

Children's Mercy Kansas City

## SHARE @ Children's Mercy

---

Manuscripts, Articles, Book Chapters and Other Papers

---

9-10-2024

### Addressing dispersion in mis-measured multivariate binomial outcomes: A novel statistical approach for detecting differentially methylated regions in bisulfite sequencing data.

Kaiqiong Zhao

Karim Oualkacha

Yixiao Zeng

Cathy Shen

Kathleen Klein

*See next page for additional authors*

Let us know how access to this publication benefits you

Follow this and additional works at: <https://scholarlyexchange.childrensmercy.org/papers>

---

#### Recommended Citation

Zhao K, Oualkacha K, Zeng Y, et al. Addressing dispersion in mis-measured multivariate binomial outcomes: A novel statistical approach for detecting differentially methylated regions in bisulfite sequencing data. *Stat Med.* 2024;43(20):3899-3920. doi:10.1002/sim.10149




This Article is brought to you for free and open access by SHARE @ Children's Mercy. It has been accepted for inclusion in Manuscripts, Articles, Book Chapters and Other Papers by an authorized administrator of SHARE @ Children's Mercy. For more information, please contact [hlsteel@cmh.edu](mailto:hlsteel@cmh.edu).

---

**Creator(s)**

Kaiqiong Zhao, Karim Oualkacha, Yixiao Zeng, Cathy Shen, Kathleen Klein, Lajmi Lakhal-Chaieb, Aurélie Labbe, Tomi Pastinen, Marie Hudson, Inés Colmegna, Sasha Bernatsky, and Celia M T Greenwood

# Addressing dispersion in mis-measured multivariate binomial outcomes: A novel statistical approach for detecting differentially methylated regions in bisulfite sequencing data

Kaiqiong Zhao<sup>1</sup>  | Karim Oualkacha<sup>2</sup>  | Yixiao Zeng<sup>3</sup> | Cathy Shen<sup>3</sup> | Kathleen Klein<sup>3</sup> | Lajmi Lakhal-Chaieb<sup>4</sup> | Aurélie Labbe<sup>5</sup> | Tomi Pastinen<sup>6</sup> | Marie Hudson<sup>3,7</sup> | Inés Colmegna<sup>7,8</sup> | Sasha Bernatsky<sup>7,8</sup> | Celia M. T. Greenwood<sup>3,9,10</sup> 

## Correspondence

Celia M. T. Greenwood, Lady Davis  
Institute for Medical Research, Jewish  
General Hospital, Montreal, QC, Canada.  
Email: [celia.greenwood@mcgill.ca](mailto:celia.greenwood@mcgill.ca)  
Kaiqiong Zhao, Department of  
Mathematics and Statistics, York  
University, Toronto, Ontario, Canada.  
Email: [kaiqiong@yorku.ca](mailto:kaiqiong@yorku.ca)

## Funding information

Canadian Institutes of Health Research,  
Grant/Award Number: MOP 130344;  
Digital Research Alliance of Canada,  
Grant/Award Number: 2541 4128;  
Genome Canada, Grant/Award Number:  
2017 (B/CB); Natural Sciences and  
Engineering Research Council of Canada,  
Grant/Award Number:  
RGPIN-2024-06287

Motivated by a DNA methylation application, this article addresses the problem of fitting and inferring a multivariate binomial regression model for outcomes that are contaminated by errors and exhibit extra-parametric variations, also known as dispersion. While dispersion in univariate binomial regression has been extensively studied, addressing dispersion in the context of multivariate outcomes remains a complex and relatively unexplored task. The complexity arises from a noteworthy data characteristic observed in our motivating dataset: non-constant yet correlated dispersion across outcomes. To address this challenge and account for possible measurement error, we propose a novel hierarchical quasi-binomial varying coefficient mixed model, which enables flexible dispersion patterns through a combination of additive and multiplicative dispersion components. To maximize the Laplace-approximated quasi-likelihood of our model, we further develop a specialized two-stage expectation-maximization (EM) algorithm, where a plug-in estimate for the multiplicative scale parameter enhances the speed and stability of the EM iterations. Simulations demonstrated that our approach yields accurate inference for smooth covariate effects and exhibits excellent power in detecting non-zero effects. Additionally, we applied our proposed method to investigate the association between DNA methylation, measured across the genome through targeted custom capture sequencing of whole blood, and levels of anti-citrullinated protein antibodies (ACPA), a preclinical marker for rheumatoid arthritis (RA) risk. Our analysis revealed 23 significant genes that potentially contribute to ACPA-related differential methylation, highlighting the relevance of cell signaling and collagen metabolism in RA. We implemented our method in the R Bioconductor package called “SOMNiBUS.”

## KEYWORDS

additive dispersion, binomial, DNA methylation, EM algorithm, measurement error, multiplicative dispersion

For affiliations refer to page 3918.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Author(s). *Statistics in Medicine* published by John Wiley & Sons Ltd.

## 1 | INTRODUCTION

This article addresses the challenge in fitting and inferring a multivariate binomial regression model for outcomes contaminated by errors and exhibiting extra-parametric variations, commonly referred to as dispersion. The primary motivation behind this work is to optimize the analysis and interpretation of high-resolution large-scale DNA methylation measures generated from the state-of-the-art bisulfite sequencing (BS-seq) protocol. DNA methylation involves the addition of a methyl group to the DNA, mostly at cytosine-phosphate-guanine (CpG) sites.<sup>1</sup> The raw data obtained from BS-seq are short sequence reads. After proper alignment and data processing, the methylation level at a single site can be summarized as a pair of binomial counts: the number of methylated reads and the total number of reads covering the site, that is, read depth. Such data possess several challenges for statistical analysis. Typically, read depth varies drastically across sites and individuals, which leads to measures with wide-ranging precision and many missing values.<sup>2</sup> Additional statistical challenges are created by the possibility of data errors, arising from excessive or insufficient bisulfite treatment or other aspects of the sequencing processes.<sup>3,4</sup> To address these challenges, it is critical to develop statistical methods that are specifically tailored to the unique structure of BS-seq data and enable accurate inference for the association patterns between DNA methylation, represented as *mis-measured binomial outcomes*, and a specific disease trait of interest.

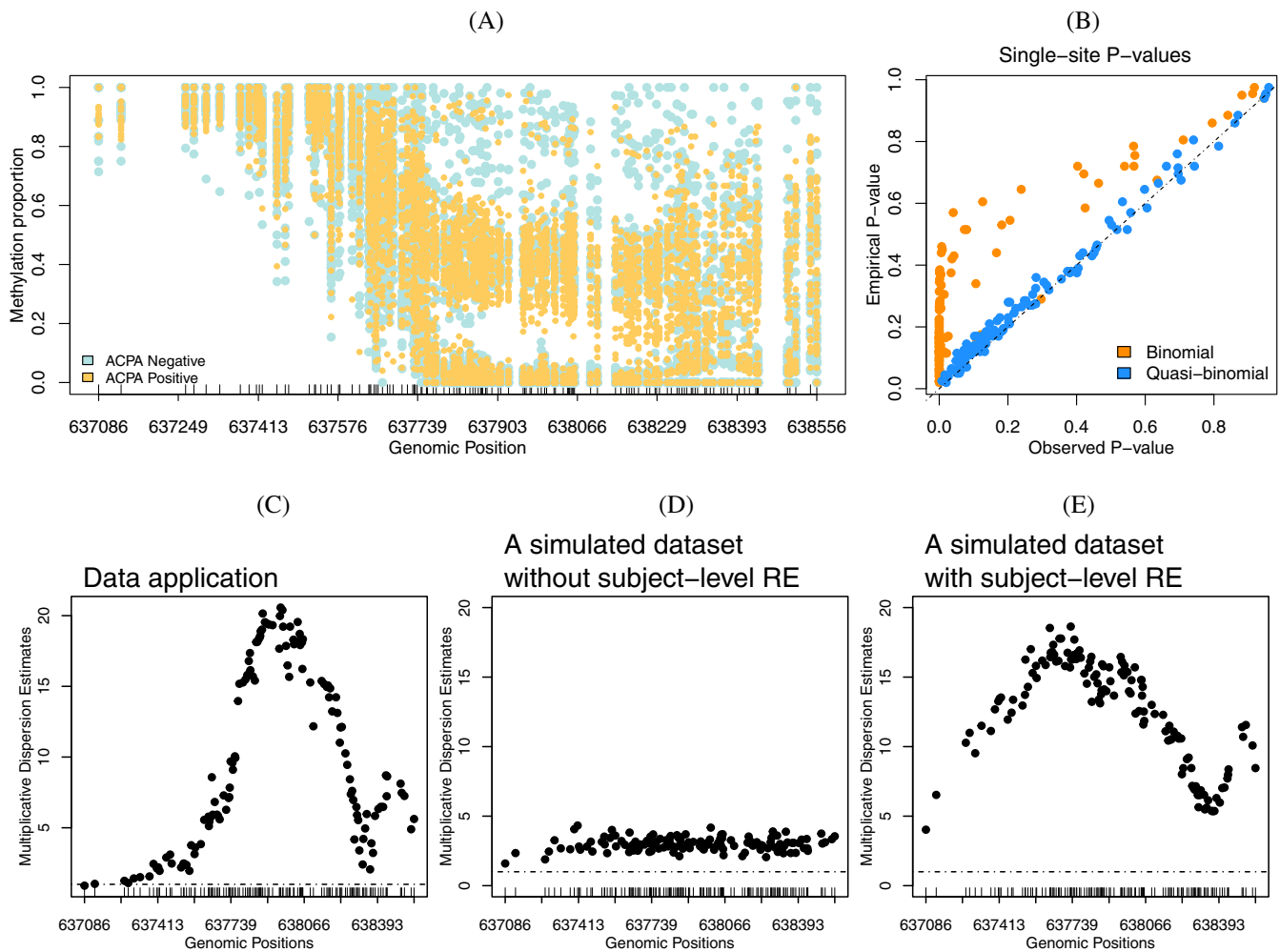
### 1.1 | Motivating dataset

In our motivating study, we aim to investigate the association between DNA methylation and the levels of anti-citrullinated protein antibodies (ACPA), a marker of rheumatoid arthritis (RA) risk that frequently appears prior to any clinical manifestations,<sup>5</sup> using asymptomatic samples drawn from the CARTaGENE cohort. Targeted-region sequencing captured the DNA methylation levels at approximately 5 million CpG sites in whole blood samples from 48 ACPA-positive and 54 ACPA-negative individuals. Using this data, previous studies<sup>6,7</sup> have examined ACPA-related methylation changes at individual CpG sites. However, region-based association studies, which deal with *multivariate binomial outcomes*, have not been explored. The motivation of region-based analyses is multi-fold. First, various studies have shown that methylation levels are strongly correlated across the genome.<sup>8,9</sup> Joint modeling of regional methylation levels allows us to borrow information from this local correlation structure, thus coping naturally with missing values or low counts, of which univariate analyses are incapable. Furthermore, many functionally relevant methylation changes have been found in genomic regions rather than individual CpGs, such as CpG islands<sup>10</sup> or genomic blocks.<sup>11</sup> These synergistic changes in methylation across a region often convey more substantial regulatory influence.<sup>12</sup> In addition, the resulting differentially methylated regions (DMR) can be subsequently explored and annotated easily by examining their overlap with other known genomic features to provide context and perspective of the potential methylation events, which helps improve the interpretability and reproducibility of the analytical results.

### 1.2 | Motivation for addressing dispersion

To detect truly differentially methylated regions without finding false associations, it is crucial to accurately account for the sources of variability across individuals. Figure 1A illustrates observed methylation proportions in a targeted region for our samples. In panel B, it can be seen that  $P$ -values testing for methylation differences, assuming a binomial mean-variance relationship, are much too small. In contrast, allowing for dispersion through a quasi-binomial model provides  $P$ -values in line with null expectation. As such, the restrictive mean-variance relationship implied by a binomial model may not adequately accommodate the data variability, thus leading to inflation of false positives.

In the context of modeling mis-measured multivariate binomial outcomes in the analysis of BS-seq data (for purified DNA samples), we have developed a method called SOMNiBUS.<sup>13</sup> SOMNiBUS utilizes a hierarchical binomial regression model and effectively addresses various challenges in BS-seq analysis, including regional testing, estimation of multiple covariate effects, adjustment for read depth variability and handling of experimental errors. Nevertheless, it is important to note that its underlying binomial assumption may be overly restrictive and is only applicable when data exhibit variability levels that are similar to those anticipated based on a binomial distribution, such as purified DNA samples from inbred animal or cell line experiments.



**FIGURE 1** Illustration of observed dispersion in a targeted region that underwent bisulfite sequencing. (A) Observed methylation proportions in one region for two groups of samples (yellow and blue); data are fully described in Section 4. (B) Single-site  $P$ -values for methylation difference between the two groups. Horizontal axis are the  $P$ -values estimated from either binomial (ignoring dispersion) or quasi-binomial (accounting for dispersion) GLMs. Vertical axis shows the empirical  $P$ -values computed from 199 permutations; the empirical  $P$ -value is a benchmark for valid statistical tests. The lower panels show estimated dispersion for each CpG site using a single-site quasi-binomial GLM, for (C) the methylation data illustrated before, and (D, E) two simulated regional methylation datasets. Specifically, data in (D) were simulated from a multiplicative-dispersion-only model ( $\phi = 3, \sigma_0^2 = 0$ ), and (E) from a model with both a multiplicative dispersion and a subject-level RE ( $\phi = 3, \sigma_0^2 = 3$ ); see Section 2 for detailed model formulations and notation definitions relating to panels (D) and (E).

In this work, we explicitly address dispersion in the modeling of mis-measured multivariate binomial outcomes. Our approach naturally accommodates more complex biological samples in methylation applications, including human samples or samples with mixed cell types. These samples often exhibit variability that deviates from a binomial distribution. To address this, we propose a novel hierarchical quasi-binomial varying coefficient mixed model. Additionally, we develop a two-stage quasi-likelihood-based expectation-maximization (EM) algorithm and provide a computationally simple method for estimating the variance of the varying coefficient estimates.

### 1.3 | Literature review

The importance of accounting for dispersion in BS-seq data has been well recognized in analysis of single CpG sites. Existing single site approaches use either *additive* overdispersion models, or *multiplicative* under- or overdispersion models to describe the variation driving the dispersion. In a multiplicative model, one includes a multiplicative scale factor,

that is, the dispersion parameter, in the variance of response. Thus, the dispersion inflates or deflates the variance estimates of the covariate effect by the multiplicative factor. Such approaches include the quasi-binomial regression model<sup>14</sup> and the beta-binomial regression model.<sup>15-17</sup> In contrast, additive methods add a subject-level random effect (RE) to capture the extra-binomial variation among individual methylation proportions, which are derived from counts of methylated and unmethylated reads. Both ABBA<sup>12</sup> and MACAU,<sup>18</sup> that use binomial mixed effect models, fall in this category. An advantage of the multiplicative approach, particularly the quasi-binomial model, is that it naturally allows for both overdispersion and underdispersion, whereas the additive model only allows overdispersion. On the other hand, the additive overdispersion approach links directly with a multilevel model and can be readily extended to analyze data with a hierarchical or clustering structure among the samples.

The challenge of accounting for dispersion when analyzing regional methylation data (ie, multivariate outcomes) is further complicated by several factors. First, even within a small genomic region, different CpG sites may exhibit different levels of dispersion and strong spatial correlation (Figure 1C). Hence, a multiplicative dispersion model with a common dispersion parameter does not adequately capture the dispersion heterogeneity across loci (Figure 1D). In addition, challenges are presented by the complex correlation structure in the regional methylation data. Apart from the spatial correlations among neighboring CpGs, there may be additional correlations among methylation measurements on the same subject. Ignoring this within-subject correlation could lead to overestimation of precision and invalid statistical tests.<sup>19</sup> One means to accommodate such a correlation structure is to add a subject-level RE. However, extra random dispersion can arise, beyond that introduced by the subject-level RE,<sup>20-22</sup> and thus, often, parametric distributions with restrictive mean-variance relations poorly describe the outcomes for individual subjects.<sup>23-25</sup> Hence, properly addressing *both* multiplicative and additive sources of dispersion in methylation data is essential for making reliable inference at the region level.

## 1.4 | Overview of the proposed approach

To overcome the limitations and challenges of existing methods (see more details in Supplementary Table S1 and Supplementary Section 1), we propose a novel approach for identifying DMRs, dSOMNiBUS (dispersion-adjusted SmOoth ModeliNg of BisUlfitE Sequencing). Our strategy explicitly accounts for all (known) sources of data variability and effectively addresses the varying degrees of dispersion across loci, thus providing accurate assessments of regional statistical significance.

Specifically, we assume that the observed methylation counts arise from an unobserved latent true methylation state compounded by errors. These true methylation counts are then described by a *quasi-binomial varying coefficient mixed model*. Such a flexible model does not require exact information about the outcome distribution but only specifies the mean and variance for the conditional distribution of the outcome given covariates and subject-specific REs. For simplicity, we assume this variance depends on the mean and a multiplicative dispersion parameter (MDP). The *combination* of subject-specific REs (ie, additive overdispersion) and multiplicative dispersion enables flexible dispersion patterns in a region (Figure 1E), which is highly plausible in methylation data (Figure 1C). In addition, this formulation entails both subject-specific (ie, conditional) and population-averaged (ie, marginal) interpretations for the varying regression coefficients in the model.

Estimating our complex model is quite challenging due to the interplay of multiple factors. First, the true outcomes are unobserved latent variables, which highlights the need to devise an EM-type algorithm<sup>26</sup> to iteratively integrate the complete data log-likelihood function over the distribution of the latent variables (E step) and compute the model parameters that maximize the integrated likelihood function (M step). However, the full likelihood function for the complete data is unavailable, requiring the development of a quasi-likelihood-based EM algorithm. In our model, the accurate definition of the quasi-likelihood (QL) function requires integrating out the subject-specific REs and the REs for smoothness regularization<sup>27</sup>; the latter helps avoid overfitting the functional coefficients. This integral, however, cannot be evaluated exactly.<sup>20,28,29</sup> A remedy is an QL analog of the Laplace approximation.<sup>30,31</sup> Adding to the complexity, unlike its parametric analogue, the Laplace-approximated quasi-likelihood (LAQL) function of our model also depends on the MDP. As a result, it is no longer a linear function of the latent variables, which poses computational difficulties in the E step; see Section 3.2 for more details.

To address these challenges, we develop a specialized two-stage EM algorithm to optimize the latent variable-dependent LAQL function of our model. In the first stage, we propose a multi-step plug-in estimator for the MDP,



which directly utilizes the contaminated data without undergoing the E step. Specifically, we first ignore the measurement error and apply a nested-optimization strategy<sup>31</sup> to maximize the LAQL function for the contaminated data, and then use its results to compute a Fletcher's moment-based MDP estimator.<sup>32</sup> We then establish the analytical relationship between this naïve MDP estimator and the MDP estimator for the true latent outcomes. Finally, the MDP estimator is obtained by plugging the naïve estimator into the established relationship. The second stage involves applying the EM algorithm to maximize the latent variable-dependent LAQL function evaluated at the MDP estimator obtained in the first stage. Such a simplified LAQL function is linear in the latent variable, so the E step is reduced to calculating the conditional expectation of the latent variable given the current estimates, for which the closed-form exact expression is available. Furthermore, we provide the variance estimator for our parameter estimates and a regional association test statistic with a simple F limiting distribution.

In summary, we develop a novel model addressing the diverse dispersion patterns in BS-Seq measures of DNA methylation in tissue samples. To our knowledge, no other methods exist that have our sensitivity to detect association with this kind of data. We develop the complex theory required for testing regional associations, and finally, we demonstrate the properties of the resulting estimators using both simulation evaluations and data applications; this comprehensive algorithm is implemented in an R Bioconductor package, SOMNiBUS.

## 2 | MODEL

We consider DNA methylation measures over a targeted genomic region from  $N$  independent samples. For each sample  $i$ ,  $i = 1, 2, \dots, N$ ,  $m_i$  represents the number of CpG sites with nonzero read depth, that is, the number of measured CpG sites. We write  $t_{ij}$  for the genomic position (in base pairs) for the  $i$ th sample at the  $j$ th CpG site,  $j = 1, 2, \dots, m_i$ . We define  $X_{ij}$  as the total number of reads aligned to CpG  $j$  from sample  $i$ . We denote the *true* methylation status for the  $k$ th read obtained at CpG  $j$  of sample  $i$  as  $S_{ijk}$ , where  $k = 1, 2, \dots, X_{ij}$ . For a single DNA strand read,  $S_{ijk}$  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, the *observed* methylation status, written as  $Y_{ijk}$  can be different from the true underlying status  $S_{ijk}$ . We define  $Y_{ijk} = 1$  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  $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \dots, Z_{iP})$ , for  $i = 1, 2, \dots, N$ .

In the presence of experimental errors, the true methylation data,  $S_{ij}$ , are unknown and one only observes  $Y_{ij}$ . We assume the following error mechanism

$$\begin{aligned} P(Y_{ijk} = 1 | S_{ijk} = 0) &= p_0, \\ P(Y_{ijk} = 1 | S_{ijk} = 1) &= p_1. \end{aligned} \quad (1)$$

Here,  $p_0$  is the rate of false methylation calls, and  $1 - p_1$  is the rate of false non-methylation calls. The error parameters  $p_0$  and  $1 - p_1$  can be estimated from raw sequencing data at CpG sites known in advance to be methylated or unmethylated.<sup>33</sup> Thus, we assume hereafter that  $p_0$  and  $p_1$  are known.

We then propose a quasi-binomial varying coefficient mixed effect model to describe the relationship between the true methylated counts,  $S_{ij}$ , and  $\mathbf{Z}_i$ . Specifically,

$$\log \frac{\pi_{ij}}{1 - \pi_{ij}} = \beta_0(t_{ij}) + \sum_{p=1}^P \beta_p(t_{ij}) Z_{ip} + u_i, \quad (2)$$

$$u_i \stackrel{iid}{\sim} N(0, \sigma_0^2) \quad (3)$$

$$\text{Var}(S_{ij} | u_i) = \phi X_{ij} \pi_{ij} (1 - \pi_{ij}),$$

where  $\pi_{ij} = \mathbb{E}(S_{ij} | u_i) / X_{ij}$  is the *individual's* methylation proportion (ie, the conditional mean),  $\beta_0(t_{ij})$  and  $\{\beta_p(t_{ij})\}_{p=1}^P$  are functional parameters for the intercept and covariate effects. In this model, each  $\pi_{ij}$  incorporates a subject-specific random intercept  $u_i$ , normally distributed with mean 0 and variance  $\sigma_0^2$ . The inclusion of  $u_i$  allows for sample heterogeneity in baseline methylation patterns, and at the same time accounts for the correlation among methylation measurements taken on the same sample. Moreover, we assume the variance of  $S_{ij}$  for individual samples to be a product of a MDP  $\phi$  and a known mean-variance function implied by a binomial distribution (ie,  $V(\pi_{ij}) = X_{ij} \pi_{ij} (1 - \pi_{ij})$ ).

Both the REs  $\mathbf{u} = (u_1, u_1, \dots, u_N)^T$  and the MDP  $\phi$  captures extra-binomial dispersion. However, they address two different aspects of dispersion:  $\mathbf{u}$  models the variation that is due to independent noise across samples, while  $\phi$  aims to relax the assumption of the conditional distribution of  $S_{ij}$  given  $\mathbf{u}$  such that it is not confined to a binomial distribution. In fact, our model generalizes the binomial-based model in Zhao et al.<sup>13</sup> by introducing both the additive dispersion term  $\mathbf{u}$  and multiplicative dispersion term  $\phi$ . Specially, imposing  $\phi = 1$  leads to an additive-dispersion-only (ADO) model and  $\sigma_0^2 = 0$  corresponds to a multiplicative-dispersion-only (MDO) model. When  $\sigma_0^2 = 0$  and  $\phi = 1$ , our model reduces to the binomial-based model in Zhao et al.<sup>13</sup>

## 2.1 | Marginal interpretations

A key feature of the mixed effect model in (2) is that the regression coefficients  $\beta_p(t_{ij})$  need to be interpreted conditional on the value of random effect  $u_i$ . For example,  $\beta_p(\cdot)$  describes how an *individual's* methylation proportions in a region depend on covariate  $Z_p$ . If one desires estimates of such covariate effects on the population average, it is more appropriate to determine the marginal model implied by (2). After applying a cumulative Gaussian approximation to the logistic function and taking an expectation over  $u_i$ , it can be shown that the marginal mean,  $\pi_{ij}^M$ , has the form

$$\pi_{ij}^M = \mathbb{E}(S_{ij})/X_{ij} \approx g\left(\sum_{p=0}^P a \beta_p(t_{ij})Z_{ip}\right), \quad (4)$$

where  $g(x) = 1/(1 + \exp(-x))$ ,  $Z_{i0} \equiv 1$ , and the constant  $a = (1 + c^2\sigma_0^2)^{-1/2}$  with  $c = \sqrt{3.41}/\pi$ ; see detailed derivations in Supplementary Appendix A (henceforth referred to as “SA A,” “SA B,” etc.). The approximation in (4) is quite accurate with errors  $\leq 0.001$ . Thus, the marginal mean induced by our mixed effect model depends on the covariates  $Z_p$  through a logistic link with attenuated regression coefficients  $a\beta_p(t_{ij})$ . Although the smooth covariate effect parameters  $\beta_p(t_{ij})$  have no marginal interpretation, they do have a strong relationship to their marginal counterparts. Hence, the results from hypothesis testing  $H_0 : \beta_p(t_{ij}) = 0$  describe the significance of the covariate effect on both the population-averaged and an individual's DNA methylation levels across a region.

Similarly, the marginal variance of  $S_{ij}$  does not coincide with its conditional counterpart as shown in (3). Our mixed effect model implies a marginal variance of  $S_{ij}$  defined as

$$\text{Var}(S_{ij}) \approx X_{ij}\pi_{ij}^*(1 - \pi_{ij}^*)\left\{\phi + \sigma_0^2 X_{ij}\pi_{ij}^*(1 - \pi_{ij}^*) + O(\phi\sigma_0^4)\right\}, \quad (5)$$

where  $\pi_{ij}^* = g\left(\sum_{p=0}^P \beta_p(t_{ij})Z_{ip}\right)$ ; see detailed derivations in SA A. Note that  $\pi_{ij}^*$  is the mean methylation proportion when setting random effects  $u_i$  to zero and is related to the marginal mean  $\pi_{ij}^M$  via  $\pi_{ij}^* = g\left(g^{-1}\left(\pi_{ij}^M\right)/a\right)$ . Equation (5) illustrates that, under the dSOMNiBUS model, the marginal variance of methylated counts at a CpG site is approximately the variance of the binomial model multiplied by a dispersion factor  $\phi_{ij}^* = \phi + \sigma_0^2 X_{ij}\pi_{ij}^*(1 - \pi_{ij}^*) + O(\phi\sigma_0^4)$ , which depends on the combined effect of  $\phi$  and  $\sigma_0^2$ . Notably, the marginal dispersion factor  $\phi_{ij}^*$  also depends on genomic position  $t_{ij}$  via the dependence of  $\pi_{ij}^*$  on  $t_{ij}$ . Consequently, our dSOMNiBUS model in (2) naturally allows dispersion levels to vary across loci, whereas a MDO model can only accommodate constant dispersion in a region, as illustrated in Figure 1D,E. It is also clear from Equation (5) that an ADO model only allows for overdispersion, and the combination of additive and multiplicative dispersion naturally accounts for both over- and underdispersion.

## 3 | ESTIMATION AND INFERENCE

### 3.1 | Laplace-approximated marginal quasi-likelihood function for the complete data

In model (2), the function parameters,  $\beta_p(t_{ij})$ , can be represented by the coefficients of chosen spline bases of rank  $L_p$ ,  $\beta_p(t_{ij}) = \sum_{l=1}^{L_p} \alpha_{pl}B_l^{(p)}(t_{ij})$ , for  $p = 0, 1, \dots, P$ . Here  $\left\{B_l^{(p)}(\cdot)\right\}_{l=1}^{L_p}$  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, we can write the conditional mean in (2) in a compact way as  $g^{-1}(\boldsymbol{\pi}) =$



$\mathbb{X}^{(B)}\boldsymbol{\alpha} + \mathbb{X}^{(1)}\mathbf{u}$ , 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  $M = \sum_{i=1}^N m_i$ ,  $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^T, \boldsymbol{\alpha}_1^T, \dots, \boldsymbol{\alpha}_p^T)^T \in \mathcal{R}^K$  with  $K = \sum_{p=0}^P L_p$ , and  $\mathbf{u} = (u_1, u_2, \dots, u_N)^T$ .  $\mathbb{X}^{(B)}$  is the spanned design matrix for  $\boldsymbol{\alpha}$  of dimension  $M \times K$ , stacked with elements  $B_l^{(p)}(t_{ij}) \times Z_{ip}$  with  $Z_{i0} \equiv 1$ .  $\mathbb{X}^{(1)}$  is a random effect model matrix of dimension  $M \times N$ , with element 1 if the corresponding CpG site in the row belongs to the sample in the column, and 0 otherwise. If we write the overall spanned design matrix  $\mathbb{X} = [\mathbb{X}^{(B)}, \mathbb{X}^{(1)}] \in \mathcal{R}^{M \times (K+N)}$  and  $\mathcal{B} = (\boldsymbol{\alpha}^T, \mathbf{u}^T)^T \in \mathcal{R}^{K+N}$ , the conditional mean can be further simplified as

$$g^{-1}(\boldsymbol{\pi}) = \mathbb{X}\mathcal{B}.$$

To impose the assumption that the true covariate effect function is more likely to be smooth than jumpy, we add a smoothness penalty<sup>34,35</sup> for each  $\beta_p(t)$ . The total amount of such penalty is an aggregate from all smooth terms, that is,

$$\mathcal{L}^{\text{Smooth}} = \sum_{p=0}^P \lambda_p \int (\beta_p''(t))^2 dt = \sum_{p=0}^P \lambda_p \boldsymbol{\alpha}_p^T \mathbf{A}_p \boldsymbol{\alpha}_p = \boldsymbol{\alpha}^T \mathbf{A}_\lambda \boldsymbol{\alpha}, \quad (6)$$

where  $\mathbf{A}_p$ 's are  $L_p \times L_p$  positive semidefinite matrices with the  $(l, l')$  element

$$\mathbf{A}_p(l, l') = \int B^{(p)''}(t) B^{(p)''}(t) dt.$$

The weights  $\lambda_p$ , that is, the smoothing parameters, are positive parameters which establish a tradeoff between the closeness of the curve to the data and the smoothness of the fitted curves.  $\mathbf{A}_\lambda$  is a  $K \times K$  positive semidefinite block diagonal matrix of the form  $\mathbf{A}_\lambda = \text{Diag}\{\lambda_0 \mathbf{A}_0, \dots, \lambda_p \mathbf{A}_p\}$ . As justified in Wahba<sup>36</sup> and Silverman,<sup>37</sup> employing such smoothing penalty during fitting is equivalent to imposing random effects for spline coefficients  $\boldsymbol{\alpha}$ ; specifically,  $\boldsymbol{\alpha}$  is assumed to follow a (degenerate) multivariate normal distribution with precision matrix  $\mathbf{A}_\lambda$ . Therefore, model (2) with penalization (6) implies the following restraint on the vector of random effects  $\mathcal{B}$ ,

$$\mathcal{B} \sim \text{MVN}(\mathbf{0}, \phi \boldsymbol{\Sigma}_\theta^-),$$

where  $\boldsymbol{\Sigma}_\theta = \text{Diag}\{\phi \mathbf{A}_\lambda, \phi / \sigma_0^2 \mathbf{I}_N\} \in \mathbb{R}^{(K+N) \times (K+N)}$ ,  $\boldsymbol{\Sigma}_\theta^-$  is the pseudoinverse of  $\boldsymbol{\Sigma}_\theta$ , and  $\boldsymbol{\Theta}^T = (\phi \lambda^T, \phi / \sigma_0^2)$  denotes the vector of distinct variance-covariance parameters associated with  $\boldsymbol{\Sigma}_\theta$ . Therefore, the integrated *marginal quasi-likelihood function* for the complete data  $\{\mathbf{S}, \mathbf{X}, \mathbf{Z}\}$  can be defined as

$$qL^M(\phi, \boldsymbol{\Theta}) = \int \exp \left\{ ql^{(\mathbf{S}|\mathcal{B})}(\mathcal{B}, \phi) - \frac{1}{2\phi} \mathcal{B}^T \boldsymbol{\Sigma}_\theta \mathcal{B} + \frac{1}{2} \log \{|\boldsymbol{\Sigma}_\theta / \phi|_+\} \right\} d\mathcal{B}, \quad (7)$$

where  $|\bullet|_+$  denotes the generalized determinant of a matrix, that is, the product of its non-zero eigenvalues, and  $ql^{(\mathbf{S}|\mathcal{B})}(\mathcal{B}, \phi)$  is the conditional log-quasi-likelihood function given the values of REs  $\mathcal{B}$ . Specifically, we follow the notion of *extended quasi-likelihood*<sup>38</sup> and define the following conditional quasi-likelihood

$$\exp \{ ql^{(\mathbf{S}|\mathcal{B})}(\mathcal{B}, \phi) \} \propto \exp \left\{ -\frac{1}{2\phi} \sum_{ij} d_{ij}(S_{ij}, \pi_{ij}) - \frac{M}{2} \log \phi \right\},$$

where

$$d_{ij}(S_{ij}, \pi_{ij}) = -2 \int_{S_{ij}/X_{ij}}^{\pi_{ij}} \frac{S_{ij} - X_{ij} \pi_{ij}}{\pi_{ij}(1 - \pi_{ij})} d\pi_{ij} \quad (8)$$

is the quasi-deviance contributed from a single observation. We use the Laplace approximation<sup>28,31,39</sup> to evaluate the integral inside the marginal quasi-likelihood (7). Let  $\hat{\mathcal{B}}_\theta$  be the value of  $\mathcal{B}$  maximizing the integrand in (7) given the values of variance components  $\boldsymbol{\Theta}$ , that is,

$$\hat{\mathcal{B}}_\theta = \text{argmax}_{\mathcal{B} \in \mathcal{R}^{K+N}} \left\{ -\frac{1}{2\phi} \sum_{ij} d_{ij}(S_{ij}, \pi_{ij}) - \frac{1}{2\phi} \mathcal{B}^T \boldsymbol{\Sigma}_\theta \mathcal{B} \right\}, \quad (9)$$

where terms not dependent on  $\mathbf{B}$  have been dropped. By writing the log of the integrand in (7) by a quadratic Taylor series expansion about  $\widehat{\mathbf{B}}_{\Theta}$ , we have

$$-2 \log [qL^M(\phi, \Theta)] \approx \frac{\sum_{ij} \widehat{d}_{ij}}{\phi} + M \log \phi + \frac{1}{\phi} \widehat{\mathbf{B}}_{\Theta}^T \Sigma_{\Theta} \widehat{\mathbf{B}}_{\Theta} + \log \left| \frac{\mathbb{X}^T \widehat{\mathbf{W}} \mathbb{X} + \Sigma_{\Theta}}{\phi} \right| - \log \left| \frac{\Sigma_{\Theta}}{\phi} \right|_+; \quad (10)$$

see detailed derivations in SA B. In Equation (10),  $\widehat{d}_{ij} = d_{ij}(S_{ij}, \widehat{\pi}_{ij})$ , where  $\widehat{\pi}_{ij} = \mathbf{g}(\mathbb{X}_{(l)} \widehat{\mathbf{B}}_{\Theta})$  with  $l$  denoting the row in the matrix  $\mathbb{X}$  corresponding to CpG  $j$  for sample  $i$ .  $\widehat{\mathbf{W}}$  is the weight matrix whose diagonal is  $X_{ij} \widehat{\pi}_{ij} (1 - \widehat{\pi}_{ij})$ , and both  $\widehat{\mathbf{B}}_{\Theta}$  and  $\widehat{\mathbf{W}}$  depend on  $\Theta$ , that is,  $\widehat{\mathbf{B}}_{\Theta} = \widehat{\mathbf{B}}_{\Theta}(\Theta)$  and  $\widehat{\mathbf{W}} = \widehat{\mathbf{W}}(\Theta)$ . The negative half of the right-hand side expression in Equation (10) constitutes the log of the LAQL function, denoted as  $\text{Laplace}(\phi, \Theta; \mathbf{S})$ .

*Remark 1.* (i) Equation (9) implies that, given the values of  $\Sigma_{\Theta}$ , the REs  $\mathbf{B}$  and  $\phi$  are orthogonal, which explains why we parametrize the covariate matrix of  $\mathbf{B}$  as  $\phi \Sigma_{\Theta}^{-1}$ . (ii) The LAQL function in (10) cannot be written as the sum of a part related only to  $\phi$  and a part related only to  $\Sigma_{\Theta}$  (or  $\Theta$ ), which implies that the maximum quasi-likelihood estimate (MQLE) for  $\Theta$  also depends on the estimate for  $\phi$  and vice versa. Thus, it is undesirable to estimate  $\Theta$  under  $\phi = 1$  and then adjust  $\phi$  based on the estimated  $\Theta$ , as in GLMs. Instead, a joint optimization is needed, that is, finding  $\text{argmax}_{\phi, \Theta} \text{Laplace}(\phi, \Theta; \mathbf{S})$ .

### 3.2 | A two-stage estimation algorithm

In the presence of experimental errors, the true methylation data,  $S_{ij}$  are unknown, and one only observes  $Y_{ij}$ , which is assumed to be a mixture of binomial counts arising from both the truly methylated and truly unmethylated reads. When  $S_{ij}$  is modeled by a parametric distribution, as in Zhao et al,<sup>13</sup> the EM algorithm<sup>26</sup> provides a computationally simple way to obtain the maximum likelihood estimate (MLE) of the smooth covariate effects based on the observed data  $Y_{ij}$ . However, this computational simplicity does not apply to our quasi-likelihood function in (7). Specifically, we can evaluate the integral inside our quasi-deviance function  $d_{ij}$  (8), and obtain  $\widehat{d}_{ij} = -2\{S_{ij} \log \widehat{\pi}_{ij} + (X_{ij} - S_{ij}) \log(1 - \widehat{\pi}_{ij}) - S_{ij} \log(S_{ij}/X_{ij}) - (X_{ij} - S_{ij}) \log(1 - S_{ij}/X_{ij})\}$ . Then, it is evident that there is a nonlinear term with respect to  $S_{ij}$  involved in the complete-data LAQL function (10), that is,

$$\{S_{ij} \log(S_{ij}/X_{ij}) - (X_{ij} - S_{ij}) \log(1 - S_{ij}/X_{ij})\} / \phi.$$

Therefore, for a trial estimate  $(\phi^*, \Theta^*)$ , the integrated likelihood (ie, the E step)

$$Q(\phi, \Theta | \phi^*, \Theta^*) = \mathbb{E}_{\mathbf{S} | \mathbf{Y}} \{ \text{Laplace}(\phi, \Theta; \mathbf{S}) | \mathbf{Y}; \phi^*, \Theta^* \}$$

involves an intractable integral/summation whose exact close-form expression is not available. To circumvent this problem, we develop a two-stage EM algorithm—first obtain  $\widehat{\phi}$  using our proposed multi-step plug-in estimator and then apply the EM algorithm to maximize the complete-data LAQL function with  $\phi$  fixed at the  $\widehat{\phi}$ .

#### 3.2.1 | Stage 1: A multi-step plug-in estimator for $\phi$

The key idea of estimating  $\phi$  without undergoing the EM iteration is to derive the induced model for  $\mathbf{Y}$  by marginalizing over the “distribution” of  $\mathbf{S}$  and then equate the MDP implied by such an induced model with the naïve analysis result obtained by treating  $\mathbf{Y}$  as error-free. Specifically, based on our assumed mean-variance relationship for  $\mathbf{S}$  in (3) and error model for the conditional mean of  $Y_{ijk}$  given  $S_{ijk}$  in (1), we can express the mean and MDP for the observed outcome  $\mathbf{Y}$  as

$$\begin{aligned} \pi_{ij}^Y &= \mathbb{E}(Y_{ij} | u_i) = \pi_{ij} p_1 + (1 - \pi_{ij}) p_0 \\ \phi_{ij}^Y &= \frac{\text{Var}(Y_{ij} | u_i)}{X_{ij} \pi_{ij}^Y (1 - \pi_{ij}^Y)} = 1 + (\phi - 1) \frac{(\pi_{ij}^Y - p_0)(p_1 - \pi_{ij}^Y)}{\pi_{ij}^Y (1 - \pi_{ij}^Y)}; \end{aligned}$$

see detailed derivations in SA C. When  $Y_{ij}$ s are truly error-free, these  $Y$ -related parameters  $\pi_{ij}^Y$  and  $\phi_{ij}^Y$  coincide with the  $\pi_{ij}$  and  $\phi$  for  $S_{ij}$  in the true model. Unlike the constant  $\phi$  for the true outcome  $\mathbf{S}$ , the induced dispersion parameter  $\phi_{ij}^Y$  varies with each CpG site, when  $\phi \neq 1$ , and its mean across all CpG sites in the dataset is

$$\overline{\phi_{ij}^Y} = \frac{1}{M} \sum_{ij} \phi_{ij}^Y = 1 + (\phi - 1) \frac{1}{M} \sum_{ij} \frac{(\pi_{ij}^Y - p_0)(p_1 - \pi_{ij}^Y)}{\pi_{ij}^Y(1 - \pi_{ij}^Y)}. \quad (11)$$

On the other hand, we can run the naïve analysis that assumes  $p_0 = 1 - p_1 = 0$ . The goal of the naïve analysis is to maximize the LAQL function  $\text{Laplace}(\phi, \Theta; \mathbf{Y})$  as defined in (10), with  $\mathbf{Y}$  replacing  $\mathbf{S}$ . Here, the search for parameters is still constrained to the model space defined in (2) and (3), and thus a constant dispersion parameter estimate  $\hat{\phi}^Y$  will be obtained. We assume that  $\hat{\phi}^Y$  is an estimate for the mean of individual dispersions,  $\overline{\phi_{ij}^Y}$ ; simulation results show that this is a reasonable assumption (Supplementary Figure S17). Therefore, once the naïve estimator  $\hat{\phi}^Y$  is obtained, we can then plug the  $Y$ -related estimates  $\hat{\phi}^Y$  and  $\hat{\pi}_{ij}^Y$  into the relationship in (11) to obtain the estimate for  $\phi$ . Specifically, we propose the following steps to compute this plug-in estimator  $\hat{\phi}$ .

- Step 1:* Use the nested-optimization strategy proposed by Wood<sup>31</sup> to obtain  $(\hat{\phi}_{Lik}^Y, \hat{\Theta}^Y) = \text{argmax}_{\phi, \Theta} \text{Laplace}(\phi, \Theta; \mathbf{Y})$ . In summary, this algorithm has an outer iteration for updating  $\Theta$  and  $\phi$  using Newton's method, with each iterative step supplemented with an inner iteration to obtain  $\hat{\mathbf{B}}_{\Theta}$  (9) corresponding to the current  $\Theta$ ; the detailed description is summarized in the SA D.
- Step 2:* First calculate a moment-like dispersion estimator,  $\hat{\phi}_P$ , based on equating Pearson's  $\chi^2$  goodness-of-fit statistic to its expectation under the model. In the presence of random effects, such an expectation equals to  $M$  minus the effective degrees of freedom (EDF) of the model,<sup>40</sup> which depends on the magnitude of  $\hat{\Theta}^Y$  and is smaller than  $K + N$ , the dimension of  $\mathbf{B}$ . We then adjust  $\hat{\phi}_P$  using Fletcher's method<sup>32</sup> to improve the stability of  $\hat{\phi}_P$ . The detailed description is summarized in the SA E. The final estimator obtained in this step is denoted as  $\hat{\phi}^Y$ .
- Step 3:* Plug the  $\hat{\pi}_{ij}^Y$  from Step 1 and  $\hat{\phi}^Y$  from Step 2 into (11), that is,

$$\hat{\phi} = (\hat{\phi}^Y - 1) \left[ \frac{1}{M} \sum_{ij} \frac{(\hat{\pi}_{ij}^Y - p_0)(p_1 - \hat{\pi}_{ij}^Y)}{\hat{\pi}_{ij}^Y(1 - \hat{\pi}_{ij}^Y)} \right]^{-1} + 1. \quad (12)$$

*Remark 2.* (i) When  $\mathbf{Y}$  is error-free, running Steps 1 and 2 yields estimates for  $\Theta$ ,  $\mathbf{B}$  and  $\phi$  in our model. (ii) Our dispersion estimator combines the quasi-likelihood (Step 1) and moment-based (Step 2) estimation. Due to the lack of orthogonality, joint optimization for  $\Theta$  and  $\phi$  should be performed to provide MQLEs,  $\hat{\phi}_{Lik}$  and  $\hat{\Theta}$  (and thus  $\hat{\pi}_{ij}$ ). However, the MQLEs, in particular, the estimator for the dispersion parameter, can be biased in finite samples.<sup>41</sup> Therefore, instead of using  $\hat{\phi}_{Lik}$ , we take a step further and compute the moment-based estimator for  $\phi$  based on  $\hat{\pi}_{ij}$ . Simulation results show that such a combined strategy significantly reduces the bias of the estimated  $\phi$ , compared to quasi-likelihood estimation alone, regardless of whether error-free or contaminated data are analyzed; see Supplementary Figures S14-S16. (iii) For regions with many  $\hat{\pi}_{ij}^Y$  less than  $p_0$  or greater than  $p_1$ , (12) can lead to negative  $\hat{\phi}$ . To better tackle these pathological cases, we calculate  $\hat{\phi}$  by averaging only those CpGs with  $p_0 \leq \hat{\pi}_{ij}^Y \leq p_1$  in (12). Such a stabilized approach produces superior results in our data application, as illustrated in Figure S1.

### 3.2.2 | Stage 2: EM iteration for $\Theta$ and $\mathbf{B}$

In the second stage, we aim to maximize the latent variable-dependent LAQL function evaluated at  $\hat{\phi}$ , that is,  $\text{Laplace}(\hat{\phi}, \Theta; \mathbf{S})$ . Dropping the terms not depending on  $\Theta$ , we have

$$\text{Laplace}(\hat{\phi}, \Theta; \mathbf{S}) = \sum_{ij} 1/\hat{\phi} \{S_{ij} \log \hat{\pi}_{ij} + (X_{ij} - S_{ij}) \log(1 - \hat{\pi}_{ij})\} + f(\hat{\phi}, \Theta), \quad (13)$$

where  $f(\hat{\phi}, \Theta) = -1/2 \left\{ \hat{\mathbf{B}}_{\Theta}^T \Sigma_{\Theta} \hat{\mathbf{B}}_{\Theta} / \hat{\phi} + \log |\mathbb{X}^T \widehat{\mathbf{W}} \mathbb{X} + \Sigma_{\Theta}| - \log |\Sigma_{\Theta}| \right\}$ . Evidently, this simplified LAQL function is linear in the latent variables  $S_{ij}$ . Therefore, its conditional expectations given the observed data  $Y_{ij}$  and trial estimate  $\Theta^*$  can be simplified as

$$Q(\Theta | \Theta^*) = \mathbb{E}_{S|Y} \left\{ \text{Laplace}(\hat{\phi}, \Theta; \mathbf{S}) | \mathbf{Y}; \Theta^* \right\} = \text{Laplace}(\hat{\phi}, \Theta; \eta^*). \quad (14)$$

### E step

In Equation (14),  $\eta^* \in \mathcal{R}^M$  are conditional expectations of  $\mathbf{S}$  given  $\mathbf{Y}$  evaluated at the trial estimate  $\Theta^*$ , and for our model, take the form

$$\eta_{ij}^* = \mathbb{E}(S_{ij} | Y_{ij}; \Theta^*) = \frac{Y_{ij} p_1 \pi_{ij}^*}{p_1 \pi_{ij}^* + p_0 (1 - \pi_{ij}^*)} + \frac{(X_{ij} - Y_{ij})(1 - p_1) \pi_{ij}^*}{(1 - p_1) \pi_{ij}^* + (1 - p_0)(1 - \pi_{ij}^*)},$$

where  $\pi_{ij}^* = g(\mathbb{X}_{(i)}, \mathbf{B}_{\Theta}^*)$ , which depends on  $\Theta^*$  via the dependence of  $\mathbf{B}_{\Theta}^*$  on  $\Theta^*$ . Calculating these conditional expectations  $\eta_{ij}^*$  constitutes the E step in Stage 2.

### M step

Each M step involves maximizing the Q function in (14) to update  $\Theta$ . Similarly, this can be achieved by the nested-optimization strategy<sup>31</sup> described in Step 1 of the first stage (Section 3.2.1 and SA D), but without updating  $\phi$ .

### E-M iteration

We iterate between the E and M steps until convergence to obtain  $\hat{\Theta}$  and  $\hat{\mathbf{B}}$ . The first  $K$  elements of the final estimates  $\hat{\mathbf{B}}$ , that is,  $\hat{\alpha}$ , yield estimates of the functional parameters  $\beta_p(t)$ , for  $p = 0, 1, \dots, P$ :  $\hat{\beta}_p(t) = \{\mathbf{B}^{(p)}(t)\}^T \{\hat{\alpha}_p\}$ , where  $t$  is a genomic position lying within the range of the input positions  $\{t_{ij}\}$ , and  $\mathbf{B}^{(p)}(t) = (B_1^{(p)}(t), B_2^{(p)}(t), \dots, B_{L_p}^{(p)}(t))^T \in \mathcal{R}^{L_p}$  is a column vector with nonrandom quantities obtained from evaluating the set of basis functions  $\{B_l^{(p)}(\cdot)\}_l$  at position  $t$ .

## 3.2.3 | Algorithm overview

The complete sequence of steps for our two-stage EM algorithm is outlined in Algorithm 1. A key feature of our algorithm is that the multiplicative dispersion parameter (also known as the scale parameter),  $\phi$ , is handled separately from the mean and variance component-based parameters  $(\mathbf{B}, \Theta)$ . Unlike the latter parameters,  $\phi$  remains fixed throughout the EM iterations. This strategy ensures that maximizing the Q function in (14) necessarily increases the quasi-likelihood for the error-prone outcomes  $\mathbf{Y}$ , with values consistently ascending at each iteration of the EM update, a fact empirically validated in Supplementary Figure S18. Detailed derivations can be found in SA F. This appendix clarifies how the Q function forms a lower bound on the log quasi-likelihood of our observed data  $\mathbf{Y}$ . By iteratively maximizing this lower bound, we edge closer to the maximum of the quasi-likelihood of  $\mathbf{Y}$ , which is otherwise challenging to maximize directly.

Another standout feature of our algorithm is how we handle the regularization parameters  $\Theta$ , which include the smoothing parameters  $\lambda_p, p = 0, 1, \dots, P$  and the variance of RE,  $\sigma_0^2$ . Instead of relying on prediction-based criteria that target minimizing model prediction error, we adopt a marginal likelihood-based approach for regularization parameter selection. This approach hinges on integrating out  $\mathbf{B}$ , thereby crafting an objective function solely dependent on  $\Theta$ . Such a formulation allows for direct optimization of a well-defined function of  $\Theta$ , specifically  $\text{Laplace}(\hat{\phi}, \Theta; \eta^*)$  as defined in Supplementary Equation (S20). Extensively studied by Wood,<sup>31</sup> this likelihood-based strategy demonstrates notably superior convergence properties in selecting regularization parameters compared to prediction-based methods.

In the second stage of Algorithm 1, each M step aims to maximize the Laplace approximated quasi-likelihood,  $\text{Laplace}(\hat{\phi}, \Theta; \eta^*)$ —the Q function in (14)—using a nested-optimization strategy, similarly employed in the first stage's Step 1. This process updates  $\rho = \log(\Theta)$  via Newton's method. Each trial  $\rho$  proposed in the outer Newton iteration requires a subsequent (inner) Newton iteration for solving penalized quasi-score equations for  $\mathbf{B}$ , as detailed in Supplementary (SA) Section 2.4.2. Within this nested-optimization framework, the negative Hessian matrix of our objective function, as discussed in SA Section 6.2.2 and supported by Wood's studies,<sup>31,42</sup> is generally positive definite—a critical aspect for the success of Newton's method, which relies on inverting the Hessian matrix at each iteration. Yet, challenges such as

**Algorithm 1.** A two-stage EM algorithm to estimate the smoothed quasi-binomial mixed model with error-prone outcomes

**Stage 1:** Calculate the plug-in estimator  $\hat{\phi}$

Step 1: run the algorithm for error-free outcomes on  $\{\mathbf{Y}, \mathbf{Z}, \mathbf{X}\}$ ; return  $\hat{\pi}^Y$ ,  $\hat{\phi}_{lik}^Y$ ,  $\hat{\mathbf{B}}^Y$ , and  $\hat{\Theta}^Y$

▷ Algorithm S1 in SA Section 2.4

Step 2: calculate Fletcher's moment estimator,  $\hat{\phi}^Y$

▷ Equation (S10)

Step 3: calculate the plug-in estimator  $\hat{\phi}$

▷ Equation (12)

**Stage 2:** E-M iterations with  $\phi$  fixed at  $\hat{\phi}$  to estimate  $\mathbf{B}$  and  $\Theta$ ; Specifically

**Initialize**  $\Theta^{(0,0)} = \hat{\Theta}^Y$ ,  $\mathbf{B}^{(0,0)} = \hat{\mathbf{B}}^Y$ ; Choose  $\varepsilon = 10^{-6}$ ; Set  $\ell = 0, s = 0$

**repeat**

• E step:  $\eta_{ij}^{(\ell)} = \mathbb{E}(S_{ij} | Y_{ij}; \mathbf{B}^{(\ell,s)})$

• M step:  $\Theta^{(\ell)} = \operatorname{argmax}_{\Theta} \operatorname{Laplace}(\hat{\phi}, \Theta; \eta_{ij}^{(\ell)})$ . Specifically

**repeat**

• Newton's update for the Laplace approximated marginal likelihood evaluated at data  $\eta_{ij}^{(\ell)}$ :

$$\rho^{(\ell,s+1)} = \rho^{(\ell,s)} - \left[ \nabla^2 \operatorname{Laplace}(\rho^{(\ell,s)}; \mathbf{B}^{(\ell,s)}, \eta^{(\ell)}) \right]^{-1} \nabla \operatorname{Laplace}(\rho^{(\ell,s)}; \mathbf{B}^{(\ell,s)}, \eta^{(\ell)}) \quad \triangleright \rho = \log(\Theta)$$

▷ details in SA Section 2.4.2

• Solve  $\mathbf{U}(\mathbf{B}; \Theta^{(\ell,s+1)}; \eta^{(\ell)}) = \mathbf{0}$  to obtain  $\mathbf{B}^{(\ell,s+1)}$

▷ details in SA Section 2.4.1

$s \leftarrow s + 1$

**until**  $\|\rho^{(\ell,s)} - \rho^{(\ell,s-1)}\|_2 < \varepsilon$

$\ell \leftarrow \ell + 1$

**until**  $\|\mathbf{B}^{(\ell,s)} - \mathbf{B}^{(\ell-1,s)}\|_2 < \varepsilon$

**return**  $\Theta^{(\ell,s)}, \mathbf{B}^{(\ell,s)}$

numerical singularities may arise, particularly at the boundary of the parameter space where elements of  $\rho$  are extremely large or effectively zero, or when certain elements of  $\mathbf{B}$  become unidentifiable.<sup>42</sup> Fortunately, the `mgcv` package<sup>43</sup> offers a robust, user-friendly nested-optimization implementation, featuring step length control and Hessian perturbation for positive definiteness,<sup>44</sup> thereby facilitating successful convergence of Newton's method to a global maximum. By directly adopting the `mgcv` package for the nested optimization, we are able to maximize  $\operatorname{Laplace}(\hat{\phi}, \Theta; \eta^*)$  with enhanced computational robustness.

In our algorithm implementation, we used natural cubic splines to parameterize the functional parameters  $\beta_p(t)$ , with their interior knots placed at the empirical quantiles of  $t_{ij}$ . Users of our software can adjust the basis dimension as part of their model-building process. We recommend keeping  $L_p$  fixed at a slightly larger size than it is believed could reasonably be. Specific choices of  $L_p$  for our data application and simulation study are detailed in Section 4.1 and Supplementary Section 5.1, respectively. Notably, the choice of  $L_p$  sets an upper limit on the flexibility of  $\beta_p(t)$ , with its actual flexibility being governed by the smoothing parameter  $\lambda_p$ . Therefore, as long as we refrain from choosing overly small basis dimensions, the exact value of  $L_p$  has minimal impact on the fitted model.<sup>35</sup>

The algorithm begins by initializing the regularization parameters, that is, setting up  $\Theta^{(0)}$  in the Algorithm S1 in Supplementary Section 2.4. We employ the implementation in the R package `mgcv`<sup>43</sup> for this initialization. The initial values of  $\lambda_p$  and  $\sigma_0^2$  are selected to achieve a rough balance between the leading diagonals of  $\mathbb{X}^T \mathbf{W} \mathbb{X}$  and  $\Sigma_{\Theta}$ . Specifically, Let  $\mathbf{a}_j$  denote the elements of  $\operatorname{diag}(\mathbf{A}_j)$  and  $\mathbf{d}_j$  denote the corresponding elements of  $\operatorname{diag}(\mathbb{X}^T \mathbf{W} \mathbb{X})$ . We set  $\lambda_j$  such that the mean of  $[\mathbf{d}_j / (\mathbf{d}_j + \lambda_j \mathbf{a}_j)] \approx 0.5$ . Similarly, we determine  $\sigma_0^2$  such that  $[\mathbf{d}_N / (\mathbf{d}_N + 1/\sigma_0^2)] \approx 0.5$ , where  $\mathbf{d}_N$  is the last  $N$  elements of  $\operatorname{diag}(\mathbb{X}^T \mathbf{W} \mathbb{X})$ . Here, the initial estimate of  $\mathbf{W}$  can be derived from the sample methylation proportions  $\hat{\pi}_{ij} = Y_{ij}/X_{ij}$ . In summary, this setup ensures that  $\Theta^{(0)}$  are neither effectively zero, nor infinity, thus preventing numerical instability.

Regarding the convergence criterion, we monitor the relative change in the estimate of  $\mathbf{B}$  between consecutive EM iterations. The algorithm stops when this relative change falls below a predefined threshold,  $\varepsilon = 10^{-6}$ . Additionally, we impose a maximum number of iterations of 500 to prevent the algorithm from running indefinitely. Owing to the robust numerical properties of the EM algorithm and the nested-optimization strategy used within the M-step to estimate  $\Theta$ , our proposed algorithm consistently converges across various settings in our data applications and simulation studies.

### 3.3 | Inference for smooth covariate effects

From the results generated by our two-stage algorithm, we proceed to compute the pointwise confidence intervals (CI) for the smoothed covariate effects  $\{\beta_1(t), \beta_2(t), \dots, \beta_p(t)\}$ , and obtain tests of hypotheses for these effects. Note that the inference is carried out conditional on the values of  $\Theta$  and  $\phi$ , that is, the uncertainty in estimating them is not accounted for. In addition, we assume negligible remainder error in the basis expansion,  $\beta_p(t_{ij}) = \sum_{l=1}^{L_p} \alpha_{pl} B_l^{(p)}(t_{ij})$ . This is a valid assumption for smoothing splines, because a sufficiently large  $L_p$  is used and  $\lambda_p$  controls the variance-bias tradeoff. Under this assumption, the region-wide test of the null hypothesis  $H_0 : \beta_p(t) = 0$  is equivalent to  $H_0 : \alpha_p = \mathbf{0}$ .

#### 3.3.1 | Confidence interval estimation

Let  $\ell$  be an iteration index in the EM algorithm. Given the sequence of  $\Theta^{(\ell)}$  obtained in each M step, the estimate  $\hat{\mathbf{B}}$  arises from iterating between calculating the expectations  $\eta^{(\ell)}$  and solving the *quasi-score* equation for  $\mathbf{B}$ , that is,

$$\mathbf{U}^{(1)}(\mathbf{B}) = \frac{1}{\hat{\phi}} \left[ \mathbb{X}^T (\eta^{(\ell)} - \mathbf{A}_X \boldsymbol{\pi}) - \boldsymbol{\Sigma}_{\Theta}^{(\ell)} \mathbf{B} \right] = \mathbf{0},$$

where  $\mathbf{A}_X \in \mathcal{R}^{M \times M}$  is a diagonal matrix with entries  $X_{ij}$ . To quantify the uncertainty of this expectation-solving (ES) estimate  $\hat{\mathbf{B}}$ , we adopt the approach in Elashoff and Ryan.<sup>45</sup> Specifically, we can reformulate the E step as an estimating equation that solves for latent variables  $\mathbf{S}$ , namely  $\mathbf{U}^{(2)}(\mathbf{S}) = \mathbf{S} - \eta^{(\ell)} = \mathbf{0}$ . Thus, this iterative procedure can be viewed as solving an augmented set of estimating equations; see SA G for details. Under this formulation, we use the established theory for estimating equations,<sup>46-48</sup> and propose a model-based variance estimator for  $\hat{\mathbf{B}}$ . Specifically, under correct specification of the first two moments of  $\mathbf{S}$ , the asymptotic variance of  $\hat{\mathbf{B}}$  equals to the observed Fisher information

$$\widehat{\text{Var}}(\hat{\mathbf{B}}) = \left[ \mathbb{X}^T (\widehat{\mathbf{W}} - \widehat{\mathbf{W}}_{\delta}) \mathbb{X} + \widehat{\boldsymbol{\Sigma}}_{\Theta} \right]^{-1} \hat{\phi}. \quad (15)$$

In (15),  $\widehat{\mathbf{W}}_{\delta}$  is a diagonal matrix with elements  $\hat{\delta}_{ij} \hat{\pi}_{ij} (1 - \hat{\pi}_{ij})$ , where

$$\hat{\delta}_{ij} = \frac{Y_{ij} p_1 p_0}{[p_1 \hat{\pi}_{ij} + p_0 (1 - \hat{\pi}_{ij})]^2} + \frac{(X_{ij} - Y_{ij})(1 - p_1)(1 - p_0)}{[(1 - p_1) \hat{\pi}_{ij} + (1 - p_0)(1 - \hat{\pi}_{ij})]^2}.$$

The detailed derivation is given in SA G. Let  $\hat{\mathbf{V}}$  denote the variance estimator in (15) and  $\hat{\mathbf{V}}_p$  be the diagonal blocks of  $\hat{\mathbf{V}}$  corresponding to  $\alpha_p$ , with dimensions  $L_p \times L_p$ . We then immediately have the estimated variance of  $\hat{\beta}_p(t)$ :  $\widehat{\text{Var}}(\hat{\beta}_p(t)) = \{\mathbf{B}^{(p)}(t)\}^T \hat{\mathbf{V}}_p \{\mathbf{B}^{(p)}(t)\}$ . Therefore, the confidence interval for  $\beta_p(t)$  at significance level  $\nu$  can be approximately estimated by  $\hat{\beta}_p(t) \pm \mathbb{Z}_{\nu/2} \sqrt{\widehat{\text{Var}}(\hat{\beta}_p(t))}$ , for any  $t$  in the range of interest, where  $\mathbb{Z}_{\nu/2}$  is  $\nu/2$  (upper-tail) quantile of a standard normal distribution.

#### 3.3.2 | Hypothesis testing for a regional zero effect

We can also construct a region-wide test of the null hypothesis  $H_0 : \beta_p(t) = 0$ . This test depends on the association between covariate  $Z_p$  and methylation levels across the region, after adjustment for all the other covariates. We propose the following region-based F statistic

$$T_p = \frac{\hat{\boldsymbol{\alpha}}_p^T \left\{ \hat{\mathbf{V}}_p \right\}^{-1} \hat{\boldsymbol{\alpha}}_p}{\tau_p},$$



where  $\{\widehat{\mathbf{V}}_p\}^{-1}$  denotes inverse if  $\widehat{\mathbf{V}}_p$  is nonsingular; for singular  $\widehat{\mathbf{V}}_p$ , the inverse is replaced by the Moore-Penrose inverse  $\{\widehat{\mathbf{V}}_p\}^-$ . Here,  $\tau_p$  is the EDF for smooth term  $\beta_p(t)$  and

$$\tau_p = \sum_{l=a_p}^{b_p} (2\mathbf{F} - \mathbf{F}\mathbf{F})_{(l,l)}, \text{ for } p = 0, 1, \dots, P,$$

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  $l$ th leading diagonal element of a matrix.  $\mathbf{F}$  is the smoothing matrix of our model, which has the form  $\mathbf{F} = (\mathbb{X}^T \widehat{\mathbf{W}} \mathbb{X} + \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}})^{-1} \mathbb{X}^T \widehat{\mathbf{W}} \mathbb{X}$ . Using the property of our plug-in estimator (Supplementary Section 2.8.1), we can conclude that, under the null hypothesis,  $T_p$  asymptotically follows a F distribution with degrees of freedom  $\tau_p$  and  $M - \tau$ , where  $\tau = \text{trace}(\mathbf{F})$ , that is,  $T_p \sim F_{\tau_p, M - \tau}$ ; see detailed derivations in SA H.

## 4 | DIFFERENTIAL METHYLATION ANALYSIS OF ACPA STATUS

We apply our new method to genome-wide targeted bisulfite sequencing data from a preclinical study on rheumatoid arthritis.<sup>6,7</sup> We compare the findings of dSOMNiBUS to those of five existing methods: BiSeq,<sup>49</sup> BSmooth,<sup>50</sup> SMSC,<sup>4</sup> dmrseq<sup>51</sup> and GlobalTest<sup>52</sup> (see Supplementary Section 3.1 for a detailed description of the five existing methods).

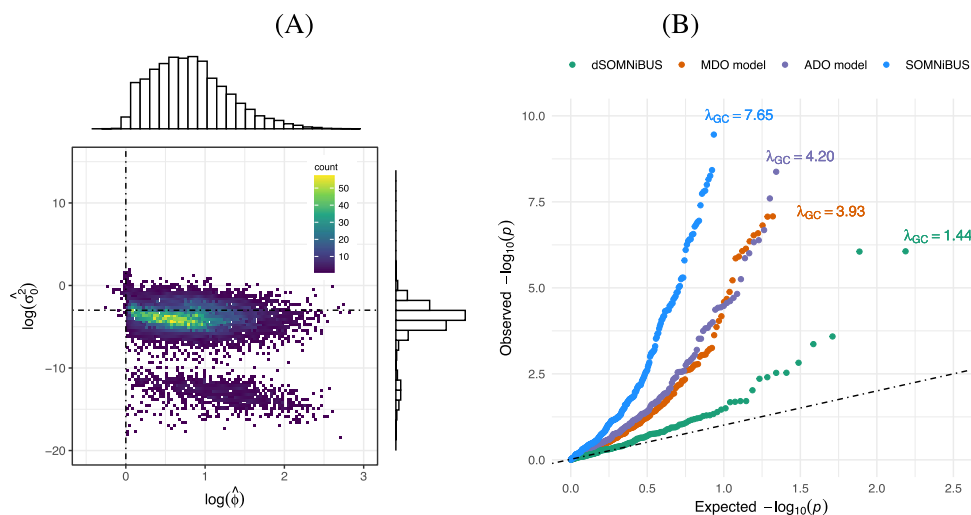
### 4.1 | ACPA dataset

In this study, participants were sampled from the CARTaGENE cohort, a population-based cohort of 43 000 subjects aged between 40 and 69 years, from Quebec, Canada. First, the serum ACPA levels were measured for 3600 randomly-sampled individuals from the CARTaGENE cohort (<https://www.cartagene.qc.ca/>), based upon which individuals were classified as either ACPA positive or ACPA negative. Then, the whole blood samples of the ACPA positive individuals, and a selected subset of age-sex-and-smoking-status-matched ACPA negative individuals were sent for Targeted Custom Capture Bisulfite Sequencing. Specifically, the sequencing used blood cell-specific immune panels that cover the majority of human gene promoters, active regulatory regions observed in blood, blood-cell-lineage-specific enhancer regions and CpGs from Illumina Human Methylation 450 Bead Chips. Cell type proportions in the blood samples were also measured at the time of the sampling. We excluded the samples who reported a diagnosis of RA before the CARTaGENE study started and samples with missing information on cell type proportions. In our final analysis, there are 48 ACPA-positive and 54 ACPA-negative subjects.

We focused on autosomal analysis only. To better translate the methylation information in the sequence of the genome into biologically relevant knowledge, we defined the gene-specific methylation regions as the first exon and 2000 base pairs upstream of each protein-coding gene. For simplicity, we focused on regions with at least 20 CpG sites, and our final analysis includes 12 569 methylation regions, covering around 1.4 million CpG sites. For details on these regions, refer to Supplementary Figure S25 for the distribution of the number of CpG sites and Supplementary Figure S26 for read depth summaries, including the first quartile, median, and third quartile. The association analyses were conducted with the adjustment for age, sex, smoking status and cell type composition. For dSOMNiBUS, we assumed two settings of data errors: (1) zero errors:  $p_0 = 1 - p_1 = 0$  and (2) non-zero errors:  $p_0 = 0.003$ ,  $p_1 = 0.9$ . The value 0.003 was reported by Prochenka et al<sup>53</sup> as insufficient Bisulfite conversion rate and 0.1 was estimated as the average excessive conversion rate from a (single-cell-type) bisulfite dataset in Hudson et al<sup>54</sup> using the method SMSC.<sup>4</sup> We used natural cubic splines to expand the smooth terms in the model, and its rank  $L_p$  was approximately equal to the number of CpGs in a region divided by 10 for  $\beta_0(t)$ , and divided by 20 for  $\beta_p(t)$ ,  $p \geq 1$ , ensuring around 10 or 20 CpGs in each piecewise polynomial. Here, the intercept  $\beta_0(t)$  is assigned a larger rank  $L_0$  to accommodate a more flexible shape compared to the covariate effects  $\beta_p(t)$  with  $p \geq 1$ . For smaller regions, we also imposed a minimum rank of 3 for all the smooth terms.

### 4.2 | Dispersion in the data

Figure 2A presents the distributions of estimated multiplicative dispersion  $\phi$  and additive dispersion  $\sigma_0^2$  for the test regions. Overall, widespread overdispersion is observed; 98.5% regions show multiplicative dispersion  $\phi$  greater than 1



**FIGURE 2** (A) Distribution of the estimated  $\phi$  and  $\sigma_0^2$  for all test regions. (B) QQ plot for regional  $P$ -values on chromosome 18, obtained from models addressing different types of dispersion. All results are under the zero-error assumption ( $p_0 = 1 - p_1 = 0$ ).

and 51.2% regions show additive dispersion  $\sigma_0^2$  greater than 0.05. The Pearson correlation coefficient between the estimated  $\phi$  and  $\sigma_0^2$  is  $-0.015$ . There exist 49.8% regions with both multiplicative dispersion  $\phi > 1$  and additive dispersion  $\sigma_0^2 > 0.05$ .

Figure 2B shows quantile-quantile (QQ) plots for the regional  $P$ -values for the effect of ACPA on the 292 regions of Chromosome 18. The results are compared among four different approaches: (1) dSOMNiBUS which models both the multiplicative and additive dispersion, (2) the MDO model, (3) the ADO model, and (4) the standard SOMNiBUS which ignores any extra-binomial variation. Genomic control values  $\lambda_{GC}$ <sup>55</sup> are also reported in Figure 2B;  $\lambda_{GC} \approx 1$  indicates correct control of type I error rate. These QQ plots reveal that, when ignoring either type of dispersion, the distribution of regional  $P$ -values is biased away from what would be expected under the null (ie, elevated genomic control values  $\lambda_{GC}$ ). The inclusion of both multiplicative and additive dispersion is important for correct type I error control.

### 4.3 | ACPA-associated differentially methylated genes

QQ plots in Figure 3 show that dSOMNiBUS is more powerful in identifying ACPA-associated DMRs than the existing methods we considered. Using Bonferroni thresholds for significance at a 5% family-wise error rate, dSOMNiBUS ( $p_0 = 0.003, p_1 = 0.9$ ) and dSOMNiBUS ( $p_0 = 1 - p_1 = 0$ ) identified 33 and 56 significant genes, respectively, with 23 overlapping (Table 1). The other approaches, on the other hand, failed to detect statistically significant signals. Supplementary Figures S2-S4 show the methylation proportions on the top three genes, *LINC01168*, *SPRED3* and *PLOD2*, for ACPA-positive and ACPA-negative subjects separately, along with the covariate effect curves estimated from our model. Within a target region, our software also identifies the subsections whose pointwise CIs do not include 0 (panel “ACPA” in Figures S2-S4). For the top three genes, these identified subsections indeed correspond to subregions with notable methylation differences between ACPA-positive and ACPA-negative subjects (panel B in Figures S2-S4), demonstrating that our method captures important underlying methylation patterns associated with the covariate of interest.

Gene ontology (GO) analysis was performed on the 23 overlapping genes to uncover functional biological concepts involved in ACPA positivity. Table S3 shows the over-represented GO terms with Benjamini-Hochberg-adjusted  $P$ -value  $< 0.1$  and Figure S5 shows the gene-concept network. Several immune-signaling pathways, such as leukotriene B4 receptor activity and non-membrane spanning protein tyrosine phosphatase activity are highlighted around the genes *LTB4R*, *RXFP3* and *DUSP22*. Metabolic pathways involved in collagen synthesis and degradation are highlighted around the genes *PLOD2* and *SLC2A8*. In summary, our findings highlight the importance of cell signaling and collagen metabolism in RA and point to significant protein-coding genes as potential contributors in ACPA-related differential methylation. These findings were made possible by our dSOMNiBUS model and were not readily available using the existing regional methylation techniques.

**TABLE 1** The 23 overlapping significant genes identified by both the dSOMNiBUS ( $p_0 = 0.003, p_1 = 0.9$ ) and dSOMNiBUS ( $p_0 = 1 - p_1 = 0$ ).

Gene	Chr	Start	End	# of CpGs	dSOMNiBUS non-zero error			BiSeq			GlobalTest			BSmooth			SMSC		
					P-value	Rank	Rank	P-value	Rank	Rank	P-value	Rank	Rank	P-value	Rank	Rank	P-value	Rank	Rank
LINC01168	10	134 773 524	134 779 406	240	1.15e-21 (1)	9.72e-68 (1)	9.72e-68 (1)	4.45e-02 (220)	4.47e-03 (47)	4.47e-03 (47)	1.50e-02 (160)	6.10e-02 (747)							
SPRED3	19	38 875 708	38 881 011	313	6.83e-14 (4)	6.09e-16 (7)	6.09e-16 (7)	8.82e-02 (475)	3.34e-02 (326)	3.34e-02 (326)	4.80e-02 (585)	2.00e-02 (240)							
PLOD2	3	145 878 681	145 879 725	94	8.21e-14 (5)	8.83e-19 (6)	8.83e-19 (6)	1.91e-01 (1207)	1.34e-04 (3)	1.34e-04 (3)	< 0.001 (1)	6.55e-01 (8134)							
SHARPIN	8	145 158 456	145 165 570	486	1.63e-13 (6)	2.76e-28 (3)	2.76e-28 (3)	3.97e-01 (3344)	3.92e-01 (4936)	3.92e-01 (4936)	7.47e-01 (9246)	6.93e-01 (8606)							
SMIM24	19	3 480 411	3 483 988	171	3.36e-13 (7)	1.48e-34 (2)	1.48e-34 (2)	1.12e-01 (638)	4.60e-03 (49)	4.60e-03 (49)	2.20e-02 (250)	5.90e-02 (724)							
PARD6G-AS1	18	77 905 051	77 906 126	64	3.01e-12 (8)	3.88e-13 (10)	3.88e-13 (10)	7.94e-01 (9806)	4.94e-01 (6286)	4.94e-01 (6286)	5.34e-01 (6645)	6.34e-01 (7844)							
LINC00987	12	9 392 255	9 393 068	48	4.39e-12 (9)	3.13e-13 (9)	3.13e-13 (9)	1.25e-01 (724)	2.73e-02 (258)	2.73e-02 (258)	2.16e-01 (2688)	5.54e-01 (6850)							
MRPL36	5	1 799 907	1 802 477	149	4.33e-10 (13)	2.63e-10 (15)	2.63e-10 (15)	5.87e-01 (6146)	3.61e-01 (4485)	3.61e-01 (4485)	1.79e-01 (2202)	2.93e-01 (3653)							
HLA-DRB6	6	32 551 888	32 552 731	58	5.17e-10 (14)	2.53e-09 (20)	2.53e-09 (20)	4.03e-01 (3412)	3.98e-01 (5012)	3.98e-01 (5012)	6.15e-01 (7619)	6.34e-01 (7845)							
PSMD5	9	123 605 025	123 606 053	77	1.36e-09 (15)	1.00e-08 (26)	1.00e-08 (26)	2.52e-01 (1733)	5.68e-02 (557)	5.68e-02 (557)	7.40e-02 (927)	1.09e-01 (1320)							
ANKRD18B	9	33 523 284	33 524 679	106	3.72e-08 (19)	4.91e-09 (24)	4.91e-09 (24)	2.33e-02 (135)	6.23e-03 (63)	6.23e-03 (63)	3.00e-03 (24)	8.80e-01 (10 881)							
LINC00336	6	33 560 786	33 561 449	42	1.93e-07 (21)	1.03e-07 (33)	1.03e-07 (33)	8.35e-03 (58)	1.83e-02 (170)	1.83e-02 (170)	4.60e-02 (559)	7.26e-01 (9014)							
RXFP3	5	33 935 796	33 939 022	210	4.31e-07 (22)	2.05e-08 (28)	2.05e-08 (28)	1.50e-03 (13)	1.35e-02 (132)	1.35e-02 (132)	2.60e-02 (303)	4.38e-01 (5416)							
SOWAHC	2	110 370 655	110 373 886	305	4.66e-07 (23)	7.55e-08 (31)	7.55e-08 (31)	6.40e-01 (7062)	7.69e-02 (745)	7.69e-02 (745)	1.12e-01 (1391)	5.80e-01 (7172)							
LTB4R	14	24 779 718	24 780 931	112	5.92e-07 (24)	1.24e-09 (17)	1.24e-09 (17)	4.00e-01 (3375)	1.46e-01 (1560)	1.46e-01 (1560)	2.37e-01 (2954)	4.00e-02 (486)							
LOC1002880	1	713 672	715 283	97	6.25e-07 (25)	2.47e-06 (51)	2.47e-06 (51)	5.14e-01 (4970)	3.32e-01 (4101)	3.32e-01 (4101)	1.88e-01 (2311)	5.00e-02 (612)							
FAM160A1	4	152 328 814	152 330 609	123	7.02e-07 (26)	2.72e-06 (53)	2.72e-06 (53)	5.59e-01 (5689)	1.46e-01 (1568)	1.46e-01 (1568)	4.39e-01 (5400)	5.59e-01 (6920)							
TTC28	22	29 075 610	29 076 614	119	1.09e-06 (27)	1.91e-07 (35)	1.91e-07 (35)	6.38e-02 (321)	1.64e-01 (1772)	1.64e-01 (1772)	9.20e-02 (1161)	2.67e-01 (3334)							
CIDEB	14	24 779 875	24 780 931	108	1.32e-06 (29)	2.78e-09 (21)	2.78e-09 (21)	3.54e-01 (2833)	1.40e-01 (1504)	1.40e-01 (1504)	2.11e-01 (2620)	3.40e-02 (414)							
DUSP22	6	290 588	292 526	111	1.48e-06 (30)	2.19e-10 (14)	2.19e-10 (14)	2.63e-01 (1845)	2.08e-01 (2317)	2.08e-01 (2317)	1.82e-01 (2233)	1.13e-01 (1371)							
SLC2A8	9	130 158 503	130 159 560	66	1.84e-06 (31)	4.59e-09 (23)	4.59e-09 (23)	7.27e-04 (10)	2.43e-02 (224)	2.43e-02 (224)	2.10e-02 (235)	2.00e-02 (241)							
HOXA4	7	27 169 737	27 171 618	138	2.99e-06 (32)	2.76e-07 (38)	2.76e-07 (38)	6.91e-01 (7919)	5.70e-01 (7304)	5.70e-01 (7304)	4.98e-01 (6160)	7.05e-01 (8755)							
SEPTIN10	2	110 371 382	110 373 886	238	3.44e-06 (33)	8.52e-07 (45)	8.52e-07 (45)	7.12e-01 (8298)	7.73e-02 (750)	7.73e-02 (750)	1.24e-01 (1529)	6.13e-01 (7594)							

Note: The rank in the DMR list identified by the corresponding approach is provided in brackets. The results from dmrseq are in Table S2.

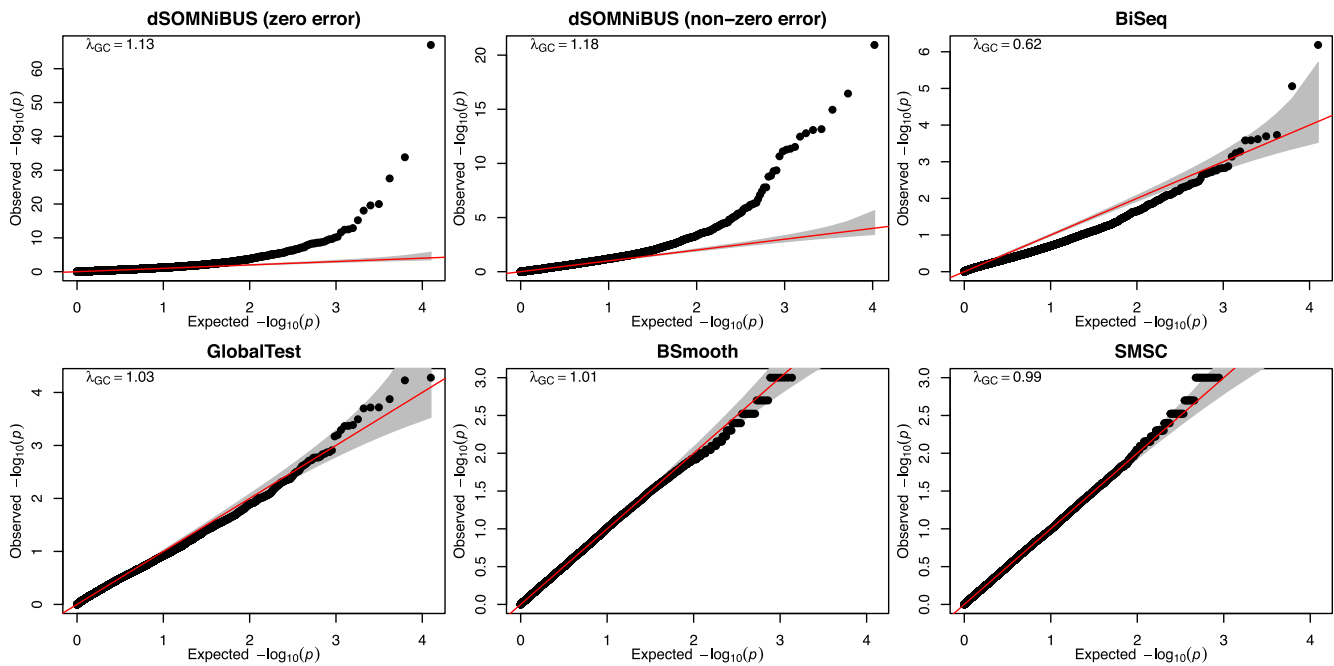


FIGURE 3 Q-Q plots for region-based  $P$ -values obtained from different methods. The results from the dmrseq method are in Table S2.

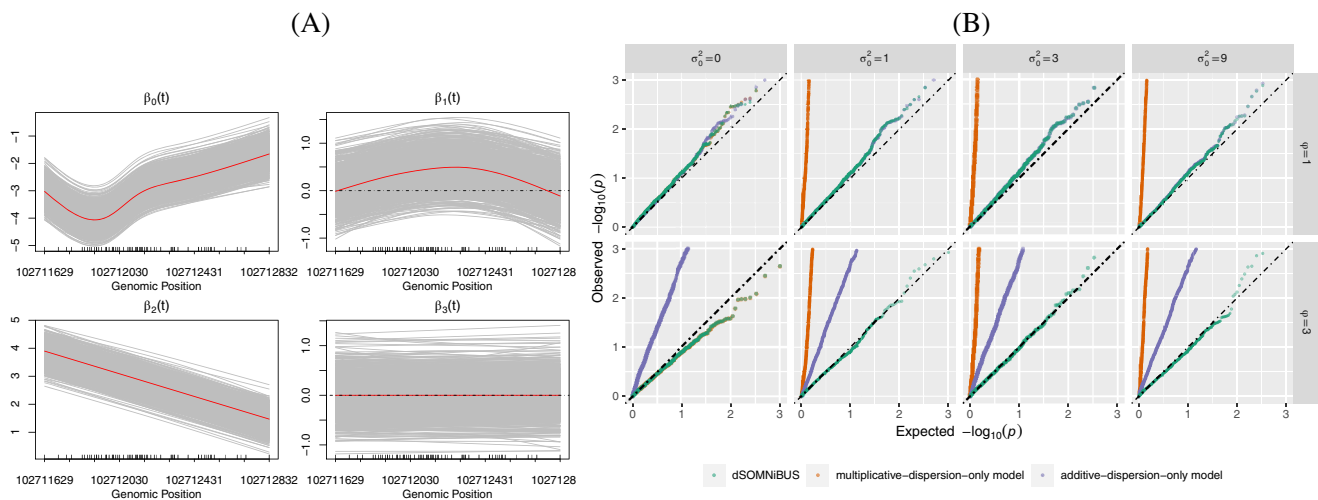
## 5 | SIMULATION STUDY

We conducted simulations to assess the proposed inference approach, and to compare the performance of our method with the five existing methods, in terms of type I error and power. The detailed descriptions of how the simulated data were generated are given in Supplementary Section 5.1. The simulated methylation regions include 123 CpGs sites. For our approach, dSOMNiBUS, we used natural cubic splines with dimension  $L_p = 5$  to parameterize the smooth terms of interest. Figure 4A presents the estimates of the functional parameters  $\beta_0(t)$ ,  $\beta_1(t)$ ,  $\beta_2(t)$  and  $\beta_3(t)$  over 1000 simulations, obtained from dSOMNiBUS. It demonstrates that the proposed method provides unbiased curve estimates for smooth covariate effects when the regional methylation counts exhibit extra-parametric variation and are measured with errors.

Figure 4B shows the QQ plots for the regional  $P$ -values under the null. The results show that ignoring the presence of additive dispersion leads to substantial estimation bias, poor CI coverage probabilities (see Figure S9) and highly inflated type I errors. Although the ADO model provides relatively accurate pointwise CIs, the distributions of its regional  $P$ -values are biased away from what would be expected under the null, when multiplicative dispersion  $\phi > 1$ . Overall, dSOMNiBUS provides pointwise CIs attaining their nominal levels, and region-based statistics whose distribution under the null is well calibrated, regardless of the types and degrees of dispersion that data exhibit. Similar results were observed when data were generated without error (Figures S10 and S11).

### 5.1 | Comparative analysis: dSOMNiBUS versus existing DMR detection methods

Figure 5 further demonstrates the performance of the proposed regional test, when compared with the existing methods GlobalTest, dmrseq, BSmooth, SMSC, and BiSeq. Here, data were simulated with error parameters  $p_0 = 0.003$  and  $1 - p_1 = 0.1$ . Figure 5A shows the distributions of  $P$ -values for the regional effect of the null covariate  $Z_3$ . Because we estimated the empirical regional  $P$ -values for BSmooth and SMSC by permutations, both methods are able to control type I errors, under all settings of  $\phi$  and  $\sigma_0^2$ . Both BiSeq and dmrseq show deflated type I error rate when  $\sigma_0^2 = 0$  and inflated type I error rate when  $\sigma_0^2 > 0$ . The distributions of  $P$ -values from GlobalTest are well calibrated when the within subject correlation  $\sigma_0^2 > 0$ , but are slightly biased away from the uniform distribution when  $\sigma_0^2 = 0$ . When  $\sigma_0^2 = 0$  and  $\phi = 3$ , dSOMNiBUS provides slightly conservative type I errors; this bias vanishes when the data were generated without error (Figure S12). Figure 5B shows the powers of the six methods for detecting DMRs under the 15 settings of



**FIGURE 4** (A) Estimates of smooth covariate effects (gray) over 1000 simulations, obtained from dSOMNiBUS. The red curves are the true functional parameters used to generate the data. Data were generated with error ( $p_0 = 0.003, p_1 = 0.9$ ), using  $\phi = 3, \sigma_0^2 = 3$  and  $N = 100$ . (B) QQ plot for regional  $P$ -values for the test  $H_0 : \beta_3(t) = 0$ , obtained from dSOMNiBUS, the MDO model and the ADO model. Data were simulated with error ( $p_0 = 0.003, p_1 = 0.9$ ), under simulation Scenario 1 (outlined in Supplementary Table S4) with  $N = 100$ . When  $\phi = 1$ , the results from dSOMNiBUS (green) and the ADO model (purple) are indistinguishable. When  $\sigma_0^2 = 0$ , the lines for the MDO model (orange) and dSOMNiBUS (green) are indistinguishable.

methylation patterns displayed in Figure S6. Here, methylation difference is defined as the maximum difference between  $\pi_1(t)$  and  $\pi_0(t)$  in the region. When data exhibit neither additive nor multiplicative dispersion, dSOMNiBUS and BSmooth provide the highest power, followed by dmrseq, BiSeq, GlobalTest, and SMSC. When  $\sigma_0^2 = 0$  and  $\phi = 3$ , BSmooth and dmrseq are more powerful than other methods. When there are correlations among methylation measurements on the same subject, that is,  $\sigma_0^2 > 0$ , dSOMNiBUS clearly outperforms the five alternative methods; this superiority remains when the data were generated without error (Figure S13). In summary, dSOMNiBUS exhibits greater power to detect DMRs, while correctly controlling type I error rates, especially when the regional methylation counts exhibit (additive) extra-binomial variation.

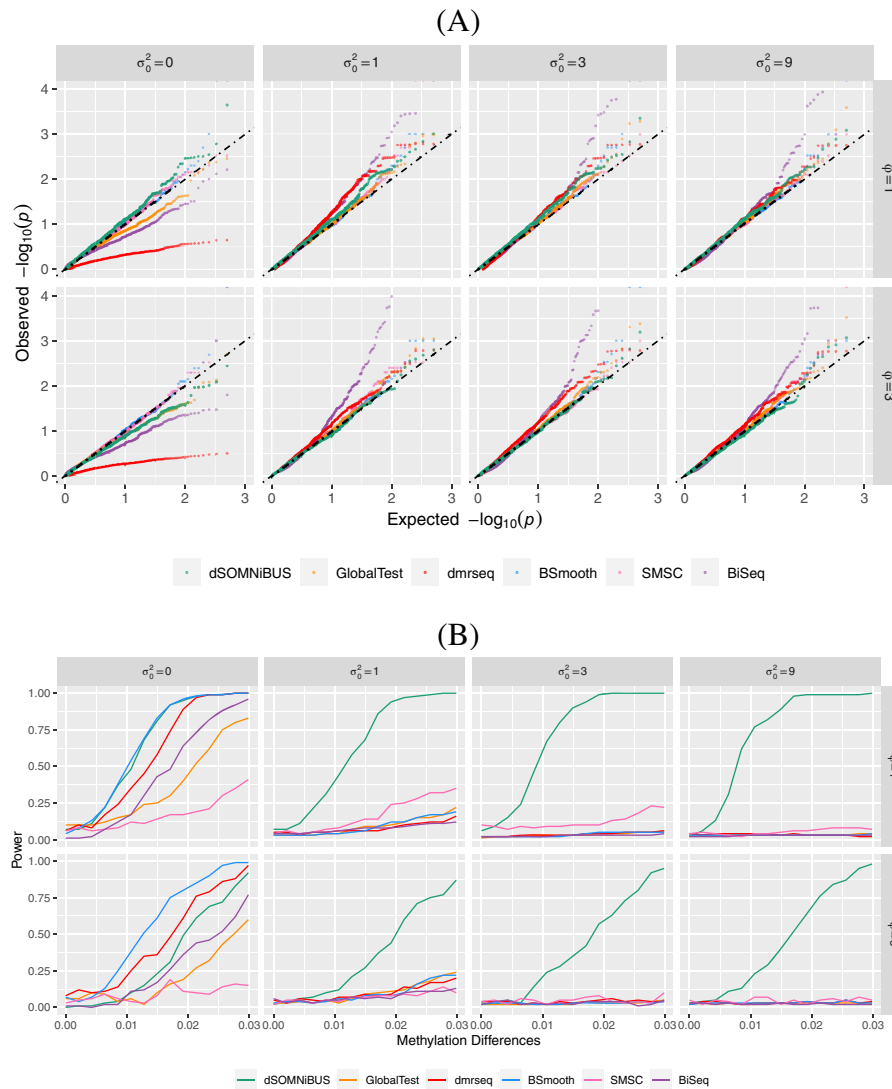
## 5.2 | Computational time

Supplementary Figure S19 presents the dSOMNiBUS computation times for estimating and inferring the functional parameters  $\beta_0(t), \beta_1(t), \beta_2(t), \beta_3(t)$ —the red curves in Figure 4A—under sample sizes of 100,300 and 500. The average runtimes, based on 10 replications, are 2.73, 49.97, and 246.45 minutes for  $N = 100, 300$  and 500, respectively. In these experiments, the spline coefficients  $\alpha$  have a dimension of  $K = \sum_{p=0}^3 L_p = 20$ . The dimension of  $\mathcal{B}$ , which incorporates both  $\alpha$  and REs  $u_i$ , is  $20 + N$ . The computational bottleneck in Algorithm 1 arises from the need to calculate the inverse of a matrix sized  $(K + N) \times (K + N)$ , an operation iteratively performed during the process of updating  $\mathcal{B}$  given  $\Theta$ . Empirical findings in Figure S19 corroborate that the computational complexity of dSOMNiBUS scales with the order of  $O((K + N)^3)$ . Despite the significant computational burden posed by larger sample sizes, experiments with increased  $N$  present enhanced statistical power (Figure S20) and greater precision in estimating  $\phi$  (Figure S21A), while maintaining the correct type I error rate (Figure S21B) and achieving nominal pointwise confidence interval coverage (Figure S21C).

## 5.3 | Sensitivity to bisulfite sequencing error parameters

To further assess the robustness of our estimation algorithm and inference methods, we conducted simulations under four different error rate scenarios: (1)  $p_0 = 0.1$  and  $1 - p_1 = 0.003$ ; (2)  $p_0 = 1 - p_1 = 0.1$ ; (3)  $p_0 = 1 - p_1 = 0.2$ ; and (4) the previously used  $p_0 = 0.003$  and  $1 - p_1 = 0.1$ . Consistent with our prior simulation and data application analyses, we





**FIGURE 5** (A) QQ plot for regional  $P$ -values for the test  $H_0 : \beta_3(t) = 0$ , obtained from different approaches. Data were simulated with error ( $p_0 = 0.003, p_1 = 0.9$ ), under simulation Scenario 1 ( $N = 100$ ). (B) Powers to detect DMRs using the six methods for the 15 simulation settings in Scenario 2 ( $N = 100$ ), calculated over 100 simulations.

operated under the assumption that the correct values of  $p_0$  and  $p_1$  were known. The results are illustrated in Supplementary Figure S22. Specifically, results demonstrate dSOMNiBUS's consistent control of type I error rate across these error scenarios, as shown in the third panel of Figure S22C. In addition, Figure S22B indicates that our 95% confidence intervals generally achieve expected coverage, with minor under-coverage observed at the boundary for  $\beta_1(t)$  at higher error rates ( $p_0 = 1 - p_1 = 0.2$ ). Moreover, the increase in error rates, as depicted in Figure S22C's second panel, leads to higher regional  $P$ -values for the non-null covariate, implying reduced power to detect DMRs in data influenced by higher error rates.

In our sensitivity analysis, we further assessed the impact of incorrectly specified  $p_0$  and  $1 - p_1$  on our inference accuracy, with detailed simulation setups in Supplementary Section 5.3. Our analysis reveals that setting  $p_0$  or  $1 - p_1$  lower than their true values leads to substantial underestimation of  $\phi$ , as demonstrated in Table S6 and Figure S23A. Conversely, setting these rates above their actual values leads to an overestimated  $\phi$ . This observed trend aligns with the behavior of our plug-in estimator, as defined in (12), which indicates that decreasing  $p_0$  or  $1 - p_1$  leads to a reduced  $\hat{\phi}$ , and vice versa.

Accurately determining  $\phi$  is critical for correctly estimating and quantifying uncertainty in  $\hat{\beta}_p(t)$ . Therefore, mis-specifying error rates—either by overstating or understating—results in biased curve estimates (Figure S24), and



poor CI coverage (Figure S23B). Notably, under-specifying  $p_0$  or  $1 - p_1$  inflates the type I error rate (Figure S1), and interestingly, boosts the power, as shown in Figure S23C and Table S5. In contrast, specifying more errors than exist leads to reduced power (Figure S23C and Table S5), highlighting the sensitivity of our inference results to the accuracy of error rate specification.

In summary, to ensure the reliability of the analytical results obtained from dSOMNiBUS, a well-grounded understanding of  $p_0$  and  $p_1$  values is required. Fortunately, this prerequisite is achievable in practice; the error parameters  $p_0$  and  $1 - p_1$  can be estimated from raw sequencing data by examining CpG sites that are known a priori to be either methylated or unmethylated.<sup>33</sup> Assuming accurate specification of  $p_0$  and  $p_1$ , our proposed inferential procedure consistently demonstrates reliable performance across various scenarios of error intensities.

## 6 | DISCUSSION

Heterogeneity in DNA methylation can be attributed to various factors beyond the trait of interest, including genetic effects, cellular heterogeneity, past exposures and environmental influences, many of which may not be directly measured. Accounting for the sources of variability across samples is crucial for identifying differentially methylated regions while avoiding false associations. This is especially important when analyzing human samples, which contain mixed cell types and exhibit significant biological variation. Methods that can achieve accurate statistical uncertainty assessment of DMR from human samples are currently lacking, particularly for sequencing-derived measures of DNA methylation.

To fill this gap, we have developed a hierarchical quasi-binomial varying coefficient mixed model, called dSOMNiBUS, for testing DMRs in BS-seq data. We applied it to investigate the association between genome-wide whole blood DNA methylation and ACPA positivity, a preclinical marker of RA risk. Among the 12 569 gene-specific methylation regions, we identified 23 genes that were determined as significant by both the dSOMNiBUS ( $p_0 = 0.003, p_1 = 0.9$ ) and dSOMNiBUS ( $p_0 = 1 - p_1 = 0$ ), using Bonferroni thresholds for significance at a 5% family-wise error rate. These 23 genes were found to be enriched in several immune signaling, collagen, and chondrocytes pathways (Figure S5). These identified ACPA-associated functional pathways further highlight five core genes *LTB4R*, *RXFP3*, *DUSP22*, *PLOD2* and *SLC2A8*. Among them, the *LTB4R* gene product is a receptor for the chemoattractant leukotriene B4, a key player in mediating inflammation.<sup>56</sup> *DUSP22* is a protein tyrosine phosphatase involved in several immune signaling pathways, and it appears to suppress autoimmunity.<sup>57</sup> *PLOD2* codes for an enzyme that catalyzes collagen cross-linking.<sup>58</sup> The gene product of *SLC2A8* is a glucose transporter, which plays a role in mediating glucose transport in chondrocytes.<sup>59</sup> Therefore, several lines of congruent evidence support our analytical results, implying that both immune signaling and collagen metabolism may play important roles in RA risk prior to the manifestation of any clinical symptoms.

These findings were made possible by the increased sensitivity of our dSOMNiBUS method and were not discovered by the 5 existing regional methylation methods considered in the Sections 4 and 5. We demonstrate that our model, which incorporates both multiplicative and additive sources of data dispersion, provides a plausible representation of realistic dispersion trends in regional methylation data. Also, we provide a formal inference for smooth covariate effects and construct a region-based statistic for the test of DMRs, where outcomes might be contaminated by errors and/or exhibit extra-parametric variations. Results from simulations show that the new method captures important underlying methylation patterns with excellent power, provides accurate estimates of covariate effects, and correctly quantifies the underlying uncertainty in the estimates. The method has been implemented in the R package SOMNiBUS, which has been published in R Bioconductor. For optimal performance with our package, we recommend targeting regions with  $\geq 20$  CpGs and an average read depth of  $\geq 10$ . Greater heterogeneity necessitates larger sample sizes for detecting subtle DNA methylation changes. We suggest referencing Figures 5B and S20 for a rough sample size calculation under desired power levels.

Our model captures dispersion in the regional count data via the combination of a subject-specific RE and a multiplicative dispersion. The latter aims to capture the extra random dispersion beyond that introduced by the subject-to-subject variation. An alternative way to add multiplicative dispersion might be to add locus-specific REs. Such model would avoid the problem of estimating  $\phi$ , but would result in a substantially increased number of REs, in which case our Laplace approximation is unlikely to provide well-founded inference.<sup>30</sup> In addition, such a model would only capture overdispersion. In contrast, our quasi-binomial mixed effect model provides an adequate representation of any kind of dispersion without much increase in computational complexity.

An extension worth exploring in the future is to model the dispersion parameter  $\phi$  as a function of covariates. For example, the methylation variation across cancer samples has been found to be higher than for normal samples.<sup>11,60</sup>

Identification of such disease-associated methylation variation changes might provide further insights into biological mechanisms. This extension would also allow modeling of the hypothesis that some individuals are more sensitive to their environment.<sup>61</sup> From the methodology point of view, our proposal of combining quasi-likelihood with random effects can be generally applied to any type of count data for a more comprehensive representation of dispersion.

## AFFILIATIONS

<sup>1</sup>Department of Mathematics and Statistics, York University, Toronto, Ontario, Canada

<sup>2</sup>Département de Mathématiques, Université du Québec à Montréal, Montreal, Quebec, Canada

<sup>3</sup>Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Quebec, Canada

<sup>4</sup>Département de Mathématiques et de Statistique, Université Laval, Quebec, Quebec, Canada

<sup>5</sup>Département de Sciences de la Décision, HEC Montréal, Montreal, Quebec, Canada

<sup>6</sup>Genomic Medicine Center, Children's Mercy, Independence, Missouri, USA

<sup>7</sup>Department of Medicine, McGill University, Montreal, Quebec, Canada

<sup>8</sup>The Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada

<sup>9</sup>Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada

<sup>10</sup>Department of Human Genetics and Gerald Bronfman Department of Oncology, McGill University, Montreal, Quebec, Canada

## FUNDING INFORMATION

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. K.Z. was supported by a Doctoral Training Award from the Fonds de Recherche du Québec - Santé (FRQS) and the McGill University Faculty of Medicine's Gerald Clavet Fellowship. K.Z. gratefully acknowledges funding via the Canadian Statistical Sciences Institute (CANSSI) through the CDPF program and the Natural Sciences and Engineering Research Council of Canada (NSERC) discovery grant RGPIN-2024-06287. We also acknowledge the Digital Research Alliance of Canada (formerly Compute Canada) Resources for Research Groups (RRG) ID 2541 and 4128.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

The method has been implemented in the R package SOMNiBUS, which has been published in R Bioconductor (<https://www.bioconductor.org/packages/release/bioc/html/SOMNiBUS.html>). The data that support the findings of this study are available from CARTaGENE. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from <https://cartagene.qc.ca/en/researchers.html> with the permission of CARTaGENE. R codes to reproduce the simulation results are provided in the public repository: [https://github.com/kaiqiong/dSOMNiBUS\\_simu](https://github.com/kaiqiong/dSOMNiBUS_simu).

## ORCID

Kaiqiong Zhao  <https://orcid.org/0000-0003-0810-8764>

Karim Oualkacha  <https://orcid.org/0000-0002-9911-079X>

Celia M. T. Greenwood  <https://orcid.org/0000-0002-2427-5696>

## REFERENCES

1. Lister R, Pelizzola M, Dowen RH, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*. 2009;462(7271):315.
2. Sims D, Sudbery I, Illott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet*. 2014;15(2):121-132.
3. Cheng L, Zhu Y. A classification approach for DNA methylation profiling with bisulfite next-generation sequencing data. *Bioinformatics*. 2013;30(2):172-179.
4. Lakhali-Chaieb L, Greenwood CM, Ouhourane M, Zhao K, Abdous B, Oualkacha K. A smoothed EM-algorithm for DNA methylation profiles from sequencing-based methods in cell lines or for a single cell type. *Stat Appl Genet Mol Biol*. 2017;16(5-6):333-347.

5. Forslind K, Ahlmén M, Eberhardt K, Hafström I, Svensson B. Prediction of radiological outcome in early rheumatoid arthritis in clinical practice: role of antibodies to citrullinated peptides (anti-CCP). *Ann Rheum Dis*. 2004;63(9):1090-1095.
6. Shao X, Hudson M, Colmegna I, et al. Rheumatoid arthritis-relevant DNA methylation changes identified in ACPA-positive asymptomatic individuals using methylome capture sequencing. *Clin Epigenetics*. 2019;11(1):110.
7. Zeng Y, Zhao K, Oros Klein K, et al. Thousands of CpGs show DNA methylation differences in ACPA-positive individuals. *Genes*. 2021;12(9):1349. doi:10.3390/genes12091349
8. Eckhardt F, Lewin J, Cortese R, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*. 2006;38(12):1378-1385.
9. Affinito O, Palumbo D, Fierro A, et al. Nucleotide distance influences co-methylation between nearby CpG sites. *Genomics*. 2020;112(1):144-150.
10. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet*. 2003;33:245.
11. Hansen KD, Timp W, Bravo HC, et al. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet*. 2011;43(8):768.
12. Rackham OJ, Langley SR, Oates T, et al. A Bayesian approach for analysis of whole-genome bisulphite sequencing data identifies disease-associated changes in DNA methylation. *Genetics*. 2017;205:1443-1458.
13. Zhao K, Ouakacha K, Lakhali-Chaieb L, et al. A novel statistical method for modeling covariate effects in bisulfite sequencing derived measures of DNA methylation. *Biometrics*. 2021;77(2):424-438.
14. Akalin A, Kormaksson M, Li S, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*. 2012;13(10):1-9.
15. Dolzhenko E, Smith AD. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. *BMC Bioinformatics*. 2014;15(1):215.
16. Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Res*. 2014;42(8):e69.
17. Park Y, Wu H. Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*. 2016;32(10):1446-1453.
18. Lea AJ, Tung J, Zhou X. A flexible, efficient binomial mixed model for identifying differential DNA methylation in bisulfite sequencing data. *PLoS Genet*. 2015;11(11):e1005650.
19. Cui S, Ji T, Li J, Cheng J, Qiu J. What if we ignore the random effects when analyzing RNA-seq data in a multifactor experiment. *Stat Appl Genet Mol Biol*. 2016;15(2):87-105.
20. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc*. 1993;88(421):9-25.
21. Molenberghs G, Verbeke G, Demétrio CG. An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Anal*. 2007;13(4):513-531.
22. Vahabi N, Kazemnejad A, Datta S. A joint overdispersed marginalized random-effects model for analyzing two or more longitudinal ordinal responses. *Stat Methods Med Res*. 2019;28(1):50-69.
23. Molenberghs G, Verbeke G, Demétrio CG, Vieira AM. A family of generalized linear models for repeated measures with normal and conjugate random effects. *Stat Sci*. 2010;25(3):325-347.
24. Molenberghs G, Verbeke G, Iddi S, Demétrio CG. A combined beta and normal random-effects model for repeated, overdispersed binary and binomial data. *J Multivar Anal*. 2012;111:94-109.
25. Ivanova A, Molenberghs G, Verbeke G. A model for overdispersed hierarchical ordinal data. *Stat Model*. 2014;14(5):399-415.
26. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodology*. 1977;39:1-38.
27. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Vol 12. Cambridge, UK: Cambridge University Press; 2003.
28. Wolfinger R. Laplace's approximation for nonlinear mixed models. *Biometrika*. 1993;80(4):791-795.
29. Rabe-Hesketh S, Skrondal A, Pickles A. Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata J*. 2002;2(1):1-21.
30. Shun Z, McCullagh P. Laplace approximation of high dimensional integrals. *J R Stat Soc B Methodol*. 1995;57(4):749-760.
31. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J R Stat Soc Series B Stat Methodology*. 2011;73(1):3-36.
32. Fletcher D. Estimating overdispersion when fitting a generalized linear model to sparse data. *Biometrika*. 2012;99(1):230-237.
33. Wreczycka K, Gosdschan A, Yusuf D, Gruening B, Assenov Y, Akalin A. Strategies for analyzing bisulfite sequencing data. *J Biotechnol*. 2017;261:105-115.
34. Parker R, Rice J. Discussion on "some aspects of the spline smoothing approach to non-parametric regression curve fitting" (by B. W. Silverman). *J R Stat Soc B Methodol*. 1985;47(1):40-42.
35. Wahba G. Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. *Approximation Theory III*. Cambridge, MA: Academic Press; 1980:905-912.
36. Wahba G. Bayesian "confidence intervals" for the cross-validated smoothing spline. *J R Stat Soc B Methodol*. 1983;45(1):133-150.
37. Silverman BW. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J R Stat Soc B Methodol*. 1985;47(1):1-21.
38. Nelder JA, Pregibon D. An extended quasi-likelihood function. *Biometrika*. 1987;74(2):221-232. doi:10.1093/biomet/74.2.221
39. Tierney L, Kadane JB. Accurate approximations for posterior moments and marginal densities. *J Am Stat Assoc*. 1986;81(393):82-86.
40. Wood SN. On p-values for smooth components of an extended generalized additive model. *Biometrika*. 2013;100(1):221-228.

41. Saha KK. Semiparametric estimation for the dispersion parameter in the analysis of over-or underdispersed count data. *J Appl Stat.* 2008;35(12):1383-1397.
42. Wood SN, Pya N, Säfken B. Smoothing parameter and model selection for general smooth models. *J Am Stat Assoc.* 2016;111(516):1548-1563.
43. Wood SN. *Generalized Additive Models: an Introduction with R.* Boca Raton, FL: CRC Press; 2017.
44. Nocedal J, Wright SJ. *Numerical Optimization.* New York: Springer; 1999.
45. Elashoff M, Ryan L. An EM algorithm for estimating equations. *J Comput Graph Stat.* 2004;13(1):48-65.
46. Lindsay B. Conditional score functions: some optimality results. *Biometrika.* 1982;69(3):503-512.
47. Heyde C, Morton R. Quasi-likelihood and generalizing the EM algorithm. *J R Stat Soc B Methodol.* 1996;58(2):317-327.
48. Small CG, Christopher G, Wang J. *Numerical Methods for Nonlinear Estimating Equations.* Vol 29. Oxford, UK: Oxford University Press; 2003.
49. Hebestreit K, Dugas M, Klein HU. Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. *Bioinformatics.* 2013;29(13):1647-1653.
50. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biol.* 2012;13(10):R83.
51. Korthauer K, Chakraborty S, Benjamini Y, Irizarry RA. Detection and accurate false discovery rate control of differentially methylated regions from whole genome bisulfite sequencing. *Biostatistics.* 2019;20(3):367-383.
52. Goeman JJ, Van De Geer SA, Van Houwelingen HC. Testing against a high dimensional alternative. *J R Stat Soc Series B Stat Methodology.* 2006;68(3):477-493.
53. Prochenka A, Pokarowski P, Gasperowicz P, et al. A cautionary note on using binary calls for analysis of DNA methylation. *Bioinformatics.* 2015;31(9):1519-1520.
54. Hudson M, Bernatsky S, Colmegna I, et al. Novel insights into systemic autoimmune rheumatic diseases using shared molecular signatures and an integrative analysis. *Epigenetics.* 2017;12(6):433-440.
55. Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* 1999;55(4):997-1004.
56. Mathis S, Jala VR, Haribabu B. Role of leukotriene B4 receptors in rheumatoid arthritis. *Autoimmun Rev.* 2007;7(1):12-17.
57. Li JP, Yang CY, Chuang HC, et al. The phosphatase JKAP/DUSP22 inhibits T-cell receptor signalling and autoimmunity by inactivating Lck. *Nat Commun.* 2014;5(1):1-13.
58. Slot AJ, Zuurmond AM, Bardoel AF, et al. Identification of PLOD2 as telopeptide lysyl hydroxylase, an important enzyme in fibrosis. *J Biol Chem.* 2003;278(42):40967-40972.
59. Goldring MB, Marcu KB. Cartilage homeostasis in health and rheumatic diseases. *Arthritis Res Ther.* 2009;11(3):1-16.
60. Schoofs T, Rohde C, Hebestreit K, et al. DNA methylation changes are a late event in acute promyelocytic leukemia and coincide with loss of transcription factor binding. *Blood.* 2013;121(1):178-187.
61. Meaney MJ, Szyf M. Environmental programming of stress responses through DNA methylation: life at the interface between a dynamic environment and a fixed genome. *Dialogues Clin Neurosci.* 2005;7(2):103.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Zhao K, Oualkacha K, Zeng Y, et al. Addressing dispersion in mis-measured multivariate binomial outcomes: A novel statistical approach for detecting differentially methylated regions in bisulfite sequencing data. *Statistics in Medicine.* 2024;43(20):3899-3920. doi: 10.1002/sim.10149