BIOMETRIC METHODOLOGY

Bayesian functional data analysis over dependent regions and its application for identification of differentially methylated regions

Suvo Chatterjee1 | Shrabanti Chowdhury2 | Duchwan Ryu3 | Sanjib Basu4

1Department of Epidemiology and Biostatistics, Indiana University, School of Public Health, Bloomington, Indiana, USA
2Department of Genetics and Genomic Sciences and Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, USA
3Department of Statistics and Actuarial Science, Northern Illinois University, Illinois, USA
4Division of Epidemiology and Biostatistics, University of Illinois Chicago, Illinois, USA

Abstract
We consider a Bayesian functional data analysis for observations measured as extremely long sequences. Splitting the sequence into several small windows with manageable lengths, the windows may not be independent especially when they are neighboring each other. We propose to utilize Bayesian smoothing splines to estimate individual functional patterns within each window and to establish transition models for parameters involved in each window to address the dependence structure between windows. The functional difference of groups of individuals at each window can be evaluated by the Bayes factor based on Markov Chain Monte Carlo samples in the analysis. In this paper, we examine the proposed method through simulation studies and apply it to identify differentially methylated genetic regions in TCGA lung adenocarcinoma data.

KEYWORDS
Bayesian smoothing splines, differentially methylated regions, dynamic weighted particle filter, functional data analysis, posterior Bayes factor, TCGA lung adenocarcinoma

1 | INTRODUCTION

DNA methylation is an epigenetic modification when a methyl group gets attached to the cytosine of cancer genome (CG) dinucleotide, forming a methyl cytosine. Differential DNA methylation has been associated with several chronic diseases such as cancer. Therefore, the detection of differential methylation has gained traction in the last decade. Among the vast literature on methods to detect differential methylation, site-wise detection methods are more developed however, recent evidence suggests that analyses based on differentially methylated regions (DMRs) may lead to substantially improved results. This has led to the development of several DMR identification methods such as Jaffe et al. (2012) and Pedersen et al. (2012).

Despite the vast development of statistical methodologies to identify DMRs, the models summarized in Robinson et al. (2014) come with several disadvantages. Most of them use traditional multivariate techniques to detect DMRs that are incapable of handling the high dimensionality, high measurement error, missing values, and high degree of correlation in methylation rates among neighboring CpG sites (Eckhardt et al., 2006), which are some of the inherent components in methylation data. Hence, the traditional multivariate methods may result in spurious statistical analysis and low-powered statistical tests by disregarding the smooth functional behavior of the methylation profiles.

To identify DMRs addressing the issues of large dimensionality, missingness, and correlation of methylation rates from neighboring CpG sites, we consider a Bayesian functional data analysis (BFDA) approach. Specifically, functional data analysis (FDA) proposed in Silverman & Ramsay (2001) views these measurements as realizations of continuous smooth dynamic functions evolving over
space or time. FDA fits smooth curves to these multivariate points represented as linear combinations of suitably chosen basis functions (Ramsay & Silverman, 2005). Such a smooth curve enables the imputation of missing values, helps in the removal of the high noise, models data with irregular time sampling schedules, and accounts for any significant correlation among the observations (Ryu et al., 2016). On similar grounds, BFDA fits a flexible Bayesian nonparametric regression model to this sequence of multivariate points and conducts statistical inference based on a Bayesian framework.

Specifically, as a nonparametric regression model, we use the natural cubic smoothing splines (NCS) which is a conventional smoothing method with consideration of boundary problem. Compared to P-splines, while the smoothing splines and P-splines show similar estimated curves (Berry et al., 2002), the smoothing splines have the lowest mean squared error (MSE) among other smoothing functions due to the interpolant properties when the smoothing parameters are fixed (Keele, 2008). Qin & Zhu (2009) empirically compared the performance of smoothing splines with P-splines, which favored smoothing splines in terms of bias and MSE. Moreover, smoothing splines is free from the knot selection problem. In P-splines, the appropriate location of knots is crucial to prevent the estimation bias (Hastie & Tibshirani, 1990). Whereas, in the virtue of fast computation, P-splines may be preferred when sample size is fairly large. Ma et al. (2017) considered the identification of DMRs with smoothing splines. Other nonparametric regression models such as the wavelet approach (Denault & Jugessur, 2021) are available, as well.

Despite the evidence that cytosine guanine phosphate (CpG) sites within a genomic region (Eckhardt et al., 2006) are correlated to each other, it should be noted that the methylation rates between genomic regions can also be dependent on each other and are solely based on the approach by which the regions were created. Although there is a vast literature of statistical models to account for this dependency (Boker et al., 2002; West & Hepworth, 1991), to provide flexibility in our modeling framework and take advantage of the functional nature of methylation profiles, we consider the roughness/smoothness and variability of the fitted nonparametric regression functions for each genomic region to be dependent on each other. Furthermore, given the issues of scalability under the Bayesian paradigm, we utilize the dynamically weighted particle filter (DWPF) (Liang, 2002; Ryu et al., 2013) for efficient computation.

To assess the performance of our proposed methodology, we performed simulation studies and compared their performance with existing popular methods. Also as an application of the proposed model, we perform DMR identification analysis over the whole genome on the the cancer genome atlas (TCGA) lung adenocarcinoma (LUAD) data. The results from both simulation studies and real-data analysis showed that our proposed approach is effective in finding the true DMRs while effectively controlling for the number of false positives.

The rest of the paper is structured as follows. Section 2 introduces the proposed novel methods to model methylation values. Section 3 shows the comparison of the simulated data for the proposed methods with Bumphunter, DMRcate, and Combp. Section 4 shows the real-data analysis and important findings upon the application of the proposed-dependent method to LUAD data. Finally, a discussion is provided in Section 5.

2 | BAYESIAN FUNCTIONAL DATA ANALYSIS

In this section, we propose a BFDA within and between windows and the identification of DMR by utilizing the Bayes factor at each window.

2.1 | Bayesian functional data analysis within a window

As the first step in our methodological pipeline, we partitioned the whole genome into regions/windows of equal size with a total of 100 CpG sites in them for downstream analysis. Specifically, we have constructed the windows based on the constraint that each window should have a genomic span of 100 kilobases (kbs). In the TCGA LUAD data, we found that genomic regions that spanned 100 kbs, on average, contained 100 CpG sites in them. Hence, following this empirical evidence, we constructed the windows by partitioning the whole genome into equally sized windows each containing 100 CpG sites. The reasoning for considering genomic windows that spanned 100 kbs comes from the evidence presented by recent studies in cervical cancer (Lando et al., 2015) which suggested that changes in the methylation pattern of a large genomic region can have a stronger effect of association with chronic disorders such as cancer as compared to smaller regions. Other studies (Binder et al., 2015; Limbach et al., 2016; Li et al., 2014) have pointed out that genomic regions spanning between 100 kbs and several megabases were necessary to properly characterize changes in methylation patterns for larger regions. Based on this evidence we chose the criteria for creating the genomic windows in our study which spanned 100 kbs.
Next, we consider the methylation data measured at each window with \( n \) CpG sites from \( m_k \) individuals in the group \( k, k = 1, \ldots, G \).

Let \( Y_{ijk} \) denote the log-transformed methylation rate, referred as \( M \)-value in Du et al. (2010), at the CpG site \( i \) from the individual \( j \) in the group \( k \) as

\[
Y_{ijk} = \log \left( \frac{\beta_{ijk} + c}{1 - \beta_{ijk} + c} \right), \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_k, \quad k = 1, \ldots, G, \tag{1}
\]

where \( \beta_{ijk} \) is the methylation rate and \( c \) is an offset value.

We utilize a Bayesian nonparametric regression in the BFDA for the sequence of measured methylation rates over CpG sites, \( Y_{1jk}, \ldots, Y_{njk}, \) from the individual \( j \) in the group \( k \). The typical features of methylation rates include high variability, nonperiodic behavior, correlation, and complex patterns over CpG sites. Regarding the differential methylation, we investigate the group mean of methylation rates at site \( i \) from individual \( j \) in the group \( k \) such that

\[
Y_{ijk} = \mu_{ijk} + \varepsilon_{ijk}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_k, \quad k = 1, \ldots, G,
\]

where \( \mu_{ijk} \) is the mean methylation rate and \( \varepsilon_{ijk} \) is the random error with zero mean and constant variance \( \sigma_{ijk}^2 \) over sites.

Further, considering the group mean as a function of CpG sites, \( x_1, \ldots, x_m \), the mean methylation rate can be modeled by

\[
\mu_{ik} = g_k(x_i) + \delta_{ik}, \quad i = 1, \ldots, n, \quad k = 1, \ldots, G,
\]

where \( g_k(x_i) \) is the group mean functional without assumption of specific functional form and \( \delta_{ik} \) is a discrepancy of mean function with zero mean and constant variance \( \sigma^2_k \) as in Higdon et al. (2004). Combining two models we have the following model for the transformed methylation rates:

\[
Y_{ijk} = g_k(x_i) + \delta_{ik} + \varepsilon_{ijk}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m_k, \quad k = 1, \ldots, G. \tag{2}
\]

In this paper, to investigate the functional pattern of methylation rates over CpG sites, we use the order of CpG site in the window, that is, \( x_i = i \), as the predictor in the model instead of its genomic coordinate and use a sufficient estimator of \( \mu_{ik}, \bar{Y}_{ik} = \frac{1}{m_k} \sum_{j=1}^{m_k} Y_{ijk} \), as a response.

Using the conventional natural cubic spline as our basis function, the Bayesian NCS take all design points as knots. Denoting \( \mu_i = x_i, i = 1, \ldots, n, \) the mean function for group \( k \) can be described by \( g_k(x) = \sum_{i=1}^{n} a_{ki} B_i(x - \mu_i) \), where \( a_{ki} \) are the coefficients of basis functions. The smoothing splines allow only one response at a unique design point. Using the group mean as the response, \( \bar{Y}_{ik} = \frac{1}{m_k} \sum_{j=1}^{m_k} Y_{ijk} \), for \( i = 1, \ldots, n \) and \( k = 1, \ldots, G \), the smoothing splines \( g_k(\cdot) \) for group \( k \) can be estimated by minimizing the following penalized residual sum of square:

\[
\sum_{i=1}^{n} \left\{ \bar{Y}_{ik} - g_k(x_i) \right\}^2 + \alpha_k \int_{u \in R} g_k''(u)^2 du, \tag{3}
\]

where \( \alpha_k \) is a given positive smoothing penalty, \( g_k''(\cdot) \) is the second-order derivative of \( g_k(\cdot) \), and \( R \) is the range of design points. Denoting \( a_k = (a_{k1}, \ldots, a_{kn})^T \) a vector of coefficients and \( B = \{ B_i(x_j - \mu_j) \}, r, s = 1, \ldots, n, \) a matrix of basis functions evaluated at design points, the penalty term can be expressed by \( \alpha_k \int_{u \in R} g_k''(u)^2 du = \alpha_k (\Omega_k + \alpha_k I)^{-1} \), where \( \Omega_k = \{ B''(u - \mu_r)B''(u - \mu_s) du \) is a matrix of integrals of squared double derivatives of basis functions. Denoting the vector of functional value of the smoothing splines as \( g_k = [g_k(x_1), \ldots, g_k(x_n)]^T = B_{\alpha_k} \), we can construct \( K \) that satisfies \( K = B^{-T} \Omega B^{-1} \), that is, \( a_k^T \Omega_k a_k = g_k^T K g_k \), under the piecewise polynomial basis function \( B_{\cdot}(\cdot) \). See Eubank (1999) for details of construction of \( K \).

As mentioned in Berry et al. (2002), Ryu et al. (2011), and Yue et al. (2012), for a Bayesian approach, we take all design points as knot points and assign a singular normal prior on \( g_k \) that has the probability density function proportional to \( (\frac{\alpha_k}{\sigma_k^2})^{(n-2)/2} \exp(-\frac{\alpha_k}{\sigma_k^2} g_k^T K g_k) \), where \( \alpha_k \) is a smoothing penalty and \( \sigma_k^2 \) is the variance of the discrepancy of the mean function. Without loss of generality, we use \( \tau_k = \frac{\sigma_k}{\sigma_k^2} \) and assign a conjugate Gamma prior, \( \tau_k \sim G(A_k, B_k) \). We assign conjugate inverse Gamma priors on \( \sigma_k^2 \) and \( \eta_{jk}^2 \), respectively, \( \sigma_k^2 \sim IG(A_k, B_k) \) and \( \eta_{jk}^2 \sim IG(A_k^*, B_k^*) \).

Denoting \( y_{jk} = (y_{1jk}, \ldots, y_{njk}) \) and \( \bar{y}_k = (\bar{Y}_{1k}, \ldots, \bar{Y}_{nk})^T \), the full conditional distributions of the parameters are given by

\[
g_k \sim N (I + \alpha_k K)^{-1} \Omega_k^{-1} (I + \alpha_k K)^{-1} \sigma_k^2 \bar{Y}_k, \quad k = 1, \ldots, G
\]

\[
\tau_k \sim G \left[ \frac{n}{2} + A_k^*, \frac{1}{2} \bar{Y}_k^T K g_k \right].
\]

\[
\sigma_k^2 \sim IG \left[ \frac{n}{2} + A_k, \frac{1}{2} \bar{y}_k^T (I + \alpha_k K)^{-1} \bar{y}_k + B_k \right].
\]

\[
\eta_{jk}^2 \sim IG \left[ \frac{n}{2} + A_k^*, \frac{1}{2} (y_{jk} - \bar{y}_k)^T (I + \alpha_k K)^{-1} (y_{jk} - \bar{y}_k) + B_k^* \right], \quad j = 1, \ldots, m_k.
\]

where \( I \) is the \( n \times n \) identity matrix. Using Gibbs sampler with \( N \) iterations after \( B \) iterations as burning time, we generate MCMC samples of \( g_k \) and other nuisance parameters.

When the design points are unequally spaced or have missing responses, we first estimate \( g_k \) without missing
values, and then estimate the components of \( g_k \) corresponding to missing values by the interpolations as Eubank (1999) and Hastie & Tibshirani (1990) suggested. In this paper, our data do not contain missing responses.

2.2 Bayesian functional data analysis between windows

When the windows of CpG sites are independent, we may apply the BFDA discussed in the previous subsection to each window. However, the windows can be associated with each other especially when they are adjacent. In this subsection, we model the dependent structure of consecutive windows by using a parameter transition model and propose to utilize the DWPF approach for efficient computation as in Ryu et al. (2013).

When two adjacent windows are dependent on each other, the corresponding group mean functions and variabilities would be associated with each other. The group means functions are utilized in the identification of DMRs. To provide more flexibility of mean functions, we consider the association of smoothing penalty rather than the mean function between windows. The variabilities are utilized in the identification of variably methylated regions (Jaffe et al., 2012). Hence, we also consider the association of variance of the discrepancy between windows with associated variabilities.

We consider a dynamic model across the sequence of dependent windows. For each window, we keep the same model but consider the involved parameters in the model to be sequentially varying by windows. Specifically, because the smoothing penalty and the variance of the discrepancy characterize the fitted curve in the window, we let those parameters take into account the dependent windows. Regarding \( T \) windows for whole genome and \( G \) groups of samples, let \( \tau_{t,k} \) denote the smoothing penalty and \( \sigma_{t,k}^2 \) denote the variance of the discrepancy for window \( t \) and group \( k \), \( t = 1, \ldots, T; k = 1, \ldots, G \). Then, the changes of parameters from window \( t-1 \) to window \( t \) can be described by the following transition models:

\[
\begin{align*}
\log(1/\tau_{t,k}) &= \log(1/\tau_{t-1,k}) + U_{t,k}, \\
\log(\sigma_{t,k}^2) &= \log(\sigma_{t-1,k}^2) + V_{t,k},
\end{align*}
\]

where \( U_{t,k} \) and \( V_{t,k} \) are Gaussian random errors with zero mean and fixed variance, respectively.

The sequential Monte Carlo (SMC) has been proved as a useful tool for inference of dynamic systems (Liu & Liu, 2001). For window \( t \), let \( \mathbf{y}_t = (\mathbf{y}_{1,t}, \ldots, \mathbf{y}_{G,t}) \) denote the vector of methylation rates, \( \lambda_t = (\tau_{1,t}, \ldots, \tau_{G,t}, \sigma_{1,t}^2, \ldots, \sigma_{G,t}^2) \) denote the vector of window-varying parameters over all groups and \( \theta \) denote the vector of other parameters specified in the previous subsection. To improve efficiency, within the Gibbs algorithm, we may use the Markov Chain Monte Carlo (MCMC) samples of \( \lambda_t \) from the marginal posterior. Suppressing the design points that are the order of CpG sites within a window, the marginal posterior can be expressed by

\[
p(\lambda_1, \lambda_2, \ldots, \lambda_{t-1}, \mathbf{y}_1, \mathbf{y}_2) \propto \int p(y_t | \lambda_0, \ldots, \lambda_t, \theta) p(\lambda_t | \lambda_{t-1}) \cdots p(\theta) \, d\theta,
\]

where \( \lambda_0 \) indicates the vector of hyperparameters for \( \lambda_1 \). Assuming a Markovian structure for \( \lambda_t \) and using the colon notation in the subscript to denote the accumulated sequence up to window \( t \) such that \( \mathbf{y}_{1:t} = (\mathbf{y}_1, \ldots, \mathbf{y}_t) \) and \( \lambda_{1:t} = (\lambda_1, \ldots, \lambda_t) \), the marginalized joint posterior distribution of the sequence of \( \lambda_1, \ldots, \lambda_t \) can be decomposed by

\[
p(\lambda_{1:t} | \lambda_0, \mathbf{y}_{1:t}) \propto p(\lambda_{1:t-1} | \lambda_0, \mathbf{y}_{1:t-1}) p(\lambda_t | \lambda_{t-1}, \mathbf{y}_t).
\]

Using SMC, we update the methylation rates at the window \( t \) to the methylation rates from the previous windows and infer the parameters at window \( t \) without inferring the parameters for previous windows again. Specifically, using sequential importance sampling (SIS), we can generate sequences of importance samples across the windows with corresponding importance weights, under the dynamic model. However, SIS may lead to the degeneration problem. To avoid the degeneration problem, we utilize the dynamically weighted importance sampling (DWIS) algorithm that reduces the variation of importance weights and increases the variation of importance samples through dynamic weighting and population control. See Liang (2002) for details of DWIS. Specifically, we use \( W \)-type move in the dynamic weighting step and the adaptive pruned-enriched population control scheme in the population control step. For window 1, as an initial population, we generate MCMC samples for all parameters with size \( N_1 \) instead of extrapolation.

The DWPF uses a combination of extrapolation and DWIS to generate a population of particles or MCMC samples of all parameters. Regarding the population size \( N_t \) at window \( t \), for \( i = 1, \ldots, N_t \), let \( \lambda_{i,t}^{(1)} \) denote the particle \( i \) of \( \lambda_t \) and \( \omega_{i,t}^{(1)} \) denote the importance weight of the particle \( i \), and \( \lambda_t = (\lambda_{1,t}^{(1)}, \ldots, \lambda_{N_t,t}^{(1)}) \) denote the population of all particles with weights \( \mathbf{W}_t = (\omega_{1,t}^{(1)}, \ldots, \omega_{N_t,t}^{(1)}) \).

In the extrapolation, we use the transition model of \( \lambda_t \) to project the particles from window \( t-1 \) to window \( t \) and transmit them to DWIS. In DWIS, we generate a new population based on the transmitted population for window \( t \). We denote the lower and upper population size control bounds as \( N_{\text{low}} \) and \( N_{\text{up}} \), respectively, and the lower and
upper limiting population sizes as \(N_{\text{min}}\) and \(N_{\text{max}}\), respectively. We also denote the lower and upper weight control bounds for all windows as \(W_{\text{low}}\) and \(W_{\text{up}}\), respectively.

We use the following algorithm to collect the particles of \(\lambda_i\) with corresponding weights \(w_t\) for window \(t\), \(t = 1, \ldots, T\):

**window 1**

Sample: Collect \(N_1\) MCMC samples of \(\lambda_1\) from the posterior distribution by applying the BFDA discussed in the previous subsection after \(B\) burning iterations, and set the MCMC samples at each iteration as \(\hat{\lambda}_1^{(i)}\) with weight \(\hat{w}_1^{(i)} = 1, i = 1, \ldots, N_1 = 20,000\). It establishes \(\hat{\lambda}_1\) and \(\hat{W}_1\).

DWIS: Generate \((\Lambda_1, W_1)\) from \((\hat{\lambda}_1, \hat{W}_1)\) using DWIS, with the marginal posterior distribution \(p(\lambda_1|y_1)\) as the target distribution.

**window 2**

Extrapolation: Generate \(\hat{\lambda}_2^{(i)}\) from \(\hat{\lambda}_1^{(i)}\), with the extrapolation operator \(q(\lambda_2|\lambda_1^{(i)}, y_1^{(i)})\), and set \(\hat{w}_2^{(i)} = w_1^{(i)} \frac{p(\lambda_2^{(i)}|y_1^{(i)}, \lambda_1^{(i)})}{p(\lambda_1^{(i)}|y_1^{(i)}, \lambda_2^{(i)})} \text{ for each } i = 1, 2, \ldots, N_1\), to establish \((\hat{\lambda}_2, \hat{W}_2)\).

DWIS: Generate \((\Lambda_2, W_2)\) from \((\hat{\lambda}_2, \hat{W}_2)\) using DWIS, with the target \(p(\lambda_2|y_1^{(i)})\).

\[
\vdots
\]

**window T**

Extrapolation: Generate \(\hat{\lambda}_T^{(i)}\) from \(\hat{\lambda}_{T-1}^{(i)}\), with the extrapolation operator \(q(\lambda_T|\lambda_{T-1}^{(i)}, y_1^{(i)})\) and set \(\hat{w}_T^{(i)} = w_{T-1}^{(i)} \frac{p(\lambda_T^{(i)}|y_1^{(i)}, \lambda_{T-1}^{(i)})}{p(\lambda_{T-1}^{(i)}|y_1^{(i)}, \lambda_T^{(i)})} \text{ for each } i = 1, 2, \ldots, N_{T-1}\), to establish \((\hat{\lambda}_T, \hat{W}_T)\).

DWIS: Generate \((\Lambda_T, W_T)\) from \((\hat{\lambda}_T, \hat{W}_T)\) using DWIS, with the target \(p(\lambda_T|y_1^{(i)})\).

At each window, the functional values of methylation rates \(g_t\) can be generated from its full conditional distribution \(p(g_t|\cdot), t = 1, \ldots, T\). See Ryu et al. (2013) for details of DWPF.

### 2.3 Identification of differentially methylated regions using the Bayes factor

We examine the differential methylation by groups at each window. In model (2), when different group means functions are desirable to model the methylation rates, we may assess the window to be differentially methylated. Otherwise, the window is not differentially methylated and one group mean function will be enough to model the methylation rates in the window. We consider the following two models \(M_1\) and \(M_2\):

- \(M_1\) : the window has one group mean function, \(G = 1\) in model (2),
- \(M_2\) : the window has more than one (say, \(K\)) group mean functions, \(G = K\) in model (2).

If \(M_1\) is preferred to model the methylation rates in the window, we may conclude that the window is not a DMR; whereas, if \(M_2\) is preferred, we can conclude that the window is a DMR.

To choose a good model for the methylation rates in the window, we utilize the posterior Bayes factor that provides more consistent results than the traditional Bayes factor (Aitkin, 1991, 1998; Jiang & Xu, 2022; Wang & Xu, 2021). Let \(\Theta_l\) denote all unknown quantities in model \(M_l\) and \(y\) denote the log-transformed methylation rates in the window with the likelihood \(p(y|\Theta_l)\), \(l = 1, 2, \ldots, K\), then the posterior Bayes factor to compare \(M_1\) and \(M_2\) can be calculated by the ratio of the marginal likelihoods as follows:

\[
BF(M_1, M_2) = \frac{\int_{\Theta_1} p(y|\Theta_1) p(\Theta_1) d\Theta_1}{\int_{\Theta_2} p(y|\Theta_2) p(\Theta_2) d\Theta_2},
\]

where \(p(\Theta_l|y)\), \(l = 1, 2, \ldots, K\), indicates the posterior density. Under the model (2), apparently \(p(y|\Theta_l)\) is given by a product of Gaussian densities. Utilizing the particles and weights for \(p(\Theta_l|y)\) discussed in the previous subsection, we obtain the marginal likelihoods by taking the weighted average of the likelihoods and calculating the Bayes factor for each window. To avoid the computational difficulty, we use the log-scaled Bayes factor. If the Bayes factor is less than a threshold value, we prefer \(M_2\) over \(M_1\) and identify the window as a DMR.

#### 2.4 Parameter values for simulation and real-data analysis

In this paper, for our simulation studies as well as real-data analysis, we consider two groups to identify differential methylation, the cancer (case) group and the normal (control) group, that is, we have \(G = 2\). In Equation (1), we set the offset \(c = 0.01\). We set the hyper-parameters for
the Gamma and inverse Gamma priors as \( A_i = 1, B_i = 1,000 \) and \( A_s = B_s = A^*_s = B^*_s = 1 \). We use the Gibbs sampler with \( N = 20,000 \) iterations and \( B = 1,000 \) iterations as burning time. In DWPF, we set the lower and upper population size control bounds as \( N_{\text{low}} = 15,000 \) and \( N_{\text{up}} = 25,000 \), respectively, and the lower and upper limiting population sizes as \( N_{\text{min}} = 10,000 \) and \( N_{\text{max}} = 30,000 \), respectively. We also set the lower and upper weight control bounds for all windows as \( W_{\text{low}} = e^{-5} \) and \( W_{\text{up}} = e^{5} \), respectively.

3 | SIMULATION STUDIES

In this section, we examine the performance of our proposed methodologies in identifying differential mean patterns under a dependent data structure through extensive simulation studies. We do so by utilizing two simulation strategies in which we benchmark their performance in comparison to existing statistical methods.

3.1 | Simulation study 1

Since our proposed method has been developed with the general aim to detect two group differential mean patterns for long sequences of data under a dependence structure, in this simulation strategy, we generate long sequences of synthetic data from random functions and induce a dependent structure between them. This general strategy of dependent data generation applies to methylation data as well as to broader settings. Briefly, we simulated 10 windows each with 75 samples (\( n_1 = 25, n_2 = 50 \)) and 100 observations in them. We obtained four simulation scenarios (\( \rho = 0, 0.3, 0.5, 0.7 \)) by varying the magnitude of dependency between these simulated windows and each scenario was replicated 100 times. As can be seen in Figure 1, we generated windows 1 and 8 with no functional difference between the group mean curves however, the other windows were generated using significant differences. The details of our simulation setting are mentioned on page 1 of the Web Appendix, Supporting information.

We evaluated the performance of our methods (BFDAM1 = independent method; BFDAM2 = dependent method) using misclassification tables with the popular Bumphunter method (Jaffe et al., 2012). Under this strategy, only Bumphunter was chosen as a comparator because from existing literature (Robinson et al., 2014) we found Bumphunter to be the only general method that is designed to model differential mean patterns for any data with sequentially dependent data structure and not just methylation microarrays.

From Figure 2, we can see that the BFDAM1 (independent) method performs the best when the consecutive windows are uncorrelated (\( \rho = 0 \)). However, its performance deteriorates with the increase in the correlation between neighboring windows. In addition, from Table 1 we can see that the BFDAM1 method performs fairly well in detecting differential mean patterns for windows 1, 2, 7, 9, and 10, but it often misclassifies windows 3, 4, 5, 6, and 8. These results point toward the fact that BFDAM1 does not account for the correlation between the windows and that is the reason we can see the misclassification rates for BFDAM1 increased as the correlation between the neighboring windows increased. In addition, we also noticed that the computation times reported by BFDAM1 were sub-optimal and raised concerns about computational scalability.

As anticipated, in the case of the BFDAM2 (dependent) method we can see from Figure 2 and Table 1 that it outperforms BFDAM1 in detecting windows with differential or non-differential mean patterns in terms of misclassification rates. Specifically, unlike BFDAM1, we see that despite the increase in correlation between neighboring windows the misclassification rates of BFDAM2 remain fairly low in all simulation scenarios. In addition, we also notice that BFDAM2 is computationally at least 5 times more efficient than BFDAM1 due to the implementation of the SMC approach.

Furthermore, from Table 1 we also observe that Bumphunter performs fairly poorly than both BFDAM1 and BFDAM2 in terms of misclassification rates. We found that Bumphunter had high sensitivity in windows with low variations while low sensitivity in windows with larger variations between the functional curves. For instance, this can particularly be seen for windows 2, 7, and 9 which presented larger variability in the mean methylation among case and control groups (Web Appendix, Figure S1). The Bumphunter method misclassifies these windows as having non-differential mean patterns in almost over 90% of the cases. In line with that, since, window 8 had a very low variability between the two group means, Bumphunter had a 100% accuracy in detecting this window as having non-differential mean patterns. The reasoning for this behavior might be due to the fact that Bumphunter does not account for modeling the variance in addition to modeling for the mean between the two groups. In addition, past studies suggest (Li et al., 2015) that Bumphunter is usually more powerful in identifying smaller windows with differential mean patterns and its performance drops significantly when windows of larger sizes are involved.

Collectively, these results suggest that in comparison to the competing methods such as BFDAM1 and Bumphunter, BFDAM2 provides the most optimal performance.
**FIGURE 1** Left to the right and first line to the second line represents window 1 to window 10. Blue color and red color indicate the control group and case group, respectively. Dotted lines describe group means functions, and circles, and crosses indicate the generated data based on the random functions. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version.

**FIGURE 2** Performance of BFDAM1 (independent) and BFDAM2 (dependent) methods. The plots show the distribution of the logarithm of the Bayes factors for Windows 1 through 10 over 100 replications in each of the four simulation scenarios. The four different simulation scenarios were the varying correlations of $\rho=0, 0.3, 0.5, 0.7$ between neighboring windows. The horizontal blue line indicates the threshold value set at $-5$. This threshold was chosen based on recommendations from previous works (Aitkin, 1991; Kass & Raftery, 1995). Any window below this threshold is detected as a window with differential mean patterns. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version.
TABLE 1 Misclassification rates of our proposed methods BFDAM1 (independent), BFDAM2 (dependent), and the competing bumphunter method.

| Method          | Window | Correlation | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | Time (min) |
|-----------------|--------|-------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----------|
| BFDAM1 (independent) | ρ = 0.0 | 0% | 0% | 0% | 3% | 20% | 15% | 0% | 2% | 0% | 4% | 112.45 |
|                 | ρ = 0.3 | 0% | 0% | 1% | 6% | 22% | 22% | 0% | 25% | 0% | 5% | 115.12 |
|                 | ρ = 0.5 | 0% | 0% | 9% | 14% | 32% | 24% | 0% | 39% | 0% | 21% | 113.58 |
|                 | ρ = 0.7 | 0% | 8% | 23% | 38% | 41% | 28% | 6% | 56% | 0% | 23% | 111.33 |
| BFDAM2 (dependent) | ρ = 0.0 | 0% | 0% | 0% | 1% | 9% | 7% | 0% | 1% | 0% | 2% | 22.49 |
|                 | ρ = 0.3 | 0% | 0% | 0% | 3% | 10% | 10% | 0% | 11% | 0% | 2% | 23.02 |
|                 | ρ = 0.5 | 0% | 0% | 0% | 3% | 10% | 10% | 0% | 17% | 0% | 9% | 22.72 |
|                 | ρ = 0.7 | 0% | 3% | 10% | 17% | 18% | 12% | 3% | 24% | 0% | 9% | 22.27 |
| Bumphunter       | ρ = 0.0 | 0% | 81% | 0% | 0% | 28% | 24% | 52% | 0% | 100% | 0% | 7.49 |
|                 | ρ = 0.3 | 0% | 89% | 1% | 0% | 33% | 36% | 61% | 0% | 100% | 0% | 7.12 |
|                 | ρ = 0.5 | 0% | 84% | 1% | 0% | 26% | 36% | 68% | 0% | 100% | 0% | 6.99 |
|                 | ρ = 0.7 | 0% | 90% | 16% | 5% | 28% | 41% | 78% | 0% | 100% | 2% | 8.25 |

Note: The misclassification rates span from Windows 1 through 10 in each of the four simulation scenarios. The four different simulation scenarios were the varying correlations of ρ = 0, 0.3, 0.5, 0.7 between neighboring windows. The reported misclassification rates are the average taken over 100 replicated datasets for each simulation scenario. The computation time depicts the average time taken to complete the analysis on each window.

by utilizing flexible non-parametric smooth functions in modeling the mean and the variance; transition models in accounting for the correlation both within and between spatial windows and an SMC approach in increasing computational efficiency.

3.2 Simulation study 2

Since our motivating data pertains to a methylation 450k microarray data and to preserve correlation patterns in real datasets, under this simulation strategy, we generated simulation data by using a real dataset and examined the performance of our proposed methods specifically under the 450k methylation microarray data setting. To achieve this, we obtained a publicly available methylation dataset. The Gene Expression Omnibus (GEO) dataset GSE66836 from (Bjaanaes et al., 2016) included DNA methylation (DNAm) profiles of 183 lung tissue samples. Given the statistical justification of using M-values for model development purposes (Du et al., 2010), all of the following data generation and analysis have been done under the M-value scale.

Specifically, for this simulation study, we selected 18 normal lung tissue samples which spanned a total of 485,577 CpG sites. The 18 methylation samples were then randomly divided into two groups (normal and tumor). Next, we partitioned the data into (i) equally spaced windows of length 80 CpG sites, (ii) equally spaced windows of length 100 CpG sites, and (iii) equally spaced windows of length 120 CpG sites. For each type of data partition, we then randomly chose 100 windows for simulation purposes. Next, for each type of data partition, we then simulated DMRs of CpGs by randomly selecting 30 windows (30%) from the total of 100 windows and added treatment effects to the M-values in the group with higher average M-values. This process was repeated 5 times for each data partition and effect size μ = (2, 3, 4, 5, 6) yielding a total of 75 simulation datasets. Finally, we applied three existing DMR detection methods (DMRcate (Peters et al., 2015), Bumphunter (Jaffe et al., 2012), comb-p (Pedersen et al., 2012)) along with our proposed (BFDAM1, BFDAM2) methods to the generated datasets and compared their results on power and false positive rate (FPR) (Table 2). All existing methods were run on their best settings as mentioned in the previous work (Mallik et al., 2019). R codes for downloading, extracting, and pre-processing the methylation data along with the codes for data generation, analysis, and benchmarking are all provided in the GitHub repository https://github.com/schatterjee30/BFDAMs.

We began our benchmarking by first validating that our proposed methods along with the other competing methods controlled for the type 1 error rate (at 5% level) when no DMRs were present in the data. To achieve this, we performed a permutation-based analysis. Briefly, to create this null setting, we selected normal lung tissue samples and divided them into two equal-sized groups. We labeled the first group as normal and the second as the tumor. Next, we partitioned the data into windows of length 100 CpG sites. We then randomly chose 100 regions (containing a total of 10,000 CpG sites) from the whole data and created 100 replicated datasets. These replicated datasets were 100 random permutations of these two groups. Then, for each permuted datasets we ran our model along
Table 2 Power and false positive rate (FPR) of the independent, dependent, and other comparator methods over 100 simulations for each of the five datasets which were created with varying effect sizes.

| Method          | Effect size | #CpGs per window = 120 | #CpGs per window = 100 | #CpGs per window = 80 |
|-----------------|-------------|------------------------|------------------------|------------------------|
|                 |             | Power      | FPR        | Power      | FPR        | Power      | FPR        |
| BFDAM1 (independent) | 2           | 83.7%      | 4.3%       | 80.7%      | 4.9%       | 77.3%      | 5.3%       |
|                 | 3           | 86.7%      | 4.2%       | 83.0%      | 4.3%       | 80.0%      | 4.7%       |
|                 | 4           | 89.2%      | 2.9%       | 87.3%      | 3.9%       | 85.0%      | 4.3%       |
|                 | 5           | 93.3%      | 2.8%       | 90.0%      | 3.1%       | 88.3%      | 3.9%       |
|                 | 6           | 96.6%      | 1.2%       | 94.0%      | 2.9%       | 91.6%      | 3.2%       |
| BFDAM2 (dependent) | 2           | 87.0%      | 4.1%       | 83.0%      | 4.4%       | 80.0%      | 4.8%       |
|                 | 3           | 89.5%      | 2.9%       | 86.7%      | 4.3%       | 83.7%      | 4.4%       |
|                 | 4           | 93.3%      | 1.7%       | 90.2%      | 2.8%       | 86.7%      | 4.1%       |
|                 | 5           | 96.5%      | 1.6%       | 93.3%      | 2.8%       | 89.9%      | 3.9%       |
|                 | 6           | 100%       | 0.0%       | 96.3%      | 1.7%       | 93.3%      | 2.8%       |
| Bumphunter      | 2           | 46.2%      | 2.1%       | 66.1%      | 2.3%       | 86.1%      | 2.5%       |
|                 | 3           | 46.5%      | 2.2%       | 66.5%      | 2.4%       | 86.5%      | 2.3%       |
|                 | 4           | 46.8%      | 2.2%       | 66.8%      | 2.4%       | 86.8%      | 2.3%       |
|                 | 5           | 47.0%      | 2.8%       | 67.0%      | 2.5%       | 87.0%      | 2.4%       |
|                 | 6           | 47.0%      | 2.8%       | 67.0%      | 2.5%       | 87.0%      | 2.4%       |
| DMRcate         | 2           | 40.1%      | 5.2%       | 46.1%      | 5.3%       | 86.1%      | 11.2%      |
|                 | 3           | 40.3%      | 5.5%       | 46.4%      | 5.8%       | 86.5%      | 11.5%      |
|                 | 4           | 40.8%      | 5.8%       | 46.9%      | 5.9%       | 86.8%      | 11.8%      |
|                 | 5           | 41.0%      | 5.8%       | 47.1%      | 5.9%       | 87.0%      | 12.0%      |
|                 | 6           | 41.0%      | 5.8%       | 47.1%      | 5.9%       | 87.0%      | 12.0%      |
| Comb-p          | 2           | 40.1%      | 0.1%       | 33.1%      | 0.0%       | 80.1%      | 2.1%       |
|                 | 3           | 40.4%      | 0.2%       | 33.5%      | 0.0%       | 80.5%      | 2.1%       |
|                 | 4           | 40.8%      | 0.2%       | 34.0%      | 0.0%       | 81.0%      | 2.5%       |
|                 | 5           | 41.0%      | 0.2%       | 34.0%      | 0.1%       | 81.0%      | 5.1%       |
|                 | 6           | 41.0%      | 0.2%       | 34.0%      | 0.1%       | 81.0%      | 5.7%       |

From Table 2, we see that our proposed method (BFDAM2) outperforms all the other methods in the majority of simulation scenarios. Specifically, in terms of FPR, we see that BFDAM1, BFDAM2, and Bumphunter show good control at the 5% nominal level in all the simulation scenarios. However, comb-p had a slightly inflated FPR when the size of the genomic windows decreased. DMRcate on the other hand showed the worst control of FPR in almost all the simulation scenarios. In terms of statistical power, we see that BFDAM2 had the best performance in maintaining a fairly high power in most simulation scenarios and it showed an increase in power as the effect sizes increased. For the competing methods, we see that their performance enhanced with the increase in effect sizes; however, this change was only in very small fractions in comparison to the relative changes in effect sizes.

We also note that the variations in BFDAM1 and BFDAM2’s power by varying window sizes remained relatively moderate in comparison to other methods which showed a drastic decrease in sensitivity as the size of the windows increased. This evidence is in line with our results from simulation study 1. In addition, evidence from previous studies which have suggested that existing DMR detection methods perform poorly when the size of DMRs increase (Li et al., 2015) is also in agreement with our findings. Taken together, these results depict that our method BFDAM2 had the most optimal performance as compared to other competing methods. Therefore, based on this evidence we only compared BFDAM2 with other competing methods in the following real-data analysis.
4 | IDENTIFICATION OF DIFFERENTIALLY METHYLATED REGIONS FOR THE CANCER GENOME ATLAS LUNG ADENOCARCINOMA DATA

4.1 | Data description

In this section, we demonstrate our proposed approach for DMR detection using the 450k methylation microarray data on LUAD obtained from TCGA portal. These data contain DNA methylation profiling using the Illumina Infinium HumanMethylation450 platform. The Infinium platform analyzed up to 482,421 CpG dinucleotides and 3,091 CpH dinucleotides, spanning gene-associated elements as well as intergenic regions. DNA samples were received, bisulfite converted, and cytosine methylation was evaluated using Illumina Infinium HumanMethylation450 microarrays. There were 254 lung tissue samples of LUAD patients and 32 normal samples with the methylation rates, denoted by $\beta$-values, which were calculated using the intensities received from the methylated and the unmethylated alleles. Methylation arrays are measured on beta-values which are computed based on the ratios of the methylated signal intensity to the sum of both methylated and unmethylated signals after background subtraction, so they range from 0 (completely unmethylated) to 1 (fully methylated). $M$-values are logit transformation of beta-values, that is, $M = \log(\frac{\beta}{1-\beta})$ and these values have been shown to have better statistical properties such as homoscedasticity (Du et al., 2010) in methylation data analysis. In the following, we present a detailed analysis of DMR detection across the whole genome (including sex chromosomes) using our proposed and three other competing approaches.

4.2 | DNA methylation data analysis

Following our proposed notations, our analysis was based on $G = 2$ or two groups (normal and cancer samples). Based on the evidence from simulation studies, first, we partitioned the CpG sites in every chromosome into windows of 100 CpG sites. Next, we performed Bayesian NCS fitting on the observed methylation data of 254 LUAD and 32 normal samples and obtained the underlying smooth functions. Web Appendix, Figure S2 depicts the functional data visualization of the observed multivariate measurements in a genome-wide manner along all autosomes chromosomes including the sex chromosomes denoted as chromosome 23 from the hereafter. In each plot, the green and red colored curves represent the smoothed mean function for the normal and cancer samples, respectively, while the blue curves denote the smoothed mean functions of all 286 samples.

We started with window 1 by performing the smoothing spline estimation and generated 20,000 MCMC particles of $\sigma^2_{1k}$ and $\tau_{1k}$, $k = 1, 2$. For the following windows, we applied our dependent method by using the transition model (3) that projects the particles from window $t$ to $t+1$, for $t=1, \ldots, T-1$. As mentioned in Section 2 to account for the additional variation for the individual $j$ in the group $k$ we add $\eta^2_{jk}$, $j=1, \ldots, m_k$, to the projected values of $\hat{\sigma}^2_{t+1,k}$ across all samples for each CpG site. We obtained Bayes factors for inference in every window as described in Section 2.3. Using a suitable threshold, we determine if the data provide sufficient evidence of the presence of two distinct groups and hence detect a window to be DMR.

Figure 3 shows the number of DMRs detected by our dependent approach BFDAM2 and other competing methods over 23 chromosomes. From Figure 3, we see that BFDAM2 alone identified 922 windows as DMRs which were identified as non-DMRs by the other methods. Also, we see that the number of DMRs uniquely identified by each competing method is much smaller than that by BFDAM2. This is compatible with our simulation results which indicate that BFDAM2 has a higher ability to detect DMRs as compared to the other comparator methods. Also, we see that 1,565 windows that have been identified as DMRs by BFDAM2 have also been identified as DMRs by three other methods which increase our confidence and support the validity of the DMR regions identified by BFDAM2.
DISCUSSION

This research is motivated by a growing body of literature focusing on epigenetic features that may be associated with the disparities in non-small cell lung cancer progression and survival outcomes. The hypermethylation of the CpG island sequences located in the promoter regions of genes is increasingly being used to study the impact of epigenetic modifications. In this paper, we proposed a BFDA model to identify, select, and jointly model differential methylation features from methylation array data. To date, the applications of the FDA in the genomic or public health domain remain very scarce and there are still a lot of uncovered areas in genomics that can make use of such a powerful and robust method. The proposed functional modeling approach for detecting DMR is parsimonious to address the large dimensionality of whole-genome sequencing and incorporates potential correlation among neighboring regions. We proposed a DWPF with Bayesian non-parametric smoothing splines for modeling individual functional patterns followed by the identification of DMRs.

We used simulation studies to compare the performance of our method with existing popular DMR detection methods. In simulation study 1, first, we proposed our independent approach of DMR detection that fits a Bayesian NCS in individual windows without taking into account the correlation among the CpG sites from neighboring windows. As our simulation results indicated although this approach outperformed the existing bumphunter method, its performance deteriorated with an increase in the correlation among the CpG sites from neighboring windows. Moreover, this approach was also challenged by high computational time. To remedy these two immediate problems associated with the independent approach, we proposed our dependent approach next. In this approach, we used transition models to account for the dependency that inherently exists between two genomic regions. We used an efficient SMC method named DWPF to get the parameter estimates of the subsequent regions without fitting the non-parametric regression function in every window. This maneuver not only made this approach computationally very efficient but also showed a very robust performance in detecting DMRs, as our simulation results indicated.

In simulation study 2, we generated the data using real TCGA LUAD data and benchmarked our proposed methods with three other competing methods. The results indicated that our proposed approach BFDAM2 not only controls for type-I error but also has a higher ability to detect DMRs in most of the simulation settings as compared to the other competing methods. This further creates a major milestone in the use of functional data modeling in genomics data.

We applied our dependent approach as well as other competing methods to identify whole genome-wide differential methylation in LUAD patients’ data from the TCGA portal. We identified several DMRs along the whole genome that were uniquely identified by BFDAM2. These biological findings can further be translated into clinical research and thus we see great promises of FDA in genomics data applications.

ACKNOWLEDGMENTS

The authors thank the Editor, Associate Editors, and anonymous reviewers. SC was supported by the Intramural Research Program of Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) of the National Institutes of Health (NIH). The results shown in this paper are in whole or part based upon data generated by the https://www.cancer.gov/tcga TCGA Research Network.

DATA AVAILABILITY STATEMENT

The Lung Adenocarcinoma data used for simulation purposes can be found on the Gene Expression Omnibus portal https://www.ncbi.nlm.nih.gov/geo/ with the accession number GSE66836. For the real-data analysis, The Cancer Genome Atlas Program Lung Adenocarcinoma data can be found by clicking on this link below. https://drive.google.com/drive/folders/1YvzjHYM6mjNgjMGQi0pjPN9C7aRHL?usp=sharing

ORCID

Suvo Chatterjee https://orcid.org/0000-0002-9771-109X
Shrabanti Chowdhury https://orcid.org/0000-0001-8429-7822
Duchwan Ryu https://orcid.org/0000-0003-2435-7496
Sanjib Basu https://orcid.org/0000-0001-5607-4058

REFERENCES

Aitkin, M. (1991) Posterior Bayes factors. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(1), 111–128.
Aitkin, M. (1998) Simpson’s paradox and the Bayes factor. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1), 269–270.
Berry, S.M., Carroll, R.J. & Ruppert, D. (2002) Functional data analysis and mixed effect models. *Journal of the American Statistical Association*, 97, 160–169.
Binder, A.M., LaRocca, J., Lesueur, C., Marsit, C.J. & Michels, K.B. (2015) Epigenome-wide and transcriptome-wide analyses reveal gestational diabetes is associated with alterations in the human leukocyte antigen complex. *Clinical Epigenetics*, 7(1), 1–12.
Bjaanæs, M.M., Fleischer, T., Halvorsen, A.R., Daunay, A., Busato, F., Solberg, S., Jørgensen, L., Kure, E., Edvardsen, H., Børresen-Dale, A.-L., Brustugun, O.T., Tost, J., Kristensen, V. & Helland, Å. (2016) Genome-wide DNA methylation analyses in lung adenocarcinomas: association with EGFR, KRAS and TP53 mutation status, gene expression and prognosis. Molecular Oncology, 10(2), 330–343.

Boker, S.M., Rotondo, J.L., Xu, M. & King, K. (2002) Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. Psychological Methods, 7(3), 338.

Denault, W.R. & Judessur, A. (2021) Detecting differentially methylated regions using a fast wavelet-based approach to functional association analysis. BMC Bioinformatics, 22, 1–15.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W.A., Hou, L. & Lin, S.M. (2010) Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics, 11(1), 1–9.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A., Haefliger, C., Horton, R., Howe, K., Jackson, D.K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K. & Beck, S. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. Nature Genetics, 38(12), 1378–1385.

Eubank, R.L. (1999) Nonparametric regression and spline smoothing. CRC Press.

Hastie, T. & Tibshirani, R. (1990) Generalized additive models. New York: Chapman and Hall.

Higdon, D., Kennedy, M., Cavendish, J.C., Cafeo, J.A. & Ryne, R.D. (2004) Combining field data and computer simulations for calibration and prediction. SIAM Journal on Scientific Computing, 26(2), 448–466.

Jaffe, A.E., Feinberg, A.P., Irizarry, R.A. & Leek, J.T. (2012) Significance analysis and statistical dissection of variably methylated regions. Biostatistics, 13(1), 166–178.

Jaffe, A.E., Murakami, P., Lee, H., Leek, J.T., Fallin, M.D., Feinberg, A.P. & Irizarry, R.A. (2012) Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. International Journal of Epidemiology, 41(1), 200–209.

Jiang, Y. & Xu, X. (2022) A two-sample test of high-dimensional significance analysis and test of high-dimensional covariance matrix. Journal of the American Statistical Association, 108(501), 111–123.

Kass, R.E. & Raftery, A.E. (1995) Bayes factors. Journal of the American Statistical Association, 89(427), 773–795.

Keele, L.J. (2008) Semiparametric regression for the social sciences. John Wiley & Sons.

Lando, M., Fjeldbo, C.S., Wilting, S.M., Snoek, B.C., Aarnes, E.-K., Forsberg, M.F., Kristensen, G.B., Steenbergen, R.D. & Lyng, H. (2015) Interplay between promoter methylation and chromosomal loss in gene silencing at 3p11.23 in cervical cancer. Epigenetics, 10(10), 970–980.

Li, D., Xie, Z., Le Pape, M. & Dye, T. (2015) An evaluation of statistical methods for DNA methylation microarray data analysis. BMC Bioinformatics, 16(1), 1–20.

Li, Y., Chen, J.A., Sears, R.L., Gao, F., Klein, E.D., Karydas, A., Geschwind, M.D., Rosen, H.J., Boxer, A.L., Guo, W., Pellegrini, M., Horvath, S., Miller, B.L., Geschwind, D.H. & Coppola, G. (2014) An epigenetic signature in peripheral blood associated with the haplotype on 17q21. 31, a risk factor for neurodegenerative tauopathy. PLOS Genetics, 10(3), e1004211.

Liang, F. (2002) Dynamically weighted importance sampling in Monte Carlo computation. Journal of the American Statistical Association, 97(459), 807–821.

Limbach, M., Saare, M., Tsrel, L., Kisand, K., Eglit, T., Sauer, S., Axelsson, T., Syvänen, A.-C., Metspalu, A., Milani, L. & Peterson, P. (2016) Epigenetic profiling in CD4+ and CD8+ T cells from Graves’ disease patients reveals changes in genes associated with T cell receptor signaling. Journal of Autoimmunity, 67, 46–56.

Liu, J.S. & Liu, J.S. (2001) Monte Carlo strategies in scientific computing, vol. 75. Springer.

Ma, P., Zhang, N., Huang, J.Z. & Zhong, W. (2017) Adaptive basis selection for exponential family smoothing splines with application in joint modeling of multiple sequencing samples. Statistica Sinica, 27(4), 1757–1777.

Mallik, S., Odom, G.J., Gao, Z., Gomez, L., Chen, X. & Wang, L. (2019) An evaluation of supervised methods for identifying differentially methylated regions in illumina methylation arrays. Briefings in Bioinformatics, 20(6), 2224–2235.

Pedersen, B.S., Schwartz, D.A., Yang, I.V. & Kechris, K.J. (2012) Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. Bioinformatics, 28(22), 2986–2988.

Peters, T.J., Buckley, M.J., Statham, A.L., Pidslay, R., Samaras, K., V Lord, R., Clark, S.J. & Molloy, P.L. (2015) De novo identification of differentially methylated regions in the human genome. Epigenetics & Chromatin, 8(1), 1–6.

Qin, G. & Zhu, Z. (2009) Local asymptotic behavior of regression splines for marginal semiparametric models with longitudinal data. Science in China Series A: Mathematics, 52(9), 1982–1994.

Ramsay, J.O. & Silverman, B.W. (2005) Functional data analysis. Springer.

Robinson, M.D., Kahraman, A., Law, C.W., Lindsay, H., Nowicka, M., Weber, L.M. & Zhou, X. (2014) Statistical methods for detecting differentially methylated loci and regions. Frontiers in Genetics, 5, 324.

Ryu, D., Li, E. & Mallick, B.K. (2011) Bayesian nonparametric regression analysis of data with random effects covariates from longitudinal measurements. Biometrics, 67(2), 454–466.

Ryu, D., Liang, F. & Mallick, B.K. (2013) Sea surface temperature modeling using radial basis function networks with a dynamically weighted particle filter. Journal of the American Statistical Association, 108(501), 111–123.

Ryu, D., Xu, H., George, V., Su, S., Wang, X., Shi, H. & Podolsky, R.H. (2016) Differential methylation tests of regulatory regions. Statistical Applications in Genetics and Molecular Biology, 15(3), 237–251.

Silverman, B.W. & Ramsay, J.O. (2001) Functional data analysis. International Encyclopedia of the Social and Behavioral Sciences. Amsterdam: Elsevier.

Wang, Z. & Xu, X. (2021) Testing high-dimensional covariance matrices via posterior Bayes factor. Journal of Multivariate Analysis, 181, 104674.

West, S.G. & Hepworth, J.T. (1991) Statistical issues in the study of temporal data: daily experiences. Journal of Personality, 59(3), 609–662.
Yue, Y.R., Speckman, P.L. & Sun, D. (2012) Priors for Bayesian adaptive spline smoothing. *Annals of the Institute of Statistical Mathematics*, 64(3), 577–613.

**SUPPORTING INFORMATION**
Web Appendix A, referenced in Sections 3.1, 3.2, and 4.2, as well as data and code are available with this paper at the Biometrics website on Wiley Online Library. Data and code for all the simulation studies and data analysis also can be found at the GitHub repository https://github.com/schatterjee30/BFDAMs. The Lung Adenocarcinoma data can be found at the Gene Expression Omnibus portal https://www.ncbi.nlm.nih.gov/geo/ with accession number GSE66836. The Cancer Genome Atlas Program Lung Adenocarcinoma data can be found by clicking on this link below. https://drive.google.com/drive/folders/1YvzjHYM6mjNjjMGQi0pjIPN9C7aRHL?usp=sharing

**How to cite this article**: Chatterjee, S., Chowdhury, S., Ryu, D. & Basu, S. (2023) Bayesian functional data analysis over dependent regions and its application for identification of differentially methylated regions. *Biometrics*, 79, 3294–3306. https://doi.org/10.1111/biom.13902