Received: 15 May 2019 | Accepted: 8 May 2020

DOI: 10.1111/biom.13307

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

A novel statistical method for modeling covariate effects in bisulfite sequencing derived measures of DNA methylation

Kaiqiong Zhao[1,5] | Karim Oualkacha[2] | Lajmi Lakhal-Chaieb[3] | Aurélie Labbe[4] | Kathleen Klein[5] | Antonio Ciampi[1,5] | Marie Hudson[5,6] | Inés Colmegna[6,7] | Tomi Pastinen[8] | Tieyuan Zhang[9] | Denise Daley[10] | Celia M.T. Greenwood[1,5,11]

[1] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada
[2] Département de Mathématiques, Université du Québec à Montrèal, Montreal, QC, Canada
[3] Département de Mathématiques et de Statistique, Université Laval, Quebec City, QC, Canada
[4] Département des Sciences de la Décision, HEC Montrèal, Montreal, QC, Canada
[5] Lady Davis Institute for Medical Research, Montreal, QC, Canada
[6] Department of Medicine, McGill University, Montreal, QC, Canada
[7] The Research Institute of the McGill University Health Centre, Montreal, QC, Canada
[8] Center for Pediatric Genomic Medicine, Children's Mercy Kansas City, Kansas City, MO, USA
[9] Department of Psychiatry, Douglas Mental Health University Institute, McGill University, Montreal, QC, Canada
[10] The Centre for Heart Lung Innovation, and Department of Medicine, University of British Columbia, Vancouver, BC, Canada
[11] Department of Human Genetics and Gerald Bronfman Department of Oncology, McGill University, Montreal, QC, Canada

Correspondence
Celia M.T. Greenwood, Department of Epidemiology, Biostatistics and Occupational Health, Gerald Bronfman Department of Oncology, Department of Human Genetics, Lady Davis Institute for Medical Research, McGill University, Montreal, QC H3A 0G4, Canada.
Email: celia.greenwood@mcgill.ca

Funding information
Compute Canada, Grant/Award Number: Research Groups (RRG) ID 2514; Canadian Institutes of Health Research, Grant/Award Number: MOP 130344; Natural Sciences and Engineering Research Council of Canada; Genome Canada, Grant/Award Number: 2017 Bioinformatics and Computational Biology (B/C); Fonds de Recherche du Québec - Santé, Grant/Award Numbers: Doctoral Training Award, FRQS-31110

Abstract

Identifying disease-associated changes in DNA methylation can help us gain a better understanding of disease etiology. Bisulfite sequencing allows the generation of high-throughput methylation profiles at single-base resolution of DNA. However, optimally modeling and analyzing these sparse and discrete sequencing data is still very challenging due to variable read depth, missing data patterns, long-range correlations, data errors, and confounding from cell type mixtures. We propose a regression-based hierarchical model that allows covariate effects to vary smoothly along genomic positions and we have built a specialized EM algorithm, which explicitly allows for experimental errors and cell type mixtures, to make inference about smooth covariate effects in the model. Simulations show that the proposed method provides accurate estimates of covariate effects and captures the major underlying methylation patterns with excellent power. We also apply our method to analyze data from rheumatoid arthritis patients and controls. The method has been implemented in R package `SOMNiBUS`.

KEYWORDS

differentially methylated region, EM algorithm, generalized additive model, next-generation sequencing, penalized regression splines

1 INTRODUCTION

Heritability is high for a wide range of human diseases (Maurano *et al.*, 2012), but only a portion of it is attributable to additive genetic variation (Ober and Vercelli, 2011). Maher (2008) suggested that environmental exposures play an important role in explaining the "missing" heritability. Plausibly, such exposures, in interaction with genetic predisposition, may lead to epigenetic modification that alters gene regulation without changing genome sequence (Jaenisch and Bird, 2003). For example, differences in epigenetic profiles may explain how risk factors like age (Horvath, 2013) and smoking (Teschendorff *et al.*, 2015) impact disease susceptibility. Consequently, examining how epigentic profiles contribute to disease development and are influenced by environmental factors can provide novel insights into disease etiology and possible therapies (Feinberg, 2007).

The most-studied epigenetic mark is DNA methylation that primarily occurs at a cytosine-guanine dinucleotide (ie, CpG site) (Lister *et al.*, 2009). Localized differential methylation is a characteristic feature of many diseases, such as diabetes (Nilsson *et al.*, 2014), Alzheimer's disease (De Jager *et al.*, 2014) and autoimmune disorders (Liu *et al.*, 2013).

Measuring large-scale DNA methylation at single nucleotide resolution is now possible owing to the development of bisulfite sequencing protocols (Frommer *et al.*, 1992), which can be implemented genome-wide or in a set of targeted regions. Targeted Custom Capture Bisulfite Sequencing (TCCBS) platforms produce DNA methylation levels for comprehensive subsets of informative CpGs. Thus, epigenomic dysregulation can be captured at a much lower cost than whole-genome bisulfite sequencing (WGBS). This approach's capacity to detect novel disease associations has been demonstrated (Allum *et al.*, 2015; Li *et al.*, 2015). In this work, we focus on analysis of predefined regions targeted by TCCBS, with the aim to identify differentially methylated regions (DMRs) that are associated with phenotypes or traits.

Methods for extracting interpretable results from the raw methylation data derived from either WGBS or TCCBS are greatly hindered by the variability in read depths, the many missing values, and the possibility of data errors. Specifically, due to the stochastic nature of sequencing and alignment, coverage—the total number of reads spanning a CpG site—varies substantially across sites and individual samples, which leads to wide-ranging precision for methylation proportions, and to many missing values. In fact, estimates of DNA methylation are correlated with read depths (Stephens *et al.*, 2016). Furthermore, the observed counts of methylated and unmethylated reads could be contaminated by errors arising from excessive or insufficient bisulfite treatment, and from misalignment of reads or other aspects of the sequencing processes. Studies show that ignoring these errors could bias inference about the associations of interest (Cheng and Zhu, 2013; Lakhal-Chaieb *et al.*, 2017).

Additionally, due to cell type specific differences in methylation levels, variability in cell type mixture proportions has a strong effect on observed levels of methylation from mixed tissue samples. This mixture, as well as factors known to alter methylation levels, such as age (Horvath, 2013), can confound associations of interest. Hence, it is essential to develop methods to adjust methylation signals for multiple covariates.

Moving in this direction, approaches have been proposed for identifying DMRs from bisulfite sequencing data; see overviews in Shafi *et al.* (2017) and Yu and Sun (2016a). Typically, to account for spatial correlations of methylation between neighboring CpG sites, strategies include Hidden Markov models (HMM) (Sun and Yu, 2016; Yu and Sun, 2016b; Shokoohi *et al.*, 2018), hierarchical models with autoregressive or random walk correlation structures (Rackham *et al.*, 2017; Korthauer *et al.*, 2019), and kernel-based smoothing methods (Hansen *et al.*, 2012; Hebestreit *et al.*, 2013; Lakhal-Chaieb *et al.*, 2017). However, none of these methods meet all the desirable objectives *simultaneously*: regional testing, estimation of multiple covariate effects, adjustment for read depth variability, and experimental errors. For example, several of the current HMM-based (Sun and Yu, 2016; Yu and Sun, 2016b) and hierarchical methods (Rackham *et al.*, 2017) only test for differential methylation between two independent groups of samples and do not allow for the adjustment of multiple covariates. Approaches using a binomial mixed model for DNA methylation analysis (Lea *et al.*, 2015; Weissbrod *et al.*, 2017) allow for multiple covariates and can capture sample correlations, but were only designed for single site analysis. BSmooth (Hansen *et al.*, 2012), a kernel-based method, detects differential methylation after converting the methylated and total counts to proportions. However, this conversion could lead to reduced power since it disregards read depth variability and fails to distinguish between noisy and accurate measurements (Rackham *et al.*, 2017). Moreover, most of the existing methods ignore experimental errors. On the other hand, the only approach accounting for data errors, the Smooth Methylation Status Call (SMSC) (Lakhal-Chaieb *et al.*, 2017), is only developed for data from a single cell type.

More importantly, most of the existing methods are of a two-stage nature (Hansen *et al.*, 2012; Hebestreit *et al.*, 2013; Lakhal-Chaieb *et al.*, 2017). Typically, they first smooth the raw methylation data for each sample separately, and then, in the second stage, they estimate covariate effects by modeling the smoothed methylation data. These per-sample smoothing strategies do not take advantage of information contained across samples and fail to fully exploit the fact that samples with similar covariate profiles (eg, disease status, cell type composition, or other phenotypes of interest) can be expected to share similar methylation patterns. In addition, separating smoothing and inference steps results in biased uncertainty estimates. In summary, it would be highly desirable to develop a general framework of analysis, which collapses smoothing and testing steps into a single step, and simultaneously addresses regional testing, estimation of multiple covariate effects, adjustment for read depth variability, and experimental errors.

In this paper, we propose such a general framework. Our strategy allows information to be shared not only between nearby CpGs, but also across samples, thus providing greater sensitivity for capturing patterns common to several samples of similar characteristics (rather than one sample).

Specifically, our approach is built on a hierarchical regression model that describes bisulfite sequencing data. We assume, as in Lakhal-Chaieb *et al.* (2017) and Cheng and Zhu (2013), that the observed read counts arise from an unobserved latent true methylation state compounded by errors. These true methylation counts are then modeled by a binomial distribution, dependent on read depth. Note that the probability parameter of this binomial distribution depends on the sample-level covariates of interest, such as cell-type mixture proportions and the trait of interest, but also nearby methylation information. To capture realistic methylation patterns across regions, we additionally allow baseline methylation levels, covariate effects, and adjustment effects to vary smoothly along genomic positions: this is done by using spines. This amounts to borrowing information from the local correlation structures between methylation levels, and allows us to remedy local information gaps due to missingness. This formulation naturally allows for any number of covariates in the model.

This article is organized as follows. Section 2 describes the proposed model along with its estimation and inference procedures. A motivating data example from a study of cases with rheumatoid arthritis (RA) and controls is described in Section 3. Simulation studies evaluating the performance of our proposed method and comparing our type I errors and power to existing methods are summarized in Section 4. The paper concludes with a discussion in Section 5.

2 | METHOD

2.1 | Notation and data

We consider DNA methylation measures over a targeted genomic region from N independent samples. Let m_i be the number of CpG sites for the ith sample, $i = 1, 2, \dots N$. We write t_{ij} for the genomic position (in base pairs) for the ith sample at the jth CpG site, $j = 1, 2, \dots, m_i$. The set of genomic positions captured in different samples do not have to be identical because each sample has an individual profile of covered CpG sites, due to read depth variability. Methylation levels at a site are quantified by the number of methylated reads and the total number of reads. We define X_{ij} as the total number of reads aligned to CpG j from sample i. The tissue samples sent for bisulfite sequencing experiments from most studies will normally be composed of a mixture of cell types. For example, common cell types are in blood: granulocytes, T cells, B cells, monocytes, neutrophils, and eosinophils; in adipose tissues: adipocyte, preadipocyte, endothelial and mural cells. Thus, the reads obtained at a CpG site are likely to capture contributions from different cell types; the true underlying methylation statuses are probably different across these X_{ij} reads. We denote the *true* methylation status for the kth read obtained at CpG j of sample i as S_{ijk}, where $k = 1, 2, \dots X_{ij}$. S_{ijk} is binary and we define $S_{ijk} = 1$ if the corresponding read is methylated and $S_{ijk} = 0$ otherwise. In the presence of experimental errors in sequencing or preprocessing, the *observed* methylation status, written as Y_{ijk}, can be distinct from the true underlying information S_{ijk}. We denote $Y_{ijk} = 1$ if the corresponding read is observed as methylated and $Y_{ijk} = 0$ otherwise. We additionally denote the *true* and *observed methylated counts* at CpG j for sample i with $S_{ij} = \sum_{k=1}^{X_{ij}} S_{ijk}$ and $Y_{ij} = \sum_{k=1}^{X_{ij}} Y_{ijk}$, respectively. Furthermore, we assume that we have the information on P covariates for the N samples, denoted as $\boldsymbol{Z_i} = (Z_{1i}, Z_{2i}, \dots Z_{Pi})$, for $i = 1, 2, \dots N$.

2.2 | Model

We built here on concepts introduced in Cheng and Zhu (2013) and Lakhal-Chaieb *et al.* (2017) to account for experimental errors. We assume that, depending on the true underlying methylation status S_{ijk}, the observed status Y_{ijk} is a Bernoulli variable with parameters p_0 or p_1, that is,

$$p_0 = \mathbb{P}(Y_{ijk} = 1 \mid S_{ijk} = 0),$$

$$p_1 = \mathbb{P}(Y_{ijk} = 1 \mid S_{ijk} = 1). \tag{1}$$

Here, these two parameters capture errors; p_0 is the rate of false methylation calls, and $1 - p_1$ is the rate of false non-methylation calls. These rates are assumed to be constant across all reads and positions. The error parameters p_0 and p_1 can be estimated by looking at raw sequencing data at CpG sites known in advance to be methylated or unmethylated (Wreczycka *et al.*, 2017). We assume hereafter that p_0 and p_1 are known. Implications of such an assumption is discussed later in the Supporting Information Section 2.2.

We then assume the true methylated counts S_{ij} follows a binomial distribution with a methylation proportion parameter π_{ij} that depends on the sample-level covariates $\boldsymbol{Z_i}$, and on nearby methylation patterns. Specifically,

$$S_{ij} \mid \boldsymbol{Z_i}, X_{ij} \sim \text{Binomial}(X_{ij}, \pi_{ij}),$$

$$g(\pi_{ij}) = \beta_0(t_{ij}) + \sum_{p=1}^{P} \beta_p(t_{ij}) Z_{pi}, \qquad (2)$$

where $g(\cdot)$ is a logit link function and $\beta_0(t_{ij})$ and $\{\beta_p(t_{ij})\}_{p=1}^{P}$ are functional parameters for the intercept and covariate effects. This amounts to assuming smoothly varying methylation levels and covariate effects on methylation levels across our targeted small genomic regions. In practice, to estimate Model (2), the functions $\beta_p(t_{ij})$ can be represented by the coefficients of a chosen spline bases of rank L_p,

$$\beta_p(t_{ij}) = \sum_{l=1}^{L_p} \alpha_{pl} B_l^{(p)}(t_{ij}), \text{ for } p = 0, 1, \dots P,$$

where $\{B_l^{(p)}(\cdot)\}_{l=1}^{L_p}$ denotes the spline basis, and $\boldsymbol{\alpha_p} = (\alpha_{p1}, \dots \alpha_{pL_p})^T \in \mathcal{R}^{L_p}$ are the coefficients to be estimated. In this way, model (2) becomes a generalized linear model (GLM), $g(\boldsymbol{\pi}) = \mathbb{X}\boldsymbol{\alpha}$, where $\boldsymbol{\pi} = (\pi_{11}, \dots \pi_{1m_1}, \pi_{21}, \dots \pi_{2m_2}, \dots \pi_{Nm_N})^T \in [0,1]^M$ with $M = \sum_{i=1}^{N} m_i$, $\boldsymbol{\alpha} \in \mathcal{R}^K$ with $K = \sum_{p=0}^{P} L_p$, and $\mathbb{X}$ is the spanned design matrix of dimension $M \times K$, stacked with elements $B_l^{(p)}(t_{ij}) \times Z_{pi}$; for detailed forms, see Supporting Information Appendix A.

To avoid over-fitting, we penalize departure from smoothness, using penalized regression splines (Wahba, 1980; Parker and Rice, 1985). Specifically, we use a comparatively large number of knots (equivalent to large L_p) and a penalization, quantified by the integrated squared curvature of the splines, is added as an extra term in the log-likelihood function (loss function),

$$\mathcal{L}^{\text{Penalization}} = \sum_{p=0}^{P} \lambda_p \int \left(\beta_p''(t)\right)^2 dt = \sum_{p=0}^{P} \lambda_p \boldsymbol{\alpha_p^T A_p \alpha_p}. \qquad (3)$$

In Equation (3), $\boldsymbol{A_p'}s$ are $L_p \times L_p$ positive semidefinite matrices with the (l, l') element $\boldsymbol{A_p}(l, l') = \int {B^{(p)}}''_l(t) {B^{(p)}}''_{l'}(t) dt$; these are fixed quantities given the specified set of basis functions. The weights λ_p, that is, the smoothing parameters, are positive parameters that establish a trade-off between the closeness of the curve to the data and the smoothness of the fitted curves. Note that there is one smoothing parameter per covariate in our model. The smoothing process across targeted regions is accomplished by adding the penalization terms in Equation (3) to the model in Equation (2).

2.3 | Estimation

2.3.1 | Penalized complete likelihood

If the true methylated counts S_{ij} were available, model (2) with penalization (3) would be estimated by maximizing the penalized log-likelihood,

$$l^{\text{complete}}(\boldsymbol{S}; \boldsymbol{\alpha}, \boldsymbol{\lambda}) = l(\boldsymbol{S}; \boldsymbol{\alpha}) - \frac{1}{2} \sum_{p=0}^{P} \lambda_p \boldsymbol{\alpha_p^T A_p \alpha_p}$$
$$= l(\boldsymbol{S}; \boldsymbol{\alpha}) - \frac{1}{2} \boldsymbol{\alpha}^T \boldsymbol{A_\lambda \alpha},$$

where $l(\boldsymbol{S}; \boldsymbol{\alpha}) = \sum_{i=1}^{N} \sum_{j=1}^{m_i} \{S_{ij} \log(\pi_{ij}) + (X_{ij} - S_{ij}) \log(1 - \pi_{ij})\}$, and $\boldsymbol{A_\lambda}$ is a $K \times K$ positive semidefinite block diagonal matrix of the form $\boldsymbol{A_\lambda} = \text{Diag}\{\lambda_0 \boldsymbol{A_0}, \lambda_1 \boldsymbol{A_1}, \dots, \lambda_P \boldsymbol{A_P}\}$. This is also the complete-data log-likelihood of the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{S}$, that is, $\log(f(S)) + \log(f(Y \mid S))$; indeed, $f(Y \mid S)$ only depends on the known error rates p_0 and p_1, and bears no information on the parameters of interest.

2.3.2 | Smoothed E-M algorithm

In practice, the true methylation data, S_{ij}, are unknown and one only observes Y_{ij}, which is a mixture of binomial counts arising from both the truly methylated and truly unmethylated reads. The EM algorithm (Dempster *et al.*, 1977) allows us to estimate model (2) based on the observed data Y_{ij}, by repeatedly replacing a trial estimate $(\boldsymbol{\alpha}^\star, \boldsymbol{\lambda}^\star)$ by a new $(\boldsymbol{\alpha}, \boldsymbol{\lambda})$, which is a maximum of the function

$$Q(\boldsymbol{\alpha} \mid \boldsymbol{\alpha}^\star) = \mathbb{E}\left\{ l^{\text{complete}}(\boldsymbol{S}; \boldsymbol{\alpha}, \boldsymbol{\lambda}) \mid \boldsymbol{Y}, \boldsymbol{\alpha}^\star \right\}$$
$$= l(\boldsymbol{\eta}^\star; \boldsymbol{\alpha}) - \frac{1}{2} \boldsymbol{\alpha}^T \boldsymbol{A_\lambda \alpha}. \qquad (4)$$

E step In Equation (4) $\boldsymbol{\eta}^\star = (\eta_{11}^\star, \dots, \eta_{1m_1}^\star, \eta_{21}^\star, \dots, \eta_{2m_2^\star}, \dots, \eta_{Nm_N}^\star)^T \in \mathcal{R}^M$ are conditional expectations of S_{ij} given Y_{ij} evaluated at the trial estimates $(\boldsymbol{\alpha}^\star, \boldsymbol{\lambda}^\star)$; in our case, these take the form

$$\eta_{ij}^\star = \mathbb{E}\left(S_{ij} \mid Y_{ij}; \boldsymbol{\alpha}^\star, \boldsymbol{\lambda}^\star\right) = \frac{Y_{ij} p_1 \pi_{ij}^\star}{p_1 \pi_{ij}^\star + p_0(1-\pi_{ij}^\star)} + \frac{\left(X_{ij} - Y_{ij}\right)(1-p_1)\pi_{ij}^\star}{(1-p_1)\pi_{ij}^\star + (1-p_0)(1-\pi_{ij}^\star)}, \tag{5}$$

with $\pi_{ij}^\star = g^{-1}(\mathbb{X}\boldsymbol{\alpha}^\star)$, which depends on $\boldsymbol{\lambda}^\star$ via the dependence of $\boldsymbol{\alpha}^\star$ on $\boldsymbol{\lambda}^\star$. Calculating these conditional expectations $\eta_{ij}^\star$ from (5) constitutes the E step of our algorithm.

M step Each M step involves maximizing the Q function in (4) to update $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$. This is a penalized (GLM) likelihood maximization problem with multiple quadratic penalties, previously studied in Wood (2011), Wood *et al.* (2016), and Wood and Fasiolo (2017). Our computational strategy for estimating smoothing parameters $\boldsymbol{\lambda}$ is a nested optimization procedure (Wood, 2011), with an outer iteration for optimizing $\boldsymbol{\lambda}$ and an inner penalized iteratively reweighted least squares (P-IRLS) iteration to estimate $\boldsymbol{\alpha}$ given the trial value of $\boldsymbol{\lambda}$ from the outer iteration.

For given values of smoothing parameters $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \dots \lambda_P)$, a unique maximizer of expression (4) is readily computed by P-IRLS; see more details in the Supporting Information Appendix B. Specifically, the outer iteration involves maximizing a restricted likelihood for smoothing parameters $\boldsymbol{\lambda}$, which is obtained by integrating $\boldsymbol{\alpha}$ out of the joint likelihood for $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$. We rely on the work done by Wood (2011) and use a Laplace approximated restricted likelihood; see more details in the Supporting Information Appendix C. As the analytical forms for derivatives and Hessians of this restricted likelihood are also available, the optimization for $\boldsymbol{\lambda}$ in the outer iteration can be readily achieved via Newton's method.

Although the combination is undoubtedly computationally complex, the nested iterations will guarantee convergence for models with properly defined likelihoods (Wood, 2011; Wood *et al.*, 2016).

E-M iteration We iterate between the E and M steps until convergence to obtain $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\lambda}}$. Given the estimates of basis coefficients $\boldsymbol{\alpha}_p$, for $p = 0, 1, \dots P$, the functional parameters $\beta_p(t)$ can be thus estimated by $\widehat{\beta_p(t)} = \{\boldsymbol{B}^{(p)}(t)\}^T\{\widehat{\boldsymbol{\alpha}_p}\}$, where t is a genomic position lying within the range of the input positions $\{t_{ij}\}$, and $\boldsymbol{B}^{(p)}(t) = (B_1^{(p)}(t), B_2^{(p)}(t), \dots B_{L_p}^{(p)}(t))^T \in \mathcal{R}^{L_p}$ is a column vector with nonrandom quantities obtained from evaluating the set of basis functions $\{B_l^{(p)}(\cdot)\}_l$ at position t.

2.4 | Inference for smooth covariate effects

To obtain a quantification of the uncertainty accompanying the smoothed EM estimates for the covariate effects $\{\beta_1(t), \beta_2(t), \dots \beta_P(t)\}$, we additionally estimate their pointwise confidence intervals (CI) in Section 2.4.1, and obtain tests of hypotheses for these effects in Section 2.4.2. This inference is carried out conditional on the values of smoothing parameter $\boldsymbol{\lambda}$; that is, the uncertainty in estimating $\boldsymbol{\lambda}$ is not accounted for. The potential bias associated with this assumption is shown to be small; see the pointwise confidence interval coverage in Figure 4 and the distribution of region-based P-values under the null in Figure 5.

2.4.1 | Confidence interval estimation

Analytical derivation for standard errors usually involves calculating the observed Fisher information for parameters $\boldsymbol{\alpha}$ from the marginal log-likelihood for $\boldsymbol{Y}$. However, in this case, a direct calculation of the observed Fisher information is analytically intractable because the observed $\boldsymbol{Y}$ follows a mixture of two binomial distributions. To circumvent this problem, we rely on the work of Louis (1982) and Oakes (1999), which showed that this Fisher information can be calculated solely from the Q function (4), without referring to the marginal distribution of $\boldsymbol{Y}$.

Theorem 1. *Under the usual regularity conditions for maximum likelihood, we have the following asymptotic results for the estimators $\widehat{\boldsymbol{\alpha}}$ obtained from the smoothed-EM algorithm,*

$$\sqrt{M}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{\mathcal{L}} \boldsymbol{MVN}_K(\boldsymbol{0}, \boldsymbol{\mathcal{I}}^{-1}), \ \ \textit{as } M \to \infty.$$

Here, K is the dimension of the spline coefficients $\boldsymbol{\alpha}$, and $\boldsymbol{\mathcal{I}} = \mathbb{E}[-\mathcal{H}_{ij}(\boldsymbol{\alpha})]$. Specifically $\mathcal{H}_{ij}(\boldsymbol{\alpha})$ has the form

$$\mathcal{H}_{ij}(\boldsymbol{\alpha}) = \mathbb{X}_{(l,)}^T\left(-X_{ij}w_{ij} + \delta_{ij}w_{ij}\right)\mathbb{X}_{(l,)} - \boldsymbol{A}_\lambda,$$

where $\mathbb{X}_{(l,)}$ is the lth row of the design matrix $\mathbb{X}$, which corresponds to the CpG j of sample i, $w_{ij} = \pi_{ij}(1-\pi_{ij})$ is the element of the weight matrix, and

$$\delta_{ij} = \frac{Y_{ij}p_1p_0}{\left[p_1\pi_{ij} + p_0(1-\pi_{ij})\right]^2} + \frac{\left(X_{ij} - Y_{ij}\right)(1-p_1)(1-p_0)}{\left[(1-p_1)\pi_{ij} + (1-p_0)(1-\pi_{ij})\right]^2}.$$

The proof of Theorem 1 is given in the Supporting Information Appendix D. Theorem 1 provides the desired variance-covariance matrix of the EM estimators $\widehat{\boldsymbol{\alpha}}$, which can be estimated using the observed Fisher information

$$\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\boldsymbol{\alpha}}) = \{-\mathcal{H}(\widehat{\boldsymbol{\alpha}})\}^{-1},$$

where $\mathcal{H}(\widehat{\boldsymbol{\alpha}}) = \sum_{i,j} \mathcal{H}_{ij}(\widehat{\boldsymbol{\alpha}})$. Let $\boldsymbol{V}$ denote this variance estimator and $\boldsymbol{V}_{\boldsymbol{p}}$ be the diagonal blocks of $\boldsymbol{V}$ corresponding to $\boldsymbol{\alpha}_p$, with dimensions $L_p \times L_p$. As $\beta(t)$ is a linear combination of coefficients $\boldsymbol{\alpha}_{\boldsymbol{p}}$, the estimated variance of $\widehat{\beta_p(t)}$ takes the form $\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\beta_p(t)}) = \{\boldsymbol{B}^{(p)}(t)\}^T \boldsymbol{V}_{\boldsymbol{p}} \{\boldsymbol{B}^{(p)}(t)\}$. Therefore, the confidence interval for $\beta_p(t)$ at significance level ν can be estimated by $\widehat{\beta_p(t)} \pm \mathbb{Z}_{\nu/2}\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\beta_p(t)})}$, for any t in the range of interest, where $\mathbb{Z}_{\nu/2}$ is $\nu/2$ (upper-tail) quantile of a standard normal distribution.

2.4.2 | Hypothesis testing for a regional zero effect

We can also construct a region-wide test of the null hypothesis

$$H_0 : \beta_p(t) = 0, \text{for any } t \text{ in the genomic interval.}$$

This test depends on the association between covariate Z_p and methylation levels across the region, after adjustment for all the other covariates, and the null hypothesis is equivalent to $H_0 : \boldsymbol{\alpha}_p = \boldsymbol{0}$. A Wald-type statistic can be naturally proposed as

$$T_p = \widehat{\boldsymbol{\alpha}_{\boldsymbol{p}}}^T \left\{\boldsymbol{V}_{\boldsymbol{p}}\right\}^{-1} \widehat{\boldsymbol{\alpha}_{\boldsymbol{p}}},$$

where $\{\boldsymbol{V}_{\boldsymbol{p}}\}^{-1}$ denotes inverse if $\boldsymbol{V}_{\boldsymbol{p}}$ is nonsingular; for singular $\boldsymbol{V}_p$, the inverse is replaced by the Moore-Penrose inverse $\{\boldsymbol{V}_{\boldsymbol{p}}\}^{-}$. If $\boldsymbol{\alpha}_{\boldsymbol{p}}$ is a vector of unpenalized coefficients, under the null hypothesis, T_p asymptotically follows a Chi-square distribution with degrees of freedom L_p. In the presence of smoothness penalization, L_p should be replaced by the effective degrees of freedom (EDF), τ_p, which depends on the magnitude of smoothing parameter λ_p and is smaller than L_p. Motivated by the work of Wood (2013), we define the EDF τ_p as

$$\tau_p = \sum_{l=a_p}^{b_p} \left(2\boldsymbol{F} - \boldsymbol{F}\boldsymbol{F}^T\right)_{(l,l)}, \text{ for } p = 0, 1, \ldots P, \qquad (6)$$

where $a_p = \sum_{m=0}^{p-1} L_m + 1$ if $p > 0$ and $a_p = 1$ if $p = 0$, $b_p = \sum_{m=0}^{p} L_m$ for any p, and $(\bullet)_{(l,l)}$ stands for the lth leading diagonal element of a matrix. In (6), $\boldsymbol{F}$ is the smoothing matrix of our model, which has the form $\boldsymbol{F} = (\mathbb{X}^T\widehat{\boldsymbol{W}}\mathbb{X} + \boldsymbol{A}_{\widehat{\lambda}})^{-1}\mathbb{X}^T\widehat{\boldsymbol{W}}\mathbb{X}$, where $\widehat{\boldsymbol{W}}$ is the weight matrix whose diagonal is $X_{ij}\ \widehat{\pi_{ij}}(1-\widehat{\pi_{ij}})$. A joint null hypothesis that evaluates the effects of multiple covariates can be defined in a similar way.

Hereafter we refer the proposed novel method including the region-wide test and the smooth covariate estimation as SOMNiBUS (SmOoth ModeliNg of BisUlfite Sequencing).

3 | METHYLATION DATA FROM AN RA STUDY

To illustrate our method, we report our analysis on data from an RH study (Hudson *et al.*, 2017). Methylation profiles of cell-separated blood samples of 22 rheumatoid arthritis (RA) patients and 21 healthy individuals were measured with custom captured targeted bisulfite sequencing. We focus on one targeted region on chromosome 4 near gene *BANK1*, which is known to show cell-type-specific methylation levels (Hillier *et al.*, 2005). Three additional targeted regions from the same data set are also analyzed in the Supporting Information Section 3. In this *BANK1* region, methylation levels are available at 123 CpG sites. There are 25 samples from circulating T cells and 18 samples from monocytes. We consider two binary covariates—RA status and cell type—and study their impact on methylation pattern in this region.

To fit SOMNiBUS, we specified error parameters $p_0 = 0.003$ and $1 - p_1 = 0.1$; the value 0.003 was reported by Prochenka *et al.* (2015) as insufficient conversion rate and 0.1 was estimated as the average excessive conversion rate in our data using the method proposed by Lakhal-Chaieb *et al.* (2017). We used cubic splines of rank $L_p = 5$ to expand the smooth terms in the model. Figure 1A shows the estimated smooth covariate effects on methylation levels in the targeted *BANK1* region. The panel "Intercept" displays the methylation pattern (on the logit scale) for control samples with the monocyte cell type. The panel "Effect of RA" displays the pattern of methylation difference (on the logit scale) between RA samples and control samples with the same cell type. This figure suggests that RA patients show slightly higher methylation levels in the middle part of the region, compared to controls. The panel "Effect of Tcell" represents the difference of methylation levels (on the logit scale) between T cell samples and monocyte samples with the same disease status. This effect curve, along with the confidence interval bands, clearly shows a highly significant increase of methylation in T cells relative to monocytes. Figure 1B displays the predicted methylation proportions in the four groups of samples, defined by cell type and RA status. Overall, Figure 1 demonstrates the

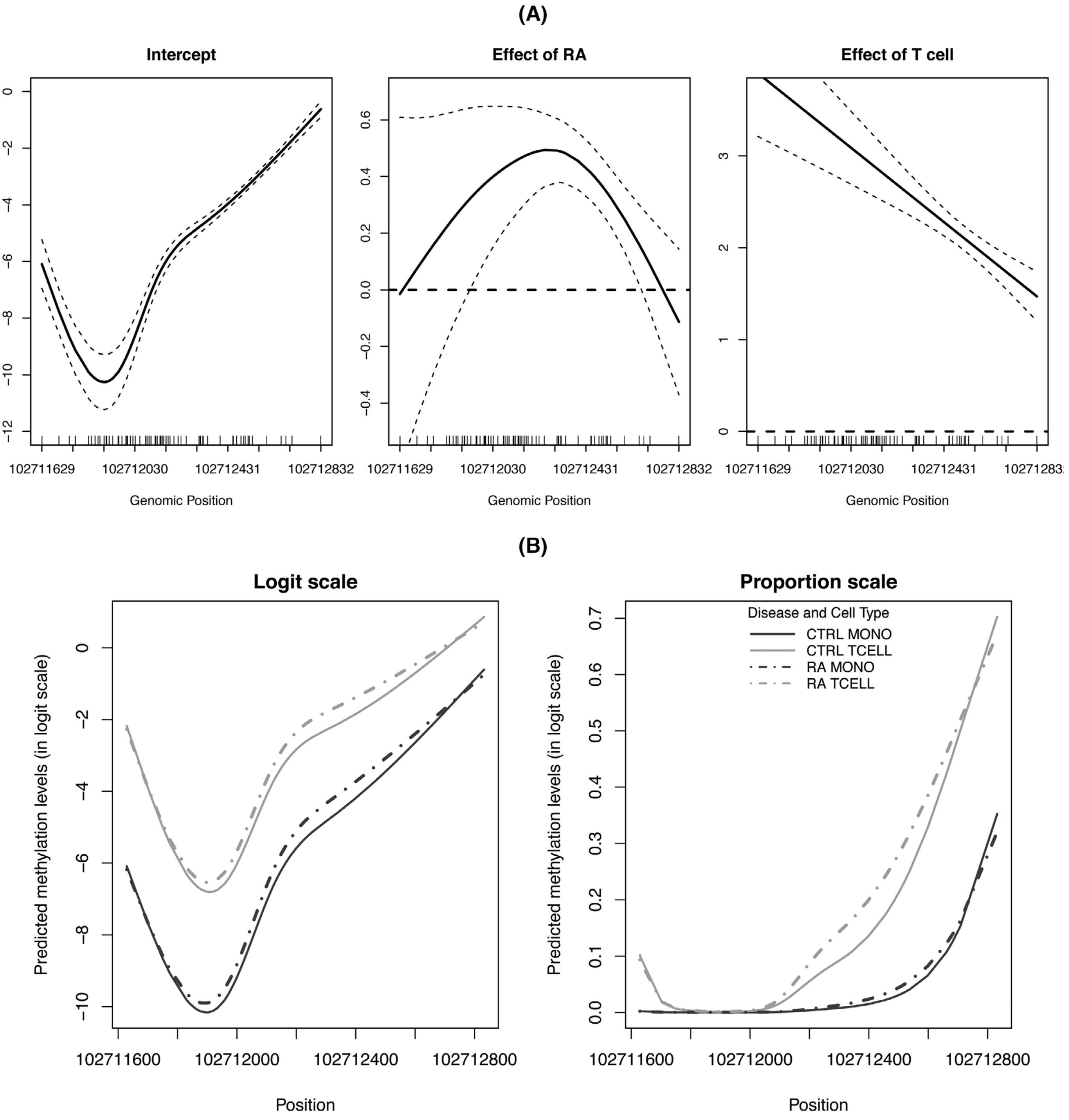

FIGURE 1 (A), The estimates (solid lines) and 95% pointwise confidence intervals (dashed lines) of the intercept, the smooth effect of RA and cell type (T cells versus monocytes) on methylation levels. (B), The predicted methylation levels in the logit scale (left) and proportion scale (right) for the four groups of samples with different disease and cell type status. The region-based *P*-values for the effect of RA status and T cell type are calculated as $1.11E-16$ and $6.37E-218$, respectively

smoothness of the fits, the ability to use multiple covariates simultaneously, and the ease of interpretation of results across the region. Region-wide tests of significance for the two covariates are highly significant (Figure 1). We also applied five alternative methods, described in Section 4; see Table S3 in the Supporting Information.

4 | SIMULATION STUDY

We conducted simulation to (a) demonstrate that the proposed inference of smooth covariate effects is valid, and to (b) compare the performance of our method with five existing methods: `BiSeq` (Hebestreit *et al.*, 2013), `BSmooth`

(Hansen *et al.*, 2012), SMSC (Lakhal-Chaieb *et al.*, 2017), dmrseq (Korthauer *et al.*, 2019) and GlobalTest (Goeman *et al.*, 2006), in terms of type I error and power. The first three methods are typical examples of two-stage analytic approaches. In the first stage, kernel smoothing (local likelihood estimation) is applied to the methylation data of each sample separately. In the second stage, the smoothed methylation data are further analyzed. Specifically, BiSeq calculates the average of Wald statistics from single-site beta regression models, while BSmooth and SMSC calculate the sum of *t*-statistics across loci; these statistics are used to test for differential methylation of a region. In contrast, dmrseq and GlobalTest are one-stage approaches that fit their models directly to the raw methylation proportions in a region. Specifically, dmrseq assesses the strength of the covariate effect using a Wald test statistic within a generalized least square regression model, whereas GlobalTest uses an improved score test in a linear regression model.

Notably, like SOMNiBUS, both GlobalTest and BiSeq are primarily tailored to targeted bisulfite sequencing data with previously identified regions, whereas BSmooth, SMSC and dmrseq are designed for WGBS data. Specifically, BSmooth and SMSC define DMRs at adjacent CpG sites with absolute t-statistics above a defined threshold. The final product from the original software of BSmooth is a list of DMRs that are ranked by the sum of t-statistics; however, BSmooth does not provide region-based *P*-values. To allow comparisons with SOMNiBUS, we estimated the empirical regional *P*-values for BSmooth by permuting the values of the covariate of interest 1000 times. When analyzing WGBS data, dmrseq first constructs candidate regions based on a user-defined cutoff of the smoothed methylation proportion differences, and then fits a generalized least squares regression model with autoregressive error structure to the transformed methylation proportions. Furthermore, the inference inside dmrseq is drawn from permutations—its approximate null distribution is generated by pooling a set of region-level statistics of many candidate regions from all permutations. To better adapt dmrseq to a single targeted region: (i) we used a small cutoff of methylation differences ($1E-5$) for detecting candidate (sub)regions, which ensures fewer CpG sites to be filtered out; (ii) we applied a relatively large number of permutations ($B = 500$) to generate a null distribution of test statistics; (iii) we reported the raw *P*-values without the multiplicity corrections. Note that in some simulations, dmrseq reported more than one DMR in the region. Therefore, for a fairer comparison, we calculated the dmrseq's *P*-value as the minimum over the reported chunks' *P*-values.

Among the five competitive methods, dmrseq, GlobalTest, and BiSeq allow adjustment for multiple covariates. SMSC is the only approach accounting for experimental errors; however, it is conceptually restricted to data from a single cell type.

4.1 | Simulation design

Our simulation design is inspired by the data example described in Section 3. Methylation regions of the same size and with the same CpG distribution as the *BANK1* region were simulated under various settings. We first generated the read depth X_{ij} by resampling with replacement the read depth values from the real data. To specify covariates Z_p and their effect curves $\beta_p(t)$, we then considered the following two scenarios.

Scenario 1 – Multiple covariates

In this case, $P = 3$ binary covariates Z_1, Z_2, and Z_3 were generated independently for each sample. Z_1 and Z_2 were simulated from Bernoulli distributions with proportions 0.51 and 0.58, which were the proportions of RA and T cell samples in the RA data set. The functional parameters for intercept and covariate effects, $\beta_0(t)$, $\beta_1(t)$, and $\beta_2(t)$, were specified to have the same shapes as seen in the *BANK1* region (Figure 1A). Covariate Z_3 was generated from a Bernoulli distribution with proportion parameter 0.5 and had zero effect on methylation, that is, $\beta_3(t) = 0$, for all *t* in the region. The inference results for the effect of the null covariate, Z_3, provide information on type I error.

Scenario 2 – Single covariate

We also considered the case of a single binary covariate ($P = 1$), generated from Bernoulli (0.5), with a variety of regional effect curves. The forms of the functional parameters $\beta_0(t)$ and $\beta_1(t)$ were specified to yield methylation proportion parameters $\pi_0(t)$ and $\pi_1(t)$ as depicted in Figure 2, where $\pi_0(t)$ and $\pi_1(t)$ denote the methylation parameters for samples with $Z = 0$ and $Z = 1$ at position *t*. As shown in Figure 2, these 14 settings of $\pi_0(t)$ correspond to varying levels of closeness between methylation patterns from the two groups.The corresponding values of $\beta_0(t)$ and $\beta_1(t)$ under these 14 settings are shown in the Supporting Information Figure S1. We defined the maximum deviation as the maximum difference between $\pi_1(t)$ and $\pi_0(t)$, for *t* in the section indicated by the dashed lines in Figure 2, where the curves of π_1 and π_0 mainly differ. Simulation scenario 2 is aimed at investigating the power for detecting DMRs at varying levels of maximum derivations.

Given the values of $\{Z_1, \dots Z_P\}$ and $\{\beta_p(t), p = 0, 1, \dots P\}$ under each setting, the true methylation counts S_{ij} were simulated from the model specified in (2). We then generated the observed methylated counts Y_{ij} according to

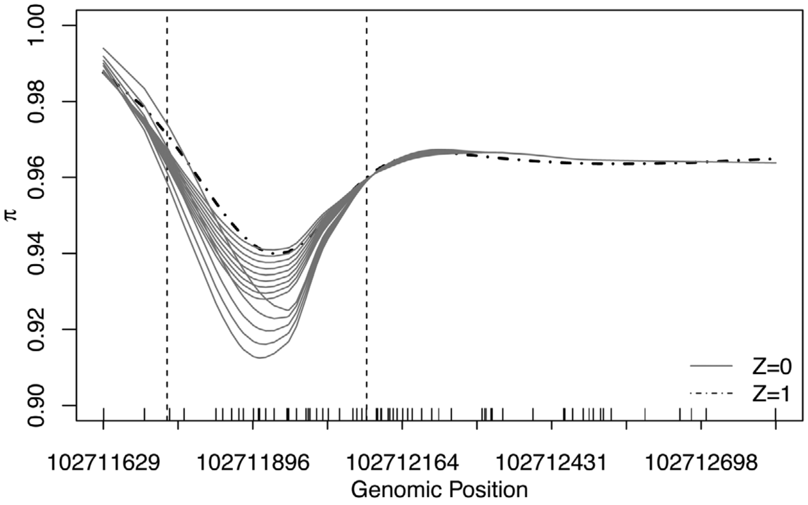

FIGURE 2 The 14 simulation settings of methylation parameters $\pi(t)$ in Scenario 2. Methylation parameters for samples with $Z = 1$ (dotted-dashed black curve) are fixed across settings, whereas the methylation parameters for samples from group $Z = 0$ (solid gray lines) vary across simulations corresponding to different degrees of closeness between methylation patterns in the two groups

Equation (1), which implies

$$Y_{ij} \mid S_{ij} \sim \text{Binomial}(S_{ij}, p_1) + \text{Binomial}(X_{ij} - S_{ij}, p_0).$$

We considered two settings for error parameters p_0 and p_1: (1) $p_0 = 0.003$ and $1 - p_1 = 0.1$, and (2) $p_0 = 1 - p_1 = 0$.

Under each scenario and setting, we generated data sets with sample sizes $N = 40, 100, 150$ and 400, each 1000 times. We then applied SOMNiBUS along with methods BiSeq, dmrseq, BSmooth, SMSC, and GlobalTest to the simulated data sets. Unless otherwise stated, default settings were used for the five alternative methods. For our approach SOMNiBUS, we used cubic splines with dimension $L_p = 5$ to parameterize the smooth terms of interest. We also assumed that the correct values of error parameters p_0 and p_1 were known, although we conducted sensitivity analyses to this assumption (see Discussion and Supporting Information Section 2.2). All simulation parameters are summarized in the Supporting Information Table S1.

4.2 | Simulation results

Figure 3 presents the estimates of the functional parameters $\beta_0(t), \beta_1(t), \beta_2(t)$ and $\beta_3(t)$ over 100 simulations, obtained from SOMNiBUS; here, data were generated under Scenario 1, with sample size $N = 40$ and error parameters $p_0 = 0.003$ and $1 - p_1 = 0.1$. It demonstrates that the proposed method provides unbiased curve estimates for all the four functional parameters in the model, and it can correctly capture both linear and nonlinear smooth covariate effects.

Figure 4 displays the empirical coverage probabilities of CIs over 1000 simulations of Scenario 1. The empirical coverage probabilities are defined as the percentage of simulations where the analytical 95% confidence interval (proposed in Section 2.4.1) covers the true value of the parameter. Overall, the coverage probabilities for $\beta_2(t)$ and $\beta_3(t)$ with linear shapes are closer to the nominal level 95% than the two nonlinear shapes for $\beta_0(t)$ and $\beta_1(t)$. This result can be expected, because nonlinear patterns require more parameters, which leads to less accurate inference results than linear patterns, given the same amount of information. When sample size is 40, the coverages for $\beta_1(t)$ tend to be less than 95%, especially at the boundaries. This may be because $\beta_1(t)$ has a nonlinear shape with relatively small effect sizes across the region, which poses extra difficulties in estimation compared to the shapes that are away from the null, such as $\beta_0(t)$. In summary, Figure 4 shows that the coverages of our 95% confidence intervals attain their nominal values in most of the simulation settings. This suggests that the proposed CI estimation approach quantifies the underlying uncertainty in the smoothed-EM estimates with reasonable accuracy, although it ignores the uncertainty from estimating the smoothing parameters.

Figures 5 and 6 further demonstrate the performance of the proposed regional test, described in Section 2.4.2. The results of type I error rate and power from our smoothed-EM method are compared to the five existing methods GlobalTest, dmrseq, BSmooth, SMSC, and BiSeq. Figure 5 shows the distributions of P-values for the regional effect of the null covariate Z_3, obtained from the six methods. Because none of GlobalTest, dmrseq, BSmooth nor BiSeq accounts for the presence of experimental errors, for a fair comparison, the simulated data used in Figure 5 were generated without error (ie, $p_0 = 1 - p_1 = 0$). The corresponding results for data generated with error are shown in the Supporting Information Figure S2. Figure 5 shows that the region-based P-values for Z_3, calculated from our smoothed-EM approach (black dots), are uniformly distributed, under all sample sizes considered. In contrast, the distributions of P-values from dmrseq, BiSeq, and

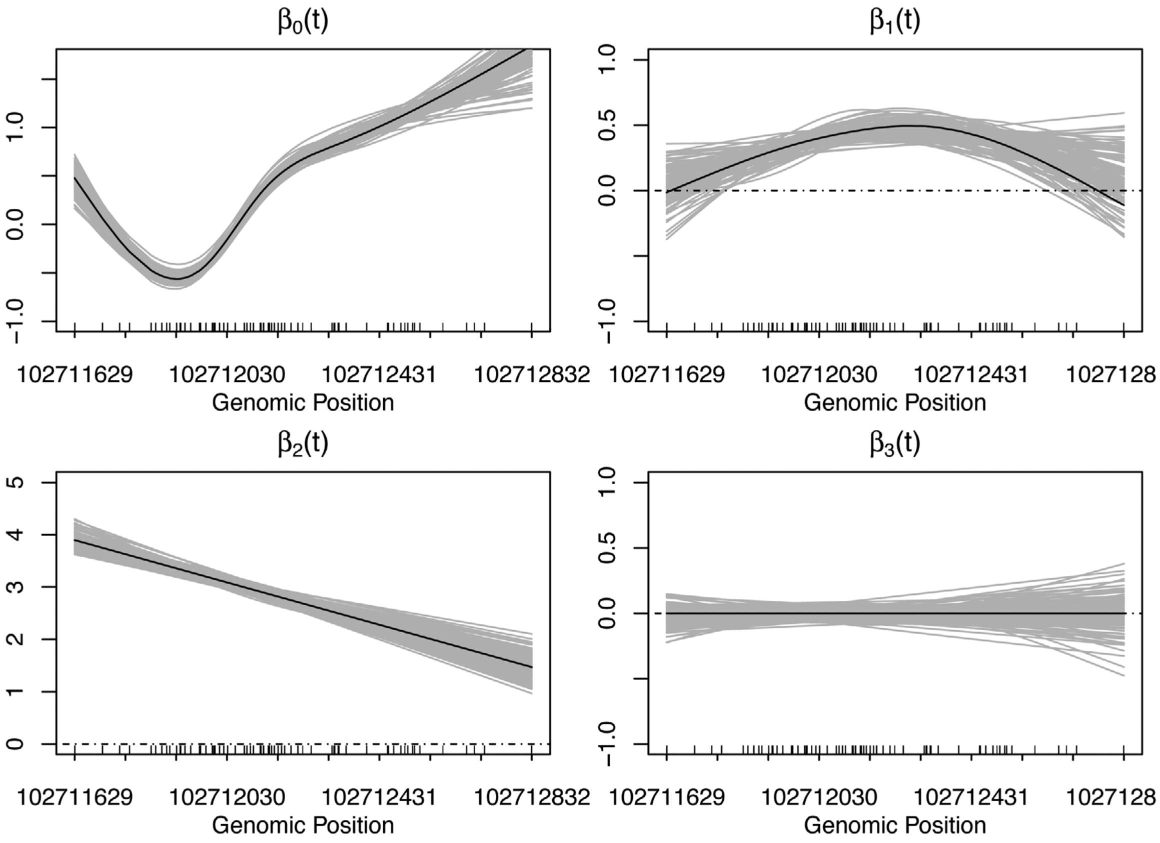

FIGURE 3 Estimates of smooth covariate effects (gray) over the 100 simulations in Scenario 1, using `SOMNiBUS`. The black curves are the true functional parameters used to generate the data. Data with sample size $N = 40$ were generated with error

`GlobalTest` are biased away from what would be expected under the null. Because the inferences for `BSmooth` and `SMSC` are drawn from permutations, both methods are able to control type I error. Similar results were observed when data were generated with error. The results demonstrate that the distribution of the SOMNiBUS region-based statistics under the null is well calibrated even at a relatively small sample size $N = 40$, indicating the proposed regional zero effect test can correctly control the type I error. Figure 6 shows the powers of the six methods for detecting DMRs under the 14 settings of methylation patterns displayed in Figure 2. In Figure 6, the left panel presents the results obtained from data with error ($p_0 = 0.003$ and $1 - p_1 = 0.1$); the right panel presents results obtained from data without error ($p_0 = 1 - p_1 = 0$). Figure 6 shows that the proposed smoothed-EM method has a higher power than the five alternative methods; this superiority is even more pronounced when the data were generated with error.

In summary, `SOMNiBUS` provides accurate estimates for smooth covariate effects; when compared with the existing methods considered here, `SOMNiBUS` exhibits greater power to detect DMRs, while correctly controlling type I error rates.

5 | DISCUSSION

Currently, there are no tools for estimating smooth covariate effects for bisulfite sequencing data. In this paper, we propose and evaluate a method, `SOMNiBUS`, that aims to fill this gap. Our contribution is threefold. First, we develop a novel model to represent the bisulfite sequencing data from multiple samples, which naturally accounts for variable read depth, experimental errors, and a mixture of cell types. Second, we provide a formal inference for smooth covariate effects across a region of interest, where outcomes may be contaminated by errors. Third, we construct a region-based statistic with a simple chi-squared limiting distribution for jointly testing multiple coefficients in the presence of penalization. Results from simulations and one real data example show that the new method captures important underlying methylation patterns, provides accurate estimates of covariate effects, and correctly quantifies the underlying uncertainty in the estimates. The method has been implemented in R package `SOMNiBUS`, which will be submitted to CRAN.

Our method assumes that the error parameters p_0 and p_1 are known and do not vary across the region of interest. Although it is conceptually feasible to estimate

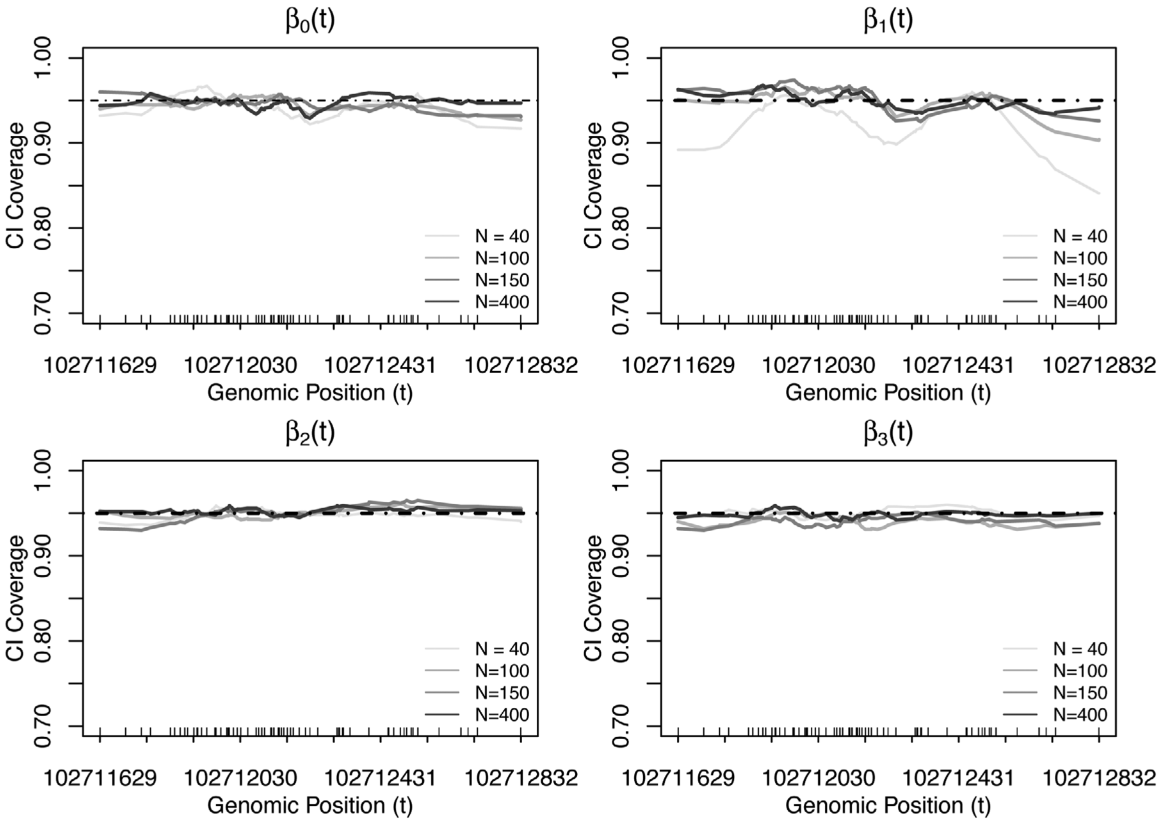

FIGURE 4 Coverage probability of confidence intervals over 1000 simulations under different sample sizes ($N = 40, 100, 150, 400$). Data were generated with error, under simulation Scenario 1

these parameters by an EM-type approach, the added computational burden in the E step would be substantial, because the complete-data likelihood is not linear in the methylated counts. Moreover, there are cases in which these parameters can actually be measured, for example by adding spike-in sequences of DNA that are known in advance to be methylated or unmethylated into the bisulfite sequencing procedure. The results from the sensitivity analyses (Supplementary Information Figures S3 and S4) show that misspecified error rates can introduce a minor bias in regional *P*-values; however, this is not likely to affect the power of our tests, as demonstrated in the Supporting Information Table S2. An extension worth exploring in the future will be to accommodate variations of p_0 and p_1 across genomic positions into our model. For example, the error rates could be modeled to depend on prior annotation information, CG content, or on the experimental quality in the test region.

Another potential limitation of our inference procedures is the treatment of the smoothing parameters as fixed, disregarding the uncertainty in estimating them. However, our simulation results show that both the confidence interval coverage at each site and the type I error rates at the region level, are close to their nominal value; hence, our compromise does not lead to a major efficiency loss. Nevertheless, this uncertainty could be accounted for by adding in our method an approximate correction, as proposed by Kass and Steffey (1989), or considering a full Bayesian inference where one could specify a prior distribution for the smoothing parameters $\boldsymbol{\lambda}$.

There is a substantial computational burden in our estimation algorithm, because the M step includes two inner iteration schemes: P-IRLS for updating smooth covariate effects, and Newton's optimization for updating smoothing parameters. A summary of runtimes for `SOMNiBUS` and the five alternative methods is displayed in the Supporting Information Figure S5. This figure shows that `SOMNiBUS` requires longer computational times than `GlobalTest`, `BSmooth`, `SMSC` and `BiSeq`, but less than `dmrseq`. Note that our proposed method, `SOMNiBUS`, is capable of estimating the effects of multiple covariates simultaneously, whereas, other methods require repeating the analysis for each covariate, which will multiply the runtimes. Our algorithm could be sped up by transforming the methylation proportions into a continuous-type variable, as in Korthauer *et al.* (2019), which allows us to replace the P-IRLS with the ordinary least square, and mitigate any instability in estimation of

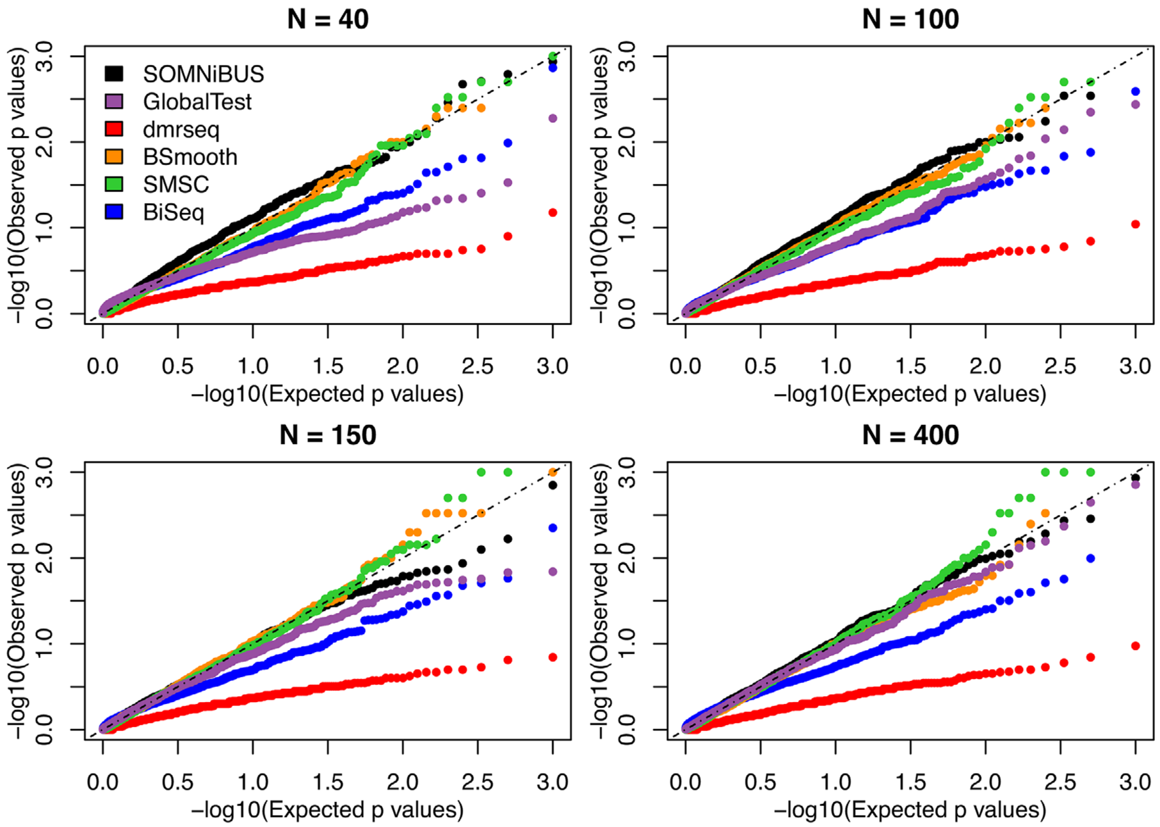

FIGURE 5 Quantile-Quantile (Q-Q) plots of the region-based P-values for the null covariate Z_3, obtained from the six methods, over 1000 simulations. Data were generated without error with a range of sample sizes ($N = 40, 100, 150, 400$), under simulation Scenario 1. Here, the Expected P-values are uniformly distributed numbers, equal to $= (1/1001, 2/1001, \ldots, 1000/1001)$.

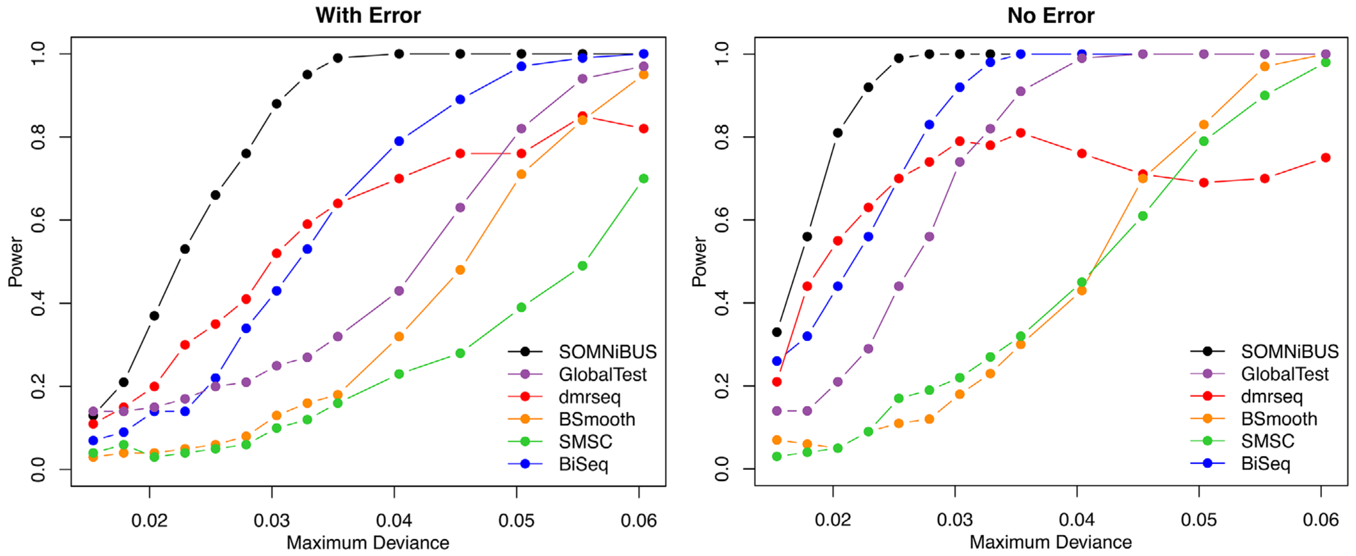

FIGURE 6 Powers to detect DMRs using the six methods for the 14 simulation settings in Scenario 2 under different levels of maximum deviation between $\pi_0(t)$ and $\pi_1(t)$, calculated over 100 simulations. (Sample size $N = 100$).

methylation levels near the boundaries (proportions of zero or one). However, transforming the count outcome into a continuous variable causes extra difficulties in the Expectation step, for which no closed-form exact expression is available.

The proposed approach is tailored to targeted bisulfite sequencing data. Another future direction is to extend our method to WGBS data. This requires first partitioning whole genome into regions or using a sliding window; optimal partitioning or choices of window sizes are challenges to be met. We recommend for the moment that algorithms such as `BSmooth` or `dmrseq` be used to find interesting regions. These regions could then be re-analyzed with `SOMNiBUS` to more comprehensively and simultaneously estimate covariate influences on methylation.

ACKNOWLEDGMENTS

This work was supported by a Genome Canada Bioinformatics and Computational Biology 2017 (B/CB) competition grant and the Canadian Institutes of Health Research MOP 130344. KZ was supported by a Doctoral Training Award from the Fonds de Recherche du Québec - Santé (FRQS), the McGill University Faculty of Medicine's Gerald Clavet Fellowship, as well as a scholarship from Canadian Statistical Sciences Institute - Collaborative Research Team Projects. KO acknowledges the Natural Sciences and Engineering Research Council of Canada and Fonds de recherch du Québec-santé grant FRQS-31110. We also acknowledge the Compute Canada Resources for Research Groups (RRG) ID 2514. We would also like to thank the editor, the associate editor, and the referee for their constructive comments that helped improve this article.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are available on request from the co-author Dr. Marie Hudson. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Kaiqiong Zhao https://orcid.org/0000-0003-0810-8764
Celia M.T. Greenwood https://orcid.org/0000-0002-2427-5696

REFERENCES

Allum, F., Shao, X., Guénard, F., Simon, M.-M., Busche, S., Caron, M., Lambourne, J., Lessard, J., Tandre, K., Hedman, A.K., Kwan, T., Ge, B., Rönnblom, L., McCarthy, M.I., Deloukas, P., Richmond, T., Burgess, D., Spector, T.D., Tchernof, A., Marceau, S., Lathrop, M., Vohl, M.-C., Pastinen, T., Grundberg, E. and Multiple Tissue Human Expression Resource Consortium, (2015) Characterization of functional methylomes by next-generation capture sequencing identifies novel disease-associated variants. *Nature Communications*, 6(1), 1–12.

Cheng, L. and Zhu, Y. (2013) A classification approach for DNA methylation profiling with bisulfite next-generation sequencing data. *Bioinformatics*, 30(2), 172–179.

De Jager, P.L., Srivastava, G., Lunnon, K., Burgess, J., Schalkwyk, L.C., Yu, L., Eaton, M.L., Keenan, B.T., Ernst, J., McCabe, C., Tang, A., Raj, T., Replogle, J., Brodeur, W., Gabriel, S., Chai, H.S., Younkin, C., Younkin, S.G., Zou, F., Szyf, M., Epstein, C.B., Schneider, J.A., Bernstein, B.E., Meissner, A., Ertekin-Taner, N., Chibnik, L.B., Kellis, M., Mill, J. and Bennett, D.A. (2014) Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nature Neuroscience*, 17(9), 1156–1163.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–22.

Feinberg, A.P. (2007) Phenotypic plasticity and the epigenetics of human disease. *Nature*, 447(7143), 433–440.

Frommer, M., McDonald, L.E., Millar, D.S., Collis, C.M., Watt, F., Grigg, G.W., Molloy, P.L. and Paul, C.L. (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences*, 89(5), 1827–1831.

Goeman, J.J., Van De Geer, S.A. and Van Houwelingen, H.C. (2006) Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3), 477–493.

Hansen, K.D., Langmead, B. and Irizarry, R.A. (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, 13(10), R83.

Hebestreit, K., Dugas, M. and Klein, H.-U. (2013) Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics*, 29(13), 1647–1653.

Hillier, L.W., Graves, T.A., Fulton, R.S., Fulton, L.A., Pepin, K.H., Minx, P., et al. (2005) Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature*, 434(7034), 724–731.

Horvath, S. (2013) DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10), 3156.

Hudson, M., Bernatsky, S., Colmegna, I., Lora, M., Pastinen, T., Klein Oros, K. and Greenwood, C.M.T. (2017) Novel insights into systemic autoimmune rheumatic diseases using shared molecular signatures and an integrative analysis. *Epigenetics*, 12(6), 433–440.

Jaenisch, R. and Bird, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33(3), 245–254.

Kass, R.E. and Steffey, D. (1989) Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association*, 84(407), 717–726.

Korthauer, K., Chakraborty, S., Benjamini, Y. and Irizarry, R.A. (2019) Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. *Biostatistics*, 20(3), 367–383.

Lakhal-Chaieb, L., Greenwood, C.M., Ouhourane, M., Zhao, K., Abdous, B. and Oualkacha, K. (2017) A smoothed EM-algorithm for DNA methylation profiles from sequencing-based methods in cell lines or for a single cell type. *Statistical Applications in Genetics and Molecular Biology*, 16(5-6), 313–331.

Lea, A.J., Tung, J. and Zhou, X. (2015) A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. *PLoS Genetics*, 11(11), e1005650.

Li, Q., Suzuki, M., Wendt, J., Patterson, N., Eichten, S.R., Hermanson, P.J., Green, D., Jeddeloh, J., Richmond, T., Rosenbaum, H., Burgess, D., Springer, N.M. and Greally, J.M. (2015) Post-conversion targeted capture of modified cytosines in mammalian and plant genomes. *Nucleic Acids Research*, 43(12), e81–e81.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Routti, V., Millar, A.H., Thomson, A.J., Ren, R. and Ecker, J.R. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271), 315–322.

Liu, Y., Aryee, M.J., Padyukov, L., Fallin, M.D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., Shchetynsky, K., Scheynius, A., Kere, J., Alfredsson, L., Klareskog, L., Ekström, T.J. and Feinberg, A.P. (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology*, 31(2), 142.

Louis, T.A. (1982) Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2), 226–233.

Maher, B. (2008) Personal genomes: the case of the missing heritability. *Nature News*, 456(7218), 18–22.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel, M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S.R., Kaul, R. and Stamatoyannopoulos, J.A. (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099), 1190–1195.

Nilsson, E., Jansson, P.A., Perfilyev, A., Volkov, P., Pedersen, M., Svensson, M.K., Poulsen, P., Ribel-Madsen, R., Pedersen, N.L., Almgren, P., Fadista, J., Rönn, T., Pedersen, B.K., Scheele, C., Vaag, A. and Ling, C. (2014) Altered DNA methylation and differential expression of genes influencing metabolism and inflammation in adipose tissue from subjects with type 2 diabetes. *Diabetes*, 63(9), 2962–2976.

Oakes, D. (1999) Direct calculation of the information matrix via the EM. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2), 479–482.

Ober, C. and Vercelli, D. (2011) Gene–environment interactions in human disease: nuisance or opportunity? *Trends in Genetics*, 27(3), 107–115.

Parker, R. and Rice, J. (1985) Discussion on "some aspects of the spline smoothing approach to non-parametric regression curve fitting" (by BW Silverman). *Journal of the Royal Statistical Society. Series B (Methodological)*, 47, 40–42.

Prochenka, A., Pokarowski, P., Gasperowicz, P., Kosińska, J., Stawiński, P., Zbieć-Piekarska, R., et al. (2015) A cautionary note on using binary calls for analysis of DNA methylation. *Bioinformatics*, 31(9), 1519–1520.

Rackham, O.J., Langley, S.R., Oates, T., Vradi, E., Harmston, N., Srivastava, P.K., Behmoaras, J., Dellaportas, P., Bottolo, L. and Petretto, E. (2017) A Bayesian approach for analysis of whole-genome bisulphite sequencing data identifies disease-associated changes in DNA methylation. *Genetics*, 205(4), 1443–1458.

Shafi, A., Mitrea, C., Nguyen, T. and Draghici, S. (2017) A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Briefings in Bioinformatics*, 19(5), 737–753.

Shokoohi, F., Stephens, D.A., Bourque, G., Pastinen, T., Greenwood, C.M. and Labbe, A. (2019) A hidden Markov model for identifying differentially methylated sites in bisulfite sequencing data. *Biometrics*, 75(1), 210–221.

Stephens, D.A., Shokoohi, F. and Aurélie, L. (2016) Hidden Markov models for identifying differentially methylated regions. 44th Annual Meeting of the Statistical Society of Canada.

Sun, S. and Yu, X. (2016) HMM-Fisher: identifying differential methylation using a hidden Markov model and Fisher's exact test. *Statistical Applications in Genetics and Molecular Biology*, 15(1), 55–67.

Teschendorff, A.E., Yang, Z., Wong, A., Pipinikas, C.P., Jiao, Y., Jones, A., Anjum, S., Hardy, R., Salvesen, H.B., Thirlwell, C., Janes, S.M., Kuh, D. and Widschwendter, M. (2015) Correlation of smoking-associated DNA methylation changes in buccal cells with DNA methylation changes in epithelial cancer. *JAMA Oncology*, 1(4), 476–485.

Wahba, G. (1980) Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. In: Cheney, W. (Ed.) *Approximation Theory III*. New York, NY: Academic Press, pp. 905–912.

Weissbrod, O., Rahmani, E., Schweiger, R., Rosset, S. and Halperin, E. (2017) Association testing of bisulfite-sequencing methylation data via a Laplace approximation. *Bioinformatics*, 33(14), i325–i332.

Wood, S.N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1), 3–36.

Wood, S.N. (2013) On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1), 221–228.

Wood, S.N. and Fasiolo, M. (2017) A generalized Fellner-Schall method for smoothing parameter optimization with application to tweedie location, scale and shape models. *Biometrics*, 73(4), 1071–1081.

Wood, S.N., Pya, N. and Säfken, B. (2016) Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516), 1548–1563.

Wreczycka, K., Gosdschan, A., Yusuf, D., Gruening, B., Assenov, Y. and Akalin, A. (2017) Strategies for analyzing bisulfite sequencing data. *Journal of Biotechnology*, 261, 105–115.

Yu, X. and Sun, S. (2016a) Comparing five statistical methods of differential methylation identification using bisulfite sequencing data. *Statistical Applications in Genetics and Molecular Biology*, 15(2), 173–191.

Yu, X. and Sun, S. (2016b) HMM-DM: identifying differentially methylated regions using a hidden Markov model. *Statistical Applications in Genetics and Molecular Biology*, 15(1), 69–81.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures, referenced in Section 2, 3, 4 and 5, are available with this paper at the Biometrics website on Wiley Online Library. Codes to replicate the simulation results in the article are deposited in the Github repository https://github.com/kaiqiong/SOMNiBUS_Simu. The R package, SOMNiBUS, implementing the proposed method is available from Github at https://github.com/GreenwoodLab/SOMNiBUS, with a user guide.

How to cite this article: Zhao K, Oualkacha K, Lakhal-Chaieb L et al. A novel statistical method for modeling covariate effects in bisulfite sequencing derived measures of DNA methylation. *Biometrics*. 2021;77:424–438. https://doi.org/10.1111/biom.13307