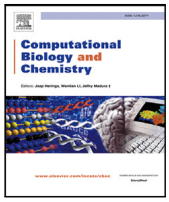
Computational Biology and Chemistry 111 (2024) 108094

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac



Research Article

A varying-coefficient model for the analysis of methylation sequencing data

Katarzyna Górczak [a,b], Tomasz Burzykowski [a,c,d], Jürgen Claesen [a,e,*]

[a] Data Science Institute, Hasselt University, Belgium
[b] Open Analytics NV, Antwerp, Belgium
[c] Department of Biostatistics and Medical Informatics, Medical University of Bialystok, Poland
[d] International Drug Development Institute (IDDI), Belgium
[e] Department of Epidemiology and Data Science, Amsterdam UMC, VU Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Methylation sequencing
CpG site
Differentially methylated regions
Beta-binomial model
Varying-coefficient model
Smoothing splines

ABSTRACT

DNA methylation is an important epigenetic modification involved in gene regulation. Advances in the next generation sequencing technology have enabled the retrieval of DNA methylation information at single-base-resolution. However, due to the sequencing process and the limited amount of isolated DNA, DNA-methylation-data are often noisy and sparse, which complicates the identification of differentially methylated regions (DMRs), especially when few replicates are available. We present a varying-coefficient model for detecting DMRs by using single-base-resolved methylation information. The model simultaneously smooths the methylation profiles and allows detection of DMRs, while accounting for additional covariates. The proposed model takes into account possible overdispersion by using a beta-binomial distribution. The overdispersion itself can be modeled as a function of the genomic region and explanatory variables. We illustrate the properties of the proposed model by applying it to two real-life case studies.

1. Introduction

DNA methylation is a well-studied epigenetic mechanism in which a methyl group is added onto a nucleotide, commonly cytosine (C) (Moore et al., 2013; Reinders et al., 2008). The majority of cytosines undergoing methylation precede guanine (G), forming a *CpG site* (Beck et al., 2022). In the last decades, studies have highlighted the importance of DNA methylation in regulating gene expression and cellular function (Hudson et al., 2017; Lister et al., 2009; Ziller et al., 2013). Moreover, patterns of this biochemical process allow characterizing many diseases, such as cancer (Shafi et al., 2018; Irizarry et al., 2009), diabetes (Lu et al., 2022; Bansal and Pinney, 2017), Alzheimer's disease, and autoimmune disorders (Hudson et al., 2017).

DNA methylation can be profiled and quantified at single-nucleotide resolution with bisulfite sequencing. Using this technique, DNA is treated with sodium bisulfite which converts unmethylated cytosine to uracil (U), while leaving methylated cytosine unchanged. The methylated and unmethylated variants of cytosine are then sequenced, quantified, and summarized as counts of methylated and unmethylated cytosines at any given site (Robinson et al., 2014). However, none of the existing methods can identify 100% of DNA methylation. Incomplete conversion, which may be affected by the quality and quantity of purified DNA, can lead to incorrect amount of methylation in a sample (Olova et al., 2018). The biases and limitations of DNA methylation have been addressed in several studies (Beck et al., 2022; Olova et al., 2018; Gong et al., 2022).

Bisulfite sequencing protocols can be implemented on a genome-wide scale or on a set of targeted regions. Both approaches have pros and cons, and selecting the correct one may often depend on the biological questions and availability of the DNA amount (Gong et al., 2022; Moser et al., 2020). Bisulfite sequencing combined with next-generation sequencing (NGS) has become more accessible and cost-efficient as the costs of NGS have dropped dramatically over recent years (Gong et al., 2022; Olova et al., 2018). Lower costs enable comparison of different biological conditions with higher power (Hebestreit et al., 2013). Moreover, a larger number of samples allows studying progressive age-related changes in DNA methylation profiles (Klein and Hebestreit, 2016; Bergman and Cedar, 2013; Maegawa et al., 2010).

Several reviews of statistical methodologies for differential methylation (DM) analysis have been published (Robinson et al., 2014; Shafi et al., 2018; Klein and Hebestreit, 2016). These statistical approaches include methods to identify differentially methylated CpG sites (DMCs) or regions (DMRs). Shafi et al. (2018) provide an extensive overview that points out important factors which should be taken into account in

* Correspondence to: Department of Epidemiology and Data Science, Amsterdam UMC, VU Amsterdam, De Boelelaan 1089A, Postcode 1081 HV, Amsterdam, The Netherlands.
E-mail address: j.claesen@amsterdamumc.nl (J. Claesen).

https://doi.org/10.1016/j.compbiolchem.2024.108094
Received 6 March 2024; Received in revised form 6 May 2024; Accepted 8 May 2024
Available online 18 May 2024

DM analysis, such as biological variation within replicates, sequencing depth, or spatial correlation between the methylation levels of the CpG sites.

Identification of DMCs focuses on estimating differences between groups of samples by using aggregated counts within each group. Such methods (like, e.g., Fisher's exact test) do not take biological variability into account, which can lead to a high number of false positives (Robinson et al., 2014; Hansen et al., 2012; Ziller et al., 2013; Klein and Hebestreit, 2016).

Although some methods directly determine DMRs (Zhao et al., 2021), the vast majority of methods are two-step approaches. First, individual statistically-significant DMCs are identified. Then they are subsequently combined into DMRs based on certain criteria (Hebestreit et al., 2013; Shafi et al., 2018; Hansen et al., 2012).

In the last decade, several smoothing-based approaches aimed at finding DMRs have been proposed (Shafi et al., 2018; Hansen et al., 2012; Zhao et al., 2021). These methods not only account for the spatial correlation between neighboring CpG sites, but also for the sequencing coverage variability and missingness of CpG sites. However, most of these methods model the smoothing function for each sample separately without combining information across samples (Zhao et al., 2021). To the best of our knowledge, there is only one method, SOMNiBUS, that directly identifies DMRs across multiple sample while estimating covariate effects (Zhao et al., 2021). However, in its basic form, this method assumes that the methylated counts are binomially-distributed which is only true in the absence of any biological or technical replicates (Shafi et al., 2018). It is possible to extend the method by applying a quasi-binomial distribution that allows taking into account overdispersion. However, the overdispersion is assumed to be constant, which may be a too restrictive assumption.

In this article, we develop a varying-coefficient model to detect DMRs while adjusting for covariates. In particular, we assume that the methylated counts follow a beta-binomial distribution, what allows accounting for possible overdispersion. Moreover, the proposed model allows making the overdispersion a function of covariates.

2. Materials and methods

2.1. Data

The proposed varying-coefficient models will be illustrated on two datasets, i.e., a rheumatoid arthritis (RA) case study (Hudson et al., 2017), and a colon cancer case study (Hansen et al., 2011).

2.1.1. Rheumatoid arthritis case study

In this study, methylation sequencing data were gathered from immune cells of 43 individuals. Next to methylation sequencing data, information on the RA-status (no RA or RA) and the type of immune-cell (monocyte or T-cell) was gathered. Fig. 1 presents the data for a region of chromosome 4 (from 102,711,629 to 102,712,032) with 123 unique methylation sites. This region is known to show cell-type specific methylation (Hillier et al., 2005). A subset of the data is part of the SOMNiBUS package (Zhao et al., 2021).

2.1.2. Colon cancer case study

Whole genome bisulfite sequencing data were gathered from matched colon cancer and normal colon of three individuals. The processed data, which includes all annotated CpG sites on human chromosome 21 and 22, is a part of the dataBsseq R package (Hansen, 2020). CpG sites with at least 2x coverage in at least two cancer and two normal samples were analyzed. Fig. 2 presents the data for chromosome 21 with 247,876 unique methylation sites.

2.2. Methodology

For an ith methylation site ($i = 1, 2, \ldots, I$), we consider information about its chromosomal position, x_i (in terms of the number of nucleotides from a starting position), the total number of reads mapped to this site, n_i, and the number of reads with a methylated cytosine at this position, Y_i. We assume that the distribution of Y_i is beta-binomial with parameters π_i and σ_i:

$$P(Y_i = y) = \frac{\Gamma(n_i+1)\Gamma(\sigma_i)\Gamma(y+\pi_i\sigma_i)\Gamma\{n_i - y + (1-\pi_i)\sigma_i\}}{\Gamma(y+1)\Gamma(n_i - y + 1)\Gamma(\pi_i\sigma_i)\Gamma\{(1-\pi_i)\sigma_i\}\Gamma(n_i+\sigma_i)}. \quad (1)$$

In particular, the mean value and variance of Y_i are given, respectively, by

$$\begin{aligned} E(Y_i) &= n_i\pi_i, \\ Var(Y_i) &= n_i\pi_i(1-\pi_i)\left\{1 + \frac{n_i - 1}{1 + 1/\sigma_i}\right\}. \end{aligned} \quad (2)$$

Thus, π_i is the methylation probability, while the scale parameter $\sigma_i > 0$ can be seen as capturing overdispersion (relative to the binomial variation given by $n_i\pi_i(1-\pi_i)$). Note that the total overdispersion is a function of σ_i and n_i. Hence, the overdispersion varies for CpG sites with different number of reads. Table S1 provides an overview of model parameters and their interpretation.

2.2.1. Varying-coefficient model

The relationship between the methylation probability, π_i, and the chromosomal location, x_i, can be flexibly estimated by using smoothing splines (Green and Silverman, 1994; Ruppert et al., 2003):

$$\begin{aligned} \text{logit}(\pi_i) = \ln\frac{\pi_i}{1-\pi_i} &= s(x_i) \\ &= \beta_0 + \beta_1 x_i + \cdots + \beta_d x_i^d + \sum_{k=1}^{K} u_k b_k(x_i), \end{aligned} \quad (3)$$

where d is the degree of the spline, K is the number of knots, $b_k(x)$ is the set of spline basis-functions, and β and u_k are the coefficients. Note that we can express (3) in the following form:

$$\begin{aligned} s(x_i) &= (1, x_i, \ldots, x_i^d, b_1(x_i), \ldots, b_K(x_i))(\beta_0, \beta_1, \ldots, \beta_d, u_1, \ldots, u_K)' \\ &= \boldsymbol{c}'_{x_i}(\boldsymbol{\beta}', \boldsymbol{u}')' = \boldsymbol{c}'_{x_i}\boldsymbol{\theta}, \end{aligned} \quad (4)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{u}')'$.

We propose to use varying-coefficient models (Hastie and Tibshirani, 1993) to assess the effect of explanatory variables on the methylation probability and to detect DMRs. Varying-coefficient models allow including multiple smoothing splines in combination with interactions between smoothing splines and covariates. In particular, we consider models of the following form:

$$\text{logit}(\pi_i) = \boldsymbol{z}'\boldsymbol{\gamma} + \sum_{j=0}^{J} s_j(x_i)\cdot z_j, \quad (5)$$

where $\boldsymbol{z}$ is the column-vector with $J+1$ coordinates (including $z_0 = 1$ as the first coordinate) corresponding to the explanatory variables describing a particular sample, $\boldsymbol{\gamma}$ is the corresponding vector of coefficients, and $s_j(x_i)\cdot z_j$ is the interaction between a smoothing spline $s_j(x_i)$ and the jth covariate (coordinate) z_j. Note that, in the case of a factor with m levels, vector $\boldsymbol{z}$ should include values of $m-1$ dummy binary covariates coding the levels of the factor. Consequently, $m-1$ separate smooth curves are fitted for such a factor.

Model (5) specifies that the effect of z_j depends on the chromosomal position x_i and takes the form described by the smooth function $s_j(x_i)$. The term $\boldsymbol{z}'\boldsymbol{\gamma}$ is included in the model to ensure centering and identifiability of the smooth functions (Wood, 2017). However, depending on the chosen parameterization of the model, the term may or may not have to be included in the model equation (see Section 2.2.5).

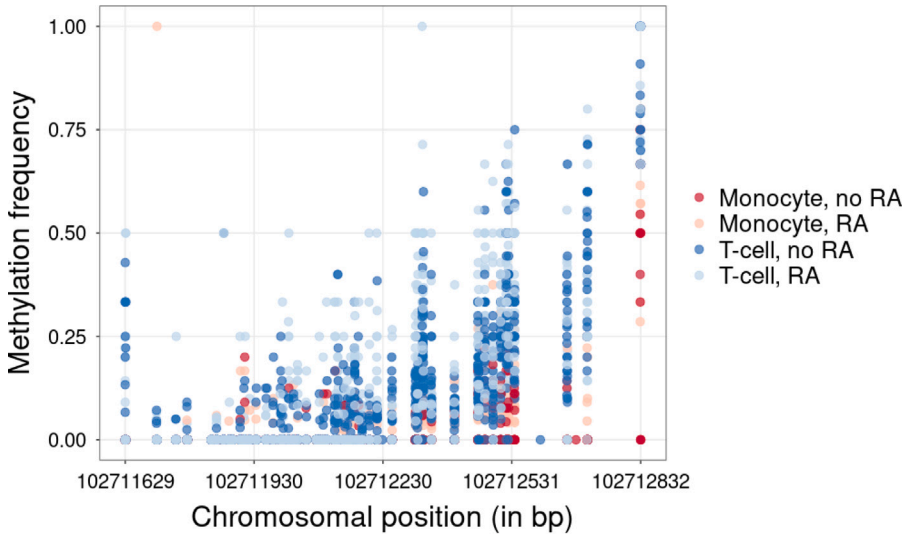


Fig. 1. Rheumatoid arthritis (RA) data. Each point represents methylation frequency for a genomic position for one sample. The points are colored based on the type of immune-cell (red: Monocyte, blue: T-cell). Transparency is added based on the RA-status (dark: no RA, light: RA).

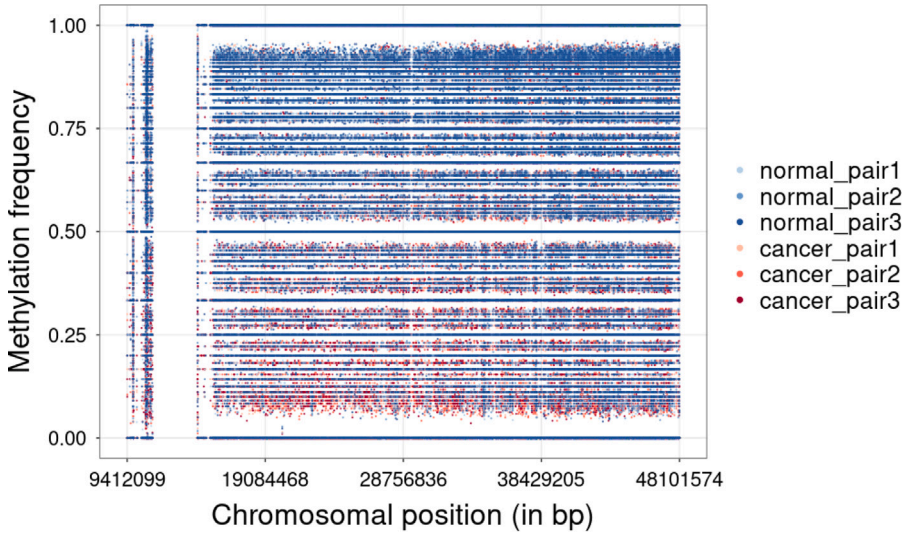


Fig. 2. Colon cancer data. Each point represents methylation frequency for a genomic position across chromosome 21 for one sample. The points are colored based on the type of sample type (red: cancer, blue: normal). Transparency is added according to the individual (light: individual 1, medium dark: individual 2, dark: individual 3).

In the case of the beta-binomial distribution (1), model (5) can be extended by making the scale parameter σ_i a function of explanatory variables (Nelder and Pregibon, 1987; Rigby and Stasinopoulos, 1996b,a). In particular, the model then takes the following form:

$$\text{logit}(\pi_i) = \boldsymbol{z}'_{\pi}\boldsymbol{\gamma}_{\pi} + \sum_{j=0}^{J_{\pi}} s_j(x_i) \cdot z_{\pi,j}, \tag{6}$$

$$\log(\sigma_i) = \boldsymbol{z}'_{\sigma}\boldsymbol{\gamma}_{\sigma} + \sum_{j=0}^{J_{\sigma}} q_j(x_i) \cdot z_{\sigma,j}, \tag{7}$$

with an obvious extension of the notation used for model (5).

2.2.2. Model fitting

Model (5) can be estimated by maximization of the penalized log-likelihood

$$l_p(\boldsymbol{\gamma}, \boldsymbol{\vartheta}, \boldsymbol{\lambda}) = l(\boldsymbol{\gamma}, \boldsymbol{\vartheta}) - \frac{1}{2}\sum_{j=0}^{J} \boldsymbol{\theta}'_j \boldsymbol{G}_j(\lambda_j)\boldsymbol{\theta}_j, \tag{8}$$

where $l(\boldsymbol{\gamma}, \boldsymbol{\vartheta})$ is the logarithm of the marginal beta-binomial likelihood (1), $\boldsymbol{\vartheta} = (\boldsymbol{\theta}'_0, \ldots, \boldsymbol{\theta}'_J)'$, $\boldsymbol{\lambda} = (\lambda'_0, \ldots, \lambda'_J)'$, and $\boldsymbol{G}_j(\lambda_j)$ is a symmetric matrix depending on hyperparameter λ_j that controls the smoothness of the estimated function $s_j(x_i)$ (Rigby and Stasinopoulos, 2005).

Maximizing the penalized log-likelihood is equivalent to maximizing the posterior log-likelihood for $\boldsymbol{\gamma}$ and $\boldsymbol{\vartheta}$, given $\boldsymbol{\lambda}$ and data $\boldsymbol{y}$ (Rigby and Stasinopoulos, 2005):

$$\log\{f(\boldsymbol{\gamma}, \boldsymbol{\vartheta}|\boldsymbol{y}, \boldsymbol{\lambda})\} = \log\{f(\boldsymbol{y}|\ \boldsymbol{\gamma}, \boldsymbol{\vartheta})\} + \log\{f(\boldsymbol{\vartheta}|\boldsymbol{\lambda})\},$$

with $\boldsymbol{\theta}_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}_j^-(\lambda_j))$, where $\boldsymbol{G}_j^-(\lambda_j)$ the generalized inverse of $\boldsymbol{G}_j(\lambda_j)$.

Maximization of the penalized log-likelihood or posterior log-likelihood yields an estimate $\widehat{\boldsymbol{\vartheta}}$ of $\boldsymbol{\vartheta}$. The smoothing parameters λ_j can be determined by various methods such as (generalized) cross-validation or (restricted) maximum likelihood estimation (Wood, 2000; Rigby and Stasinopoulos, 2005).

The variance–covariance matrix $\boldsymbol{V}_{\vartheta}$ of $\widehat{\boldsymbol{\vartheta}}$ can be estimated by using the expression of the variance–covariance matrix of the posterior density of $\boldsymbol{\vartheta}$ (Wood, 2013).

Fitting the model without penalization reduces the formula (8) to the logarithm of the marginal beta-binomial likelihood.

By using the obvious extension of the notation used in (8), the penalized log-likelihood for model (6)–(7) is defined as follows:

$$l_p(\boldsymbol{\gamma}, \boldsymbol{\vartheta}, \boldsymbol{\lambda}) = l(\boldsymbol{\gamma}, \boldsymbol{\vartheta}) - \frac{1}{2}\sum_{j=0}^{J_\pi} \boldsymbol{\theta}'_{\pi,j} \boldsymbol{G}_{\pi,j}(\lambda_{\pi,j})\boldsymbol{\theta}_{\pi,j} - \frac{1}{2}\sum_{j=0}^{J_\sigma} \boldsymbol{\theta}'_{\sigma,j} \boldsymbol{G}_{\sigma,j}(\lambda_{\sigma,j})\boldsymbol{\theta}_{\sigma,j}.$$

2.2.3. Simultaneous confidence bands

Based on the estimated model (5) or (6)–(7), smooth functions describing the change of the methylation probability across chromosomal positions can be estimated, plotted, and used for finding DMRs. As point-wise confidence intervals (CIs) do not capture the joint uncertainty in the estimation of the functions across many positions, we propose to use simultaneous confidence bands (CBs) for this purpose (Ruppert et al., 2003). These CBs are corrected for multiplicity and account for the serial correlation. A 100(1-α)% simultaneous CB for $s(x)$ is defined as follows:

$$\hat{s}(x) \pm c_{1-\alpha} \times \widehat{\text{st.dev}}\,\{\hat{s}(x) - s(x)\}$$

where $\hat{s}(x)$ denotes the estimated smoothed curve, $\widehat{\text{st.dev}}\,\{\hat{s}(x) - s(x)\}$ denotes the estimated standard deviation of the curve, and $c_{1-\alpha}$ is the $1-\alpha$ quantile of the random variable

$$\sup_{x\in\chi}\left|\frac{\hat{s}(x) - s(x)}{\widehat{\text{st.dev}}\{\hat{s}(x) - s(x)\}}\right| \approx \max_{1\le i\le I}\left|\frac{\boldsymbol{c}'_{x_i}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)}{\widehat{\text{st.dev}}\{\hat{s}(x_i) - s(x_i)\}}\right|, \tag{9}$$

where $\boldsymbol{c}_{x_i}$ and $\boldsymbol{\theta}$ are defined in (4).

The quantile can be found by, first, simulating multiple values from an approximate multivariate normal distribution

$$\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \sim \mathcal{N}\left(\mathbf{0}, \hat{\boldsymbol{V}}_\theta\right),$$

where $\hat{\boldsymbol{V}}_\theta$ is the estimated variance–covariance matrix of $\hat{\boldsymbol{\theta}}$. Subsequently, for each simulated value, the maximum statistic, defined on the right-hand-side of (9), is computed. Finally, the so-obtained sample of the maximum statistics is used to estimate $c_{1-\alpha}$ by selecting the appropriate rank-statistic.

2.2.4. Detection of differentially methylated regions

DMRs between two groups of interest, defined by values u and v of an explanatory continuous variable (covariate) z_j, can be found by deriving the difference curve

$$\hat{\Delta}(u - v, x_i) = \hat{s}_j(x_i)\cdot u - \hat{s}_j(x_i)\cdot v = (u - v)\boldsymbol{c}'_{x_i}\hat{\boldsymbol{\theta}}_j, \tag{10}$$

where $\hat{\boldsymbol{\theta}}_j = \left(\hat{\boldsymbol{\beta}}'_j, \hat{\mathbf{u}}'_j\right)'$ are the estimated coefficients that yield $\hat{s}_j(x_i)$. Variance of the difference curve at x_i can be estimated by

$$\widehat{\text{Var}}\{\hat{\Delta}(u - v, x_i)\} = (u - v)^2 \boldsymbol{c}'_{x_i}\hat{\boldsymbol{V}}_{\theta_j}\boldsymbol{c}_{x_i},$$

where $\hat{\boldsymbol{V}}_{\theta_j}$ is the estimated variance–covariance matrix of $\hat{\boldsymbol{\theta}}_j$.

For a factor, the difference curve for two levels, represented in vector $\boldsymbol{z}$ by two dummy covariates z_j and z_k, say, is given by

$$\Delta(j, k, x_i) = \hat{s}_j(x_i) - \hat{s}_k(x_i) = \boldsymbol{c}'_{x_i}\left(\hat{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}_k\right), \tag{11}$$

where $\hat{\boldsymbol{\theta}}_j$ and $\hat{\boldsymbol{\theta}}_k$ are the estimated coefficients that yield $\hat{s}_j(x_i)$ and $\hat{s}_k(x_i)$, respectively. Its variance can be estimated by

$$\widehat{\text{Var}}\{\Delta(j, k, x_i)\} = \left(\boldsymbol{c}'_{x_i}, -\boldsymbol{c}'_{x_i}\right)\hat{\boldsymbol{V}}_{\theta_j,\theta_k}\left(\boldsymbol{c}'_{x_i}, -\boldsymbol{c}'_{x_i}\right)'.$$

Subsequently, a simultaneous confidence band for the estimated difference curve is obtained by the approach outlined in Section 2.2.3. DMRs are identified by the chromosomal positions at which the simultaneous CB around the estimated difference curve excludes 0.

2.2.5. Rheumatoid arthritis case study

For the RA study introduced in Section 2.1.1, we considered varying-coefficient models with two explanatory factors, 'RA' (with 0 and 1 indicating, respectively, control and RA samples) and 'Cell-type' (with 0 and 1 indicating, respectively, monocyte and T-cell), and their interaction.

We used five knots to make the results comparable with the ones from SOMNIBUS package. In the primary analysis, we used the penalized log-likelihood (8); as a sensitivity analysis, we also fitted the models without penalization.

We implemented the models by using the R-packages *mgcv* (Wood, 2017) and *gamlss* (Rigby and Stasinopoulos, 2005) (the code can be found in the Supplementary Materials). Both packages offer various ways to specify the models. The most verbose manner is to use variables 'RA' and 'Cell-type' as factors with non-ordered levels. In this case, a separate smooth function for each of the four groups of observations is used to describe the dependence of the methylation probability on the chromosomal position. A potential disadvantage of this implementation is a relatively large number of parameters that have to be estimated. Additionally, in this case, differences between levels of a factor have to be derived from the estimated curves (see Section 2.2.4).

A potentially more parsimonious implementation is obtained by considering variables 'RA' and 'Cell-type' as factors with ordered levels. In that case, for each factor, a difference curve between a reference level and the other level is directly estimated.

Note that both implementations require inclusion of the term $\boldsymbol{z}'\boldsymbol{\gamma}$ in the model to ensure centering and identifiability of the smoothing curves.

A third approach, which we adopted, is to treat variables 'RA' and 'Cell-type' as numeric (continuous). In that case, smooth curves describing the main effect of each factor, as well as their interaction, are directly estimated. Moreover, the centering of the smooths by adding $\boldsymbol{z}'\boldsymbol{\gamma}$ to the model terms is no longer necessary, yielding a more parsimonious model. As a result, the following model is obtained:

$$\begin{aligned} Y_i &\sim \text{BetaBinomial}(n_i, \pi_i, \sigma_i), \\ \text{logit}(\pi_i) &= s_0(x_i) + s_1(x_i)\text{RA} + s_2(x_i)\text{Cell-type} + s_3(x_i)\{\text{RA}\times\text{Cell-type}\}, \\ \sigma_i &\equiv \sigma, \end{aligned} \tag{12}$$

where 'RA' and 'Cell-type' are dummy binary covariates indicating, respectively, RA and T-cells. In this model, $s_0(x_i)$ is a smooth function describing the change of the methylation probability across chromosomal positions for individuals without RA and cells derived from monocytes, $s_1(x_i)$ and $s_2(x_i)$ are the main effects of 'RA' and 'Cell-type', respectively, and $s_3(x_i)$ is the interaction effect.

Note that, in model (12), the scale parameter, σ_i, is assumed to be constant. To check this assumption, the following extended model was also considered:

$$\begin{aligned} Y_i &\sim \text{BetaBinomial}(n_i, \pi_i, \sigma_i), \\ \text{logit}(\pi_i) &= s_0(x_i) + s_1(x_i)\text{RA} + s_2(x_i)\text{Cell-type} + s_3(x_i)\{\text{RA}\times\text{Cell-type}\}, \\ \log(\sigma_i) &= q_0(x_i) + q_1(x_i)\text{RA} + q_2(x_i)\text{Cell-type} + q_3(x_i)\{\text{RA}\times\text{Cell-type}\}. \end{aligned} \tag{13}$$

In this model, σ_i is a function of the chromosomal position and the covariates.

For comparison purposes, we also considered the binomial model corresponding to (12):

$$\begin{aligned} Y_i &\sim \text{Binomial}(n_i, \pi_i), \\ \text{logit}(\pi_i) &= s_0(x_i) + s_1(x_i)\text{RA} + s_2(x_i)\text{Cell-type} + s_3(x_i)\{\text{RA}\times\text{Cell-type}\} \end{aligned} \tag{14}$$

2.2.6. Colon cancer case study

For the cancer study introduced in Section 2.1.2, we considered a varying-coefficient model with three explanatory factors, 'Type' (with 0 and 1 indicating, respectively, colon cancer and normal colon samples), 'Pair2' (with 0 and 1 indicating, respectively, matched samples of

Table 1
Results of the estimation of the "Binomial" and "Beta-Binomial" models. edf — empirical degrees of freedom; ref.df — reference degrees of freedom.

Smooth terms	With interaction							
	Binomial				Beta-Binomial			
	edf	ref.df	F-value	*p*-value	edf	ref.df	F-value	*p*-value
$s_0(x_i)$	2.466	2.861	56.884	<2e−16	1.184	1.337	12.467	7.49e−5
$s_1(x_i)$RA	3.352	3.790	5.777	0.002	2.000	2.000	0.442	0.643
$s_2(x_i)$Cell-type	4.964	4.985	70.508	<2e−16	4.903	4.989	14.831	<2e−16
$s_3(x_i)$\{RA × Cell-type\}	2.000	2.000	1.940	0.143	2.244	2.445	0.048	0.958
σ	Est.:	NA			Est.:	1.039		
Smooth terms	Without interaction							
	Binomial				Beta-Binomial			
	edf	ref.df	F-value	*p*-value	edf	ref.df	F-value	*p*-value
$s_0(x_i)$	2.460	2.853	111.88	<2e−16	1.036	1.071	47.019	<2e−16
$s_1(x_i)$RA	3.261	3.699	15.13	<2e−16	2.000	2.000	2.225	0.108
$s_2(x_i)$Cell-type	4.963	4.984	179.66	<2e−16	4.895	4.991	30.080	<2e−16
σ	Est.:	NA			Est.:	1.040		

patient 2) and 'Pair3' (with 0 and 1 indicating, respectively, matched samples of patient 3).

The following model is fitted:

$$\begin{aligned} Y_i &\sim \text{BetaBinomial}(n_i, \pi_i, \sigma_i), \\ \text{logit}(\pi_i) &= s_0(x_i) + s_1(x_i)\text{Type} + s_2(x_i)\text{Pair2} + s_3(x_i)\text{Pair3}, \\ \sigma_i &\equiv \sigma, \end{aligned} \tag{15}$$

where 'Type', 'Pair2' and 'Pair3' are dummy binary covariates indicating, respectively, colon type, patient 2 and patient 3 samples. The inclusion of 'Pair2' and 'Pair3' covariates adjusts for a potential individual-patient clustering (matching) effect. In case of a dataset with more than three patients, the use of an individual-patient random effect could be considered; model (15) can be seen as a fixed-effect counterpart. In this model, $s_0(x_i)$ is a smooth function describing the change of the methylation probability across chromosomal positions for individuals with normal colon samples, and $s_1(x_i)$, $s_2(x_i)$, and $s_3(x_i)$ are the main effects of 'Type', 'Pair2' and 'Pair3', respectively.

Following the approach recommended in the SOMNiBUS package, for each region, we used at most 50 knots, which is the number approximately equal to the number of the unique CpGs in the analyzed region (at most 1000) divided by 20. In the primary analysis, we fitted the model without penalization; as a sensitivity analysis, we also fitted them by using the penalized log-likelihood (8).

3. Results

In this section, we present the results of the analysis of the case studies. To present the salient features of the proposed approach, in Section 3.1, we discuss the analysis of the RA-study in more detail. In Section 3.2, we apply the varying-coefficient model to the colon-cancer data, with most of the results presented in the Supplementary Materials.

3.1. Rheumatoid arthritis case study

In the first step, we analyze the RA-study data with models (12) and (14). In the remainder of the text, we refer to those models as the "Binomial" and "Beta-Binomial" model, respectively. We focus on the results obtained by fitting the models using the penalized log-likelihood (8).

The upper part of Table 1 presents the results of the "Binomial" and "Beta-Binomial" models. In both models, the interaction term is not significant at the 5% significance level, while the methylation probability does seem to vary with the chromosomal position. In the "Binomial" model, the effect of both the 'RA' and 'Cell-type' covariate is statistically significant, whereas the effect of 'RA' is not significant in the "Beta-Binomial" model. In the latter case the estimate of σ_i indicates overdispersion. For instance, for a CpG site with 50 reads, the variance increases, as compared to the binomial one, by 49/(1+1/1.039) = 0.2497, i.e., 24.97%.

Any model-based test, however, should be conditional on an evaluation of the fit of the model to the data. Fig. 3 presents the worm plots, detrended Q-Q plots (van Buuren and Fredriks, 2001) for the "Binomial" and "Beta-Binomial" models. For a well-fitting model, the plot should be a flat line, indicating that the data follow the assumed model-based distribution. The plot in panel A of Fig. 3 clearly deviates from such a line. In particular, many points (residuals) fall outside the 95% point-wise CIs (indicated by the black dotted lines), and the slope of the fitted curve (indicated by the red solid line) is positive, suggesting that the variance is underestimated. Thus, the worm plot indicates that the binomial distribution is not suitable for the RA dataset. The worm plot of the "Beta-Binomial" model (see panel B of Fig. 3) is much closer to a flat line (except of some deviation in the tails), suggesting that the model fits the data better.

In view of the non-significance of the interaction term, the "Beta-Binomial" model can be simplified by dropping this term. The lower part of Table 1 presents the results of the model; for completeness, results for the simplified "Binomial" model are also shown. The exclusion of the interaction term does not change the conclusions: the effect of 'RA' remains statistically not significant in the "Beta-Binomial" model, while the effect of 'Cell-type' retains its significance. This is an important difference as compared to the "Binomial" model, which consistently suggests the significance of the RA effect. Thus, adjusting for the overdispersion implies a qualitative difference in the conclusions drawn from the models.

Fig. 4 (panels A–C) presents the estimated smoothed terms for the simplified "Beta-Binomial" model for the logit of the methylation probability with 95% point-wise CIs (indicated by the dark gray area) and simultaneous 95% CBs (indicated by the light gray area). It is clear that the latter, which accounts for the multiple-testing issue resulting from consideration of many chromosomal positions, are wider than the former. Thus, the use of point-wise CIs might lead to an inflated occurrence of type I errors in identification of DMRs.

Panel A of Fig. 4 shows a clear, almost linear effect of chromosomal position; its simultaneous 95% CBs exclude 0 across all positions. The CBs for 'RA' (panel C) include 0 across all positions and indicate, in agreement with the results presented in the lower part of Table 1, no effect of RA on the probability. On the other hand, the CBs for 'Cell-type' (panel B) indicate a statistically significant, positive effect, i.e., an increase of the methylation probability for the T-cell samples as compared to monocyte samples, across the entire chromosome except of narrow region between 102,711,919 and 102,711,961. It is worth noting that the 95% point-wise CIs would suggest a positive effect also in that narrow region.

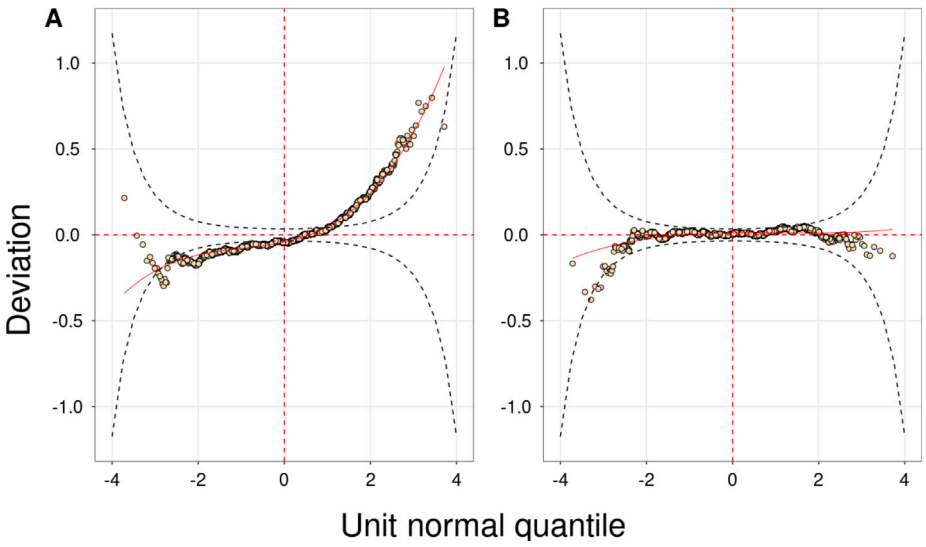


Fig. 3. The worm plot of the residuals for the "Binomial" (panel A) and "Beta-Binomial" (panel B) models. Black dashed lines indicate the 95% point-wise confidence intervals. The red solid line is a smooth curve fitted to the points of the worm. If the distributional assumption of the model is suitable for the analyzed data, the solid line should correspond to the horizontal red dashed line. This is approximately the case for the plot in panel B, but not in panel A.

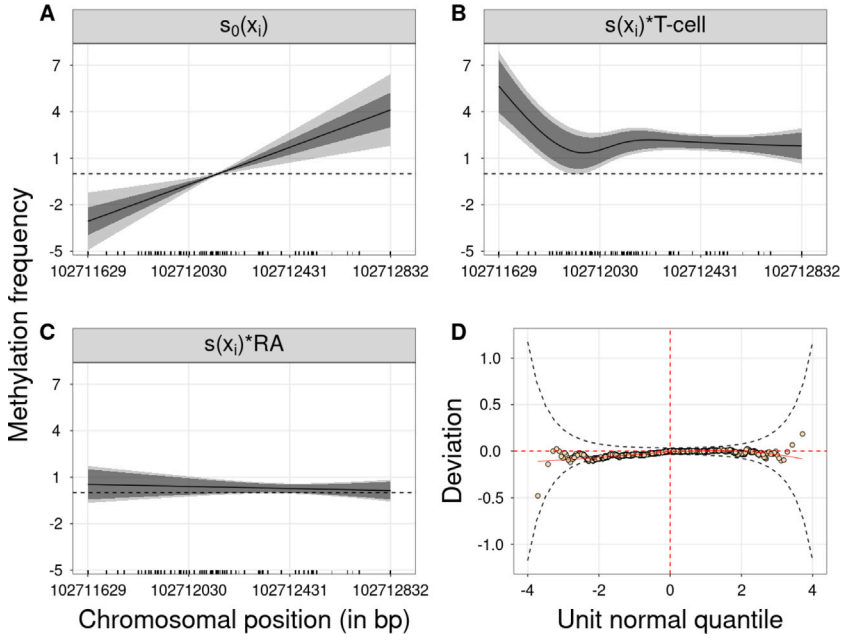


Fig. 4. Estimated smoothed curves for the "Beta-Binomial" model without the interaction term. Panels A-C: estimates of the smoothed splines (solid black lines) for the various model terms for the logit of the methylation probability, with the point-wise (dark gray) and simultaneous (light gray) 95% confidence bands. The vertical black bars at the bottom of the plot indicate the methylated genomic positions. Panel D: worm plot for the model (black dashed lines indicate the 95% point-wise confidence intervals).

Panel D of Fig. 4 shows the worm plot of the model. The plot indicates that the model fits the data well: the points (residuals) fall within the 95% CI (black dotted lines) and the fitted curve (red solid line) is approximately straight and centered around zero.

The effect of cell-type can also be seen in panel A of Fig. 5 that presents the estimated dependence of the methylation probability on the chromosomal position for the four groups of samples (indicated by different shades of red and blue). In particular, the curves for the monocyte samples (in red; dark for control subjects and light for RA subjects) differ markedly from the curves for the T-cell samples (in blue; dark for control subjects and light for RA subjects). Also, there is a clear separation between the 95% simultaneous CBs for the two smoothed curves of monocyte samples and the two curves of T-cell samples (panel B of Fig. 5). The lack of the effect of RA is reflected by the overlap of the 95% simultaneous CBs for the control and the RA samples within the strata defined by the cell-type status.

The "Beta-Binomial" model (12) assumes that the scale parameter, σ_i, is constant. However, the assumption may be too stringent. To remove it, σ_i can be modeled as a function of explanatory variables. Towards this aim, the extended model (13) can be considered.

The upper part of Table 2 presents the results for model (13). The worm plot (panel A of Fig. 6) shows that the model does fit the data reasonably well.

Results presented in the upper part of Table 2 suggest that the interaction terms are statistically significant neither for the methylation probability nor for the scale parameter. Thus, the terms may be

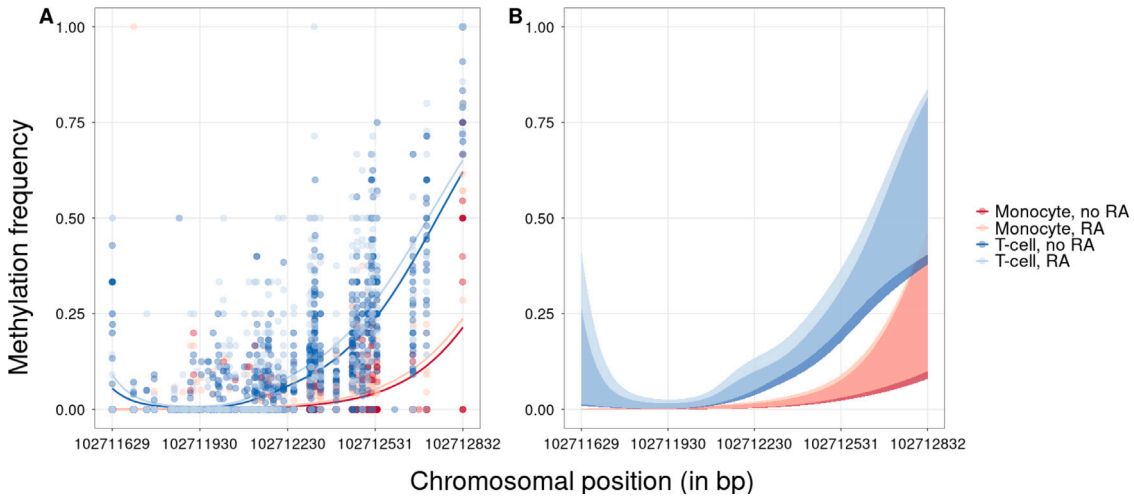


Fig. 5. Panel A: estimated curves (solid lines) for the methylation probability for the four groups of samples (indicated by different colors), with data points. The points and estimated curves are colored based on the type of immune-cell (red: Monocyte, blue: T-cell). Transparency is added based on the RA-status (dark: no RA, light: RA). Panel B: Simultaneous 95% confidence bands for the smoothed curves from panel A.

Table 2
Results of the estimation of the "Beta-Binomial" model with the scale-parameter modeled. edf — empirical degrees of freedom; ref.df — reference degrees of freedom.

Smooth terms	With interaction terms							
	Model for logit(π)				Model for log(σ)			
	edf	ref.df	F-value	*p*-value	edf	ref.df	F-value	*p*-value
Intercept	1.000	1.000	17.478	2.99e−5	1.000	1.000	8.657	0.003
RA	2.000	2.000	0.718	0.488	2.483	2.765	1.806	0.208
Cell-type	4.911	4.994	16.755	<2e−16	3.924	4.208	4.959	0.001
RA × Cell-type	2.022	2.044	0.091	0.915	2.000	2.000	1.363	0.256
Smooth terms	Without interaction terms							
	Model for logit(π)				Model for log(σ)			
	edf	ref.df	F-value	*p*-value	edf	ref.df	F-value	*p*-value
Intercept	1.000	1.000	44.738	<2e−16	1.000	1.000	33.091	<2e−16
RA	2.000	2.000	2.707	0.067	2.753	3.060	1.704	0.159
Cell-type	4.911	4.994	29.162	<2e−16	4.044	4.365	6.256	2.43e−5

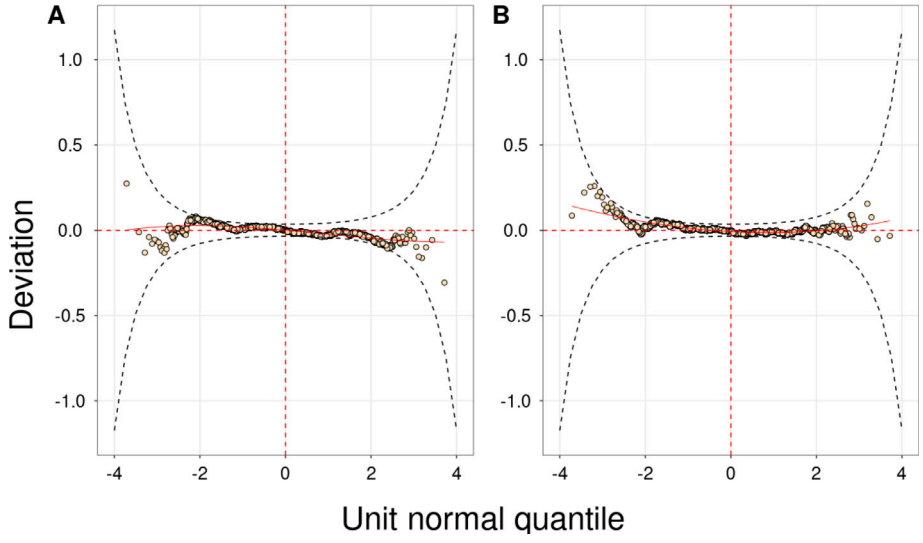


Fig. 6. The worm plot of the residuals for the "Beta-Binomial" model with the scale-parameter modeled: with interaction (panel A) and without interaction (panel B). Black dashed lines indicate the 95% point-wise confidence intervals. The red solid line is a smooth curve fitted to the points of the worm. If the distributional assumption of the model is suitable for the analyzed data, the solid line should correspond to the horizontal red dashed line. This is approximately the case for both panels.

dropped from the model. The worm plot (panel B of Fig. 6) suggests a satisfactory fit of the simplified model to the data. Results for the model are presented in the lower part of Table 2.

In particular, the results for the scale parameter indicate that overdispersion statistically significantly varies across chromosomal positions and differs between the monocyte and T-cell samples. Thus, the

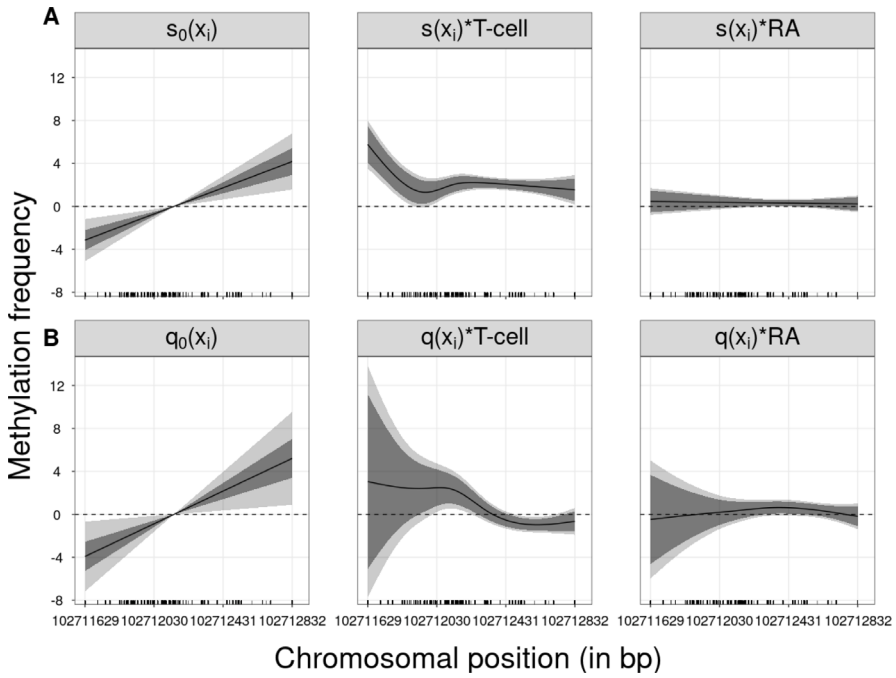


Fig. 7. Estimated smoothed splines for the "Beta-Binomial" model without the interaction terms and the scale-parameter modeled. Panel A: estimates of the smoothed splines (solid black lines) for the various model terms for the logit of the methylation probability, with the point-wise (dark gray) and simultaneous (light gray) 95% confidence bands. Panel B: estimates of the smoothed splines (solid black lines) for the various model terms for the scale parameter, with the point-wise (dark gray) and simultaneous (light gray) 95% confidence bands. The vertical black bars at the bottom of the plot indicate the methylated genomic positions.

constant-σ_i assumption used for the "Beta-Binomial" models presented in Table 1 may not to be valid. Removing this assumption, however, does not change the conclusions regarding the methylation probability, i.e., the presence of a statistically significant effect of chromosomal position and cell-type.

Fig. 7 presents the estimated smoothed splines corresponding to the results presented in the lower part of Table 2. Panel A of the figure presents the smoothed curves for the logit of the methylation probability. They are very similar to the curves obtained for the "Beta-Binomial" model that assumed a constant σ_i and no interaction (see panels A–C of Fig. 4). As compared to panel C of Fig. 4, a slight narrowing of the simultaneous 95% CBs for the RA effect in Fig. 7 can be observed, which almost results in excluding 0 in a small region around position 102,712,351. This is in agreement with the marginal non-significance ($p = 0.067$) of the test for the RA effect reported in the lower part of Table 2 for π.

The slight gain in precision might be attributed to a more precise specification of the variance-structure. Panel B of Fig. 7 presents the smoothed curves for the scale parameter. A clear, almost linear effect of chromosomal position can be seen; its simultaneous 95% CBs exclude 0 across all positions. The bands for the RA effect include 0 across all positions and indicate, in agreement with the results presented in the lower part of Table 2, no effect of RA on overdispersion. On the other hand, the simultaneous CBs for 'Cell-type' indicate a statistically significant effect in a small region around position 102,712,110 and in a small region around position 102,712,591.

Fig. 8 presents the estimated dependence of the methylation probability on the chromosomal position for the four groups of samples (indicated by different colors) for the "Beta-Binomial" model presented in the lower part of Table 2. The plots are similar to the graphs presented in Fig. 4 for the model assuming a constant scale parameter and illustrate the effect of chromosomal position and the cell-type status of a sample.

For completeness, we compared the "Binomial" and "Beta-Binomial" models to the results of SOMNiBUS which is, to the best of our knowledge, the only method that directly identifies DMRs across multiple sample while estimating covariate effects (Zhao et al., 2021). This method assumes that the methylated counts follow the binomial distribution. Note that the "Binomial" model (14) resembles the model used in SOMNiBUS (Zhao et al., 2021). There are two differences, however. First, SOMNiBUS does not report simultaneous CBs, but it returns point-wise CIs around the estimated curves. Second, SOMNiBUS accounts for experimental errors, which may potentially result in incorrect methylation read counts, by using a smoothed Expectation–Maximization algorithm (Zhao et al., 2021).

Fig. 9 provides the estimated smoothed profiles obtained for model (14) by using SOMNiBUS. There are only negligible differences between them and the profiles (not shown) corresponding to the results presented in the upper part of Table 1 for the same model. Hence, the overall conclusion (see Table 3) remains the same, i.e., a statistically significant effect of the chromosomal position, RA, and cell-type on the probability of methylation. Fig. 10 presents the estimated dependence of the methylation probability on the chromosomal position for the four groups of samples (indicated by different colors). A clear effect of cell-type is visible. Additionally, an RA effect within the genomic region between 102,712,100 and 102,712,780, more pronounced for the T-cell samples, can be seen. However, as argued in Section 3.1, the conclusions are questionable given the issues with the fit of the binomial model to the data.

SOMNiBUS also allows using a quasi-binomial distribution. The resulting smoothed profiles are, necessarily, the same as those shown in Figs. 9 and 10 for the binomial distribution. The estimated value of the overdispersion parameter ϕ is equal to 1.471 (see Table 3). Note that, in this case, the overdispersion factor is assumed to be constant and unrelated to the number of reads, unlike in (2) for the beta-binomial distribution. Adjusting for overdispersion does not change the significance of the RA effect.

Additionally, we compared the effect of the penalization on the final statistical decisions and the shapes of the estimated smoothed curves (see Section 2.2.2). Figure S1 in the Supplementary Materials presents the estimated smoothed terms for the model without interaction fitted with (panels A and B) and without penalization (panels C and D),

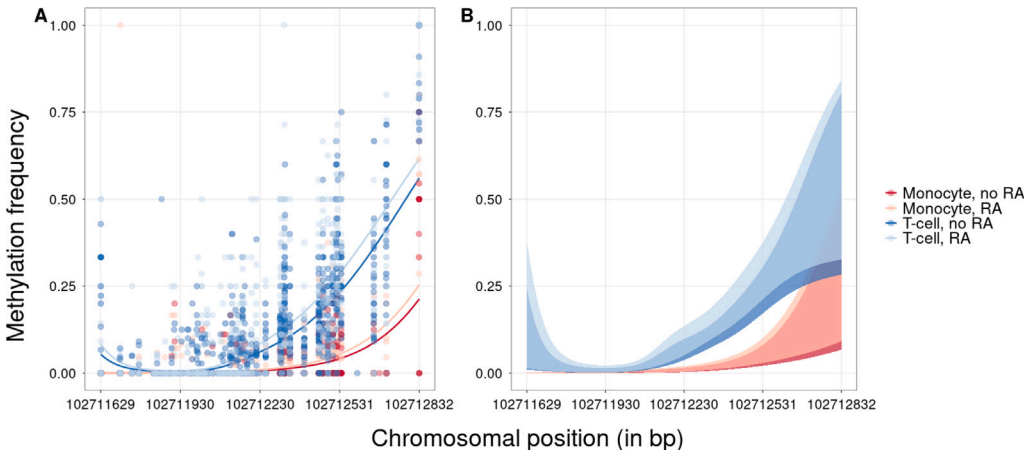


Fig. 8. The "Beta-Binomial" model without the interaction terms and the scale-parameter modeled. Panel A: estimated curves (solid lines) for the methylation probability for the four groups of samples (indicated by different colors), with data points. The points and estimated curves are colored based on the type of immune-cell (red: Monocyte, blue: T-cell). Transparency is added based on the RA-status (dark: no RA, light: RA). Panel B: Simultaneous 95% confidence bands for the smoothed curves in panel A.

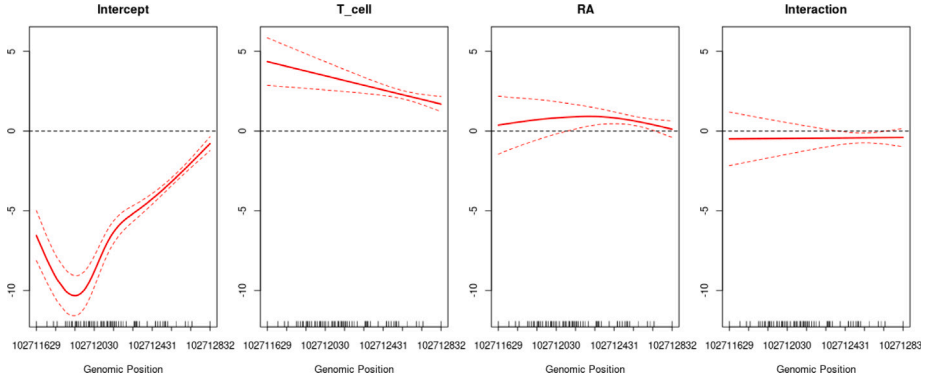


Fig. 9. The estimates of the smoothed splines (solid red lines) for the various model terms for the logit of the methylation probability, with the 95% point-wise confidence intervals (dashed red lines), obtained for the SOMNiBUS binomial model with the interaction term.

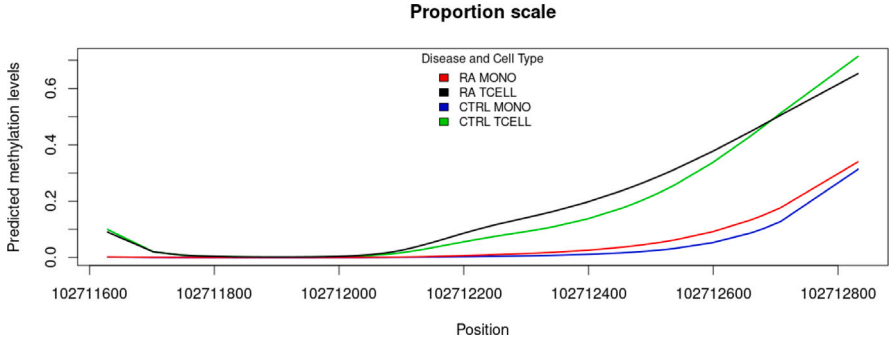


Fig. 10. The estimated curves showing the dependence of the methylation probability on the chromosomal position for the four groups of samples for the SOMNiBUS binomial model with interaction. Four colors are used to distinguish the different groups of sample (red-black for RA Mono/T-cell and blue–green for no RA Mono/T-cell.).

Table 3
Results of the estimation of the "Binomial" and "Quasi-Binomial" models. edf — empirical degrees of freedom.

Smooth terms	With interaction					
	Binomial			Quasi-Binomial		
	edf	Chi.sq	*p*-value	edf	F-value	*p*-value
Position	3.999	343.813	<2e−16	3.999	58.933	<2e−16
T_cell	2.005	302.315	<2e−16	2.003	102.493	<2e−16
RA	3.964	31.366	3.05e−06	3.820	5.469	0.0004
Interaction	2.000	7.283	2.62e−02	2.001	2.430	0.0880
ϕ	Est.:	NA		Est.:	1.471	

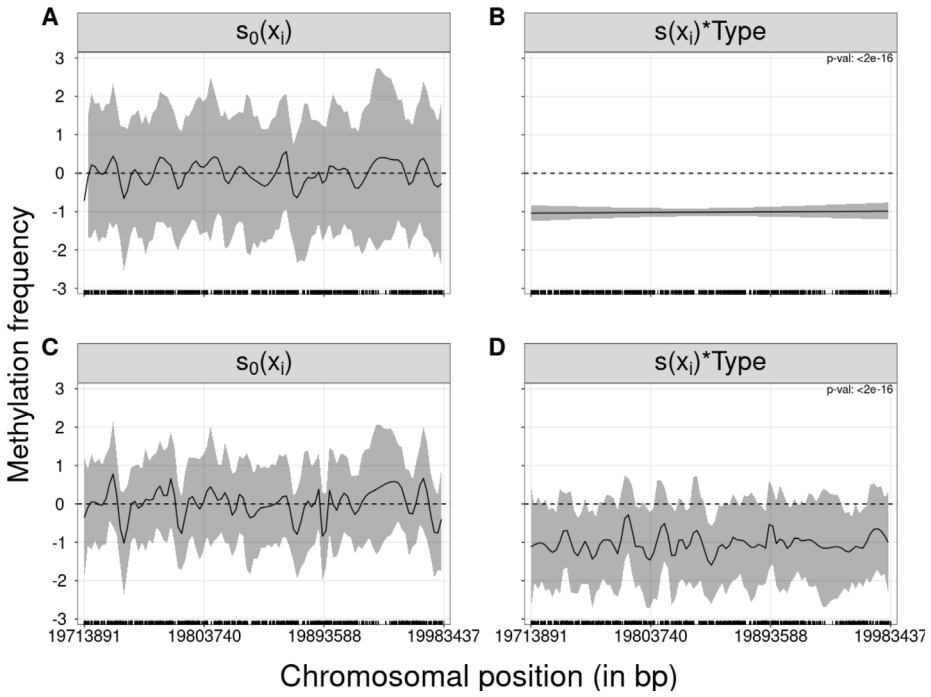


Fig. 11. Estimated smoothed curves for the model for region 20. Estimates of the smoothed splines (solid black lines) for the various model terms for the logit of the methylation probability, with the simultaneous (gray) 95% confidence bands for the model with (panels A-B) and without a penalty (panels C-D), respectively. The vertical black bars at the bottom of the plot indicate the methylated genomic positions.

respectively. There is not much change in the estimated smooth curves, while there is a loss of precision for the non-penalized analysis reflected in the wider 95% point-wise CIs (indicated by the dark gray area) and simultaneous 95% CBs (indicated by the light gray area). This is expected, given the increased dimension of the parameter space without penalization. Nevertheless, regardless of whether a penalty was applied or not, the overall positive effect for T-cell remains statistically significant (panels A and C), while there is no statistically significant effect of RA (panels B and D).

3.2. Colon cancer case study

We applied the varying-coefficient model (15) to the colon cancer data set. We focused on chromosome 21. Fitting the model to the data on the full chromosome is practically infeasible due to the computational complexity. Therefore, we have selected a sliding window approach by splitting the chromosome into 249 non-overlapping subregions, each containing at most 1000 unique methylation sites. Other sliding window approaches are possible (for instance, choosing windows that partially overlap) as long as the number of methylation sites is not too large; we will not discuss them here. There is a clear gap, without any data, from position 11,188,105 to 14,338,444 (see Fig. 2). We performed the analysis separately for the regions before and after the gap. In the following, we focus on the 245 regions after the gap.

In contrast to the RA case study, the regions are wide and contain many large gaps. This may pose an issue when fitting the model. Penalization applied in estimation of the splines causes them to be excessively smooth, often resulting in straight lines. This can be seen in Fig. 11 that presents the estimated smoothed curves for region 20, respectively with (panels A and B) and without (panels C and D) penalization. The same behavior can be seen in Figure S2 in the Supplementary Materials for region 121. For this reason, we decided to focus on the results of the non-penalized analysis, as it may offer a more detailed information about DMRs (perhaps at the cost of a reduced precision). Note that, of the 245 regions after the 11,188,105–14,338,444 gap, the non-penalized model-fitting algorithm did not converge for 57 regions.

Fig. 12 shows worm plots for the models for the first 64 regions; the plots for the remaining regions are shown in Supplementary Figures S3–S5. For most of the regions (e.g., regions 1, 2, 7, or 8), the worm plots are close to a flat line (except of some deviation in the tails) indicating a good fit. There are some plots (see, e.g., Figure S3: regions 83 and 121, Figure S4: regions 138 and 161) that show points (residuals) falling outside the 95% point-wise CIs (indicated by the black dotted lines). In general, however, for 155 (83.5%) regions, for which the model converged, the smoothed worm plot indicates a good fit.

Fig. 13 presents the estimated smoothed terms for model (15) with simultaneous 95% CBs (indicated by the gray area) for regions 1–64; for the remaining regions, the smoothed terms are shown in Supplementary Figures S6–S8. For 155 of the regions (after applying the Benjamini–Hochberg multiple-testing adjustment), the effect of 'Type' was statistically significant (at the 5% significance level) in the model. For those regions, the CBs for the effect indeed exclude, in some subregion(s), the value of 0, in accordance with the statistically significant result of the model-based test. For most of the regions the effect is negative, i.e., it suggests a decrease of the methylation probability for the cancer samples as compared to normal samples.

4. Conclusion

Smoothing-based approaches can precisely and accurately estimate methylation profiles in bisulfite sequencing data while accounting for biological variability, spatial correlation of the neighboring CpG sites, and the irregular spacing of methylation sites across genomic locations. In this article, we introduced a varying-coefficient model for estimating smooth effects of explanatory variables and detecting DMRs. Systematic differences between the estimated smoothers are formally assessed by simultaneous CBs which account for model uncertainty, multiplicity, and serial correlation.

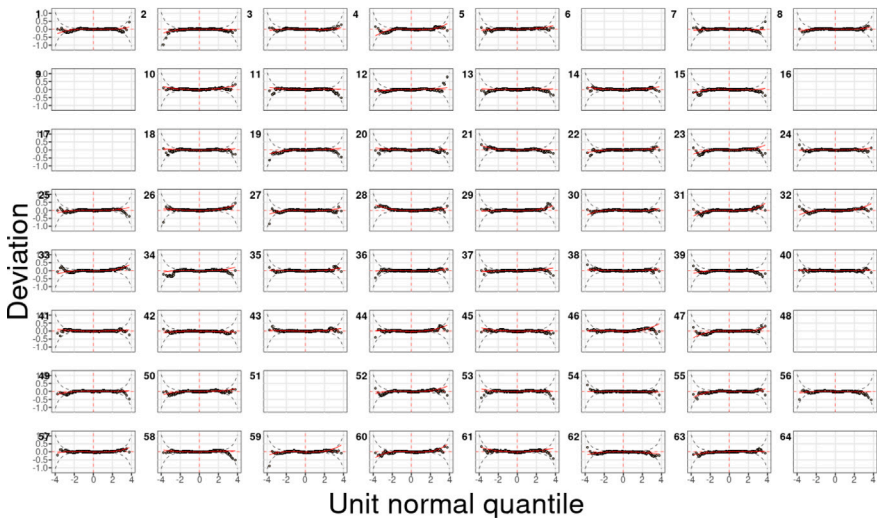


Fig. 12. The worm plots of the residuals for regions 1–64. Black dashed lines indicate the 95% point-wise confidence intervals. The red solid line is a smooth curve fitted to the points of the worm. If the distributional assumption of the model is suitable for the analyzed data, the solid line should correspond to the horizontal red dashed line. Empty subplots indicate regions for which the model did not converge.

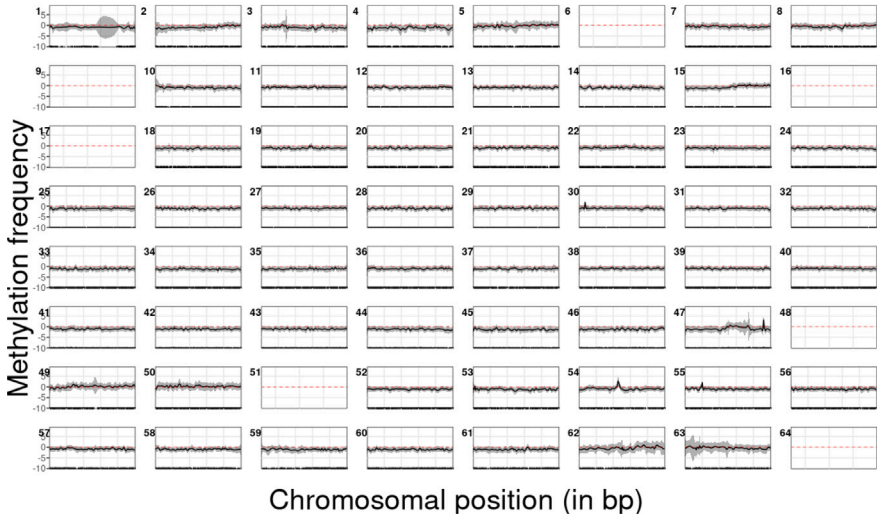


Fig. 13. Estimates of the smoothed splines (solid black lines) for the model term 'Type' for the logit of the methylation probability, with the simultaneous (gray) 95% confidence bands, for regions 1–64. The vertical black bars at the bottom of the plot indicate the methylated genomic positions. Empty subplots indicate regions for which the model did not converge.

We showed that the assumption that methylation counts are distributed according to a binomial distribution is likely not valid. In particular, in the RA study, the data exhibited a clear overdispersion. Proper modeling of the variance structure is important from a mean-structure-estimation point of view in case of binary or count data, because of the mean–variance link. The model that we propose assumes that the methylation data follow the beta-binomial distribution. Thus, it facilitates accounting for overdispersion. Moreover, the scale parameter, σ_i, can be modeled as a function of explanatory variables. In this respect, the proposed approach is more flexible than, for instance, the quasi-binomial model. This has beneficial consequences regarding the estimation and testing of the effects of explanatory variables on the methylation probability, as illustrated in the analyzed case study.

The proposed varying-coefficient model can be applied to targeted methylation sequencing data as well as whole-genome methylation sequencing data. However, in the latter case, the method requires a sliding window approach due to computational complexity. Selecting a window size may not be a trivial task: using a larger size reduces the number of windows, but wider windows not only increase computation time, but may also include larger gaps between successive CpG sites. Another question is how many CpG sites should be included in a window? (Hansen et al., 2012) A larger number of sites per window may require a larger number of knots, increasing the computation time.

CRediT authorship contribution statement

Katarzyna Górczak: Writing – original draft, Visualization, Software, Methodology, Formal analysis. **Tomasz Burzykowski:** Writing – review & editing, Methodology, Conceptualization. **Jürgen Claesen:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing interest

None.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiolchem.2024.108094.

References

Bansal, A., Pinney, S.E., 2017. DNA methylation and its role in the pathogenesis of diabetes. Pediatr Diabetes 18 (3), 167–177.

Beck, D., Maamar, M.B, Skinner, M.K., 2022. Genome-wide CpG density and DNA methylation analysis method (MeDIP, RRBS, and WGBS) comparisons. Epigenetics 17, 518–530.

Bergman, Y., Cedar, H., 2013. DNA methylation dynamics in health and disease. Nat. Struct. Mol. Biol. 20 (3).

Gong, T., Borgard, H., Zhang, Z., Chen, S., Gao, Z., Deng, Y., 2022. Analysis and performance assessment of the whole genome bisulfite sequencing data workflow: Currently available tools and a practical guide to advance DNA methylation studies. Small Methods 6.

Green, P.J., Silverman, B.W., 1994. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach, vol. 58, Chapman & Hall.

Hansen, K.D., 2020. Bsseqdata: Example whole genome bisulfite data for the bsseq package. R package version 0.26.0.

Hansen, K.D., Langmead, B., Irizarry, R.A., 2012. BSmooth: From whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. 13 (R83).

Hansen, K.D., Timp, W., Bravo, H.C., Sabunciyan, S., Langmead, B., McDonald, O.G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Zhang, K., Irizarry, R.A., Feinberg, A.P., 2011. Increased methylation variation in epigenetic domains across cancer types. Nature Genet. 43, 768–775.

Hastie, T., Tibshirani, R., 1993. Varying-coefficient models. J. R. Stat. Soc.. Ser. B 55 (4), 757–796.

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

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

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

Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J.B., Sabunciyan, S., Feinberg, A.P., 2009. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nature Genet. 41 (2).

Klein, H.U., Hebestreit, K., 2016. An evaluation of methods to test predefined genomic regions for differential methylation in bisulfite sequencing data. Brief. Bioinform. 17 (5), 796–807.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B., Ecker, J.R., 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462 (19).

Lu, S., Wang, J., Kakongoma, N., Hua, W., Xu, J., Wang, Y., He, S., Gu, H., Shi, J., Hu, W., 2022. DNA methylation and expression profiles of placenta and umbilical cord blood reveal the characteristics of gestational diabetes mellitus patients and offspring. Clin. Epigenetics 14 (69).

Maegawa, S., Hinkal, G., Kim, H.S., Shen, L., Zhang, L., Zhang, J., Zhang, N., Liang, S., Donehower, L.A., Issa, J.P.J., 2010. Widespread and tissue specific age-related DNA methylation changes in mice. Genome Res. 20, 332–340.

Moore, L.D., Le, T., Fan, G., 2013. DNA methylation and its basic function. Neuropsychopharmacology 38, 23–38.

Moser, D.A., Müller, S., Hummel, E.M., Limberg, A.S., Dieckmann, L., Frach, L., Pakusch, J., Flasbeck, V., Brüne, M., Beygo, J., Klein-Hitpass, L., Kumsta, R., 2020. Targeted bisulfite sequencing: A novel tool for the assessment of DNA methylation with high sensitivity and increased coverage. Psychoneuroendocrinology 120.

Nelder, J.A., Pregibon, D., 1987. An extended quasi-likelihood function. Biometrika 74, 221–232.

Olova, N., Krueger, F., Andrews, S., Oxley, D., Berrens, R.V., Branco, M.R., Reik, W., 2018. Comparison of whole-genome bisulfite sequencing library preparation strategies identifies sources of biases affecting DNA methylation data. Genome Biol. 19.

Reinders, J., Delucinge, V.C., Theiler, G., Chollet, D., Descombes, P., Paszkowski, J., 2008. Genome-wide, high-resolution DNA methylation profiling using bisulfite-mediated cytosine conversion. Genome Res. 18, 469–476.

Rigby, R.A., Stasinopoulos, D.M., 1996a. In: Hardle, W., Schimek, M.G. (Eds.), Statistical Theory and Computational Aspects of Smoothing. Heidelberg: Physica, pp. 215–230, Chapter Mean and dispersion additive models.

Rigby, R.A., Stasinopoulos, D.M., 1996b. A semi-parametric additive model for variance heterogeneity. Stat. Comput. 6, 57–65.

Rigby, R.A., Stasinopoulos, D.M., 2005. Generalized additive models for location, scale and shape,(with discussion). Appl. Stat. 54, 507–554.

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

Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge University Press.

Shafi, A., Mitrea, C., Nguyen, T., Draghici, S., 2018. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. Brief. Bioinform. 19 (5), 737–753.

van Buuren, S., Fredriks, M., 2001. Worm plot: A simple diagnostic device for modelling growth reference curves. Stat. Med. 20, 1259–1277.

Wood, S.N., 2000. Modelling and smoothing parameter estimation with multiple quadratic penalties. J. R. Stat. Soc. Ser. B Stat. Methodol. 62, 413–428.

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

Wood, S.N, 2017. Generalized Additive Models: An Introduction with R, second ed. Chapman and Hall/CRC.

Zhao, K., Oualkacha, K., Lakhal-Chaieb, L., Labbe, A., Klein, K., Ciampi, A., Hudson, M., Colmegna, I., Pastinen, T., Zhang, T., Daley, D., Greenwood, C.M.T., 2021. A novel statistical method for modeling covariate effects in bisulfite sequencing derived measures of DNA methylation. Biometrics 77 (2), 424–438.

Ziller, M.J., Gu, H., Müller, F., Donaghey, J., Tsai, L.T.Y., Kohlbacher, O., De Jager, P.L., Rosen, E.D., Bennett, D.A., Bernstein, B.E., Gnirke, A., Meissner, A., 2013. Charting a dynamic DNA methylation landscape of the human genome. Nature 500.