BIOINFORMATICS

Vol. 00 no. 00 2016
Pages ???

CMeth: a Bayesian semiparametric model for differential methylation analysis

Sergei Lebedev*, Roman Chernyatchik, Oleg Shpynov

JetBrains BioLabs, St. Petersburg, Russia

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: DNA methylation is an important epigenetic modification shown to be involved in cell differentiation, gene expression and diseases. Whole-genome bisulfite sequencing (WGBS) is an experimental protocol for obtaining base-resolution DNA methylation data. DNA methylation studies typically perform a number of WGBS experiments with a goal of identifying distinct differences between the two biological conditions, for example normal and tumor samples. The comparison of WGBS data is complicated by different sources of variability ranging from bisulfite conversion failures to sequencing errors to biological variation. This calls for accurate statistical models, which will enable accurate and robust comparison of DNA methylation from WGBS data.

Results: We have developed CMeth, a new tool for comparison of WGBS experiments. CMeth is based on a Bayesian extension of the hidden Markov model (HMM) which incorporates distances between consecutive cytosines into the inference process. CMeth is capable of comparing both replicated and unreplicated experiments and allows to control FDR (false discovery rate) in the predictions it produces. Our results show that CMeth is on par with existing methods in terms of sensitivity and offers improved specificity, which is especially relevant in the low coverage setting.

Availability and implementation: CMeth has been implemented in Kotlin and is available at https://github.com/JetBrains-Research/cmeth.

Contact: sergei.lebedev@jetbrains.com

1 INTRODUCTION

DNA methylation is a chemical modification of a DNA, resulting from the addition of a methyl group to a cytosine. DNA methylation has been shown to play an essential role in critical biological processes, including cell differentiation, regulation of gene expression and diseases (?).

Whole-genome bisulfite sequencing (WGBS) is an experimental protocol for measuring DNA methylation at nucleotide resolution. A key step in WGBS is bisulfite treatment of DNA, designed to promote epigenetic information to the sequence level. As a result of bisulfite treatment unmethylated cytosines undergo a conversion to uracils, while methylated cytosines remain unchanged. During amplification and sequencing, the uracils are read out as thymines. This allows to detect methylation events during read mapping: reads carrying an unmethylated cytosine will have a C/T mismatch against the reference genome. Ideally the reads covering each cytosine in the reference genome should all either carry a C or a T. However, for real WGBS this is rarely the case due to both biological and technical variation. WGBS experiments typically work with multiple cells which may differ slightly in methylation status of each individual cytosine contributing to the biological variation. Technical variation may be caused by bisulfite conversion failures, sequencing or mapping errors. Thus, the methylation status of a cytosine is usually characterized by its methylation level — an estimate of the probability that a cytosine is methylated.

DNA methylation studies typically perform a number of WGBS experiments with a goal of identifying distinct differences between the two biological conditions. For instance, cancer studies are interesting in detecting regions differentially methylated between normal and tumor samples.

A number of methods have been developed for detecting differential methylation from WGBS data (?). Among the first methods applied to methylome comparison were the classical statistical hypothesis testing procedures: Fisher's exact test (?) and t-test (?). Both tests do not account for the biological variability inherent for WGBS data, effectively overestimating the number of differences between the biological conditions being compared.

Another group of DMR detection methods is based on the beta-binomial distribution. The number of reads confirming methylation status of a particular cytosine is assumed to follow a binomial distribution with a beta distributed methylation proportion. The beta-binomial is a natural choice for replicated WGBS data, because it allows to capture the biological variability across replicates in its beta component. The tool MOABS (?) implements a hierarchical Bayesian model, which uses Empirical Bayes to estimate the prior distribution of the methylation proportion from all available samples. MOABS then uses the credible methylation difference (CDIF) statistic to discern differentially methylated cytosines. DSS (?) extends the hierarchical model of MOABS by assuming a group-specific log-normal distribution for the variance parameter of the beta-binomial distribution. The differentially methylated cytosines (DMCs) are determined by testing the means of the beta-binomial distributions for equality.

Recently introduced methylation analysis suite Bisulfighter (?) contains a tool ComMet, which implements a hidden Markov model (HMM) for DMR detection. ComMet doesn't use the beta-binomial distribution, but instead resorts to pseudo counts to regularize the parameter of the binomial

*To whom correspondence should be addressed.

distribution. An interesting feature of ComMet is the use of the between-cytosine distance distribution in the HMM architecture. Each of the three model states (no change, hypo- and hypermethylation) is paired with one (or two) "gap" states. The choice between the number of "gap" states is up to the user. The DMRs are identified via a dynamic programming algorithm which optimizes posterior log-odds ratio for a region to be differentially methylated.

Here we present CMeth, a new tool for accurate comparison of WGBS experiments. CMeth is based on the Bayesian extension of the hidden Markov model, which we call a semiparametric switching HMM. Similarly to ComMet, our model incorporates the distances between the cytosines into the inference process. Unlike ComMet, however, the appropriate number of "gap" states isn't fixed and is determined from the data as part of the inference process. CMeth is designed to work with both replicated and non-replicated experiments. Bayesian formulation of the model allows to account for biological variability directly instead of pooling the replicates into a single sample. CMeth implements FDR (false discovery rate) control for the predictions it produces. Bechmarks on simulated and real data and show that CMeth is on par with existing methods in terms of sensitivity and offers improved specificity, which is especially relevant when the sequencing depth is low.

2 METHODS

2.1 Data representation

We represent WGBS results for the r-th sample by a list of three-tuples: $x_{tr} \doteq (k_{tr}, n_{tr}, d_{tr})$, ordered by genomic position $t \in \{1, \ldots, T\}$. Each sample corresponds to one of the two compared biological conditions. The total number of samples is denoted $R \geq 2$. The quantities k_{tr} and n_{tr} are methylated and total coverage and d_{tr} is the distance between the t-th and $(t-1)$-th cytosines. We only consider cytosines covered by at least a single read, that is $\forall r, t\,(n_{tr} > 0)$.

2.2 Model overview

We propose the semiparametric switching hidden Markov model for DMR detection from bisulfite sequencing data. Our model is abstract w.r.t. the number of states used. Here to simplify the presentation we use four states

$$S = \{(\mathrm{U}_1, \mathrm{U}_2), (\mathrm{M}_1, \mathrm{U}_2), (\mathrm{U}_1, \mathrm{M}_2), (\mathrm{M}_1, \mathrm{M}_2).\}$$

The label $(\mathrm{U}_1, \mathrm{U}_2)$, abbreviated by U, marks cytosines unmethylated in both biological conditions, $(\mathrm{M}_1, \mathrm{M}_2)$ or M — methylated cytosines, $(\mathrm{M}_1, \mathrm{U}_2)$ and $(\mathrm{U}_1, \mathrm{M}_2)$ — differentially methylated cytosines, abbreviated by $\uparrow$ and $\downarrow$ respectively.

The model can be deconstructed into two parts: the nonparametric distance mixture and the binomial switching hidden Markov model. The distance mixture groups distances between consecutive observations into an unspecified number of distance clusters. The switching hidden Markov model uses distance clusters to refine the dependence structure between the states of consecutive cytosines.

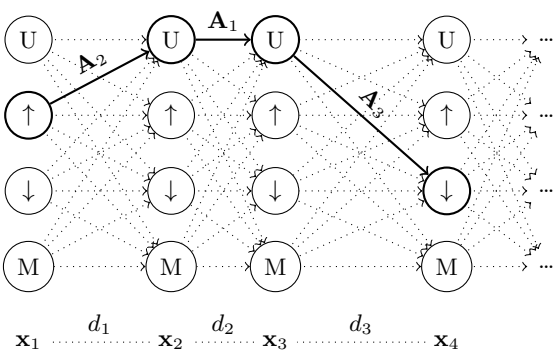

Figure 1. Trellis diagram for the switching hidden Markov model. Hidden state nodes represent methylation difference at step t. The choice of the transition probability matrix $\mathbf{A}$ for each edge in the graph is guided by the distance mixture.

2.3 Nonparametric distance mixture

The motivation for incorporating distances into the inference process comes from the fact that consecutive cytosines in the observations may not be consecutive in the genome. That is, the cytosines may be separated by a number of other nucleotides.

The distribution of distances estimated from real genomes (Figure ??, Supplementary Figure ??) suggests that using the distances in the model directly is impractical due to the large number of parameters to be estimated. We group similar distances together using a geometric Dirichlet process (DP) mixture model (?). An attractive feature of the DP mixtures is that they don't require fixing the number of clusters beforehand. The appropriate number of clusters is determined from the data as part of the inference process.

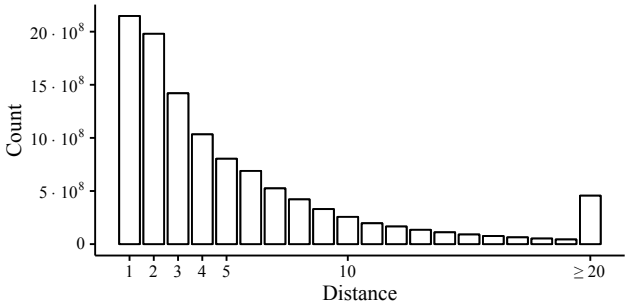

Figure 2. Histogram of distances between consecutive cytosines on the first chromosome of the hg19 reference genome.

DNA methylation is often studied relative to the cytosine context: CG, CHG or CHH, where H means any nucleotide except G. We use a shifted geometric distribution to account for the fact that the distance between consecutive cytosines can be either non-zero or strictly positive depending on the context. If a cytosine is in CG context then $d \geq 1$ and the shift must be 1. If a cytosine is in CHG or CHH context then $d \geq 0$ and the shift is 0. The p.m.f. of a geometric distribution shifted by y with success probability q_c is given by

$$p(d|q_c; y) = (1 - q_c)^{d-y} q_c,$$

where $c \in \mathbb{N}$ is the index of the mixture component or distance cluster. A natural choice for the prior distribution of the q_c is the Beta distribution as it is conjugate to the geometric distribution and has support interval of $[0, 1]$.

$$q_c \sim \text{Beta}\left(\alpha_0^{(\mathbf{q})}, \beta_0^{(\mathbf{q})}\right).$$

2.4 Binomial switching hidden Markov model

We assume that the number of reads confirming methylation k_{tr} at some genomic position t follows a binomial distribution with parameters n_{tr} and the "true" methylation proportion p_{ir}. The index $i \in \{1, \ldots, S\}$ is an unobserved state label. States of consecutive cytosines form a first-order Markov chain with multiple transition probability matrices $\mathbf{A}_c$ indexed by a distance cluster c. We use the term switching to account for the fact that a chain can switch a transition probability matrix at t-th observation with probability defined by the distance mixture.

Our prior assumptions are common for Bayesian hidden Markov models (?). For initial state probabilities π and state transition probabilities $\mathbf{A}_c$ we use symmetric Dirichlet prior

$$\pi \sim \text{SymDir}\left(\omega_0^{(\pi)}\right) \qquad \mathbf{A}_{ci} \sim \text{SymDir}\left(\omega_0^{(\mathbf{A})}\right),$$

and for the state-specific "true" methylation proportion we assume the Beta distribution prior.

$$p_{ir} \sim \text{Beta}\left(\alpha_0^{(\mathbf{p})}, \beta_0^{(\mathbf{p})}\right)$$

2.5 Design matrix

Recall that each state label $S = \{\text{U}, \uparrow, \downarrow, \text{M}\}$ can be represented as a pair of labels from a smaller set $\{\text{U}, \text{M}\}$. Components of such pair describe the methylation status of a cytosine in each of the biological conditions. For example, $\text{U} \equiv (\text{U}_1, \text{U}_2)$ and $\uparrow \equiv (\text{H}_1, \text{H}_2)$, or, if we view the pairs in terms of the corresponding model parameters,

$$p_{\text{U}} = (p_{U_1}, p_{U_2}) \qquad p_{\uparrow} = (p_{H_1}, p_{U_2}).$$

Note that the second components of p_{U} and $p_{\uparrow}$ both reference the parameter p_{U_2} and therefore must be the same. We use a design matrix $\mathbf{D}$ to enforce these constraints. The matrix assigns each sample-state pair an index in the parameter vector.

The design matrix enforcing a single value for p_{U_2} in the above example is given below. Elements marked with an asterisk (*) correspond to the p_{U_2} component of the parameters p_{U} and $p_{\uparrow}$. The same index in the design matrix ensures that the parameters p_{U_2} for both states are equal.

$$\mathbf{D} = \begin{array}{c} \begin{array}{cccc} \text{U} & \uparrow & \downarrow & \text{M} \end{array} \\ \left[\begin{array}{cccc} 1 & 2 & 1 & 2 \\ 3^* & 3^* & 4 & 4 \end{array}\right] \end{array} \begin{array}{l} \\ \text{condition 1} \\ \text{condition 2} \end{array} \qquad \mathbf{p} = \left[\begin{array}{c} p_{U_1} \\ p_{U_2} \\ p_{M_1} \\ p_{M_2} \end{array}\right] \begin{array}{l} 1 \\ 2 \\ 3^* \\ 4 \end{array}$$

Design matrix can be used to introduce arbitrary constraints on the model parameters. For instance, we might argue that the distribution of the "true" methylation proportion in the M state is a characteristic of the biological condition and force the replicates to share the corresponding parameters. The following design matrix does exactly that for the case of two biological conditions each having two replicates.

$$\mathbf{D} = \begin{array}{c} \begin{array}{cccc} \text{U} & \uparrow & \downarrow & \text{M} \end{array} \\ \left[\begin{array}{cccc} 1 & 2^* & 1 & 2^* \\ 3 & 2^* & 3 & 2^* \\ 4 & 4 & 5^+ & 5^+ \\ 6 & 6 & 5^+ & 5^+ \end{array}\right] \\ \begin{array}{cccc} \text{U}_1 & \text{M}_1 & \text{U}_1 & \text{M}_1 \\ \text{U}_2 & \text{U}_2 & \text{M}_2 & \text{M}_2 \end{array} \end{array} \begin{array}{l} \\ \text{condition 1, rep. a} \\ \text{condition 1, rep. b} \\ \text{condition 2, rep. a} \\ \text{condition 2, rep. b} \\ \\ \end{array}$$

An asterisk (*) marks p_{M_1} component of the parameters $p_{\uparrow}$ and p_{M}, while a plus (+) — p_{M_2} component of $p_{\downarrow}$ and p_{M}.

The matrix $\mathbf{D}$ can be thought of as an equivalence relation on the set $\{1, \ldots, R\} \times S$. The number of equivalence classes denoted E is the maximum index in the design matrix

$$E = \max_{r \le R,\, i \in S} \mathbf{D}_{ri}. \tag{1}$$

2.6 Inference and parameter learning

Given the model described above and the matrix of bisulfite sequencing results $\mathbf{x}$, our goal is to compute the posterior distribution of model parameters, and to infer the hidden state labels for each cytosine.

For convenience we introduce two latent indicator variables.

- The indicator w_{ct} is 1 if the t-th distance was generated by the c-th distance cluster and 0 otherwise.
- The indicator z_{it} is 1 if the t-th observation was generated by the i-th state and 0 otherwise.

2.6.1 Mean-field variational inference

As with many Bayesian models, exact posterior inference for the semiparametric switching HMM is intractable. We resort to the mean-field variational method (?), which assumes independence between model parameters and latent variables. This assumption allows to lower bound the marginal log-likelihood

$$\log p(\mathbf{x}) \ge \log q(\mathbf{x}) = \sum_{\mathbf{w}} \sum_{\mathbf{z}} \int q(\mathbf{w}, \mathbf{z}) q(\boldsymbol{\Theta}) \log \frac{p(\mathbf{x}, \mathbf{w}, \mathbf{z}, \boldsymbol{\Theta})}{q(\mathbf{w}, \mathbf{z}) q(\boldsymbol{\Theta})}.$$

The inference is then performed by iteratively maximizing the lower bound $\log q(\mathbf{x})$ with respect to latent variables $\mathbf{w}$, $\mathbf{z}$ and model parameters $\boldsymbol{\Theta}$.

Below we describe a mean-field variational algorithm for jointly estimating the parameters of the distance mixture and binomial switching HMM. Our algorithm is based on the truncated stick-breaking representation of the DP mixture (?). The point of truncation is to limit the maximum number of clusters in the DP mixture with a value C and then seek the best variational approximation to the true untruncated posterior distribution. Mean-field inference assumes a fully factorized posterior distribution, which in the case of the DP-mixture implies the independence of stick weights (and effectively the mixing coefficients) from distance cluster assignments. To overcome this unrealistic assumption we integrate out stick weights from the posterior distribution (?).

The joint distribution over the parameters and latent indicator variables can be written as

$$p(\boldsymbol{\Theta}, \mathbf{w}, \mathbf{z}) = p(\mathbf{w}, \mathbf{q}) p(\mathbf{z}, \mathbf{A}, \pi | \mathbf{w}), \tag{2}$$

and the family of variational approximation we consider is given by

$$q(\boldsymbol{\Theta}, \mathbf{w}, \mathbf{z}) = q(\mathbf{w}, \mathbf{z}) q(\mathbf{q}) q(\mathbf{A}, \pi). \tag{3}$$

Note that we do not assume the independence between distance cluster assignments $\mathbf{w}$ and state assignments $\mathbf{z}$.

We now describe the specific updates for each variational parameter. The complete derivation of updates is available in Section ?? of the Supplementary Data.

1. The update equations for the parameters of the per-distance cluster success probabilities are given by

$$\alpha_c^{(\mathbf{q})} = \alpha_0^{(\mathbf{q})} + \sum_{t=2}^{T} \mathbb{E}\left[w_{ct}\right] \tag{4}$$

$$\beta_c^{(\mathbf{q})} = \alpha_0^{(\mathbf{q})} + \sum_{t=2}^{T} \mathbb{E}\left[w_{ct}\right] d_t. \tag{5}$$

2. For the parameters of the initial and transition probabilities the updates are

$$\omega_i^{(\pi)} = \omega_0^{(\pi)} + \mathbb{E}\left[z_{i1}\right] \tag{6}$$

$$\omega_{cij}^{(\mathbf{A})} = \omega_0^{(\mathbf{A})} + \sum_{t=2}^{T} \mathbb{E}\left[w_{ct} z_{i(t-1)} z_{jt}\right]. \tag{7}$$

3. The update equations for the parameters of the methylation proportions for each equivalence class are

$$\alpha_e^{(\mathbf{p})} = \alpha_0^{(\mathbf{p})} + \sum_{r=1}^{R}\sum_{i=1}^{S} \mathbb{I}(\mathbf{D}_{ri} = e) \sum_{t=1}^{T} \mathbb{E}\left[z_{it}\right] k_{tr} \tag{8}$$

$$\beta_e^{(\mathbf{p})} = \beta_0^{(\mathbf{p})} + \sum_{r=1}^{R}\sum_{i=1}^{S} \mathbb{I}(\mathbf{D}_{ri} = e) \sum_{t=1}^{T} \mathbb{E}\left[z_{it}\right] (n_{tr} - k_{tr}). \tag{9}$$

2.6.2 Forward-backward algorithm The expectations $\mathbb{E}\left[w_{ct}\right]$, $\mathbb{E}\left[z_{it}\right]$ and $\mathbb{E}\left[w_{ct} z_{i(t-1)} z_{jt}\right]$ can be computed using a modification of the forward-backward algorithm for Bayesian hidden Markov models (??). The forward-backward algorithm computes two auxiliary variables

$$\alpha_i(t) \doteq p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, z_{it})$$
$$\beta_i(t) \doteq p(\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, \ldots, \mathbf{x}_T | z_{it}),$$

using the following recursive rules[1]

$$\alpha_i(1) = \widetilde{\pi}_i \mathbf{B}_{i1}$$

$$\alpha_i(t) = \sum_{j=1}^{S} \alpha_j(t-1) \sum_{c=1}^{C} p(\mathbf{w}^{\neg t}) p(d_t | \widetilde{q}_c) \widetilde{\mathbf{A}}_{cji} \mathbf{B}_{it}$$

$$\beta_i(T) = 1$$

$$\beta_i(t-1) = \sum_{j=1}^{S}\sum_{c=1}^{C} p(\mathbf{w}^{\neg t}) p(d_t | \widetilde{q}_c) \widetilde{\mathbf{A}}_{cij} \mathbf{B}_{jt} \beta_j(t),$$

where $p(\mathbf{w}^{\neg t}) p(d_t | \widetilde{q}_c)$ is the posterior probability of switching the transition probability matrix to $\widetilde{\mathbf{A}}_c$ at observation t and $\mathbf{B}_{it}$ is the probability of the observation $\mathbf{x}_t$ being generated in state i.

$$\mathbf{B}_{it} \doteq \prod_{e=1}^{E}\prod_{r=1}^{R} p(k_{rt} | n_{rt}, \widetilde{p}_e)^{\mathbb{I}(\mathbf{D}_{ri} = e)}.$$

The conditional $p(w_{ct} | \mathbf{w}^{\neg t})$ is approximated using a second-order Taylor expansion (?). Derivation is given in Section ?? of the Supplementary Data.

Intuitively $\mathbb{E}\left[w_{ct} z_{i(t-1)} z_{jt}\right]$ is the expected number of transitions from state i to state j via distance cluster c, which we can compute using $\alpha_i(t)$ and $\beta_i(t)$ as follows

$$\mathbb{E}\left[w_{ct} z_{i(t-1)} z_{jt}\right] \propto \alpha_i(t-1) p(\mathbf{w}^{\neg t}) p(d_t | \widetilde{q}_c) \widetilde{\mathbf{A}}_{cij} \mathbf{B}_{jt} \beta_j(t) \doteq \xi_{cij}(t). \tag{10}$$

The remaining expectations can then be computed in terms of $\xi_{cij}(t)$.

$$\mathbb{E}\left[z_{it}\right] = \sum_{c=1}^{C}\sum_{j=1}^{S} \xi_{cij}(t) \propto \alpha_i(t) \beta_i(t) \tag{11}$$

$$\mathbb{E}\left[w_{ct}\right] = \sum_{i=1}^{S}\sum_{j=1}^{S} \xi_{cij}(t) \propto p(\mathbf{w}^{\neg t}) p(d_t | \widetilde{q}_c) \sum_{i=1}^{S}\sum_{j=1}^{S} \alpha_i(t-1) \widetilde{\mathbf{A}}_{cij} \mathbf{B}_{jt} \beta_j(t) \tag{12}$$

2.6.3 Convergence Each iteration of mean-field variational inference is theoretically guaranteed to increase the lower bound on the marginal log-likelihood (?). Thus the algorithm converged once the the change in the lower bound between the iterations is less than a specified threshold. Unfortunately, this technique is not applicable to our model due to the use of approximation for computing the conditional $p(w_{ct} | \mathbf{w}^{\neg t})$. We rely on the ACVB criterion suggested by ? to ensure convergence.

Algorithm 1 Variational inference for semiparametric switching HMM

For all equivalence classes and distance clusters initialize the parameters $\alpha_e^{(\mathbf{p})}$, $\beta_e^{(\mathbf{p})}$ and $\alpha_c^{(\mathbf{q})}$, $\beta_c^{(\mathbf{q})}$ with method of moments estimates from KMeans++ (?) clusters. Initialize remaining parameters with $\pi_i = \mathbf{A}_{cij} = \frac{1}{S}$.
Repeat until convergence:

1. For each state and distance cluster,
   - Compute the expectations $\mathbb{E}\left[w_{ct}\right]$, $\mathbb{E}\left[z_{it}\right]$ and $\mathbb{E}\left[w_{ct} z_{i(t-1)} z_{jt}\right]$ using Eq. ??, Eq. ?? and Eq. ??.
   - Update the parameters of initial and transition probabilities using Eq. ?? and Eq. ??.
2. For each distance cluster update the parameters of the geometric distribution using Eq. ?? and Eq. ??.
3. For each equivalence class update the parameters of methylation proportions using Eq. ?? and Eq. ??.

2.6.4 FDR control and Q-value estimation The work of ? introduced an optimal procedure for HMM-dependent hypothesis testing. The procedure is a thresholding rule based on the posterior probability of the null hypothesis being true. ? show that under some conditions the procedure is optimal in the sense that it controls the FDR (false discovery rate) at level α and has the smallest FNR (false non-discovery rate) among all valid FDR procedures.

The Q-value of a hypothesis test is the minimum FDR at which the test may be called significant (?). Instead of fixing the level α beforehand one can directly estimate the Q-value of each hypothesis test. The FDR can then be controlled at level α using a Q-value threshold $\hat{q}(p_t) \leq \alpha$. We apply the optimal procedure developed by ? to estimate Q-values.

[1] Tilde marks the exponents of the expected values of the natural logarithms of the parameters, i.e. $\widetilde{\theta} \doteq e^{\mathbb{E}[\log \theta]}$.

Algorithm 2 Q-value estimation

1. Let $p_1, \ldots, p_T$ be posterior probabilities of H_0 being true.
2. Let $p_{[1]} \leq \ldots \leq p_{[T]}$ be posterior probabilities of H_0 sorted in increasing order.
3. For each $t \in \{1, \ldots, T\}$ set

$$\hat{q}\left(p_{[t]}\right) = \frac{1}{t} \sum_{k=1}^{t} p_{[k]}.$$

4. For each $t \in \{T-1, T-2, \ldots, 1\}$ set

$$\hat{q}\left(p_{[t]}\right) = \min\left\{\hat{q}\left(p_{[t]}\right), \hat{q}\left(p_{[t+1]}\right)\right\}.$$

2.6.5 DMR detection Having found the optimal parameters via Algorithm ??, we calculate for each cytosine the posterior probability of the null hypothesis that the methylation status of the cytosine is the same in the two biological conditions being compared.

$$p(H_0|\mathbf{x}, \mathbf{\Theta}) = p(U \vee M|\mathbf{x}, \mathbf{\Theta}) = p(U|\mathbf{x}, \mathbf{\Theta}) + p(M|\mathbf{x}, \mathbf{\Theta}) \quad (13)$$

We then use these posterior probabilities to estimate Q-values as described in Algorithm ??. The DMRs for a fixed level α can be reconstructed by joining together consecutive cytosines with a Q-value less then or equal to α.

3 DATASETS

3.1 Simulated data

To evaluate the sensitivity and specificity of CMeth results we resort to a simulation study.

Our simulation method is loosely based on DNemulator (?), a tool for simulating bisulfite converted reads. We approximate the distribution of methylation levels with a Markov chain over five methylation ranges. Each methylation range is characterized by its mean methylation level and cytosine context-specific frequency (see Table ??; Section ?? of the Supplementary Data).

Range	CG	CHG	CHH
[0, 0]	0.090	0.775	0.865
(0, 0.12]	0.005	0.105	0.070
(0.12, 0.5]	0.100	0.108	0.059
(0.5, 0.8]	0.185	0.004	0.001
(0.8, 1]	0.620	0.008	0.005

Table 1. Methylation ranges and their frequencies estimated from chr22 of the WGBS data for human ESCs, GEO Series GSE16256 (?)

Having the methylated ranges we construct bisulfite converted reads for the two samples as follows.

1. Using chromosome 22 of the hg19 reference genome as a template we independently assign each cytosine a random methylation range with a probability defined by the cytosine context.
2. We then construct a fixed number of non-overlapping regions with known methylation difference. For each region we first choose its length from the negative binomial distribution, then randomly pick its location in the genome, and finally re-assign the methylation ranges in the simulated region to the randomly chosen ranges. To simplify the validation we assume that all cytosines in the region have the same methylation range. The DMR score of a region is the absolute difference between the indices of the methylation ranges in each of the two samples. For example, the DMR score of methylation ranges $[0, 0]$ and $(0.5, 0.8]$ is 3. The regions with DMR score 4 were considered trully differentially methylated, while the regions with zero DMR score were considered trully similar.
3. Finally, we simulate bisulfite converted reads by sampling random genomic fragments of fixed length. Methylation status of each cytosine in the read is determined by a Bernoulli trial using the mean methylation level of the corresponding methylation range state as parameter.

3.2 Real data

To assess CMeth performance on real data, we have used the WGBS data from two methylation studies.

3.2.1 Fibroblast differentiation data The work of ? explored the methylation differences between the human ESCs and fetal fibroblasts. The data, GEO series GSE16256, includes both biological and technical replicates for each biological condition. Reads were pre-processed and aligned to the hg19 reference genome using bwa-meth pipeline (?). Subsampling was done using the view command of Samtools 1.3.

3.2.2 Atherosclerosis data ? consider a comparison between normal aortic tissue and atherosclerotic lesion, GEO series GSE46401 (?). The study provides microarray data for multiple donor-matched pairs, obtained from the Illumina HumanMethylation450 BeadChip array, and WGBS data for one of the donor-matched pairs. Sequencing data are available pre-preprocessed and aligned to the hg19 reference genome. Microarray data, available in the form of CSV files, were analyzed with minfi Bioconductor package (?).

4 RESULTS

4.1 Simulation results

We evaluated the performance of CMeth on the dataset simulated as described in Section ?? varying the mean length of simulated regions and sequencing depth. Simulated reads were aligned to the hg19 reference genome using bwa-meth pipeline. Results were compared against ComMet, MOABS and DSS. We used default command line argument values for each method; relevant details are available in Section ?? of the Supplementary Data. The performance of each method w.r.t. correctly identifying all cytosines in the simulated region as differentially or similarly methylated was summarized using specificity and sensitivity (Table ??).

	Method	(5x, 500bp)	(10x, 500bp)	(20x, 1000bp)
Sensitivity	CMeth	93.28	98.65	99.90
	ComMet	99.06	99.10	99.90
	DSS	53.28	76.91	89.56
	MOABS	86.56	95.35	99.62
Specificity	CMeth	100.0	99.64	99.25
	ComMet	88.82	92.51	95.85
	DSS	100.0	100.0	100.0
	MOABS	99.81	99.82	99.88

Table 2. Simulation results for different sequencing depths and mean region lengths.

In terms of sensitivity, performance of CMeth was comparable to ComMet and MOABS. However, CMeth consistently achieved better specificity, outperforming both ComMet and MOABS when the sequencing depth is low. This feature of CMeth might be especially attractive for large scale methylation studies which typically trade off the number of replicates for low sequencing depth. DSS was the most conservative among the compared methods, demonstrating the lowest sensitivity with the highest specificity. MOABS and ComMet had similar performance on the simulated data, but ComMet consistently discovered more false DMCs than other methods, detecting differential methylation even in regions with zero DMR score (Figure ??).

We emphasize that our simulation method does not account for technical and biological variation inherent for real-world WGBS data, thus the presented results can be thought of as best-case analysis.

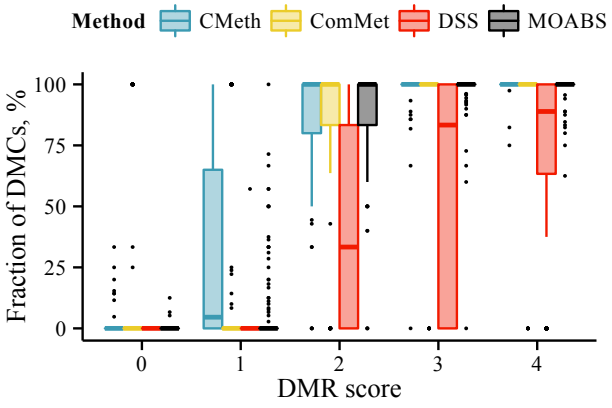

Figure 3. Fraction of DMCs relative to the number of covered cytosines in a region identified by each method over all regions at sequencing depth 20x.

4.2 Real data application

4.2.1 Replicate consistency To further investigate the properties of the CMeth approach w.r.t. existing methods we considered a comparison between technical and biological replicates of the human fibroblast differentiation dataset. For simplicity we restricted the comparison to cytosines in CG context on the first chromosome of the hg19 reference genome. For all methods except MOABS FDR was controlled at level $\alpha = 10^{-4}$.

We started by comparing each of the available technical replicates for the 1 and 2 biological replicates of the human ESCs. The fraction of DMCs reported by each method is summarized in Table ?? and in Supplementary Table ??. We hypothesized that the fraction of cytosines differentially methylated between the replicates should be low, since the samples correspond to the same biological condition and the difference if any is due to the technical variation. The output of CMeth, DSS and MOABS supported our hypothesis. These three methods identified a small ($< 1\%$) fraction of DMCs. ComMet, on the other hand, resulted in unrealistically large estimates, declaring almost $> 10\%$ of the input in $1_a/1_b$ and $2_a/2_b$ and $> 40\%$ in $2_a/2_c$ and $2_b/2_c$ as DMCs. The latter can be the result of 2_c having the lowest coverage (1.16 ± 0.02) among the replicates (Supplementary Table ??). This suggest that MOABS predictions might be unstable in the presence of very low coverage.

We then pooled together the technical replicates for both ESCs and fibroblasts and performed a comparison within the biological replicates, again seeing only a marginal number of differences for CMeth, DSS and MOABS and the opposite for ComMet (Supplementary Table ??).

Next we set to assess the robustness of each method to low sequencing depth. We subsampled the aligned reads for all biological replicates to 75%, 50%, and 25% coverage and performed both between- and within- replicate comparisons for ESCs and fibroblasts. Results were summarized in terms

	$1_a/1_b$	$2_a/2_b$	$2_a/2_c$	$2_b/2_c$
CMeth	0.00%	0.01%	0.02%	0.02%
ComMet	12.57%	12.19%	42.29%	41.12%
DSS	—	—	—	—
MOABS	0.14%	0.12%	0.01%	0.02%

Table 3. Fraction of DMCs in CG context on the first chromosome of the hg19 reference genome between the technical replicates of the human ESCs. FDR controlled at level $\alpha = 10^{-4}$. Biological replicates are identified with a digit (1 or 2), technical replicates — with a character (a, b or c).

of the cytosine status changes between the full and subsampled data. It is reasonable to assume that a small number of new DMCs might appear by chance due to the random nature of the subsampling procedure. Thus the output of a method on subsampled data is expected to deviate slightly from that on full data. A large number of deviations, however, might indicate the lack of robustness. CMeth, DSS, and MOABS produced few new DMCs in the low coverage setting, as evident from Figure ?? and Supplementary Figure ??. As previously, ComMet overestimated the number of DMCs with more new DMCs appearing at lower sequencing depth.

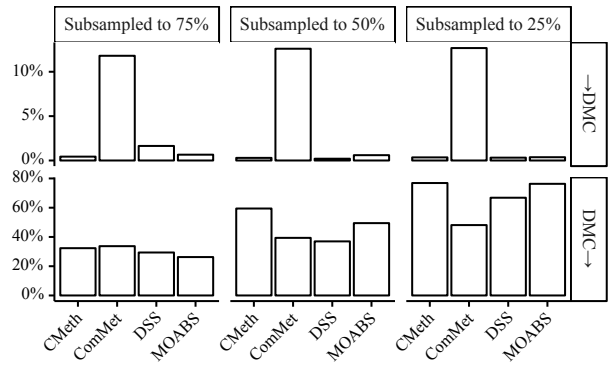

Figure 4. Fraction of newly "discovered" (→DMC) and lost (DMC→) DMCs in the comparisons of subsampled data on the first chromosome of the hg19 reference genome. Comparisons were performed between the biological replicates of the human ESCs and human fibroblasts.

Then we looked at the fraction of DMCs loosing DMC status after subsampling (Figure ??). Being the most conservative method, DSS demonstrated the highest rejection rate on subsampled data, followed by CMeth, MOABS and ComMet. The performance of CMeth was on par with MOABS, albeit MOABS exhibited a slightly lower rejection rate than CMeth, which is likely to be a consequence of low specificity of MOABS.

Finally we compared the human ESCs to human fibroblasts using both biological replicates. Following ? we used Fisher's exact test as a baseline method. ComMet was excluded from comparison because it doesn't support replicated data. Interestingly, we found that even though MOABS accepts replicates as input, it doesn't include them into the model and instead combines into a single file prior to analysis.

As expected, all of the methods discerned more DMCs than Fisher exact test with FDR controlled at the same level, see Table ??. We then overlapped the resulting DMCs (Figure ??) and discovered significant agreement between the predictions produced by different methods. Interestingly, MOABS reported more DMCs than any other method. We argue that this is a consequence of ignoring the biological variation by pooling the replicates. CMeth also produced a relatively large number of DMCs, but its predictions were more in line with that of DSS and Fisher exact test (Supplementary Figure ??).

Method	FET	CMeth	DSS	MOABS
# of DMCs	47,778 (2.91%)	194,012	60,739	219,267

Table 4. Number of DMCs in CG context on the first chromosome of the hg19 reference genome in the comparison of the human ESCs and human fibroblasts.

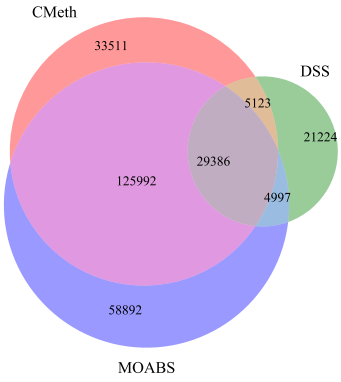

Figure 5. Venn diagram of the DMCs found by CMeth, DSS and MOABS on the first chromosome of the hg19 reference genome in the comparison of the human ESCs and human fibroblasts.

4.2.2 Atherosclerosis data We compared the WGBS data for normal aortic tissue and atherosclerosis lesion against the Illumina HumanMethylation450 BeadChip array data, which contains 15 biological replicates for both conditions. As the microarray data does not allow for DMR detection we focused on evaluating the differences between the predictions for individual cytosines.

The DMCs reported by dmpFinder from the minfi package were intersected with the ones identified by CMeth at FDR $\leq 10^{-4}$. To our surprise the degree of agreement between the results produced by the two methods was low. Specifically, CMeth and minfi identified 99,052 and 20,874 DMCs, the number of commonly identified DMCs however was only 1,463. We hypothesized that the disagreement is not caused by reduced power of the analysis methods and is instead explained by data heterogeneity. And indeed the distribution of the methylation levels obtained from WGBS data diverges from the one obtained from microarray data (Figure ??).

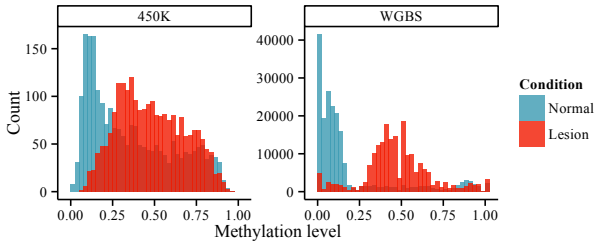

Figure 6. The histogram of methylation levels obtained from Illumina HumanMethylation450 BeadChip array complemented by the corresponding methylation levels from WGBS data.

The analysis of ? identified 7 genes containing differentially methylated cytosines and showed by qRT-PCR that these genes are also differentially expressed between normal aortic tissue and atherosclerosis lesion. We attempted to confirm the methylation differences found by ?. To do so, we counted the fraction of DMCs in promoters regions, exons and introns for each of the 7 genes (Table ??) and compared our results to genomic contexts reported by ?. For all genes except PLAT the majority of differences occurred in the expected context. The fraction of differentially methylated cytosines in each context however is marginal, with HOXA6 and HOXA9 being an exception.

To conclude our analysis we focused on the specific cytosines within the atherosclerosis-linked genes reported to be differentially methylated. For each cytosine we collected the methylated and total coverage in both samples and DMC status as predicted by CMeth (Supplementary Table ??). CMeth predictions were consistent with ? for 5 out of 10 cytosines. For the remaining cytosines available WGBS data does not exhibit any differences.

Gene	Promoter	Exons	Introns
HOXA6	*27.65%	9.24%	0%
HOXA9	2.67%	*5.73%	*36.7%
PDGFA	0.24%	0%	*2.17%
PLAT	2.13%	*1.03%	2.60%
PRRX1	0%	0%	*2.36%
PXDN	0%	*2.69%	2.82%
MIR23b	*12.84%	0%	0%

Table 5. Fraction of the DMCs in promoter regions, exons and introns for each of the 7 differentially expressed genes. The asterisk (*) marks the context of the DMC as reported by ?.

5 DISCUSSION

In this article, we described CMeth, a new tool for comparison of WGBS experiments. The major contributions of this work are twofold. First we proposed a Bayesian extension of the hidden Markov model, which incorporates distances between consecutive observations into the inference process. Unlike ComMet our model is semiparameteric in a sense that it doesn't group the distances into a fixed number of clusters; instead, an appropriate number of clusters is determined during inference. Second, we used the work of ? on dependent hypothesis testing to develop a Q-value estimation procedure for methylation differences; Q-values can then be thresholded independently of CMeth to control FDR at an arbitrary level.

The model of CMeth implicitly assumes the linearity of spatial dependencies along the genome, that is, the more distant a cytosine is from its neighbour, the less it affects the neighbours methylation state. While this might be true for some regions of the genome, in general, the validity of the assumption is questionable. A more realistic model of dependence might incorporate the conformation structure of the DNA molecule instead of linear genomic distances. Recent advances in the field allow to obtain high-resolution DNA conformation data via methods such as Hi-C (?). Another direction of work is related to the inference of the semiparameteric switching HMM. The quality of posterior estimates can be improved by applying ideas from tensor decomposition methods (?), while the time and memory requirements might benefit from exploring stochastic gradient approaches (???).

Funding: This work was supported by JetBrains.

1 MODEL DESCRIPTION

Bisulfite sequencing allows to detect methylation events for a cytosine in the reference genome by counting the number of mapped reads carrying either C (methylated cytosine) or T (bisulfite converted unmethylated cytosine). Thus for each cytosine t in the sample s we have k_{tr} — methylated coverage and n_{tr} — total coverage.

Given a two biological conditions each represented by a number of samples we want to find regions differentially methylated between two conditions. To do so we first want to assign each cytosine a label from the set $\{U, \uparrow, \downarrow, M\}$, where U and M mark cytosines methylated or unmethylated in both biological conditions, and $\uparrow$, $\downarrow$ — differentially methylated cytosines.

We use the framework of probabilistic models to infer the hidden state labels from bisulfite sequencing results $\mathbf{x}$.

1.1 Nonparamteric distance mixture

$$
\begin{aligned}
p(\mathbf{v}|\lambda) &= \prod_{c=1}^{\infty} \text{Beta}\left(v_c|1,\lambda\right) = \prod_{c=1}^{\infty} \frac{(1-v_c)^{\lambda}}{\text{B}(1,\lambda)} \\
p(\mathbf{w}) &= \mathbb{E}_{\mathbf{v}}\left[\prod_{t=2}^{T} \text{Cat}\left(w_t|\kappa\right)\right] \quad \kappa_c \doteq v_c \prod_{d=1}^{c-1}(1-v_d) \\
p(\mathbf{q}) &= \prod_{c=1}^{\infty} \text{Beta}\left(q_c|\alpha_0^{(\mathbf{q})},\beta_0^{(\mathbf{q})}\right) = \prod_{c=1}^{\infty} \frac{q_c^{\alpha_0^{(\mathbf{q})}-1}(1-q_c)^{\beta_0^{(\mathbf{q})}-1}}{\text{B}(\alpha_0^{(\mathbf{q})},\beta_0^{(\mathbf{q})})} \\
p(\mathbf{d}|\mathbf{z},\mathbf{q}) &= \prod_{t=2}^{T}\prod_{c=1}^{\infty} \text{Geom}\left(d_t|p_c\right)^{w_{ct}} \\
&= \prod_{t=2}^{T}\prod_{c=1}^{\infty} \left[(1-q_c)^{d_t-y} q_c\right]^{w_{ct}}
\end{aligned}
$$

1.2 Binomial switching hidden Markov model

$$
\begin{aligned}
p(\pi|\omega_0^{(\pi)}) &= \text{SymDir}\left(\pi|\omega_0^{(\pi)}\right) = \frac{\Gamma(\omega_0^{(\pi)}S)}{\Gamma(\omega_0^{(\pi)})^S}\prod_{i=1}^{S}\pi_i^{\omega_0^{(\pi)}-1} \\
p(\mathbf{A}|\omega_0^{(\mathbf{A})}) &= \prod_{c=1}^{C}\prod_{e=1}^{E} \text{SymDir}\left(\mathbf{A}_{ci}|\omega_0^{(\mathbf{A})}\right) = \prod_{c=1}^{C}\prod_{e=1}^{E}\frac{\Gamma(\omega_0^{(\mathbf{A})}S)}{\Gamma(\omega_0^{(\mathbf{A})})^S}\prod_{i=1}^{S}\mathbf{A}_{ci}^{\omega_0^{(\mathbf{A})}-1} \\
p(p_e|\alpha_0^{(\mathbf{p})},\beta_0^{(\mathbf{p})}) &= \prod_{e=1}^{E} \text{Beta}\left(p_e|\alpha_0^{(\mathbf{p})},\beta_0^{(\mathbf{p})}\right) = \prod_{e=1}^{E}\frac{p_e^{\alpha_0^{(\mathbf{p})}-1}(1-p_e)^{\beta_0^{(\mathbf{p})}-1}}{\text{B}(\alpha_0^{(\mathbf{p})},\beta_0^{(\mathbf{p})})} \\
p(\mathbf{k}|\mathbf{n},\mathbf{p}) &= \prod_{e=1}^{E}\prod_{r=1}^{R}\prod_{i=1}^{S}\prod_{t=1}^{T}\left[\text{Binom}\left(k_{tr}|n_{tr},p_e\right)\right]^{\mathbb{I}(\mathbf{D}_{ri}=e)z_{it}} \\
&= \prod_{e=1}^{E}\prod_{r=1}^{R}\prod_{i=1}^{S}\prod_{t=1}^{T}\left[\binom{n_{tr}}{k_{tr}} p_{\mathbf{D}_{ri}}^{k_{tr}}(1-p_{\mathbf{D}_{ri}})^{n_{tr}-k_{tr}}\right]^{\mathbb{I}(\mathbf{D}_{ri}=e)z_{it}}
\end{aligned}
$$

The joint log-likelihood of model parameters $\mathbf{\Theta}$, latent indicator variables $\mathbf{w}$ and $\mathbf{z}$ and observations $\mathbf{x}$ can be factorized into two terms

$$\log p(\mathbf{\Theta},\mathbf{w},\mathbf{z},\mathbf{x}) = \log p(\mathbf{d},\mathbf{w},\mathbf{q}) + \log p(\mathbf{k},\mathbf{n},\mathbf{z},\mathbf{A},\pi|\mathbf{w}). \tag{1}$$

Here the first term corresponds to the joint log-likelihood of between-cytosine distances and geometric mixture parameters

$$\log p(\mathbf{d},\mathbf{w},\mathbf{q}) = \log p(\mathbf{d}|\mathbf{w},\mathbf{q}) + \log p(\mathbf{w}) + \log p(\mathbf{q}), \tag{2}$$

while the second term is the joint log-likelihood of methylation counts and parameters of the switching hidden Markov model

$$\log p(\mathbf{k},\mathbf{n},\mathbf{z},\mathbf{A},\pi|\mathbf{w}) = \log p(\mathbf{k}|\mathbf{n},\mathbf{z},\mathbf{p}) + \log p(\mathbf{z}|\mathbf{w},\pi,\mathbf{A}) + \log p(\mathbf{A}) + \log p(\pi) + \log p(\mathbf{p}). \tag{3}$$

Note that the first term doesn't contain stick-weights. Following ? we integrate out stick-weights $\mathbf{v}$ from the joint log-likelihood.

$$p(\mathbf{w},\mathbf{v}) = p(\mathbf{w}|\mathbf{v})p(\mathbf{v}) = \prod_{c=1}^{\infty}\prod_{t=1}^{T} v_c^{w_{ct}} \prod_{d=1}^{c-1}(1-v_d)^{w_{dt}} \frac{(1-v_c)^{\lambda-1}}{\mathrm{B}(1,\lambda)}$$

$$= \prod_{c=1}^{\infty} \frac{v_c^{T_c}(1-v_c)^{\lambda+T_{>c}-1}}{\mathrm{B}(1,\lambda)} = \prod_{c=1}^{\infty} \lambda \frac{\Gamma(1+T_c)\Gamma(\lambda+T_{>c})}{\Gamma(1+\lambda+T_{\geq c})} \mathrm{Beta}\left(v_c|\alpha_c,\beta_c\right), \tag{4}$$

where we've denoted $T_c = \sum_{t=1}^{T} w_{ct}$ and $T_{>c} = \sum_{d=c+1}^{C} T_c$. Integrating w.r.t. $\mathbf{v}$ we get

$$p(\mathbf{w}) = \int p(\mathbf{w},\mathbf{v})d\mathbf{v} = \prod_{c=1}^{\infty} \lambda \frac{\Gamma(1+T_c)\Gamma(\lambda+T_{>c})}{\Gamma(1+\lambda+T_{\geq c})} \tag{5}$$

2 DERIVATION OF THE MEAN-FIELD VARIATIONAL ALGORITHM

We apply mean-field variational method (?) for approximate posterior inference. The mean-field approximation assumes the following factorization of the joint posterior distribution[2]

$$q(\mathbf{w},\mathbf{q},\mathbf{z},\pi,\mathbf{A},\mathbf{p}) = q(\mathbf{w},\mathbf{z})q(\mathbf{q},\pi,\mathbf{A},\mathbf{p}). \tag{6}$$

The algorithm consists of iterating two steps, E for expectation and M for maximization, until convergence. Below we derive both steps for the proposed model.

2.1 E-step

The E-step computes the variational approximation to the posterior distribution of the latent variables $\mathbf{w}$ and $\mathbf{z}$. Taking the expecation w.r.t. all parameters except $\mathbf{w}$ and $\mathbf{z}$ and then omitting the terms independent of either $\mathbf{w}$ or $\mathbf{z}$ we get

$$\log q^*(\mathbf{w},\mathbf{z}) \propto \mathbb{E}\left[\log p(\mathbf{w})\right] + \mathbb{E}\left[\log p(\mathbf{d}|\mathbf{w},\mathbf{q})\right] + \mathbb{E}\left[\log p(\mathbf{z}|\mathbf{w},\pi,\mathbf{A})\right] + \mathbb{E}\left[\log p(\mathbf{k}|\mathbf{n},\mathbf{z},\mathbf{p})\right]. \tag{7}$$

The first two terms in the expression above came from distance mixture part of the joint log-likelihood. Expanding the second term

$$\mathbb{E}\left[\log p(\mathbf{d}|\mathbf{w},\mathbf{q})\right] = \sum_{c=1}^{C}\sum_{t=2}^{T} w_{ct}\left((d_t - y)\mathbb{E}\left[\log(1-q_c)\right] + \mathbb{E}\left[\log q_c\right]\right). \tag{8}$$

Expanding the last two terms, which correspond to the binomial switching HMM part of the joint log-likelihood, we get

$$\mathbb{E}\left[\log p(\mathbf{z}|\mathbf{w},\pi,\mathbf{A})\right] = \sum_{i=1}^{S} z_{i1}\mathbb{E}\left[\log \pi_i\right] + \sum_{j=1}^{S}\sum_{c=1}^{C}\sum_{t=2}^{T} w_{ct} z_{i(t-1)} z_{jt} \mathbb{E}\left[\log \mathbf{A}_{cij}\right]$$

$$\mathbb{E}\left[\log p(\mathbf{k}|\mathbf{n},\mathbf{z},\mathbf{p})\right] \propto \sum_{e=1}^{E}\sum_{r=1}^{R}\sum_{i=1}^{S} \mathbb{I}(\mathbf{D}_{ri} = e) \sum_{t=1}^{T} z_{it}\left(k_{tr}\mathbb{E}\left[\log p_e\right] + (n_{tr} - k_{tr})\mathbb{E}\left[\log(1-p_e)\right]\right)$$

We now focus on the first term which is a little trickier to compute. Rewriting $p(\mathbf{w})$ in terms of the individual distributions of component assignments

$$\mathbb{E}\left[\log p(\mathbf{w})\right] = \sum_{c=1}^{C}\sum_{t=1}^{T} \mathbb{E}_{\mathbf{w}^{\neg t}}\left[\log p(w_{ct}|\mathbf{w}^{\neg t})\right] = \sum_{c=1}^{C}\sum_{t=1}^{T} \mathbb{E}_{\mathbf{w}^{\neg t}}\left[\log p(w_{ct}|\mathbf{w}^{\neg t})\right]$$

The conditional $p(w_{ct}|\mathbf{w}^{\neg t})$ is given by

$$p(w_t|\mathbf{w}^{\neg t}) = \frac{p(\mathbf{w})}{\mathbf{w}^{\neg t}} = \frac{\Gamma(1+T_c)\Gamma(\lambda+T_{>c})}{\Gamma(1+\lambda+T_{\geq c})} \frac{\Gamma(1+\lambda+T_{\geq c}^{\neg t})}{\Gamma(1+T_c^{\neg t})\Gamma(\lambda+T_{>c}^{\neg t})} \tag{9}$$

$$= \frac{(1+T_c^{\neg t})}{(1+\lambda+T_{\geq c}^{\neg t})} \prod_{d=1}^{c-1} \frac{(\lambda+T_{>c}^{\neg t})}{1+\lambda+T_{\geq d}^{\neg t}}, \tag{10}$$

where the last step is due to the fact that $T_{>d}^{\neg t}$ is different from T_d only for $d < c$.

[2] We use $q(\theta)$ to denote the variational approximation to the the true distribution $p(\theta)$ of parameter θ.

Following ?, we approximate the conditional using second-order Taylor expansion

$$\log p(w_{ct}|\mathbf{w}^{\neg t}) \approx \log\left(1+\mathbb{E}\left[T_c^{\neg t}\right]\right) + \frac{1}{2}\left(1+\mathbb{E}\left[T_c^{\neg t}\right]\right)^{-2}\mathbb{V}\left[T_c^{\neg t}\right] - D_c + \sum_{d=1}^{c-1}\log\left(\lambda+\mathbb{E}\left[T_{>d}^{\neg t}\right]\right) + \frac{1}{2}\left(1+\mathbb{E}\left[T_{>d}^{\neg t}\right]\right)^{-2}\mathbb{V}\left[T_{>d}^{\neg t}\right] - D_d \tag{11}$$

$$D_c = \log\left(1+\lambda+\mathbb{E}\left[T_{\geq c}^{\neg t}\right]\right) + \frac{1}{2}\left(1+\lambda+\mathbb{E}\left[T_{\geq c}^{\neg t}\right]\right)^{-2}\mathbb{V}\left[T_{\geq c}^{\neg t}\right]. \tag{12}$$

The remaining expectations can be approximated by Gaussian distributions with means and variances given by

$$\begin{aligned} \mathbb{E}\left[T_{>c}\right] &= \sum_{t=1}^{T}\sum_{d=c+1}^{C} z_{dt}q(z_{dt}) & \mathbb{V}\left[T_{>c}\right] &= \sum_{t=1}^{T}\sum_{d=c+1}^{C} z_{dt}q(z_{dt})\sum_{d=1}^{c} z_{dt}q(z_{dt}) \\ \mathbb{E}\left[T_{\geq c}\right] &= \mathbb{E}\left[T_{>c}\right] + \mathbb{E}\left[T_c\right] & \mathbb{V}\left[T_{\geq c}\right] &= \sum_{t=1}^{T}\sum_{d=c}^{C} z_{dt}q(z_{dt})\sum_{d=1}^{c-1} z_{dt}q(z_{dt}). \end{aligned} \tag{13}$$

2.2 M-step

To derive the M-step for parameter θ we take the expectation w.r.t. all parameters except θ and then omit the terms independent of θ. For parameters with conjugate priors the resulting expression is of the same form as the prior.

2.2.1 Distance mixture parameters

Success probabilities

$$\begin{aligned} \log q^*(\mathbf{q}) &\propto \mathbb{E}\left[\log p(\mathbf{d}|\mathbf{w},\mathbf{q})\right] + \log p(\mathbf{q}) \\ &\propto \sum_{t=2}^{T}\sum_{c=1}^{C}\mathbb{E}\left[w_{ct}\right]\left(d_t\log(1-q_c)+\log q_c\right) + \sum_{c=1}^{C}(\alpha_0^{(\mathbf{q})}-1)\log q_c + (\beta_0^{(\mathbf{q})}-1)\log(1-q_c) \\ &\propto \sum_{c=1}^{C}(\alpha_0^{(\mathbf{q})}+\sum_{t=2}^{T}\mathbb{E}\left[w_{ct}\right]-1)\log q_c + (\beta_0^{(\mathbf{q})}+\sum_{t=2}^{T}\mathbb{E}\left[w_{ct}\right]d_t-1)\log(1-q_c) \\ &\propto \sum_{c=1}^{C}\mathrm{Beta}\left(q_c|\alpha_c^{(\mathbf{q})},\beta_c^{(\mathbf{q})}\right) \end{aligned}$$

which gives us the update

$$\alpha_c^{(\mathbf{q})} = \alpha_0^{(\mathbf{q})} + \sum_{t=2}^{T}\mathbb{E}\left[w_{ct}\right] \tag{14}$$

$$\beta_c^{(\mathbf{q})} = \alpha_0^{(\mathbf{q})} + \sum_{t=2}^{T}\mathbb{E}\left[w_{ct}\right]d_t \tag{15}$$

2.2.2 Switching hidden Markov model parameters

Initial state probabilities

$$\log q^*(\pi) \propto \mathbb{E}\left[\log p(\mathbf{z}|\mathbf{w},\pi,\mathbf{A})\right] + \mathbb{E}\left[\log p(\pi)\right] \propto \sum_{i=1}^{S}\left(\omega_0^{(\pi)}+\mathbb{E}\left[z_{i1}\right]-1\right)\log\pi_i,$$

which gives us the update

$$\omega_i^{(\pi)} = \omega_0^{(\pi)} + \mathbb{E}\left[z_{i1}\right]. \tag{16}$$

State transition probabilities

$$\begin{aligned}\log q^*(\mathbf{A}) &\propto \mathbb{E}\left[\log p(\mathbf{z}|\mathbf{w},\pi,\mathbf{A})\right] + \mathbb{E}\left[\log p(\mathbf{A})\right] \\ &\propto \sum_{c=1}^{C}\sum_{i=1}^{S}\sum_{j=1}^{S}\left(\omega_0^{(\mathbf{A})} + \sum_{t=2}^{T}\mathbb{E}\left[w_{ct}z_{i(t-1)}z_{jt}\right] - 1\right)\log \mathbf{A}_{cij}.\end{aligned}$$

which gives us the update

$$\omega_{cij}^{(\mathbf{A})} = \omega_0^{(\mathbf{A})} + \sum_{t=2}^{T}\mathbb{E}\left[z_{i(t-1)}z_{jt}w_{ct}\right]. \tag{17}$$

Methylation rates

$$\begin{aligned}\log q^*(\mathbf{p}) &\propto \mathbb{E}\left[\log p(\mathbf{k}|\mathbf{n},\mathbf{z},\mathbf{p})\right] + \mathbb{E}\left[\log p(\mathbf{p})\right] \\ &\propto \sum_{e=1}^{E}\sum_{r=1}^{R}\sum_{i=1}^{S}\mathbb{I}(\mathbf{D}_{ri}=e)\sum_{t=1}^{T}\mathbb{E}\left[z_{it}\right]\left(k_{tr}\log p_e + (n_{tr}-k_{tr})\log(1-p_e)\right) \\ &\qquad + \left(\alpha_0^{(\mathbf{p})}-1\right)\log p_e + \left(\beta_0^{(\mathbf{p})}-1\right)\log(1-p_e) \\ &\propto \sum_{e=1}^{E}\left(\alpha_0^{(\mathbf{p})} + \sum_{r=1}^{R}\sum_{i=1}^{S}\mathbb{I}(\mathbf{D}_{ri}=e)\sum_{t=1}^{T}\mathbb{E}\left[z_{it}\right]k_{tr} - 1\right)\log p_e \\ &\qquad + \left(\beta_0^{(\mathbf{p})} + \sum_{r=1}^{R}\sum_{i=1}^{S}\mathbb{I}(\mathbf{D}_{ri}=e)\sum_{t=1}^{T}\mathbb{E}\left[z_{it}\right](n_{tr}-k_{tr}) - 1\right)\log(1-p_e),\end{aligned}$$

which gives us the update

$$\alpha_e^{(\mathbf{p})} = \alpha_0^{(\mathbf{p})} + \sum_{r=1}^{R}\sum_{i=1}^{S}\mathbb{I}(\mathbf{D}_{ri}=e)\sum_{t=1}^{T}\mathbb{E}\left[z_{it}\right]k_{tr} \tag{18}$$

$$\beta_e^{(\mathbf{p})} = \beta_0^{(\mathbf{p})} + \sum_{r=1}^{R}\sum_{i=1}^{S}\mathbb{I}(\mathbf{D}_{ri}=e)\sum_{t=1}^{T}\mathbb{E}\left[z_{it}\right](n_{tr}-k_{tr}). \tag{19}$$

3 METHYLATION RANGES

We approximate empirical distribution of methylation levels with a discerete distribution over methylation ranges. To choose the breaks for the ranges we used bisulfite sequencing data for the human ESCs (?). We pooled together both biological replicates and calculated the distribution of methylation levels for each cytosine on the chr22 (Supplementary Figure ??).

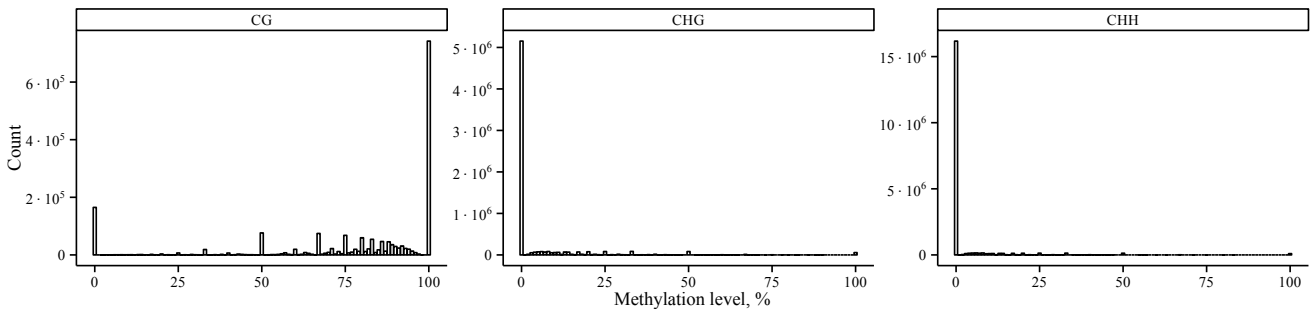

Figure 7. Per cytosine context histograms of methylation levels estimated from the 22-th chromosome of the human ESCs.

For cytosines in CG context the histogram exhibits peaks in the areas close to zero and one. Cytosines in CHG and CHH contexts remain largely unmethylated, both histograms have peaks around zero. Thus 0 and 1 methylation levels are good

candidates to be included in the list of breaks. Another interesting feature of the histograms is a peak around 0.5, which can be observed for all contexts. Thus we included 0.5 as the third break. To obtain the remaining two breaks we removed extreme 0 and 1values from the data and for the remaining data estimated the median methylation level separately for each cytosine context. For CG the median methylation level was 0.8, CHG — 0.13, CHH — 0.11. The medians for CHG and CHH were close, so we decided to average them into a single break 0.12.

The methylation ranges corresponding to the chosen breaks are [0, 0], (0, 0.12], (0.12, 0.5], (0.5, 0.8], (0.8, 1].

4 COMMAND LINE ARGUMENTS

4.1 ComMet

ComMet v1.1 from the Bisulfighter analysis suite was executed with default arguments, which assume the absence of the log-odds ratio scores thresholding for the DMRs identified by the dynamic programming algorithm.

4.2 MOABS

MOABS v1.2.9 was executed with default arguments using BED formatted inputs, as suggested by the MOABS authors in the MOABS Google Group post.

```
mcomp -p 16 -r sample1.bed -r sample2.bed -c output.txt
```

Although MOABS manual mentions the DMR detection feature it does not give the details of obtaining DMRs using MOABS, thus we used the data from the comparison file output.txt. Cytosines were considered differentially methylated if the absolute value of the CDIF statistic was at least 0.2, as suggested by the manual.

4.3 DSS

DSS v2.4.1 was downloaded from Bioconductor and applied as follows

```
run_dss <- function(input_path1, input_path2, dmp_path, dmr_path) {
    BS1 <- makeBSseqData(list(read.csv(input_path1, sep = "\t")), "1")
    BS2 <- makeBSseqData(list(read.csv(input_path2, sep = "\t")), "2")
    dml <- callDML(BS1, BS2, equal.disp = T)
    dmr <- callDMR(dml, p.threshold = 1e-4)
    write.table(dml, dmp_path, sep = "\t", quote = F)
    write.table(dmr, dmr_path, sep = "\t", quote = F)
}
```

The equal.disp argument is required because the simulated data is not replicated, which implies that the variances within the two compared conditions are equal to zero.

5 SUPPLEMENTARY FIGURES

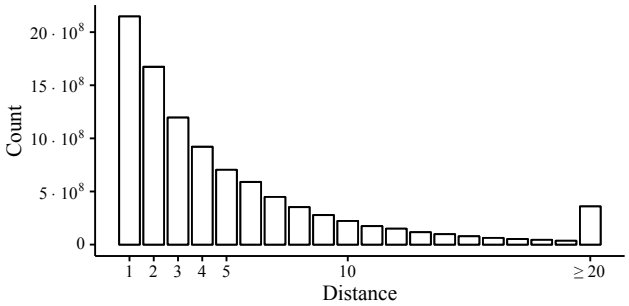

Figure 8. Histogram of distances between consecutive cytosines on the first chromosome of the mm10 reference genome.

6 SUPPLEMENTARY TABLES

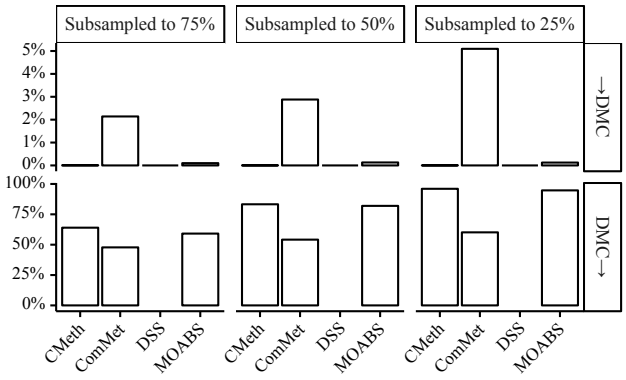

Figure 9. Fraction of DMCs newly "discovered" (→DMC) and lost (DMC→) in the comparisons of subsampled data on the first chromosome of the hg19 reference genome. Comparisons were performed within the biological replicates of the human ESCs and human fibroblasts.

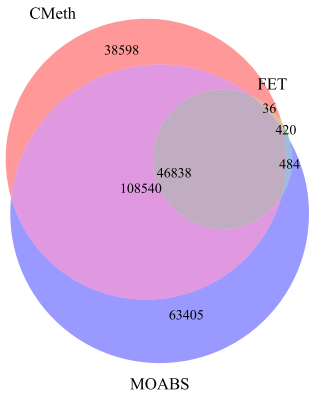

Figure 10. Venn diagram of the DMCs found by CMeth, FET (Fisher's Exact Test) and MOABS on the first chromosome of the hg19 reference genome in the comparison of the human ESCs and human fibroblasts.

Replicate	Mean	SD
1_a	5.79	86.96
1_b	5.37	83.48
2_a	7.66	100.34
2_b	17.60	444.75
2_c	1.16	5.59

Table 6. Coverage mean and standard deviation on the first chromosome of the hg19 for technical replicates of the human ESCs (?).

	$1_a/2_a$	$1_a/2_b$	$1_a/2_c$	$1_b/2_a$	$1_b/2_b$	$1_b/2_c$
CMeth	0.05%	0.04%	0.03%	0.05%	0.04%	0.01%
ComMet	13.77%	13.50%	40.68%	13.41%	13.12%	40.09%
DSS	—	—	—	—	—	—
MOABS	0.21%	0.20%	0.01%	0.17%	0.15%	0.02%

Table 7. Fraction of DMCs in CG context on the first chromosome of the hg19 reference genome between the technical replicates of the human ESCs (?). Biological replicates are identified with a digit (1 or 2), technical replicates — with a character (a, b or c).

	ESC	IMR90
CMeth	0.09%	0.02%
ComMet	18.33&	2.08%
DSS	0.00%	—
MOABS	0.21%	0.18%

Table 8. Fraction of differentially methylated and covered cytosines in CG context on the first chrosomosome of the hg19 reference genome between the biological replicates of the human ESCs and fibroblasts (?). FDR controlled at level $\alpha = 10^{-4}$.

Gene	ID	Lesion	Normal	Status
HOXA6	cg19816811	$^{2}/_{17}$	$^{26}/_{32}$	DMC
HOXA9	cg03217995	$^{5}/_{19}$	$^{33}/_{39}$	DMC
HOXA9	cg16913789	$^{0}/_{10}$	$^{4}/_{29}$	—
HOXA9	cg25188395	$^{0}/_{11}$	$^{1}/_{17}$	—
PDGFA	cg14496282	$^{0}/_{3}$	$^{1}/_{9}$	—
PLAT	cg01419713	$^{9}/_{17}$	$^{9}/_{28}$	—
PRRX1	cg21914290	$^{11}/_{25}$	$^{43}/_{50}$	DMC
PXDN	cg07608848	$^{5}/_{10}$	$^{4}/_{22}$	—
PXDN	cg15796818	$^{6}/_{12}$	$^{2}/_{25}$	DMC
MIR23b	cg00351472	$^{14}/_{39}$	$^{2}/_{28}$	DMC

Table 9. Methylation and total coverage for individual DMCs within the genes linked to atherosclerosis. The status column contains CMeth prediction at FDR $\leq 10^{-4}$.