recent findings indicate that many of important DNA methylation changes can occur in other regions and DNA methylation needs to be examined on a genome-wide scale. In this article, we apply the wavelet-based functional mixed model methodology to analyze the high-throughput methylation data for identifying differentially methylated loci across the genome. Contrary to many commonly-used methods that model probes independently, this framework accommodates spatial correlations across the genome through basis function modeling as well as correlations between samples through functional random effects, which allows it to be applied to many different settings and potentially leads to more power in detection of differential methylation.

Results: We applied this framework to three different high-dimensional methylation data sets (CpG Shore data, THREE data and NIH Roadmap Epigenomics data), studied previously in other works. A simulation study based on CpG Shore data suggested that in terms of detection of differentially methylated loci, this modeling approach using wavelets outperforms analogous approaches modeling the loci as independent. For the THREE data, the method suggests newly detected regions of differential methylation, which were not reported in the original study.

Availability and implementation: Automated software called WFMM is available at https://biostatistics.mdanderson.org/SoftwareDownload. CpG Shore data is available at http://rafalab.dfci.harvard.edu. NIH Roadmap Epigenomics data is available at http://compbio.mit.edu/roadmap.

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

Contact: jefmorris@mdanderson.org

1 Introduction

DNA methylation is an important epigenetic mechanism that is involved in the regulation of gene expression and plays a central role in normal biology and diseases. There has been much attention in profiling DNA methylation and discovering its alterations over the past decade to improve our understanding of complex diseases such as cancer. Much early methylation research focused on *CpG islands* introduced by Bird *et al.* (1987). CpG islands are genomic regions containing high frequency of CpG sites, which contain cytosine and guanine connected by a phosphodiester bond. They frequently occur in the promoter region of genes, and thus have been thought to be most relevant methylation sites in suppressing RNA expression. However, recent evidence required us to rethink this belief. For example, Irizarry *et al.* (2009) showed that most methylation alterations in colon cancer occur not in CpG islands, but in sequences up to 2 kb distant from CpG islands, which they term *CpG island shores*. This was discovered by examining DNA methylation on a comprehensive genome-wide scale via high-throughput

microarray-based methylation data rather than just focusing on CpG islands. This suggests that traditional approaches focused on a certain regions of the genome such as CpG islands or promoter regions are likely to miss important findings and genome-wide studies on DNA methylation are preferred.

Recent advancements in high-throughput assessment of DNA methylation using microarray-based or sequencing-based approaches make genome-wide methylation studies possible. For example, Irizarry *et al.* (2008) developed comprehensive high-throughput arrays for relative methylation (CHARM) which measure methylation level at more than 2 million probes for each sample and Lister *et al.* (2009) developed whole-genome bisulfite sequencing (WGBS) for measuring methylation level at every CpG site in the genome. Reviews and comparisons on several high-throughput methylation technologies can be found in Irizarry *et al.* (2008), Laird (2010) and Wilhelm-Benartzi *et al.* (2013). However, these data are high-dimensional and complex, and so there is a need to develop new methods to analyze these genome-level methylation data that are flexible enough to capture their complex structure, computationally efficient enough to scale up to their enormous size, and yielding rigorous statistical inference allowing one to identify differentially methylated loci in the genome while adjusting for multiple testing, e.g. by controlling the false discovery rate (FDR) or experiment-wise error rate (EER).

In differential methylation analysis, the most commonly used approaches examine the relationship between the phenotype of interest and methylation of a given individual locus independently across the genome (Barfield *et al.*, 2012; Bibikova *et al.*, 2011; Touleimat and Tost, 2012; Wang *et al.*, 2012; Wettenhall and Smyth, 2004; Zackay and Steinhoff, 2010 and many more). However, methylation in nearby CpG loci tends to be highly correlated (Leek *et al.*, 2010). Therefore, statistical analyses based on independence assumptions across loci are likely to be inefficient.

To deal with this dependence issue, some methods have been developed that use the idea of bump-hunting to account for correlation of nearby loci (Irizarry *et al.*, 2009; Jaffe *et al.*, 2012). In particular, Jaffe *et al.* (2012) first estimate the differential methylation individually at each genomic location. Then the estimates are smoothed using loess (Cleveland, 1979) and candidate regions of differential methylation are generated as contiguous regions where the absolute values of smoothed estimates are greater than a predetermined threshold. Finally, permutation techniques are applied to assess statistical significance of these regions. This approach has great advantages over approaches modeling loci as independent in that they borrow strength across genome by grouping neighboring probes into a region and focusing on region-based detection of differential methylation, and can account for multiple testing. They have successfully applied their approach and identified differentially methylated regions (DMRs). However, these methods also have disadvantages. As pointed out in Jaffe *et al.* (2012), they can have trouble in finding small differentially methylated regions involving a small number of loci. This is problematic since such small regions may correspond to suppressed gene expression and thus be functionally important. Also, care must be taken in specification of the predetermined threshold for defining a region or grouping probes, as these tuning parameters can strongly affect the results. The loess smoothing approach is not spatially adaptive when used with global bandwidths, so can result in the attenuation of local features that can reduce power to detect them. Additionally, the current software is limited to paired or independent group-level analyses and simple linear regression, which may not be sufficient to answer some research questions of interest in methylation studies or handle certain experimental designs inducing correlation between arrays.

Morris and Carroll (2006) introduced a method, called wavelet-based functional mixed models (WFMM), that can model complex, irregular functional data within a functional mixed model framework, which generalizes linear mixed models to the setting of functional responses. This framework allows general design matrices that can accommodate group effects, linear regression and various types of interactions, plus random effect levels to model correlation between functions induced by experimental design, e.g. multiple functions from the same subject or cluster. Multiple factors can be simultaneously modeled, e.g. allowing us to adjust for batch effects. They used wavelet basis functions to represent functions, which allows the method to capture local within-function correlations and thus adaptively borrow strength within the function. They have standalone executable software to run the general model.

In this paper, we will apply this method to the analysis of genome-wide probe-level methylation data and explore its properties. Modeling is done using a fully Bayesian approach that has significant inferential advantages, yielding statistical inference to flag genomic loci as differentially methylated while accounting for multiple testing using experiment-wise error rate and/or false discovery rate criteria. By modeling methylation profiles as functions and considering both spatial correlation given a function and correlation among samples, this approach can borrow strength across neighboring probes as well as across samples. The implementation of the method is relatively straightforward by using automated software called WFMM freely available at https://biostatistics.mdanderson.org/SoftwareDownload. Although this article focuses on genome-wide methylation studies, the framework is more general and can be used for analyzing other high-throughput genomic data sets observed for many locations spaced through the genome, including copy number and tiling transcriptome arrays.

The remainder of the article is organized as follows. Section 2 introduces three methylation data sets we analyze and the preprocessing procedures we employ. Section 3 describes the wavelet-based functional mixed model and Bayesian approaches for identifying differentially methylated loci. We present numerical results in Section 4 and conclude with a discussion of general usefulness of the framework and its potential for many other data sets.

2 Application

In this section, we briefly describe three different methylation data sets we analyze and describe how they were preprocessed.

2.1 CpG shore data

Irizarry *et al.* (2009) examined DNA methylation of 13 colorectal cancers and 13 matched normal mucosa on a genome-wide scale using the CHARM microarrays described in Irizarry *et al.* (2008) to investigate methylation differences between normal and colon tumor samples. This study showed that most DNA methylation alterations in colon cancer occur in sequences up to 2 kb distant from CpG islands or promoter regions, which they term 'CpG island shores', rather than in the promoter or CpG island regions themselves. They standardized log ratios from the CHARM microarrays using suitable control probes from unmethylated regions so that the average log ratio in the control probes was zero. Subsequently, Aryee *et al.* (2011) developed an improved preprocessing procedure, which can be implemented using the R package **charm** (version 2.8.0). In our analysis, the CHARM microarray data were normalized and pre-processed using the **charm** package. The methylation log ratios were first normalized using the control probes. Then the normalized log ratios were

transformed into percentage methylation estimates, which range between zero and one. As a result, percentage methylation values were given for a total of 2 162 406 probes for each sample. For our modeling, we applied the logit-transform to the percentage methylation values, which made them symmetric and approximately normal tailed. We will use the logit-transformed data to compare methylation differences between the normal and tumor samples.

2.2 THREE data

While the application to the CpG Shore data is focused on identifying methylation differences in a paired group comparison, in other settings one is interested in detecting methylation differences associated with certain continuous variables in a regression context. For example, Lee *et al.* (2012) studied differential methylation associated with gestational age at birth. They identified three differentially methylated regions associated with gestational age using a bump hunting method introduced in Jaffe *et al.* (2012). We obtained a dataset from this study (Roadmap Study) by Lee *et al.* (2012) which included comprehensive methylation profiles of 1 569 888 probes on 141 newborns from the Baltimore Tracking Health Related to Environmental Exposures (THREE) study and gestational age at birth (the 'best obstetrical estimate'; Apelberg *et al.*, 2007). We obtained permission from the Roadmap Study and THREE Study investigators for this publication. In this study, CHARM hybridization and processing were performed across 5 different days, which raises the potential for batch effects. Our modeling approach will enable us to adjust for these batch effects by modeling these processing dates as covariates.

2.3 NIH roadmap epigenomics data

The NIH Roadmap Epigenomics Consortium generated large collection of human epigenomics for primary cells and tissues (Kundaje *et al.*, 2015). By performing the integrative analysis of histone modification, DNA accessibility, DNA methylation and RNA expression of all the reference epigenomes generated by the consortium, they provided comprehensive map of the human epigenomic landscape and demonstrated the central role of epigenomic information for understanding gene regulation and human disease. In our analysis, we focused on DNA methylation profiled using whole-genome bisulfite sequencing (WGBS) and applied the WFMM method to identify DMRs between heart and digestive tissues.

3 Methods

3.1 Wavelet-based functional mixed models

In this section, we briefly overview the wavelet-based functional mixed model (WFMM) approach introduced by Morris and Carroll (2006) and describe how it can be applied to comprehensive genome-wide methylation analysis.

Suppose we observe N methylation profiles $Y_i(t), i = 1, \ldots, N$, on a common grid of chromosomal locations $\mathbf{t} = (t_l; l = 1, \ldots, T)$. For instance, $Y_i(t)$ is the logit-transformed percentage methylation value of the subject i on the location t, and $\mathbf{t}$ is a vector of chromosomal locations where methylation levels are measured. A functional mixed model at these locations in matrix form is given by

$$\mathbf{Y} = X\mathbf{B} + \sum_{h=1}^{H} Z_h \mathbf{U}_h + \mathbf{E}, \quad (1)$$

where $\mathbf{Y}$ is an $N \times T$ matrix containing the logit-transformed methylation values, $\mathbf{B}$ is a $p \times T$ matrix of functional fixed effects with corresponding $N \times p$ design matrix X, $\mathbf{U}_h$ is an $M_h \times T$ matrix of functional random effects at level h with corresponding $N \times M_h$ design matrix Z_h, and $\mathbf{E}$ is an $N \times T$ matrix of residual errors. For example, in the CpG Shore data, X was set to be a 26×2 matrix with the columns indicating normal or tumor group. The corresponding $\mathbf{B}$ was a $2 \times T$ matrix with each row giving the mean methylation profile for each group. One may want to use other parameterizations of the design matrix. However, in general, linear models are invariant to parameterizations and that is also true in our setting if the priors are vague without much shrinkage. We also introduced a random effect function for each subject to take the within-subject correlation between paired normal and tumor samples into account. This leads to 13 random effect functions, and thus Z_1 was a 26×13 matrix with the columns indicating the subjects. In the THREE data, X was a 141×6 matrix with the first column being the gestational age and the other columns indicating different processing dates. In this case, B_{1l} indicates the functional linear coefficient for the gestational age at the location l adjusted for the block effects, and the remaining 5 columns indicate separate functional intercepts for each block.

This model has great flexibility in many aspects. The fixed effects can be mean functions, functional main effects, functional linear coefficients for continuous covariates, and any interactions among these factors. This model also allows multiple levels ($h = 1, \ldots, H$) of random effect functions, which enables this model to capture various multi-level correlation structures for different strata based on the experimental design. We assume that $\mathbf{U}_h$ follows a matrix normal distribution, $\mathbf{U}_h \sim \mathcal{MN}(P_h, Q_h)$, where P_h and Q_h are the $M_h \times M_h$ between-function and $T \times T$ within-function covariance matrices, respectively (Morris and Carroll, 2006). We also assume that $\mathbf{E} \sim \mathcal{MN}(R, S)$ with $N \times N$ between-function covariance R and $T \times T$ within-function covariance S.

It would be convenient to fit each column of the model (1) separately, which would be equivalent to fitting separate linear mixed models to each probe. However, this approach, effectively assuming Q_h and S are diagonal matrices, does not capture correlation or borrow strength across the genome, and likely would be inefficient. This would be an example of an independent probe-by-probe analysis that is prevalent in the literature. However, nearby probes tend to generally be in correlated blocks (Leek *et al.*, 2010). Although correlations within blocks need to be taken into account, the problem is that the blocks are generally unknown in advance.

To fit the model (1) while taking correlations across t into account, Morris and Carroll (2006) used a *basis function transform approach*, which transforms the observed functions from the data space into a basis space, fits the basis space version of the model, and then transform results back to the original data space for inference. Among other possibilities, wavelets were chosen because of their fast calculation, local support and whitening property, which makes them ideal for modeling very large functional data sets for which the functions are characterized by local features like spikes and change points, and in cases like this where we expect correlation within blocks of probes but do not know a priori what those blocks are. They are extensively used in signal processing applications because of their sparsity and denoising properties. Wavelets have also been applied in many genome-wide studies to detect histone modification enrichments (Mitra and Song, 2012), identify nucleosome position (Nguyen *et al.*, 2013) and identify genetic variants associated with chromatin accessibility (Shim and Stephens, 2015). When the data are equally spaced, wavelet coefficients can be computed in linear time O(T) using the discrete wavelet transform. When unequally spaced, the lifting scheme of Sweldens (1996) can be used,

or wavelets for equally spaced data could be used without accounting for the unequal spacing. In the latter approach, the domain of the wavelet basis functions are effectively $l = 1, \ldots, T$ instead of $t = t_1, \ldots, t_T$. Others have found that for genome-level copy number data, this yields similar results as the lifting scheme so was preferred for computational reasons (Hsu *et al.*, 2005; Sardy *et al.*, 1999). Thus, we will use that approach here.

Applying the DWT to each row of $\mathbf{Y}$, we obtain a $N \times T^*$ matrix of wavelet basis coefficients $\mathbf{Y}^*$, which is considered as the raw data in the wavelet domain. This transformation can be represented as matrix multiplication by a $T \times T^*$ wavelet transform matrix Φ', $\mathbf{Y}^* = \mathbf{Y}\Phi'$. The corresponding wavelet space model can be derived by right matrix multiplication of both sides of the model (1) by Φ':

$$\mathbf{Y}^* = X\mathbf{B}^* + \sum_{h=1}^{H} Z_h \mathbf{U}_h^* + \mathbf{E}^*, \tag{2}$$

where $\mathbf{B}^* = \mathbf{B}\Phi'$, $\mathbf{U}_h^* = \mathbf{U}_h\Phi'$ and $\mathbf{E}^* = \mathbf{E}\Phi'$ are the wavelet space analogs to $\mathbf{B}$, $\mathbf{U}_h$ and $\mathbf{E}$ with columns indexing basis coefficients instead of function locations. Based on the linearity of the transform, we can show that $\mathbf{U}_h^* \sim \mathcal{MN}(P_h, Q_h^*)$ and $\mathbf{E} \sim \mathcal{MN}(R, S^*)$, where $Q_h^* = \Phi Q_h \Phi'$ and $S^* = \Phi S \Phi'$. We assume that Q_h^* and S^* are diagonal with heterogeneous diagonal elements. This parsimonious specification of Q_h^* and S^* can accommodate flexible assumptions on Q_h and S in the data space since $Q_h = \Phi' Q_h^* \Phi$ and $S = \Phi' S^* \Phi$, which are generally not diagonal. Using wavelets, this structure indeed captures within-function (between-probe) correlation, and allows this correlation to vary across different locations of functions as illustrated in Morris and Carroll (2006), which enables us to adaptively borrow strength from nearby genomic locations in estimation and inference. This parsimonious specification also makes the fitting of the model (2) separable in each column of $\mathbf{Y}^*$ as follows:

$$\mathbf{Y}_k^* = X\mathbf{B}_k^* + \sum_{h=1}^{H} Z_h \mathbf{U}_{h,k}^* + \mathbf{E}_k^*, \quad k = 1, \ldots, T^*, \tag{3}$$

where $\mathbf{Y}_k^*$, $\mathbf{B}_k^*$, $\mathbf{U}_{h,k}^*$ and $\mathbf{E}_k^*$ are the kth columns of $\mathbf{Y}^*$, $\mathbf{B}^*$, $\mathbf{U}_h^*$ and $\mathbf{E}^*$ respectively. The parsimonious structure requires only T^* parameters to estimate for each covariance matrix instead of $T(T+1)/2$ parameters and makes the fitting parallel, which helps in making the method scalable to large whole-genome data sets.

The model (3) is fit using a fully Bayesian approach with a Markov chain Monte Carlo simulation. A vague proper prior was imposed on the B_{ik}^*, the k-th wavelet coefficient for the ith fixed effect: $B_{ik}^* \sim N(0, \tau_{ik})$ with a big value of τ_{ik}. Here we focused on benefit of accounting for intra-genome correlation via basis functions. Thus, vague proper priors were used so that the difference between several approaches is only the effective intra-genomic covariance assumptions used. These vague priors also make sense if a pre-thresholding step is performed on the basis coefficients modeled, e.g. by joint compression as described in Morris *et al.* (2011). In other cases, one may wish to encourage sparsity in the fixed effect estimation through sparsity priors (See Section 8 in the Supplementary Materials). The fitting of (3) yields posterior samples in the wavelet domain. These posterior samples are then transformed back to the data space using the fast inverse discrete wavelet transform (IDWT) for subsequent Bayesian inference. The relevant inference procedures are described in the next section.

We have automated software which implements the WFMM described above and can handle sufficiently large data sets, freely available at https://biostatistics.mdanderson.org/SoftwareDownload. Users only need to provide $\mathbf{Y}$ and specify X and Z_h if necessary. Default choices for other wavelet and MCMC specification can be all automatically computed by the software. Details for the input and output can be found in an accompanying documentation.

The WFMM has been further adapted and extended to handle partially missing data (Morris *et al.*, 2006), model image data (Morris *et al.*, 2011), perform robust regression insensitive to outliers (Zhu *et al.*, 2011), model nonstationary time series (Martinez *et al.*, 2013), regress one function on another (Meyer *et al.*, 2015) and allow one to classify individuals based on their functional data (Zhu *et al.*, 2012).

3.2 Bayesian inference for methylation data

Here, we describe how to perform Bayesian inference using the posterior samples from the WFMM model. First, we introduce a joint credible band approach to detect significantly differentially methylated loci while controlling the experiment-wise error rate. Second, we define a global genome-wide test using the joint credible bands. Finally, we describe another approach introduced by Morris *et al.* (2008) which looks for differences of some minimum effect size δ and controls the Bayesian false discovery rate (FDR).

3.2.1 Experiment-wise error rate

Identifying significant sites over the whole genome involves a multiple-testing problem where the experimental error rate control is often desirable. A simultaneous credible band idea illustrated in Ruppert *et al.* (2003) can be used to detect significant locations while controlling the experiment-wise error rate. Let $\mathbf{B}_j = (B_{j1}, \ldots, B_{jT})'$ be the j-th fixed effect function on the grid $\mathbf{t} = (t_1, \ldots, t_T)$ which indicates differential methylation. Suppose we have G posterior samples of $\mathbf{B}_j$, denoted by $\{\mathbf{B}_j^{(g)}, g = 1, \ldots, G\}$, where $\mathbf{B}_j^{(g)} = (B_{j1}^{(g)}, \ldots, B_{jT}^{(g)})'$. Let $m(B_{jl})$ and $\hat{\text{sd}}(B_{jl})$ be the posterior mean and standard deviation of B_{jl} computed from its G posterior samples. Assuming approximate posterior normality, we can construct a $(1-\alpha)100\%$ simultaneous credible band of which credible set at location t_l is given by

$$[m(B_{jl}) - c_\alpha \text{sd}(B_{jl}), m(B_{jl}) + c_\alpha \text{sd}(B_{jl})],$$

where c_α is the $(1-\alpha)$ sample quantile of

$$\max_{l=1,\ldots,T} \left| \frac{B_{jl}^{(g)} - m(B_{jl})}{\text{sd}(B_{jl})} \right|, \quad g = 1, \ldots, G,$$

and $\text{sd}(B_{jl}) = \hat{\text{sd}}(B_{jl})/A(G, \rho_{\text{mcmc}})$. Here ρ_{mcmc} is an estimate of the lag autocorrelation in the samples, and A is the bias correction factor described in Anderson (1971), p441-451. If the lag autocorrelations are not too high (e.g. <0.95), then use of $\hat{\text{sd}}(B_{jl})$ for $\text{sd}(B_{jl})$ should give similar answers.

Based on the simultaneous credible bands computed on a range of different values of α, we define simultaneous band scores (SimBaS) at each location t_l as the minimum level α where the $(1-\alpha)\%$ simultaneous credible band excludes zero at t_l. We flag any genomic location t_l as significant at level α if the corresponding $\text{SimBaS}(t_l)$ is less than or equal to α. This adjusts for multiple testing by strictly controlling the experiment-wise error rate. Note that we expect simultaneous credible bands to be tighter when correlation across t is modeled than if probes were modeled as independent, and thus accounting for within-function correlation through basis functions is expected to yield greater power for detecting differences.

3.2.2 Global test

One can construct a global test to see if there exists any methylation difference across the whole genome. We define the global

genome-wide test score as the minimum value of SimBaS across the genome. If this is less than α, then we can strongly conclude there is a significant difference somewhere in the genome while controlling experiment-wise error rate.

3.2.3 Bayesian FDR

In discovery studies, experiment-wise error rate protection (like Bonferroni) is frequently considered too conservative, resulting in too many missed discoveries, and thus false discovery rate control is used as an alternative criterion commonly used for multiple testing adjustment. Morris *et al.* (2008) introduced an approach for controlling the Bayesian FDR in the functional regression setting. Suppose we are interested in identifying differentially methylated locations that have at least δ difference between normal and tumor groups. Denote the mean difference between the two groups at location t_l by D_l. Given posterior samples of D_l, we can compute the point-wise posterior probability of at least δ difference at each genomic site as $\hat{p}_j(t_l) = \Pr\{|D_l| > \delta|\mathbf{Y}\} = \sum_{g=1}^{G} I\{|D_l^{(g)}| > \delta|\}/G$, where I is the indicator function. As pointed out in Morris *et al.* (2008), the quantity $1 - \hat{p}_j(t_l)$ can be interpreted as the estimate of the local FDR for location t_l. We flag a site t_l as significant if $\hat{p}_j(t_l) > \phi_\alpha$, where ϕ_α is a threshold and α is a global FDR-bound. The threshold ϕ_α is chosen so that on average we expect less than α% of flagged sites are false positives. This can be done in the following way: We sort $\hat{p}_j(t_l)$ in descending order to obtain $\{\hat{p}_{(l)}, l = 1, \ldots, T\}$, find $\nu = \max[l^* : (l^*)^{-1}\sum_{l=1}^{l^*}(1 - \hat{p}_{(l)}) \leq \alpha]$, and set $\phi_\alpha = \hat{p}_{(\nu)}$. We refer this approach as to Bayesian FDR method hereafter.

The quantity δ reflects practical significance in a sense that it controls how large difference should be interpreted as practically meaningful. For example, in the data sets we analyze in this article, methylation levels were measured as % methylation which ranges from 0 to 1. Suppose that we have posterior samples of the mean difference between normal and tumor groups in the original % scale. If one sets $\delta = 0.1$, for instance, then only the locations that have at least 10% methylation differences are considered as practically meaningful in terms of differential methylation.

Our approach flags differential methylation at the probe level, while borrowing strength from nearby probes by modeling correlations across the genome. One may want to require several consecutive probes flagged to call a differentially methylation region (DMR). To make our results more comparable to DMR in Irizarry *et al.* (2009) and Lee *et al.* (2012), we define a DMR as a set of at least m consecutive significant sites flagged by our methods.

4 Results

4.1 Simulation results

In this section, we performed a simulation study, rigorously simulating data based on the CpG Shore data to see how reliably the WFMM method finds differential methylation. Recall that the goal in the CpG Shore data was to discover differentially methylated genomic locations between the normal and tumor groups. For the simulation, we generated 26 virtual methylation profiles of 75 069 probes from chromosome 3. Among 75 069 probes, the true mean methylation levels for 73 168 of the probes were made identical for the normal and tumor samples. For the remaining 1901 probes, which were flagged as differentially methylated by Irizarry *et al.* (2009), the mean methylation levels were made to differ between tumor and normal according to the estimated sample means in the CpG shore data. Correlation across probes and paired samples was induced based on the empirical correlation matrix. More details can be bound in the supplementary materials.

The functional mixed model was fitted by either using wavelet basis functions to capture the between-probe correlation (Wave-mixed) or modeling independently across probes (Indep-mixed). We found that the Wave-mixed model outperformed the Indep-mixed model in terms of discovery of differentially methylated regions while avoiding false discoveries (see supplementary materials for ROC curves and the area under the ROC curve.)

Table 1 summarizes a comparison of these models with the choice of global FDR $\alpha = 0.01$. The column No indicates the number of probes flagged as significant and the column FPR shows false positive rates among them. The column $\text{FNR}_{0.1}$ is the false negative rates among the true positive probes which have the mean difference between the two groups stronger than or equal to 0.1. For all values of δ and the SimBaS, the Wave-mixed models found more probes and shows less $\text{FNR}_{0.1}$ than the independent models at two different noise levels ($s = 0.1$ and $s = 0.2$), which clearly supports our premise that modeling correlations through the wavelet transform can improve power in detection of differential methylation.

We also compared the methods with a variant of the bump hunting method which can be implemented using the **charm** package. The bump hunting method showed higher $\text{FNR}_{0.1}$ and FPR than the other methods when $\delta = 0.05$ or 0.1. Although we compare the WFMM method and the bump hunting method here, there are caveats of differences between the two methods that do not make them immediately comparable. First of all, the bump hunting method looks for larger regions of differential methylation while the WFMM methods focus on probe level detection. They borrow strength nearby probes in different manner. The bump hunting method smoothes the estimated effect within prespecified probe groups while the WFMM method captures correlation structure through the wavelet transform. Our FDR method aims to identify probes showing at least δ change with high probability. Therefore, any probe showing methylation difference close to δ or smaller would not be detected in the FDR method. On the other hand, the bump hunting method does not consider the effect size.

We have also computed the false positive rate for the null case where the two groups have the same mean for all probes to check reliability of the methods. It turned out that it was zero for all cases.

4.2 Analysis of CpG shore data

In this section, we present results obtained by applying the WFMM method to the CpG Shore data. Recall that we modeled dependency between paired samples through random effect functions for each subject. This functional mixed model was fitted by either Wave-mixed or Indep-mixed as in simulation. For paired data like we have here, an alternative to random effects would be to compute pairwise tumor-normal differences for each probe and model these 13 paired

Table 1. Simulation result

		$s = 0.1$			$s = 0.2$		
δ	Model	No	FPR	$\text{FNR}_{0.1}$	No	FPR	$\text{FNR}_{0.1}$
0.05	Indep-mixed	916	0.000	0.321	857	0.000	0.363
	Wave-mixed	957	0.000	0.286	904	0.000	0.326
0.1	Indep-mixed	387	0.000	0.711	351	0.000	0.738
	Wave-mixed	413	0.000	0.692	373	0.000	0.722
SimBaS	Indep-mixed	46	0.000	0.966	27	0.000	0.980
< 0.1	Wave-mixed	112	0.000	0.916	68	0.000	0.949
	Bump hunting	239	0.121	0.884	386	0.142	0.816

difference functions as the responses. For comparison, the pairwise differences on logit-scale were also fitted by either using wavelet basis functions (Wave-paired) or modeling independently across probes (Indep-paired). Both the Indep-mixed and Indep-paired models were implemented using the WFMM software by specifying wavelet in the wavelet structure as 'none' with vague proper priors. In the Wave-mixed and Wave-paired models, we chose the Daubechies wavelet with vanishing fourth moments for the wavelet transformation. In each case, we modeled each chromosome separately but pooled final results together for genome-wide inference and multiple testing adjustment.

We first performed the global test for each model using the minimum SimBaS across the genome. The minimum SimBaS was less than 0.001 for all models and thus, we proceeded to identify differentially methylated probes. For the SimBaS criterion, we flagged a probe as significant if the corresponding SimBaS was less than or equal to 0.1. For the Bayesian FDR method, we used 0.05, 0.1 and 0.2 for the minimum practical effect size δ, which targeted the probes showing at least 5% and 10%, and 20% difference in methylation, respectively. The global FDR bound α was set to be 0.01. For comparison purpose, we defined DMRs as the sets of at least 3 consecutive significant probes. These DMRs were then compared with DMRs found from a variant of the bump hunting method which can be implemented using the **charm** package. In particular, dmrFinder function was used with paired = TRUE and cutoff = 0.995 specifications and the minimum length of a DMR was set to be also 3 in this approach. For the bump hunting method, DMRs with q-values less than 0.01 were flagged as significant. Results are summarized in Table 2. The columns no.probes and no.DMRs indicate the number of flagged probes and the number of identified DMRs. The column New is the number of DMRs which were not identified in the bump hunting method and the column Missed is the number of DMRs which were not flagged in each method, but were identified in the bump hunting method and have at least δ difference between the two groups on average. The locations of flagged DMRs are also plotted in the supplementary materials.

4.2.1 Wavelets versus independent probes

The Wave-mixed model flagged more significant probes and DMRs than the Indep-mixed model, as expected since the wavelet-based method borrows strength from nearby probes when flagging a particular locus as differentially methylated. Irizarry *et al.* (2009) validated nine DMRs using bisulfite pyrosequencing, which showed evidence of differential methylation in these DMRs. With $\delta = 0.05$, the Indep-paired model missed one of them while the Wave-mixed model identified all of them. For $\delta = 0.1$, the Indep-paired model failed to identify four of them. On the other hand, the Wave-mixed model missed only three DMRs. The wavelet-based models also flagged many more probes and DMRs than independent models when using the strict experiment-wise error rate control underlying the SimBaS criterion. As expected in Section 3.2, simultaneous credible bands were much tighter in the wavelet-based models compared to the independent models, confirming that the modeling of between-probe correlation leads to significantly tighter simultaneous bands and greater power to detect differences. For example, the average width of the 95% simultaneous credible band for the fixed effect function in the Wave-mixed model was 3.26 while it was 4.05 in the Indep-mixed model. Similarly, the average band width in the Wave-paired model was 4.04 while it was 4.97 in the Indep-paired model.

Table 2. CpG Shore data

δ	Model	no.probes	no.DMRs	New	Missed
0.05	Indep-mixed	52 912	3456	1610	186 (16%)
	Wave-mixed	58 971	3987	1970	105 (9%)
0.1	Indep-mixed	14 691	957	261	613 (56%)
	Wave-mixed	18 494	1304	362	468 (43%)
0.2	Indep-mixed	1091	66	4	28 (47%)
	Wave-mixed	1574	95	3	19 (32%)
SimBaS	Indep-mixed	176	5	0	1157 (100%)
< 0.1	Wave-mixed	1603	123	5	1068 (92%)
	Indep-paired	58	2	0	1160 (100%)
	Wave-paired	359	37	0	1132 (97%)

The columns no.probes and no.DMRs indicate the number of flagged probes and the number of identified DMRs. The column New is the number of DMRs which were not identified in the bump hunting method and the column Missed is the number of DMRs which were not flagged in each method, but were identified in the bump hunting method and have at least δ effect size on average.

4.2.2 Paired versus mixed modeling

To compare paired and mixed modelings, we focused on the SimBaS criterion since the Bayesian FDR depends on choice of δ, which has distinct interpretation for the paired and mixed models. However, the Wave-mixed model consistently flagged more probes and DMRs than the Wave-paired model when they were compared for several values of δ on the same logit scale (data not shown). Using the SimBaS criterion, the mixed models found more probes and DMRs than the paired models as they obtained tighter simultaneous credible bands by modeling random effects. This suggests that, at least for these data, the mixed model was more powerful in detecting differential methylation than modeling pairwise differences.

4.2.3 WFMM versus bump hunting

The bump hunting method (Jaffe *et al.*, 2012) found a total of 1162 DMRs and these DMRs were compared with DMRs from each model. With $\delta = 0.05$, the Wave-mixed model found many DMRs which were not flagged by the bump hunting method. On the other hand, with $\delta = 0.1$, many of DMRs from the bump hunting method were not identified in the WFMM method. As discussed in Section 4.1, there are caveats inherent in any comparison with the bump hunting method.

4.3 Analysis of THREE data

The CpG Shore data showed enormous numbers of differentially methylated regions between the tumor and normal groups, as the signals in the data were relatively large and the primary focus was on paired group comparison. In this section, we applied the WFMM method to the THREE data which has much more subtle signals. We focus on detecting differentially methylated probes associated with the gestational age in the linear regression context.

Recall that the design matrix X is a 141×6 matrix. The first column of X is the gestational age at birth of which association with the methylation is the main interest. The remaining 5 columns indicate different processing dates reflecting potential batch effects. We applied the DWT to each methylation profile using the Daubechies wavelet with vanishing fourth moments. This WFMM model was compared with an independent model implemented by the WFMM software with the 'none' wavelet specification. The independent model is essentially equivalent to performing linear regression at each probe independently, ignoring the spatial correlation among probes. For each model, we identified differentially methylated

probes using either SimBaS or the Bayesian FDR method. The global FDR bound α was set to be 0.05. We used $\delta = 0.004$, which for a difference in 25 days of gestational age corresponds to difference of 0.1 on logit-scale. The Supplementary Materials provide methylation difference on the raw 0–1 scale corresponding to 0.1 difference on logit-scale for various baseline values. Roughly speaking, the FDR method with $\delta = 0.004$ was designed to detect probes showing at least 2% methylation changes in 25 days of gestational age. Based on the flagged probes, we defined DMRs as the sets of at least 5 consecutive significant probes. These DMRs were then compared with DMRs reported in Lee *et al.* (2012).

Results are summarized in Table 3. The WFMM method identified a larger number of significant probes than the independent model for each criterion, again demonstrating more power to detect differentially methylated loci. Although we failed to call the DMR near the gene *RAPGEF2* on chromosome 4 reported in Lee *et al.* (2012) because of our 5 consecutive probe criterion, two probes within the DMR were flagged as significant by SimBaS as shown in Figure 1c. The SimBaS criterion also suggests two new DMRs from the WFMM model that were not reported in Lee *et al.* (2012). All DMRs identified by the SimBaS from the WFMM are listed in sup plementary materials and plotted in Figure 1. The FDR method found 10 new DMRs including the two new DMRs flagged by the SimBaS criterion. The other 8 new DMRs which were not flagged by the SimBaS are plotted in the supplementary materials.

In Figure 1, the DMRs from Lee *et al.* (2012) and some of newly identified DMRs from the WFMM are plotted. For each panel, individual raw scale methylation values are plotted in the top box. Samples were split into 6 equal sized bins by the gestational age and methylation levels are differently colored according to their bins. The colored solid lines are the average methylation level within each bin. The legend in Figure 1c shows the average gestational age in days within each bin. Gray lines in the middle boxes represent the estimated age effect function. The point-wise and simultaneous 95% credible bands are also plotted in dotted and solid black lines respectively. Red points indicate significant probes flagged by SimBaS. The bottom boxes show SimBaS. The panels (a–c) correspond to the three DMRs from Lee *et al.* (2012) and vertical dashed black lines represent the boundaries of them. Vertical solid black lines represent boundaries identified by the WFMM method.

Note that the DMRs from the WFMM tend to be shorter than the DMRs from Lee *et al.* (2012) using the bump hunting method (Jaffe *et al.*, 2012). One potential advantage of the bump hunting method is that it smoothes the effects. However, at the same time, the identified DMRs tend to include lower magnitude differences near their boundaries that could be false discoveries. In Figure 1a, it seems that some locations near the DMR boundaries from Jaffe *et al.* (2012) tend to have weak evidence of differential methylation. On the other hand, the WFMM method seems to better localize the DMRs, focusing on the regions with strong differential methylation. In Figure 1c, we can see that changes of methylation levels across gestational age bins for this reported DMR from Jaffe *et al.* (2012) tend to be weak. This region was not claimed as a DMR by the WFMM method. The DMRs flagged by WFMM but not in Jaffe *et al.* (2012) (Fig. 1d–f) appear to show stronger changes of methylation levels across the bins compared to those in Figure 1c. Since these were discovered using the very rigorous experiment-wise error rate criterion of SimBaS, these regions are likely to be true positives, although further validations are needed to assess whether this is indeed true. It would be interesting to further investigate that these DMRs can be useful indicator of developmental changes in newborns.

Table 3. THREE data

	Model	no.probes	no.DMRs	New	Missed
$\delta = 0.004$	Indep	4203	3	1	1
	WFMM	9415	12	10	1
SimBaS< 0.1	Indep	339	0	0	3
	WFMM	713	4	2	1

4.4 Analysis of NIH roadmap epigenomics data

In the CpG Shore and THREE data sets, DNA methylation was profiled using CHARM microarrays. Although application of the WFMM method to these two data sets revealed usefulness of the method for genome-wide microarray data, the framework is more general and can be used for analyzing more comprehensive high-throughput genomic data sets. In this section, we applied the WFMM method to the NIH Roadmap Epigenomics data where DNA methylation was profiled using whole-genome bisulfite sequencing (WGBS).

In this article, we focused on finding DMRs between heart and digestive tissues. In the NIH Roadmap Epigenomics data, there are four heart tissues whose DNA methylation was profiled using the WGBS: right atrium, left ventricle, right ventricle and aorta. The data set includes WGBS methylation profile of 6 digestive tissues: fetal intestine small, fetal intestine large, small intestine,

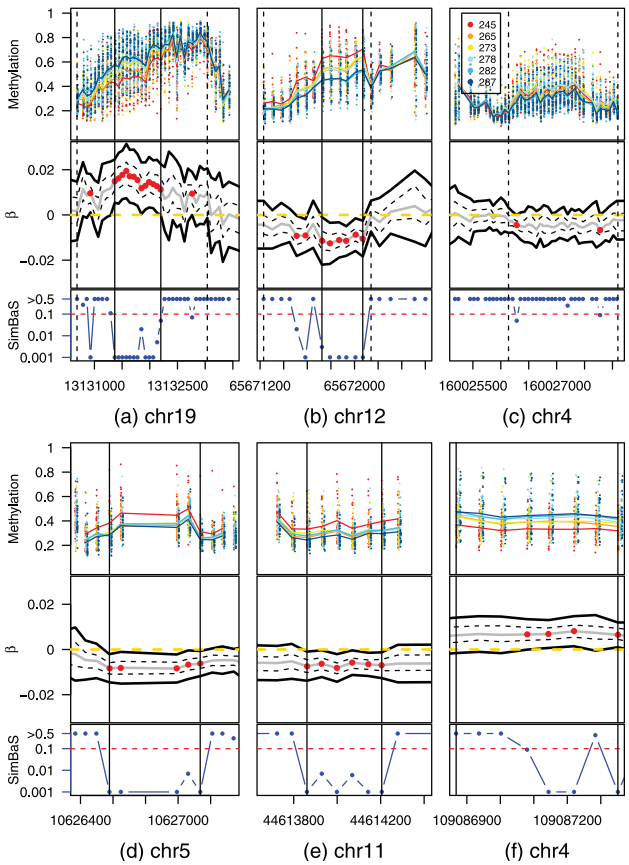

Fig. 1. Methylation plots (top), plots of estimated gestational age effects (middle) and SimBaS plots (bottom) for the DMRs from Lee *et al.* (2012) (a–c) and new DMRs by SimBaS (d, e) and a new DMR by the FDR method (f) from WFMM method

sigmoid colon, oesophagus and gastric. Here, we focused on chromosome 19 with 2 114 752 loci measured to demonstrate applicability of the WFMM method to the WGBS data. The design matrix X was a 10×2 matrix with the columns indicating heart or digestive tissue. We applied the DWT to each methylation profile using the Daubechies wavelet with vanishing fourth moments. For each model, we identified differentially methylated probes using either SimBaS or the Bayesian FDR method. The global FDR bound α was set to be 0.05. No DMR was found using SimBaS, which use experiment-wise error rate control. However, the FDR method found a total of 215, 39 and 9 DMRs having at least length 5 with $\delta = 0.05$, 0.1 and 0.2 respectively. The 9 DMRs which were flagged with $\delta = 0.2$ are listed and plotted in the supplementary materials.

5 Discussion

We have demonstrated how to use the wavelet-based functional mixed model to analyze high-throughput methylation data. Application of this method to two methylation data sets studied in Irizarry *et al.* (2009) and Lee *et al.* (2012) indicated that the wavelet-based functional mixed models outperformed independent models (i.e. modeling probes independently) in terms of detection of differentially methylated loci. Their resulting simultaneous credible bands were much tighter than those from independent models as correlations across the genome have been taken into account in the wavelet-based mixed models. Our simulation studies indicated that even when imperfectly accounting for between-probe correlation using wavelet-based modeling, we gain increased power to detect differentially methylated loci.

Although the method is complex, it is relatively straightforward to implement using the automated software called WFMM freely available at https://biostatistics.mdanderson.org/SoftwareDownload. This software can be generally applied to any complex functional data sampled on a fine grid, not just methylation data, and so can be readily applied to other genome-wide data including copy number and tiling transcriptome arrays. The method is computationally intensive, but the software is optimized so that it can handle very large data sets. For example, in THREE data, the chain of 1000 MCMC iterations for chromosome 1 of 134 386 probes with 141 samples took about 3 hours to run on a single processor and it took about 1 hour for chromosome 23 including 59114 616 probes. The method can be fitted using parallel processing and thus can be run on different cores or clusters. Given a cluster with at least 25 cores, the entire analysis can run in a matter of hours, with further speed-ups possible with more cores. Given the time and expense required to generate one of these data sets, this time frame is reasonable, especially considering the automated nature of the code makes the method easy to run. The WFMM appears to be a promising approach to perform genome-wide analyses of methylation data.

Acknowledgements

For Roadmap Study CHARM 2.0 data, the authors thank the investigators Hwajin Lee, PhD, Andrew E Jaffe, PhD, Andrew P. Feinberg, MD, M. Daniele Fallin, PhD, Jason I. Feinberg, and Shannon Brown, PhD as well as Eirikur Briem and Unner Unnsteinsdottir. For THREE study data generation, the authors thank the participants in the THREE study, Hopkins Labor and Delivery staff and THREE study investigators Benjamin Apelberg, PhD, Ellen Wells, PhD, Lynn Goldman, M.D., Rolf Halden, PhD and Frank Witter, M.D. The authors also thank Rafael A Irizarry, PhD and Andrew E Jaffe, PhD for providing helpful comments and suggestions.

Funding

National Institutes of Health (R01ES017646 to Fallin and Feinberg, R01CA107304, R01CA160736 and R01CA178744 to Morris); Johns Hopkins Bloomberg School of Public Health, The THREE study (Goldman); the Maryland Cigarette Restitution Program Research Grant (Halden); National Institute of Environmental Health Sciences (1R01ES015445 to Halden); Heinz Family Foundation (Goldman); National Science Foundation (DBI 1550088 to Morris).

Conflict of Interest: none declared.

References

Anderson,T.W. (1971) *The Statistical Analysis of Time Series*. Wiley, New York.

Apelberg,B.J. *et al.* (2007) Determinants of fetal exposure to polyfluoroalkyl compounds in Baltimore, Maryland. *Environ. Sci. Technol.*, **41**, 3891–3897.

Aryee,M.J. *et al.* (2011) Accurate genome-scale percentage DNA methylation estimates from microarray data. *Biostatistics*, **12**, 197–210.

Barfield,R.T. *et al.* (2012) CpGassoc: an R function for analysis of DNA methylation microarray data. *Bioinformatics*, **28**, 1280–1281.

Bibikova,M. *et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**, 288–295.

Bird,A. *et al.* (1987) Non-methylated CpG-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. *EMBO J.*, **6**, 999–1004.

Cleveland,W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, **74**, 829–836.

Hsu,L. *et al.* (2005) Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, **6**, 211–226.

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

Irizarry,R.A. *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.

Jaffe,A.E. *et al.* (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.*, **41**, 200–209.

Kundaje,A. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.

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

Lee,H. *et al.* (2012) DNA methylation shows genome-wide association of NFIX, RAPGEF2 and MSRB3 with gestational age at birth. *Int. J. Epidemiol.*, **41**, 188–199.

Leek,J.T. *et al.* (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, **11**, 733–739.

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

Martinez,J.G. *et al.* (2013) A study of Mexican free-tailed bat syllables: Bayesian functional mixed models for nonstationary acoustic time series. *Journal of the American Statistical Association*, **108**, 514–526.

Meyer,M.J. *et al.* (2015) Bayesian function-on-function regression for multilevel functional data. *Biometrics*, **71**, 563–574.

Mitra,A. and Song,J. (2012) Waveseq: A novel data-driven method of detecting histone modification enrichments using wavelets. *PLoS One*, **7**, e45486.

Morris,J.S. and Carroll,R.J. (2006) Wavelet-based functional mixed models. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **68**, 179–199.

Morris,J.S. *et al.* (2006) Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: a case study. *J. Am. Stat. Assoc.*, **101**, 1352–1364.

Morris,J.S. *et al.* (2008) Bayesian analysis of mass spectrometry proteomics data using wavelet based functional mixed models. *Biometrics*, **64**, 479–489.

Morris,J.S. *et al.* (2011) Automated analysis of quantitative image data using isomorphic functional mixed models, with application to proteomics data. *Ann. Appl. Stat.*, **5**, 894–923.

Nguyen,N. *et al*. (2013) A wavelet-based method to exploit epigenomic language in the regulatory region. *Bioinformatics*, btt467.

Ruppert,D. *et al*. (2003) *Semiparametric Regression*. Cambridge University Press, New York.

Sardy,S. *et al*. (1999) Wavelet shrinkage for unequally spaced data. *Stat. Comput.*, **9**, 65–75.

Shim,H. and Stephens,M. (2015) Wavelet-based genetic association analysis of functional phenotypes arising from high-throughput sequencing assays. *Ann. Appl. Stat.*, **9**, 665–686.

Sweldens,W. (1996) The lifting scheme: a custom-design construction of biorthogonal wavelets. *Appl. Comput. Harmonic Anal.*, **3**, 186–200.

Touleimat,N. and Tost,J. (2012) Complete pipeline for infinium human methylation 450 K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics*, **4**, 325–341.

Wang,D. *et al*. (2012) IMA: An R package for high-throughput analysis of illumina 450 K infinium methylation data. *Bioinformatics*, **28**, 729–730.

Wettenhall,J.M. and Smyth,G.K. (2004) limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics*, **20**, 3705–3706.

Wilhelm-Benartzi,C.S. *et al*. (2013) Review of processing and analysis methods for DNA methylation array data. *Br. J. Cancer*, **109**, 1394–1402.

Zackay,A. and Steinhoff,C. (2010) MethVisual-visualization and exploratory statistical analysis of DNA methylation profiles from bisulfite sequencing. *BMC Res. Notes*, **3**, 337.

Zhu,H. *et al*. (2011) Robust, adaptive functional regression in functional mixed model framework. *J. Am. Stat. Assoc.*, **106**, 1167–1179.

Zhu,H. *et al*. (2012) Robust classification of functional and quantitative image data using functional mixed models. *Biometrics*, **68**, 1260–1268.