Grosser and Metzler *BMC Bioinformatics* (2020) 21:115
https://doi.org/10.1186/s12859-020-3438-5

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Modeling methylation dynamics with simultaneous changes in CpG islands

Konrad Grosser and Dirk Metzler*

Abstract

Background: In vertebrate genomes, CpG sites can be clustered into CpG islands, and the amount of methylation in a CpG island can change due to gene regulation processes. Thus, single regulatory events can simultaneously change the methylation states of many CpG sites within a CpG island. This should be taken into account when quantifying the amount of change in methylation, for example in form of a branch length in a phylogeny of cell types.

Results: We propose a probabilistic model (the IWE-SSE model) of methylation dynamics that accounts for simultaneous methylation changes in multiple CpG sites belonging to the same CpG island. We further propose a Markov-chain Monte-Carlo (MCMC) method to fit this model to methylation data from cell type phylogenies and apply this method to available data from murine haematopoietic cells and from human cell lines. Combined with simulation studies, these analyses show that accounting for CpG island wide methylation changes has a strong effect on the inferred branch lengths and leads to a significantly better model fit for the methylation data from murine haematopoietic cells and human cell lines.

Conclusion: The MCMC based parameter estimation method for the IWE-SSE model in combination with our MCMC based inference method allows to quantify the amount of methylation changes at single CpG sites as well as on entire CpG islands. Accounting for changes affecting entire islands can lead to more accurate branch length estimation in the presence of simultaneous methylation change.

Keywords: Methylation, CpG islands, Sequence evolution, Markov-chain Monte Carlo, Reversible jump

Background

Epigenetic processes of DNA-methylation and - demethylation are strongly associated with differential gene expression and are essential during phenotypic development in mammals [28]. The most frequent form of methylation is the attachment of the methyl group at the fifth carbon position on a CpG site, that is, a cytosine nucleotide followed by a guanine nucleotide [7, 26, 28].

Regions in which more than 50% of the sites are either G or C are called CpG islands if the number of CpGs is greater than 60% of the expected number of CpG sites by random order [26, 28]. These regions are typically between a few hundred and two thousand base pairs in length [28]. CpG islands are involved in the regulation of gene transcription [5]. Comparisons of methylation states have been commonly applied and proved as a fruitful avenue of analysis of cell haematopoiesis [2, 33]. Pairwise comparison between cell types in different stages of differentiation or comparison between malignant and healthy cells during cancer development have provided insight into areas of transcription [2] and enabled inference of missing methylation states.

[4] have adapted phylogenetic methods to account for the tree-shaped genealogy of cell types when analyzing methylation changes during haematopoiesis. The branch lengths of the genealogy, representing expected numbers of methylation changes per site, were inferred via likelihood maximization. A common simplifying assumption in phylogenetics is that sequence positions evolve independently of each other [20]. Analogously, [4] assume

*Correspondence: metzler@bio.lmu.de
Department of Biology, Ludwigs-Maximilians Universität München, Großhaderner Straße 2, 82152 Planegg, Germany

that the methylation processes at all CpG sites are, conditioned on the genealogy, stochastically independent of each other. This model assumption is violated when, for example, methylation frequencies change in an entire CpG island in the course of gene regulation [5, 28].

The CpG methylation-demethylation model of [29] is based on the roles of the different DNA methylases in methylation maintenance and de-novo methylation (see also [1]). Also [16] take into account that methylation dynamics of CpG sites depend on the methylation rate in the surrounding DNA region, as well as other factors such as chromatin marks. [22] extended the model of [29] to account for the fact that the activity of enzymes that maintain methylation or lead to de-novo methylation of CpG sites depends on the methylation status of neighboring CpG sites. Further extensions of this model have been applied by [21] and [19] to disentangle the roles of the different DNA methylases. Meyer and Lacey [24] proposed another generalization of the model of [29], in which the effect of the methylation state of a CpG sites on other CpG sites depends on the distance between the sites. Other recent attempts at modeling neighboring methylation states made use of Ising models known from statistical mechanics [17]. There sequences in genomic regions are assumed to be subject to a model based likelihood governed by few parameters, resulting in an accurate characterization of methylation frequencies [17].

The models mentioned above have been very useful for the understanding of molecular mechanisms and function of DNA methylation. For the analysis of genome-wide methylation data and phylogenetic analyses thereof, however, the size of data sets necessitates simpler models that are compatible with efficient algorithms.

Our approach is to combine a model for independent methylation state evolution in CpG sites (single-site events, SSEs) with simultaneous changes of methylation states and rates within CpG islands (island-wide events, IWEs). The purpose of IWEs in our model is to cover variations in methylation frequencies among CpG islands as well as in time, that is among different cell types or other taxa in the tree of interest. An interpretation of an IWE could be, for example, that a CpG island in the promoter region of a gene becomes widely methylated or demethylated as part of gene regulation. For computational tractability, however, we neglect functional constraints in gene regulation that could lead to correlations between methylation rates among some CpG islands. We aim to both take simultaneous methylation changes into account and to propose a model that is at the same time simple enough for inference on large datasets with data from several cell types. With this, we aim to fill the gap in the literature where models have so far either just considered evolution of single sites [4], or concerned themselves with inference on smaller scales [19, 21, 24].

Like [4] we assume that CpG sites are affected by SSEs that change the state between unmethylated, methylated and partially methylated. We assume that the equilibrium probabilities of these three states and thus also the transition rates between them depend on the CpG island to which the site belongs. Conditioned on the specific rates within a CpG island, SSEs occur independently of each other. Some sequence evolution models developed for phylogenetic analyses allow that mutation rates change at random time points, see e.g. [15]. Here, we adapt this approach to CpG methylation-demethylation dynamics. We assume that CpG islands are at certain rates affected by events – the aforementioned IWEs – that change the equilibrium probabilities for all sites in the CpG island. We further assume that these events can simultaneously change the methylation states of some CpG sites in the island such that the new equilibrium probabilities apply immediately to the sites in the affected CpG island. We refer to our model allowing for both SSEs and IWEs as IWE-SSE model.

We have implemented a reversible-jump MCMC inference scheme [11, 13, 30] to fit this model to Reduced Restricted Bisulfite Sequencing (RRBS) methylation data [2, 23]. We validated the accuracy of this scheme in a simulation study. With RRBS data procured from mouse haematopoiesis [2, 4] we demonstrate that accounting for IWEs can lead to significantly different estimations of branch lengths of cell type genealogies. We further apply the inference scheme based on the IWE-SSE model to experimental data monitoring the differentiation of human germ cells in vitro [32]. We there compare it with an inference scheme without IWEs and with the `lyne` software package to gain independent evidence that considering IWEs improves branch estimates.

Methods

Structure of our methylation-demethylation model

We take into account that several CpGs can form a CpG island, which can be affected by CpG island wide events (IWEs), in which methylation probabilities change and some of the CpG sites in the CpG island can simultaneously change their state at the same time (see Fig. 1). Different CpG islands, however, are assumed to evolve independently of each other. In addition to IWEs we allow for single-site events (SSEs), which change the methylation states of single CpG sites within CpG islands. Following [4] we distinguish three possible states $\{u, p, m\}$ of a CpG site, denoting unmethylated, partially methylated and methylated sites. When analyzing methylation sequencing data (see "Application to the murine haematopoietic system" and "Application to in vitro cell differentiation data" sections) we classify a site as unmethylated if it is methylated in less than 10% of the reads overlapping the site, partially methylated if it is detected as methylated in

10 to 80% of the reads, and methylated if it is methylated in more than 80% of the reads.

For a branch with h IWEs set $t_0 = 0$, let t_{h+1} be the branch length, and let $t_1, \ldots, t_h$ be the branch-length distances of the IWEs to the parent node with

$$t_0 \le t_1 \le t_2 \le \cdots \le t_h \le t_{h+1}. \tag{1}$$

Here we use $h + 1$ as index of the branch length, since if there are h IWEs between a node and its parent, the resulting number of time intervals along the branch is $h + 1$. For each CpG position and each open interval (t_k, t_{k+1}) there is a rate matrix Q_k for the transitions between the states u, p, m. Before we specify the details of our model for the IWE process ("Process of island-wide events (IWEs)" section) we first define our model for SSEs between given IWEs.

Process of single-site events (SSEs)

Here we specify our model assumptions for SSEs affecting CpG sites within a CpG islands between two IWEs at time points t_k and t_{k+1}. In the open interval (t_k, t_{k+1}), the methylation dynamics of CpGs of the same island are independent of each other and the matrix P_k of transition probabilities

$$P_{k;i,j} = \Pr(X_{t_{k+1}} = j | X_{t_k} = i) \tag{2}$$

between the methylation states X_{t_k} and $X_{t_{k+1}}$ of a CpG at time points t_k and t_{k+1} can be calculated with the matrix exponential

$$P_k = \exp(Q_k \cdot (t_{k+1} - t_k)). \tag{3}$$

In analogy to the F81 sequence evolution model [9] we focus here on rate matrices Q_k that can be expressed as

$$Q_k = R \cdot \begin{bmatrix} -\pi_p - \pi_m & \pi_p & \pi_m \\ \pi_u & -\pi_u - \pi_m & \pi_m \\ \pi_u & \pi_p & -\pi_u - \pi_p \end{bmatrix}, \tag{4}$$

where $\pi_u + \pi_p + \pi_m = 1$, and each CpG has its own random rate factor $R \in \mathrm{R}_{\ge 0}$. Starting from any state u, p or m the form of Q_k implies that in the case of an event the state in the site becomes u, p or m with probabilities π_u, π_p, and π_m, respectively. Note that this does not exclude the possibility that the state after the event is the same as before at the affected site. Since the three probabilities sum up to 1, the rate factor R directly gives the expected number of events per time unit.

The probabilities (π_u, π_p, π_m) form the equilibrium distribution of Q_k. Further, for fixed R the transition probabilities $(P_k)_{i,j}$ fulfill

$$(P_k)_{i,j} = \left(1 - e^{R \cdot (t_k - t_{k+1})}\right) \cdot \pi_j, \tag{5}$$

if $j \neq i$, and

$$(P_k)_{i,j} = \pi_j + (1 - \pi_j) \cdot e^{R \cdot (t_k - t_{k+1})} \tag{6}$$

otherwise. We assume that in the root of the genealogy each CpG island samples the equilibrium probability triple (π_u, π_p, π_m) from a uniform distribution (that is Dirichlet(1,1,1)) independently of all other CpG islands. Like in F81 and related models, the time scaling in our models can be interpreted as follows. At each CpG site with the respective rate R, events occur that let the CpG sample a new state u, p or m according to the probabilities (π_u, π_p, π_m). We refer to these events as SSEs.

Process of island-wide events (IWEs)

In each IWE a new triple of equilibrium methylation frequencies (π'_u, π'_p, π'_m) is sampled from a uniform distribution, and Q_t is updated accordingly for time points t after the IWE. Furthermore, we allow that CpG sites of an island are methylated or demethylated simultaneously in an IWE in a way such that the expected frequencies of the states u, p and m match the new equilibrium distribution (π'_u, π'_p, π'_m) right after the IWE. To specify the transition probability matrix M_k in an IWE at a time point t_k, we distinguish two cases. In the first case one of the new expected frequencies is larger and the other two are smaller after the IWE. If, without loss of generality, $\pi'_u > \pi_u, \pi'_p < \pi_p$ and $\pi'_m < \pi_m$, then the transition matrix is

$$M_k = \begin{bmatrix} 1 & 0 & 0 \\ \frac{\pi_p - \pi'_p}{\pi_p} & \frac{\pi'_p}{\pi_p} & 0 \\ \frac{\pi_m - \pi'_m}{\pi_m} & 0 & \frac{\pi'_m}{\pi_m} \end{bmatrix}.$$

In the other case, one of the new expected frequencies is smaller and both others are larger. If, again w.l.o.g., $\pi'_u < \pi_u, \pi'_p > \pi_p$ and $\pi'_m > \pi_m$, the matrix of transition probabilities is

$$M_k = \begin{bmatrix} \frac{\pi'_u}{\pi_u} & \frac{\pi'_p - \pi_p}{\pi_u} & \frac{\pi'_m - \pi_m}{\pi_u} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Note that $(\pi_u, \pi_p, \pi_m) \cdot M_k = (\pi'_u, \pi'_p, \pi'_m)$ holds in both cases. For given IWEs at time points $t_1, \ldots, t_h$ between time points t_0 and t_{h+1}, the transition matrix between the states $\{u, p, m\}$ at time t_0 and the states at time t_{h+1} is

$$P_0 \cdot \prod_{k=1}^{h} M_k P_k. \tag{7}$$

Branch length in the iWE-SSE model

For R in equation 4 we assume an "invariant+gamma" model [10, 34]. That is, R is 0 with probability r, and with probability $1 - r$ the value of R comes from a discretized gamma distribution with 3 categories, expectation value 1 and a shape parameter α. The probability to be in each respective rate category, conditional on not being an invariant site, is 1/3.

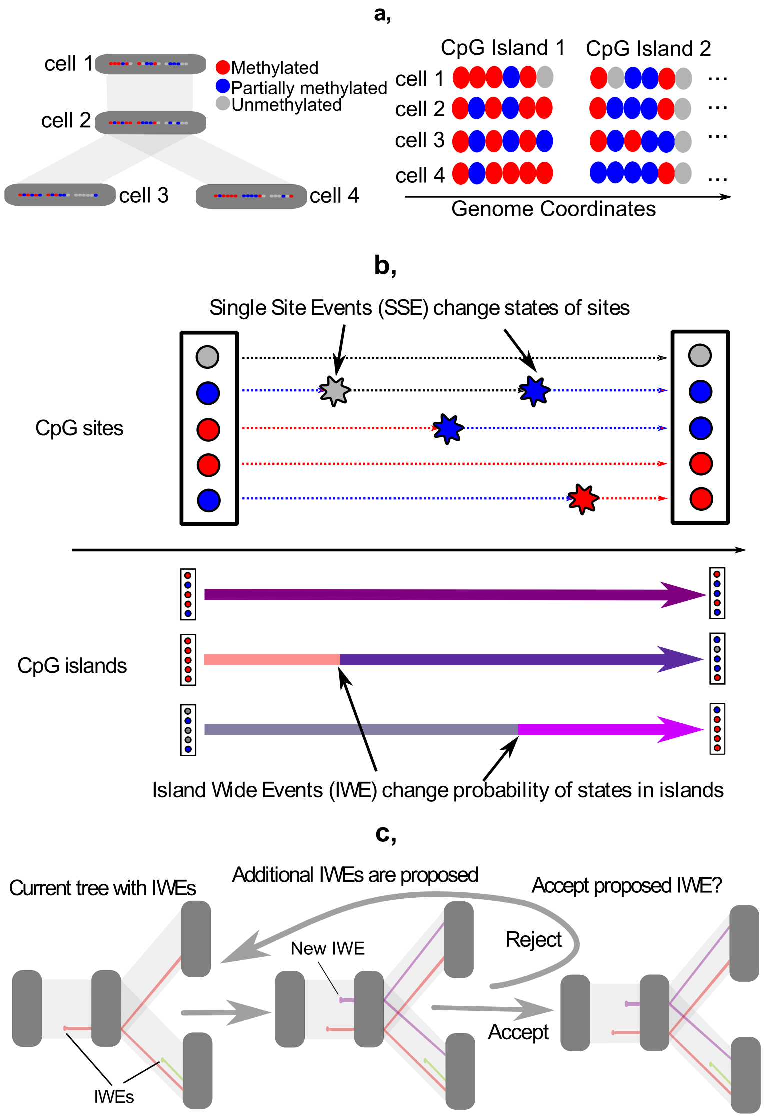

Fig. 1 a We consider cells to have evolved like species in a lineage tree. Instead of DNA we consider at methylation data, where we categorize CpG sites as unmethylated, partially methylated and methylated. Sites are contained within CpG islands, which we assume to evolve independently. **b** The IWE-SSE model is two stage model that considers both events that concern individual sites (SSEs), as well as events that change the probability of methylation states (IWEs) in whole islands during evolution. **c** Location of IWEs and their corresponding probabilities are inferred by inserting proposed IWEs in a tree, and then accepting the proposal with a Metropolis Hastings acceptance probability

Note that an expected fraction of

$$\pi_u^2 + \pi_p^2 + \pi_m^2 \geq 1/3 \tag{8}$$

of the SSEs will not change the current state of the CpG, due to the probability of an SSE being in states u, p and m respectively being $\pi_u, \pi_p$and π_m, and the probability of a switch to this state happening being again $\pi_u, \pi_p$and π_m. (The lower bound of 1/3 follows considering that $\pi_u + \pi_p + \pi_m = 1$ constitutes a plane in 3D space whose closest distance to the origin is $3^{-1/2}$ and $\pi_u^2 + \pi_p^2 + \pi_m^2$ is the square of the euclidean distance between (π_u, π_p, π_m) and the origin.)

We assume $\mathbb{E}R = 1$, which implies that our time unit is the expected number of SSEs per CpG (not conditioned on R but averaged over the possible values of R). In the following, branch lengths $B := (l_1, l_2, \ldots, l_k)$ will refer to this time scaling.

We assume that IWEs occur independently at each CpG island at rate μ and change the parameters values π_u, π_p and π_m on the CpG island. For a branch of length l we obtain an expected number of l SSEs per site and of $\mu \cdot l$ IWEs per CpG island. This implies that the branch length l can also be expressed as

$$l = \frac{E[S+W]}{n \cdot \mu + \sum_{i=1}^{n} n_i}, \tag{9}$$

where n is the number CpG islands, n_i is the number of CpG sites on CpG island i with $1 \leq i \leq n$ and the random variables S and W are the numbers of SSEs and IWEs on a branch of length l. Note here that S counts all SSEs, including those that do not change the methylation state of the site.

Likelihood calculations

We summarize the global model parameters as $\theta := (r, \alpha, \mu)$. As we assume that CpG islands evolve independently of each other, we obtain $\Pr_{\theta,B}(D) = \prod_i \Pr_{\theta,B}(D_i)$, where D_i is the data from CpG island i and B is the vector of branch lengths of the tree. For CpG island i let W_i be the configuration of IWEs and the mutation model parameters π_u, π_p, π_m around them. That is, W_i can be written as

$$W_i = (t_1, \pi_{u1}, \pi_{p1}, \pi_{m1}, \ldots, t_{n_i}, \pi_{un_i}, \pi_{pn_i}, \pi_{mn_i}), \qquad (10)$$

where the t_k parameter refers to the position of an IWE and $\pi_{uk}, \pi_{pk}, \pi_{mk}$ refer to its associated equilibrium frequencies. Here n_i constitutes the total number of IWEs that happen in the tree in island i. Conditioned on the configuration W_i, the CpGs within the island become independent and we obtain $\Pr_{\theta,B}(D_i|W_i) = \prod_j \Pr_{\theta,B}(D_{ij}|W_i)$, where D_{ij} is the data from the j-th CpG in CpG island i.

$$\Pr_{\theta,B}(D_{ij}|W_i) = \sum_x \Pr_{\theta,B}(D_{ij}|W_i, R_{ij} = x) \cdot \Pr_{\theta,B}(R_{ij} = x) \qquad (11)$$

where x is iterated over the four possible values of the rate factor R_{ij} for the CpG position. $\Pr_{\theta,B}(R_{ij} = x)$ is r for $x = 0$ and $(1-r)/3$ for the three values of x that are possible according to the discretized gamma distribution. To calculate $\Pr_{\theta,B}(D_{ij}|W_i, R_{ij} = x)$ we used a recursive scheme derived from Felsenstein's pruning algorithm [8, 10]. For this, let $D_{ij}^{(b)}$ be the part of D_{ij} that stems from the descendants of branch b. Let there be h IWEs on branch b affecting island i, with time intervals indexed by k between 1 to $h+1$. For an island indexed i and a CpG site indexed j, any branch b, state $y \in \{u, p, m\}$ and $k \geq 1$ we now define the partial likelihood $\omega_{k,b}(y)$

$$\omega_{k,b}(y) := \Pr_{\theta,B}(D_{ij}^{(b)}|W_i, R_{ij}, y), \qquad (12)$$

where $D_{ij}^{(b)}$ is the partial data, and y denotes the state that CpG site j is in just *before* IWE k (or the child node of b if $k = h+1$, where h is the number of IWEs on b affecting island i). For $k \geq 0$ let $\vec{\omega}_k$ be the column vector $(\omega_{k,b}(u), \omega_{k,b}(p), \omega_{k,b}(m))^T$. Let $\vec{\omega}_0$ be defined accordingly, but given that the state of CpG i in the parent node of the branch b is y. With the transition probability matrices P_k and M_k as defined in "Process of single-site events (SSEs)" and "Process of island-wide events (IWEs)" sections we obtain

$$\vec{\omega}_0 = P_0 \cdot \left(\prod_{j=1}^{k-1} M_j \cdot P_j \right) \cdot \vec{\omega}_k \qquad (13)$$

for any $k \in \{1, \ldots, h+1\}$. The case $k = h+1$ is sufficient for likelihood calculations, but the formula is also used for other values of k for updating likelihoods when M_{k-1} and P_{k-1} are changed in an MCMC step, see online appendix section B.1.

If the child node of b is a tip (an external node that is not the root) of the genealogy, we obtain $\omega_{h+1,b}(y) = 1$ if $y \in \{u, p, m\}$ is the state of the jth CpG site at the child node, and otherwise $\omega_{h+1,b}(y) = 0$. If b ends in a node with two daughter branches b' and b'', we obtain

$$\omega_{h+1,b}(y) = \omega_{0,b'}(y) \cdot \omega_{0,b''}(y) \qquad (14)$$

for all $y \in \{u, p, m\}$. In our application examples below, all methylation states are known not only for the tips of the tree, but also for the internal nodes and the root. In this case equation (14) holds only if y is the state of the jth CpG site at b's child node, and otherwise $\omega_{h+1,b}(y) = 0$. For the branch r that starts in the root we apply $\Pr_{\theta,B}(D_{ij}|W_i, R_{ij} = x) = \pi_{z,r} \cdot \omega_{0,r}(z)$, where z is the state of the CpG in the root node and $\pi_{z,r}$ is its probability according to the equilibrium distribution in the root.

MCMC implementation

To approximate $\Pr_{\theta,B}(D_i)$ we have to average the conditional probabilities $\Pr_{\theta,B}(D_i|W_i)$ over possible configurations of W_i. For this we apply a Metropolis-Hastings MCMC method [14, 30]. Given the current configuration of W_i in the MCMC procedure, the proposed W_i' for the next step can either lack one of the IWEs in W_i or have an additional IWE on some branch (see Fig. 1). Let l be the length of a branch b. As the IWE locations according to W_i are a priori a Poisson point process with intensity μ, the prior probability that W_i includes n IWEs on branch b is $\mathrm{Pois}_{\mu \cdot l}(n) = (\mu l)^n e^{-\mu l}/n!$. When the proposed W_i' differs from the current W_i by an additional IWE on branch b and n is the current number of IWEs on this branch, the Metropolis-Hastings acceptance probability is the minimum of 1 and

$$\frac{\Pr_{\theta,B}(D_i|W_i') \cdot \mathrm{Pois}_{\mu l}(n+1)}{\Pr_{\theta,B}(D_i|W_i) \cdot \mathrm{Pois}_{\mu l}(n)} = \frac{\Pr_{\theta,B}(D_i|W_i')}{\Pr_{\theta,B}(D_i|W_i)} \cdot \frac{\mu \cdot l}{n+1} \qquad (15)$$

(see online appendix B.1). If, conversely, W_i' with $n+1$ IWEs on b is the current state, and W_i with one IWE less on b is proposed, the acceptance probability is the minimum of 1 and the inverse of any side of Eq. 15.

For the branch lengths we apply Metropolis-Hastings acceptance steps on the log scale. If $\ell = \log(l)$ is the (natural) logarithm of the current length of a branch, the proposed $\ell' = \log(l')$ is drawn from a Gaussian mixture proposal distribution with density $g_\ell(\ell')$ that is centered around ℓ. The proposal distribution is symmetric, that is $g_\ell(\ell') = g_{\ell'}(\ell)$. The prior distribution of ℓ is a normal distribution. Let $p(\ell)$ denote its density (see online appendix A). When a new log length ℓ' is proposed for a branch of log length ℓ, we obtain the acceptance probability

$$\alpha(\ell', \ell) = \min\left\{1, \frac{L_D(l')}{L_D(l)} \frac{l'^n}{l^n} \cdot e^{-\mu N(l'-l)} \cdot \frac{p(\ell')}{p(\ell)}\right\}, \tag{16}$$

where n is the current number of IWEs on the branch, μ is the rate of IWEs, N is the number of CpG islands, $L_D(l')$ and $L_D(l)$ are the conditional probabilities of the data given the proposed and the current trees.

In further Metropolis-Hastings steps, the methylation state frequencies (π_u, π_p, π_m) for any CpG island can be updated. Further information about priors and proposal densities can be found in online appendix A.

Null model without iWEs

We test our model against a null model without IWEs. In the null model we still assume that each island has distinct equilibrium frequencies, which are sampled at the root from a Dirichlet(1,1,1) distribution and do not change during sequence evolution. When new branch lengths are sampled, the acceptance probability (16) in this model simplifies to

$$\min\left\{1, \frac{L_D(l')}{L_D(l)} \cdot \frac{p(l')}{p(l)}\right\}. \tag{17}$$

The parameters of the null model are the logarithms of branch lengths, the logarithm of the shape parameter of the gamma distribution of site specific rate factors, the fraction of invariant sites, and for each CpG island the equilibrium probabilities at the root states.

Application to the murine haematopoietic system

We tested our approach with methylation data that were gained by [2] with RRBS from murine cells at various stages of haematopoiesis (Fig. 2). The data overlap most murine CpG islands and consist of reads that are 36 base pairs long. To associate information of reads with CpG islands we used the mmp9 mapping of CpG islands from the USCL genome browser [18]. We sampled 2000 CpG islands at random, 1970 of which contained reads overlapping CpGs within the island. CpGs were categorized as unmethylated (u), partially methylated (p) or methylated (m) if less than 0.1, between 0.1 and 0.8, or more than 0.8 of the reads were detected as methylated. For Fig. 3 we categorized whole CpG islands as unmethylated if more than 50% of its CpG sites were in state u, or as methylated if more than 50% of its CpG sites were m. All other CpG islands were classified as partially methylated.

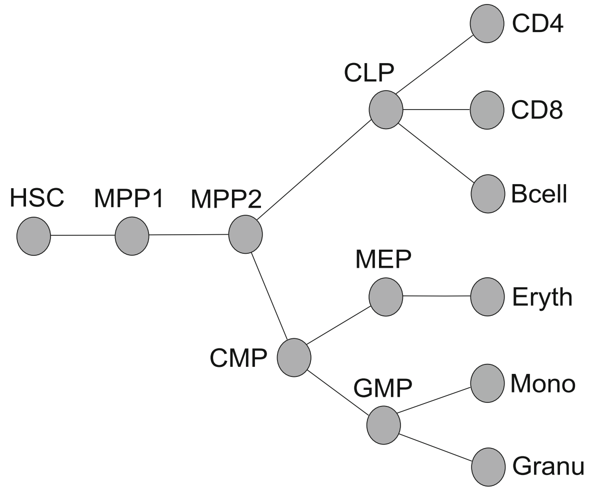

Fig. 2 Genealogy of haematopoietic cell stages [2, 4]

Application to in vitro cell differentiation data

In [32] human epiblast-like cells (hEpiLCs) were differentiated to primordial germ cells (hPGCs) over a twelve day period. Data about methylation levels on days 1, 2, 3 and 4 were present for hEpiLCs, as well as data for hPGCs on days 4, 5, 8 and 12. Reads were assigned to islands using mmp9 mapping and the hg37 mapping from the USCL genome browser. We randomly sampled 200 CpG islands and categorized CpG sites as explained in "Application to the murine haematopoietic system" section. We then

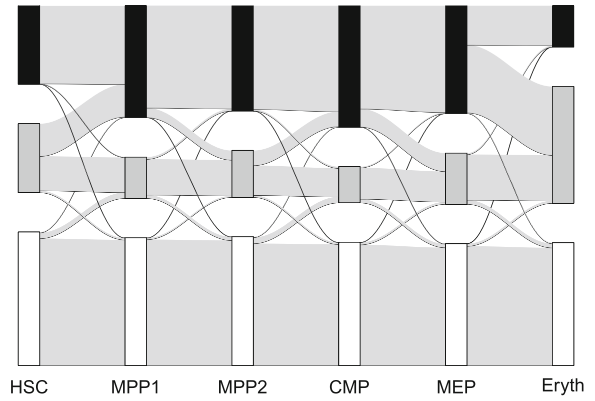

Fig. 3 Change of Island Methylation States. States of islands are categorized as methylated (black, top) if more than 50% of sites are in state *m* and as unmenthylated (white, bottom) if more than 50% of sites are in state *u*. Otherwise, islands are categorized as partially methylated (grey, middle). Vertical rectangles are proportional in size to the respective number of islands in each state for each cell type. Light grey transitions have a width proportional to the relative amount islands that transition between the states indicated by the rectangles

inferred parameters based on the SSE model, based on the IWE-SSE model and with lyne as well. To evaluate how well the branch lengths inferred with each of the models reflected the temporal distance between the samples, we fitted a linear regression model without intercept between estimated branch lengths and sampling time differences (using the lm command in R [25]) and computed the fraction of explained variance (adjusted R^2) as indicator of model fit. We adjusted ground truth time differences for the fact that in vivo the differentiation of hPGCs takes up to 10 days instead of 4 as it did in the induced transition cells in the in vitro environment [32].

Application to single cell methylation data

To evaluate whether the IWE-SSE inference program behaves notably different on single cell data as opposed to bulk data, we used a dataset procured by [35].

Therein human zygotes and cells at the two cell stage where sequenced and treated with bisulfite to produce single cell data about their methylation states. Since the cell samples were not taken from the same individual, we decided to pool the zygotes into an ancestral mean population and to then infer branch lengths between this mean population and the single cells in the 2 cell stadium. To pool the zygote data we calculated the mean methylation of each site across samples and then categorized this mean into unmethylated, partially methylated or methylated. For the cells in the 2 cell stadium we categorized sites as either methylated or unmethylated depending on their methylation state. We used 100 randomly chosen CpG islands.

Additional methods to estimate branch length

To evaluate the performance of our inference method, we compared it to two alternative approaches to estimate branch lengths. The first alternative was to use the Hamming distance which is often applied in hierarchical clustering. The second approach we took was to compute the mean methylation of the islands in each sequence. We then computed the euclidean distance between the vectors of mean values.

Simulation study

To assess the accuracy of our MCMC implementation, we simulated 150 data sets, each consisting of 100 islands. The numbers of CpG sites in the islands were chosen randomly from a uniform distribution between 10 and 400. At the start of each simulation we sampled the logarithm of branch lengths, the logarithm of the shape parameter α, the invariant probability, and the IWE rate μ from their priors (online appendix A). For the root node we sampled equilibrium frequencies from a Dirichlet(1,1,1) distribution. Then we sampled IWEs uniformly positioned along branches, where the number of sampled IWEs on a branch was Poisson distributed with mean μNl, where N is the number of CpG islands and l is the branch length. The equilibrium frequencies associated with an IWE were sampled from a Dirichlet(1,1,1) distribution. We generated the sequence at the root node by drawing each state in each island from the equilibrium frequency at the root node in this island. Sequences in the other nodes were generated iteratively going from the root to the tips of the cell lineage tree using transition probabilities between states as detailed in Process of island-wide events (IWEs).

We then used our inference method on the generated sequences to find posterior distributions of the simulated data sets with known ground truths sampled from priors. Here the MCMC runs were started from the means of the priors for all parameters other than the number of IWEs, where we started without IWEs to avoid long convergence times in the case of many misplaced IWEs in the initial configuration. We used a burn-in of 10^5 Metropolis Hasting steps. In addition to our full model we also fitted a null model without IWEs to the data.

Test for cpG-island-wide events (IWEs)

To test the relevance of IWEs for the data of [2], we simulated 150 data sets according to this null model using 1970 islands with the same number of CpG sites as in the restricted data set we used for initial inference. These simulations were conducted with the same procedure as in the simulation study, with the starting parameters being sampled from the posterior distribution of the null model and the IWE rate being restricted to 0. We then fitted the full model with IWEs to these simulated sequences and estimated posterior number of IWEs inferred in the adapted model.

Comparison with the r software package lyne

We conducted further simulations to compare the results of our IWE-SSE inference to results obtained with the R package lyne, which is associated to [4] and is available from the Kostka lab website[1]. We simulated branch lengths for the simplest tree that is accepted as input by lyne. This directed tree consists of a root whose sole offspring has two offspring nodes (tips). The length of the branch adjacent to the root was drawn from a log-normal distribution with mean $\exp(-2)$ and $\sigma = 2$. Another log-normally distributed value was simulated and used the (equal) lengths of the branches adjacent to the tips, using the same distribution. The other model parameters were sampled from their priors (online appendix A). We performed 150 simulations according to the lyne model and 150 simulations following the IWE-SSE model. For the Capra-Kostka model simulations we assumed 10,000 CpG sites, which were for the purposes of IWE-SSE inference

[1] https://www.kostkalab.net/software.html

randomly sampled into islands of uniformly distributed random sizes between 10 and 200 CpG sites (but the island structure had no influence on methylation states). The IWE-SSE model simulations used 100 islands with sizes between 10 and 200 CpG sites. In silico data produced in this way was analyzed with both Kostka's `lyne` package and our implementation of the IWE-SSE inference method.

Results

Application to methylation data from haematopoietic cells

We applied our method to 1970 randomly chosen CpG islands from the methylation data of murine haematopoietic cells [2, 4]. Using first the null model without IWEs ($\mu = 0$), we obtained a very long branch between MEP and Eryth indicating many changes in methylation of single CpGs (Fig. 4). Note that branch lengths are proportional to the expected number of SSEs or, in other words, to the product of cell divisions and SSE rate per cell division. We generated data following this fitted null model in 150 simulations and estimated parameters according to the model with IWEs for the null model simulations. Estimated total numbers of IWEs never exceeded 30 in any of these inferences and the inferred percentage of islands carrying an IWE along an edge was a most 0.07%. When we analyzed the data set of [2] with the IWE model ($\mu \geq 0$), the minimum number of IWEs after the burn-in period of 10^6 Metropolis-Hasting steps was 3488, and we inferred high levels of enrichment of IWEs on all branches (Fig. 5).

With the null model we estimated branch lengths similar to estimated lengths in the literature on these branches, e.g. between MEP and Eryth 4.56 units by Capra and Kostka, compared to a distribution mean of 5.94 SSE units with our null model. Here, an SSE unit refers to the expected number of SSEs per CpG, whereas Capra an Kostka's unit refers to the expected number of methylation state changes per CpG. As at least a third of the SSEs do not change the state of a CpG, and $5.94 \cdot 2/3 = 3.96$, our estimation of the length of the MEP-Eryth branch is smaller than that of Capra and Kostka, but the values are not directly comparable because the model of Capra and Kostka is more general than our null model without IWEs.

When we allowed for IWE events, we found considerably less variation among the inferred branch lengths (Fig. 5). Regarding the number of IWEs, the formation of the first multipotent progenitor cells from haematopoietic stem cells and the formation of erythrocytes showed an increased frequency of such events, explaining the methylation changes between MEP and Eryth by simultaneous methylation changes in IWEs rather than by many independent single-site events.

Evidence that iWE rate vary among branches

In the tree that we inferred with the IWE model (Fig. 5), the estimated numbers of IWEs vary among the branches more than the branch lengths. Indeed, credibility intervals of the log-transformed numbers of IWEs per branch length unit (Fig. 6) suggest that the IWE rate is substantially increased during the transitions from HSC to MPP1, from MPP1 to MPP2 and from MEP to Eryth. This is indicative of pronounced regularly activity along these transitions in particular (see also Fig. 3).

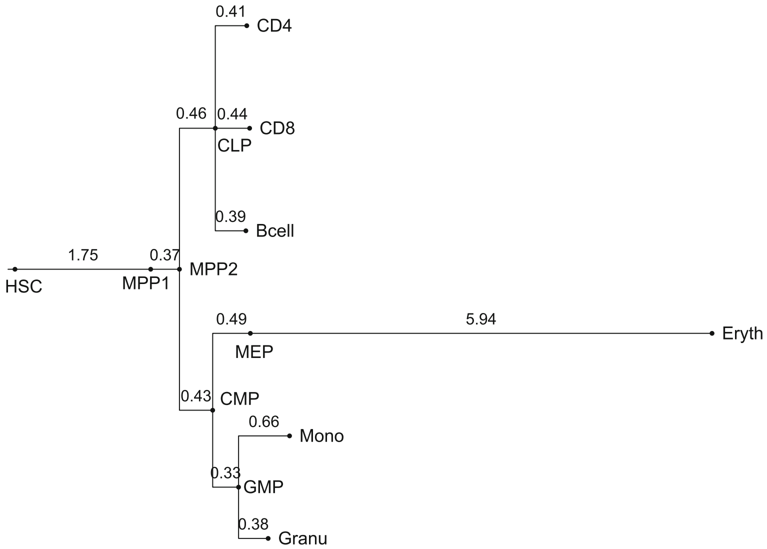

Fig. 4 Estimates without IWEs. Tree resulting from estimates without modeling IWEs. The logarithmic branch lengths are the means of the MCMC samples after a burn in phase of 10^6 steps

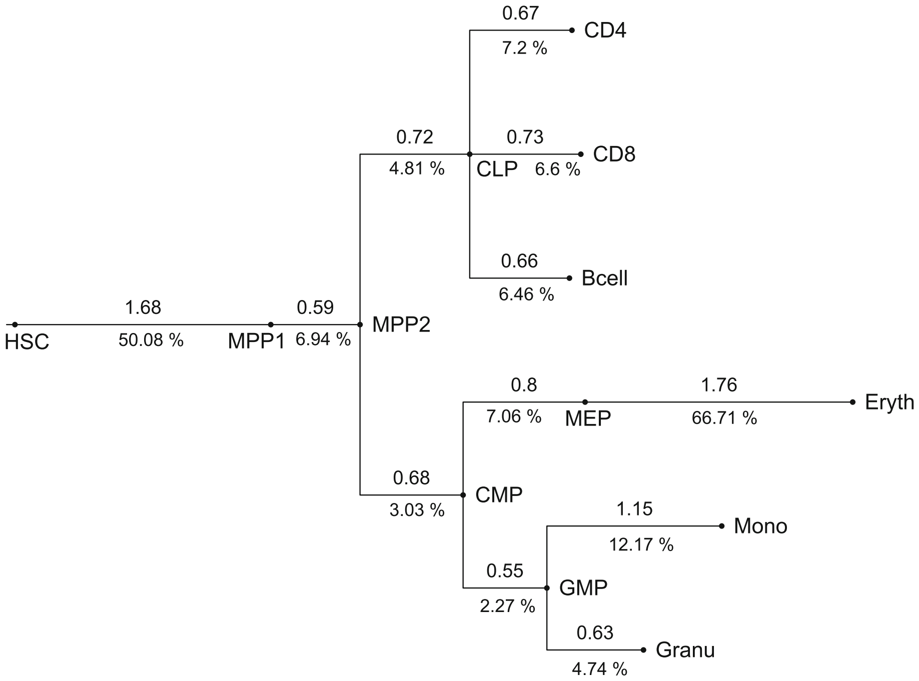

Fig. 5 Estimates IWE-SSE model. Tree resulting from estimates modeling IWEs. The logarithmic branch lengths (above branches) and percentages of CpG islands affected by IWEs (below branches) are the means of the MCMC samples after a burn in phase of 10^6 steps

Application to in vitro cell differentiation data

We applied our method to 200 randomly chosen CpG islands from the methylation data of human cell cultures [32]. Here we found that branch lengths in the EpiLCs were overestimated by all three inference approaches, whether we were using adjusted or experimental time spans for this stage (Fig. 7). The R^2 values (that is, fractions of explained variance) we found with branch lengths were 0.40, 0.24 and 0.18 for the IWE-SSE, SSE only and `lyne` inference respectively. Fits improved, with the IWE-SSE model again giving the best results, after adjusting for the induced differentiation taking less time than the

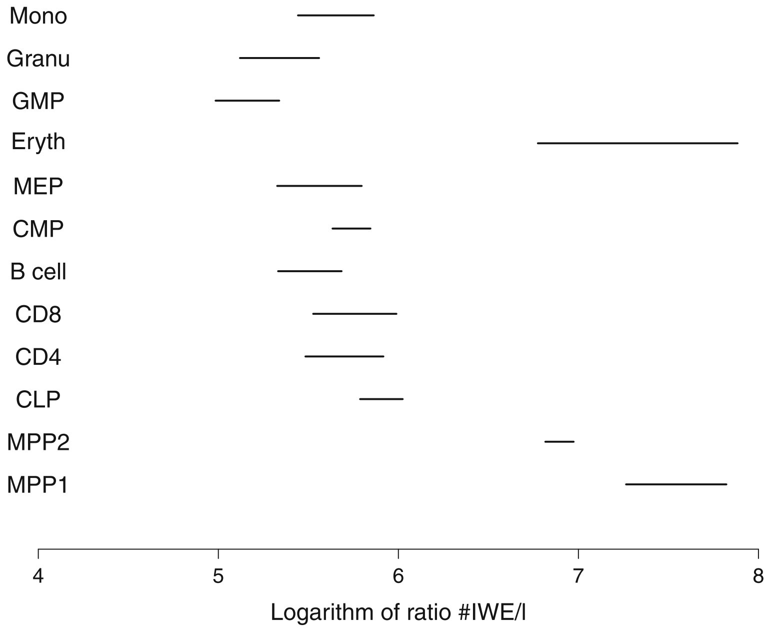

Fig. 6 Multiple testing corrected 95% intervals of the ratio of estimated number of events to the estimated branch length

in vivo differentiation by multiplying the EpiLC branch lengths by 2.5 (since differentiation takes 10 days under natural conditions instead of 4 days). The R^2 values we found with adjusted branch lengths were 0.80, 0.67 and 0.57 for the IWE-SSE, SSE only and lyne inference respectively. Nonetheless there was still considerable deviation from linearity, with the lengths in the EpiLC stages being overestimated for all inferences. We further applied the nonparametric Hamming distance and the euclidean mean difference for this analysis as described in "Additional methods to estimate branch length" section (see online appendix Fig. 7). Using the same linear comparison approach to a linear model without intercept, the adjusted R^2 were 0.55 and 0.68 for Hamming distance and euclidean mean distance respectively.

Application to single cell embryonic data

we applied our method to single cell data procured by [35]. We found that the estimated branch lengths were all very similar in length, as expected from their phylogenetic position (online appendix Fig. 8).

Simulation experiments

To validate the accuracy of our inference method, we simulated 150 data sets with parameters values drawn from the prior distributions (see MCMC implementation and online appendix A). Each of the simulated data sets contained 100 CpG islands with sizes varying uniformly between 10 and 400. For each of the simulated data sets we inferred the posterior distribution of the parameters. In Fig. 8 we compare the MCMC-sampled parameter values and credibility intervals to the actual parameter values underlying the simulations. To validate our implementation we computed the 95% credibility intervals and verified that the ground truth was within these intervals in approximately 95% of the cases. This was done for individual branch lengths, all branch lengths, the rate of IWEs and the shape parameter of rate heterogeneity. Indeed, credibility intervals overlapped the true branch length in 93 to 98% of the cases. Overall, 95% of the credibility intervals contained the true value. True values were in the credibility intervals in 96% of the cases for IWE rates and in 94% for the shape parameter.

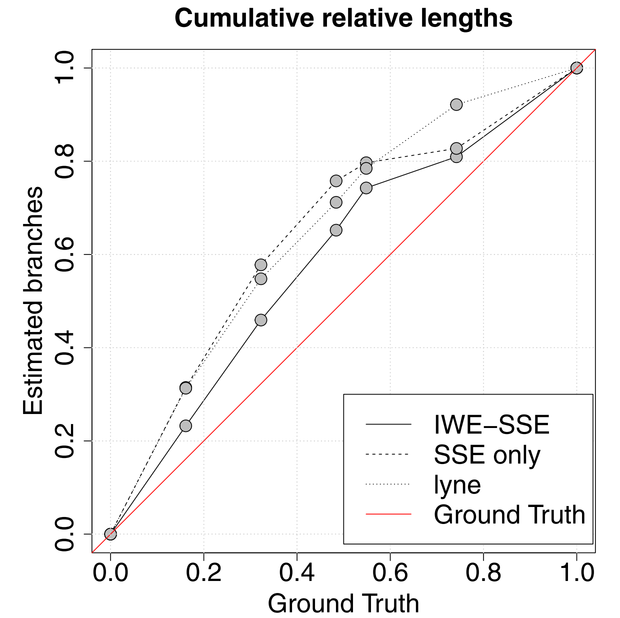

Fig. 7 Cumulative relative length of estimates compared to known ground truth during the transition of hEpiLCs on day 1 to hPGCs on day 12. Branches in the hEpiLC stage are underestimated by all three inference tools, even though they were adjusted to match in vivo conditions

Comparison with `lyne`

When data was simulated according to the model of [4], we observed a slight bias for overestimation in the IWE-SSE inference (online appendix Fig. 6). The lyne estimates showed no obvious bias but were less accurate for longer branches. When data was produced with the IWE-SSE model, lyne substantially over-estimated branch lengths, especially for the leaf-adjacent branches, while IWE-SSE inference was very accurate (Fig. 9).

Discussion

Our simulation results and the application example with methylation states of haematopoietic cells suggest that the possibility of CpG-island wide methylation changes should be taken into account when analysing methylation dynamics. For the single-site methylation changes (SSEs) we assume in our current model that the new methylation state (unmethylated, partially methylated, or methylated) is independent of the state before the SSE. A possible extension of our model would be to allow for the SSEs and for state-changes within IWEs the class of models proposed by [4], who consider all reversible 3 × 3 rate matrices for the three states. One could further consider Ising model based constraints on regions or islands [17], since the resulting likelihoods can also be calculated in linear time and the number of parameters per region is just three. The efficiency of our method, however, relies on the conditional independence of CpG sites given the IWE state of the CpG island because this allows us to carry out MCMC steps only for entire CpG islands and calculate marginal likelihoods for the CpG sites with an efficient dynamic programming approach. Allowing that SSEs depend additionally on the current states of neighboring CpG sites would require separate MCMC steps for all CpG sites. This would substantially slow down our MCMC method, especially if more than just the two directly neighbored CpGs have to be evaluated in each of these MCMC steps [21, 22, 24].

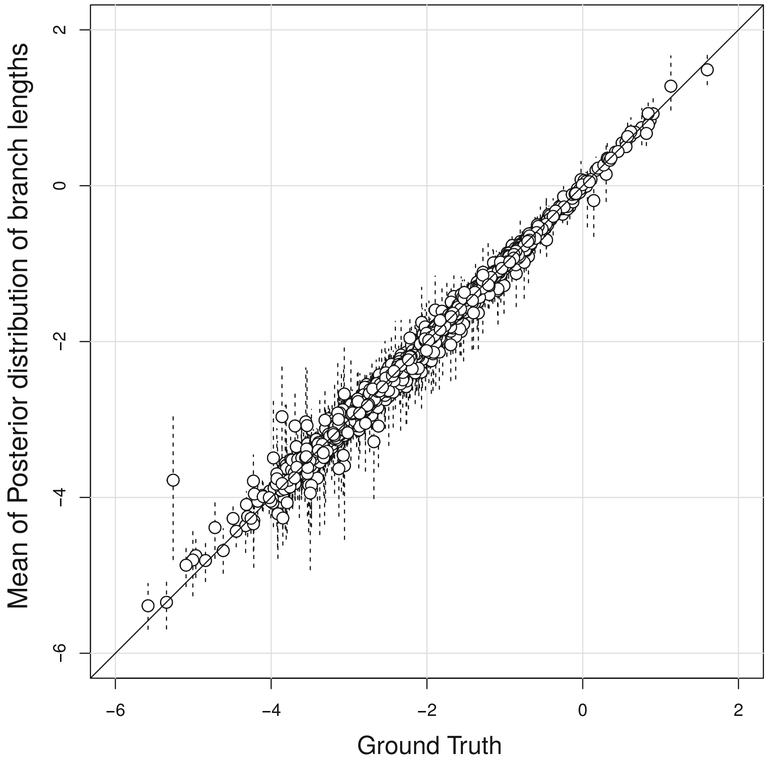

Fig. 8 Comparison between estimated logarithms of branch lengths and target values in simulation study. Dotted lines indicate 95% credibility intervals

Even though we assumed a priori a constant IWE rate in our model, we obtained clear evidence that the number of IWEs per branch length unit (which summarizes expected numbers of IWEs and SSEs) varies among the branches (Fig. 6). Further, our results suggest that overall methylation frequencies vary among the branches (Fig. 3). Also this is not explicitly taken into account in our model, as we assume that IWEs have their probabilities sampled from the same Dirichlet distribution across the tree. However, compound Poisson based models [15] of genome wide change are natural extensions to our framework. Thus, we could allow for genome-wide events that modify the IWE rate and the parameters of the Dirichlet distribution from which the methylation state distribution are sampled in IWEs. An alternative approach, in analogy to some relaxed molecular-clock models in phylogenetics [6], would be to assume that IWE rates or other parameters are sampled from a prior distribution independently for each branch.

Further comparisons of the performance of the IWE-SSE model on data procured in vitro experiments of germ cell differentiation suggested that inference with the IWE-SSE model is more accurate than inference without IWEs or inference with the `lyne` software package, as estimates were closer to a linear model of ground truth time spans. However all inferences produced a deviation from proportionality, by overestimating branch lengths in the early stages of the experiment when cells were in the EpiLC stage.

We went on to further estimate branch lengths in this datasets using the Hamming distance and an euclidean mean based distance. Both were outperformed by the IWE-SSE inference tool.

In the application example with the data of [2], the tree topology and the methylation states at the internal nodes were given. Our computational approach for the IWE-SSE model can also be adapted to reconstruct genealogies when methylation states are given only for the tips of the tree and combined with methods to explore possible tree topologies [10]. A potential application area could then be the inference of genealogies of cells sampled from neoplasms, e.g. to reconstruct the growth and mutation history of cancer clones [3, 31]. Accounting for IWEs may not only improve the accuracy of inferred cell genealogies but also allow for a better detection of aberrant methylations, which are a known hallmark of cancer [12]. The best possible data for reconstructing cell genealogies from methylation patterns would obviously be single-cell methylation data. To our knowledge, however, it is not yet possible to generate such data. Therefore the possibility of inferring single cell genealogies from long-read methylation data, which are now

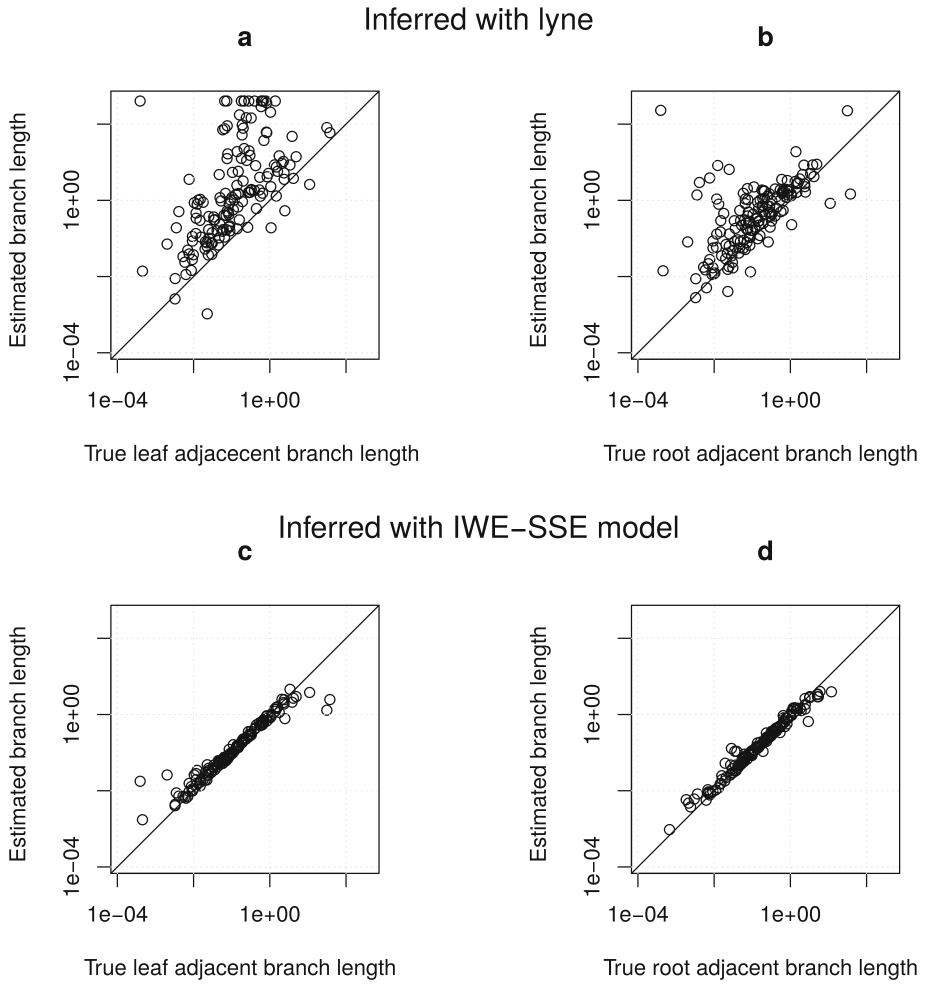

Fig. 9 Comparison between inferred branch lengths by `lyne` (**a** and **b**) and the IWE-SSE inference (**c** and **d**) when data was simulated according to the IWE-SSE model. Plots a and c refer to the branches that are adjacent to the leaves and plots b and d to the branches that are adjacent to the root

becoming available [27], constitutes a promising avenue of research.

Conclusion

We found that the model with CpG-island wide methylation rate changes (IWEs) fit the methylation data from murine haematopoietic cells significantly better than a model without IWEs. Furthermore, the IWE-SSE model detected certain developmental phases in haematopoiesis that many CpG islands were affected by IWEs, which may indicate enhanced activity in gene regulation. Our simulations comparing the inference methods based on the IWE-SSE model to inference method that assume independence between CpG sites show the necessity of modeling simultaneous changes of methylation states in CpG islands in addition to single-site changes. This view is further affirmed by comparison between ground truth times in cultured cells with estimated branch lengths. The IWE-SSE model produces a better proportional fit than do models not accounting for IWEs.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3438-5.

Additional file 1: Online appendix.

Abbreviations

Eryth: nucleated erythrocyte; hEpiLCs: human epiblast-like cells; hPGCs: human primordial germ cells; HSC: haematopoietic stem cells; IWE: island-wide events (changing methylation frequencies in a CpG island); MCMC: Markov Chain Monte Carlo; MEP: megakaryocyte-erythroid progenitor cells; MPP1/2: multipotent progenitors 1/2; SSE: single-site events (changing methylation at a single CpC site)

Acknowledgements

We thank Tobias Altmiks for testing an early version of our inference program.

Authors' contributions

Paper was written by KG. and DM. The inference program was written by KG. with contributions from DM. Data analyses and simulation studies were performed by KG. The project was designed and supervised by DM. All authors read and approved the final manuscript.

Funding

This project was funded by the German Science Foundation DFG through the Collaborative Research Consortium SFB 1243 (project A16). Funding body played no role in the design of the study or analysis or interpretation of data or in writing the manuscript.

Availability of data and materials
An implementation of the methods presented in this article is freely available as C++ source code on https://github.com/statgenlmu/IWEPoissonPaper under the terms of the GNU general public license (GPL v3). The murine hematopoietic data procured by [2] can be found under the accession number GSE38557. The human hematopoietic data procured by [32] can be found under the accession number GSE86586. The human single cell lines procured by [35] can be found under the accession number GSE81233.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 14 August 2019 Accepted: 2 March 2020
Published online: 18 March 2020

References
1. Arand J, Spieler D, Karius T, Branco MR, Meilinger D, Meissner A, Jenuwein T, Xu G, Leonhardt H, Wolf V, Walter. J. In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. PLoS Genet. 2012;8(6):e1002750.
2. Bock C, Beerman I, Lien W-H, D Smith ZD, Gu H, Boyle P, Gnirke A, Fuchs E, Rossi DJ, Meissner A. DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. Mol Cell. 2012;47(4):633–47.
3. Brocks D, Assenov Y, Minner S, Bogatyrova O, Simon R, Koop C, Oakes C, Zucknick m, Lipka db, Weischenfeldt J, et al. Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. Cell Rep. 2014;8(3):798–806.
4. Capra JA, Kostka D. Modeling DNA methylation dynamics with approaches from phylogenetics. Bioinformatics. 2014;30(17):i408–14.
5. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011;25(10):1010–22.
6. Drummond AJ, Ho SY, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006;4(5):e88.
7. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet. 2006;38(12):1378.
8. Felsenstein J. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Hum Genet. 1973;25(5):471.
9. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. J Mol Evol. 1981;17(6):368–76.
10. Felsenstein J. Inferring Phylogenies, volume 2. Sunderland: Sinauer associates; 2004.
11. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995;82(4):711–32. https://doi.org/10.1093/biomet/82.4.711.
12. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74.
13. Hastie DI, Green PJ. Model choice using reversible jump Markov chain Monte Carlo. Statistica Neerlandica. 2012;66(3):309–38. https://doi.org/10.1111/j.1467-9574.2012.00516.x.
14. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970;57(1):97–109.
15. Huelsenbeck JP, Larget B, Swofford D. A compound Poisson process for relaxing the molecular clock. Genetics. 2000;154(4):1879–92.
16. Jeltsch A, Jurkowska RZ. New concepts in DNA methylation. Trends Biochem Sci. 2014;39(7):310–8.
17. Jenkinson G, Abante J, Feinberg AP, Goutsias J. An information-theoretic approach to the modeling and analysis of whole-genome bisulfite sequencing data. BMC Bioinformatics. 2018;19(1):87.
18. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002;12(6):996–1006.
19. Kyriakopoulos C, Giehr P, Lück A, Walter J, Wolf V. A hybrid HMM approach for the dynamics of DNA methylation. In: Hybrid Systems Biology. Springer; 2019. p. 117–31. https://doi.org/10.1007/978-3-030-28042-0_8.
20. Liò P, Goldman N. Models of molecular evolution and phylogeny. Genome Res. 1998;8(12):1233–44. https://doi.org/10.1101/gr.8.12.1233.
21. Lück A, Giehr P, Nordström K, Walter J, Wolf V. Hidden markov modelling reveals neighborhood dependence of dnmt3a and 3b activity. IEEE/ACM Trans Comput Biol Bioinformatics. 2019;16(5):. https://doi.org/https://doi.org/10.1109/tcbb.2019.2910814.
22. Lück A, Giehr P, Walter J, Wolf V. A stochastic model for the formation of spatial methylation patterns. In: Feret J, Koeppl H, editors. Computational Methods in Systems Biology. Cham: Springer; 2017. p. 160–178.
23. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res. 2005;33(18): 5868–77.
24. Meyer KN, Lacey MR. Modeling Methylation Patterns with Long Read Sequencing Data. IEEE/ACM Trans Comput Biol Bioinform. 2018;15(4): 1379–89.
25. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2018. https://www.R-project.org/.
26. Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc Natl Acad Sci. 2006;103(5):1412–7.
27. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. 2017;14(4):407–10.
28. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet. 2013;14(3):204.
29. Sontag LB, Lorincz MC, Luebeck EG. Dynamics, stability and inheritance of somatic DNA methylation imprints. J Theor Biol. 2006;242(4):890–9. https://doi.org/10.1016/j.jtbi.2006.05.012.
30. Sorensen D, Gianola D. Likelihood, Bayesian, and MCMC methods in quantitative genetics. New York: Springer-Verlag; 2002.
31. Sottoriva A, Spiteri I, Shibata D, Curtis C, Tavaré S. Single-molecule genomic data delineate patient-specific tumor profiles and cancer stem cell organization. Cancer Res. 2013;73(1):41–9.
32. von Meyenn F, Berrens RV, Andrews S, Santos F, Collier AJ, Krueger F, Osorno R, Dean W, Rugg-Gunn PJ, Reik W. Comparative principles of dna methylation reprogramming during human and mouse in vitro primordial germ cell specification. Dev Cell. 2016;39(1):104–15.
33. Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, Whitaker JW, Tian S, Hawkins RD, Leung D, Yang H, Wang T, Lee AY, Swanson SA, Zhang J, Zhu Y, Kim A, Nery JR, Urich MA, Kuan S, Yen CA, Klugman S, Yu P, Suknuntha K, Propson NE, Chen H, Edsall LE, Wagner U, Li Y, Ye Z, Kulkarni A, Xuan Z, Chung WY, Chi NC, Antosiewicz-Bourget JE, Slukvin I, Stewart R, Zhang MQ, Wang W, Thomson JA, Ecker JR, Ren B. Epigenomic analysis of multilineage differentiation of human embryonic stem cells. Cell. 2013;153(5):1134–48.
34. Yang Z. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: Approximate methods. J Mol Evol. 1994;39(3):306–14. https://doi.org/10.1007/BF00160154.
35. Zhu P, Guo H, Ren Y, Hou Y, Dong J, Li R, Lian Y, Fan X, Hu B, Yun Gao Y, et al. Single-cell dna methylome sequencing of human preimplantation embryos. Nat Genet. 2018;50(1):12–9.

Publisher's Note