Modeling methylation dynamics with simultaneous changes in CpG islands

Konrad Grosser and Dirk Metzler*
Faculty of Biology, Division of Evolutionary Biology
Ludwig-Maximilians-Universität München
Großhaderner Str. 2, 81152 Planegg-Martinsried, Germany

May 14, 2019

Abstract

Motivation: Probabilistic models for methylation dynamics of CpG sites are usually based on sequence evolution models that assume indepedence between sites. In vertebrate genomes, CpG sites can be clustered in CpG islands, and the amount of methylation in a CpG island can change due to gene regulation processes. We propose a probabilistic 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 method to fit this model to methylation data from cell type phylogenies and apply this method to available data from murine haematopoietic cells.
Results: Branch lengths in cell phylogenies show the amount of changes in methylation in the development of one cell type from another. We 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.
Availability: 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 (GPLv3).

1 Introduction

DNA methylation is a common epigenetic process (Smith and Meissner, 2013). It is considered essential during phenotypic development in mammals and strongly associated with differential gene expression. The most frequent form of methylation is the attachment of the methyl group at the fifth carbon position on cytosine nucleotides that are followed by guanine nucleotide (Saxonov et al., 2006; Eckhardt et al., 2006; Smith and Meissner, 2013). Together, this configuration is called a CpG site.

Regions in which more than 50% of 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 (Saxonov et al., 2006; Smith and Meissner, 2013). These regions are typically between a few hundred and two thousand base pairs in length (Smith and Meissner, 2013). CpG islands are involved in the regulation of transcription (Deaton and Bird, 2011).

Comparisons of methylation states have been commonly applied and proved as a fruitful avenue of analysis of cell haematopoiesis (Bock et al., 2012; Xie et al., 2013). 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 (Bock et al., 2012) and enabled inference of missing methylation states. Capra and Kostka (2014) have adapted phylogenetic methods to account for the tree-shaped genealogy of cell types when analyzing methylation changes during haematopoiesis.

*for correspondence: metzler@bio.lmu.de

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. In an analogous manner, Capra and Kostka (2014) assume 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 (Deaton and Bird, 2011; Smith and Meissner, 2013).

Some sequence evolution models allow that mutation rates change at random time points, see e.g. Huelsenbeck et al. (2000). Here, we adapt this approach to CpG methylation-demethylation dynamics and allow that methylation frequencies in CpG islands can change during island-wide events (called IWEs throughout this text) at random time points. Furthermore, we allow that CpG sites that belong to the same island can simultaneously be methylated or demethylated in an IWE.

We implemented a reversible-jump MCMC inference scheme (Green, 1995; Sorensen and Gianola, 2002; Hastie and Green, 2012) to fit this model to RRBS methylation data (Meissner et al., 2005; Bock et al., 2012). We validate the accuracy of this scheme in a simulation study. With RRBS data procured from mouse haematopoiesis (Bock et al., 2012; Capra and Kostka, 2014) we demonstrate that accounting for IWEs can lead to significantly different estimations of branch lengths of cell type genealogies.

2 Methods

2.1 Methylation model

We assume that several CpGs can form a CpG islands, 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 simulaneously change their state at this time point. Different CpG islands, however, evolve independently of each other. Following Capra and Kostka (2014) 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 section 2.5) we consider a site 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. 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. 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 $P_{k;i,j} = P(X_{t_{k+1}} = j | X_{t_1} = i)$ 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 $P_k = \exp(Q_k \cdot (t_{k+1} - t_k))$. In analogy to the F81 sequence evolution model (Felsenstein, 1981) we focus here on rate matrices Q_k that can be expressed as

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

where $\pi_u + \pi_p + \pi_m = 1$, and each CpG has its own random rate factor $R \in \mathrm{R}_{\geq 0}$. This implies that (π_u, π_p, π_m) is a reversible equilibrium of Q_k and for fixed R the transition probabilities $(P_k)_{i,j}$ take the form $\left(1 - e^{R \cdot (t_k - t_{k+1})}\right) \cdot \pi_j$ for $j \neq i$ and $\pi_j + (1 - \pi_j) \cdot e^{R \cdot (t_k - t_{k+1})}$ for $j = i$. We assume that in the root of the genealogy each CpG island samples a distribution (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 will refer to these events as single-site events (SSEs) in the following. Note that an expected fraction of $\pi_u^2 + \pi_p^2 + \pi_m^2 \geq 1/3$ of the SSEs will not change the current state of the CpG. 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 if n is the number CpG islands and n_i 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, we obtain

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

Note here that S also counts all SSEs, including those that do not change the methylation state of the site.

In an 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. Without loss of generality, assume $\pi'_u > \pi_u$, $\pi'_p < \pi_p$ 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 $P_0 \cdot \prod_{k=1}^{n} M_k P_k$.

For R we assume an "invariant+gamma" model (Yang, 1994; Felsenstein, 2004). 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.

2.2 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. If we condition 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. $\Pr_{\theta,B}(D_{ij}|W_i)$ is a weighted average of $\Pr_{\theta,B}(D_{ij}|W_i, R_{ij} = x)$, where x is iterated over the four possible values of the rate factor R_{ij} for the CpG position. To calculate $\Pr_{\theta,B}(D_{ij}|W_i, R_{ij} = x)$ we used a recursive scheme derived from Felsenstein's pruning algorithm (Felsenstein, 1973, 2004). For this, let $D_{ij}^{(b)}$ be the part of D_{ij} that stems from the offspring of branch b. For focal b, i and j, and any $y \in \{u, p, m\}$ and $k \geq 1$ we now define the partial likelihood $\omega_{k,b}(y)$ to be the conditional probability of the partial data $D_{ij}^{(b)}$, given that CpG site j is in methylation state y just before IWE k (or the child node if $k = h+1$, where h is the number of IWEs on b affecting island i), and given the current states of θ, B, W_i and R_{ij}. For $k \geq 0$ be $\vec{\omega}_k$ 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 section 2.1 we obtain $\vec{\omega}_0 = P_0 \cdot \left(\prod_{j=1}^{k-1} M_j \cdot P_j\right) \cdot \vec{\omega}_k$ 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 2.1.

If the child node of b is a tip of the genealogy, we obtain $\omega_{h+1,b}(y) = 1$ if $y \in \{u, p, m\}$ is the state of the focal CpG site at the child node, and otherwise $\omega_{m+1,b}(y) = 0$. If b is a node with two daughter nodes b' and b'', we obtain

$$\omega_{h+1,b}(y) = \omega_{0,b'}(y) \cdot \omega_{0,b''}(y) \tag{2}$$

for all $y \in \{u, p, m\}$. In our application example below, all methylation states are known not only for the tips of the tree but also for the internal nodes. In this case equation (2) holds only if y is the state of the focal CpG site at b's child node, and otherwise $\omega_{h+1,b}(y) = 0$. For the branch r that starts in the root (HSC in our example below) 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 note and $\pi_{z,r}$ is its probability according to the equilibrium distribution in the root.

2.3 MCMC implementation

To approximate $\Pr_{\theta,B}(D_i)$ we have to average $\Pr_{\theta,B}(D_i|W_i)$ over possible configurations of W_i. For this we apply a Metropolis-Hastings MCMC method (Hastings, 1970; Sorensen and Gianola, 2002). 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. 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

$$\frac{\Pr_{\theta,B}(D_i|W_i') \cdot \mathrm{Pois}_{\mu l}(n+1)}{\Pr_{\theta,B}(D_i|W_i) \cdot \mathrm{Pois}_{\mu l}(n)} = \frac{\Pr_{\theta,B}(D_i|W_i')}{\Pr_{\theta,B}(D_i|W_i)} \cdot \frac{\mu \cdot l}{n+1} \tag{3}$$

(see online appendix 2.1). 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 equation 3.

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 1). When a new log length ℓ' is proposed for a branch of log length ℓ, we obtain the acceptance probability

$$\alpha(\ell', \ell) = \min\left\{1, \frac{L_D(l')}{L_D(l)} \frac{l'^n}{l^n} \cdot e^{-\mu N(l'-l)} \cdot \frac{p(\ell')}{p(\ell)}\right\}, \tag{4}$$

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 the online appendix 1.

2.4 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 (4) in this model simplifies to

$$\min\left\{1, \frac{L_D(l')}{L_D(l)} \cdot \frac{p(l')}{p(l)}\right\}. \tag{5}$$

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.

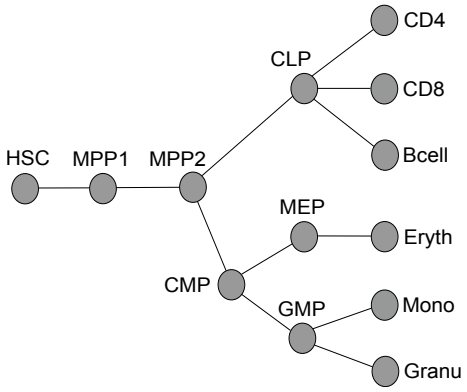

Figure 1: Genealogy of haematopoietic cell stages (Bock et al., 2012; Capra and Kostka, 2014).

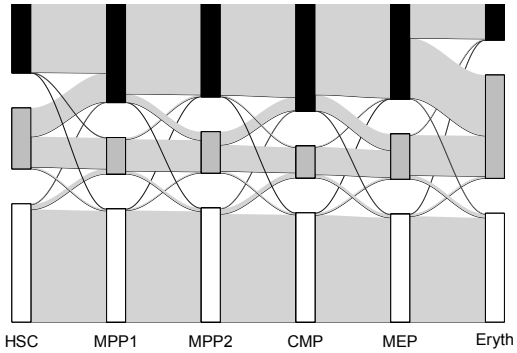

Figure 2: Change of Island Methylation States. States of islands are categorized as methylated (black, top) if more than 50% of sites are in state m and as unmenthylated (white, bottom) if more than 50% of sites are in state u. Otherwise, islands are categorized as partially methylated (grey, middle). 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.

2.5 Data

We tested our approach with methylation data that were gained by Bock et al. (2012) with reduced restricted bisulfite sequencing (RRBS) from murine cells at various stages of haematopoiesis (Figure 1). 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 UCL genome browser (Kent et al., 2002). 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 then 0.1, between 0.1 and 0.8, or more than 0.8 of the reads were detected as methylated. For Fig. 2 we categorized whole CpG islands as unmethylated if more than 50 % of its CpG sites were in state u, or as methylated if if more than 50 % of its CpG sites were m. All other CpG islands were classified as partially methylated.

2.6 Simulation study

We assess the accuracy of our MCMC implementation in a simulation study. We simulated 150 data sets, each consisting of 100 islands. The number of CpG sites in each islands were chosen randomly from

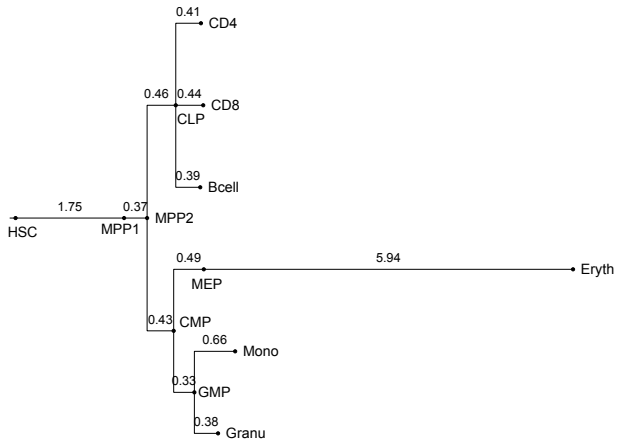

Figure 3: Tree resulting from estimates without modeling IWEs. The logarithmic branch lengths are the means of the MCMC samples after a burn in phase of 10^6 steps.

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 our priors. 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 $\mu N l$, 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. Given a vector of equilibrium frequencies and positions for IWEs occurring at an island we calculated transition probabilities between states as detailed in the model specifications.

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 where generated iteratively going from the root to the tips of the cell lineage tree.

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.

2.7 Test for CpG-island-wide events (IWEs)

In addition to our full model we also fitted a null model without IWEs to the data. To test the relevance of IWEs for the data of Bock et al. (2012), 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.

3 Results

3.1 Application to methylation data from haematopoietic cells

We applied our method to 1970 randomly chosen CpG islands from the methylation data of murine haematopoitic cells (Bock et al., 2012; Capra and Kostka, 2014). 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. 3). Note that branch lengths are proportional to the expected number of changes of events on the branch or, in other words, to the product of cell divisions and the rate of

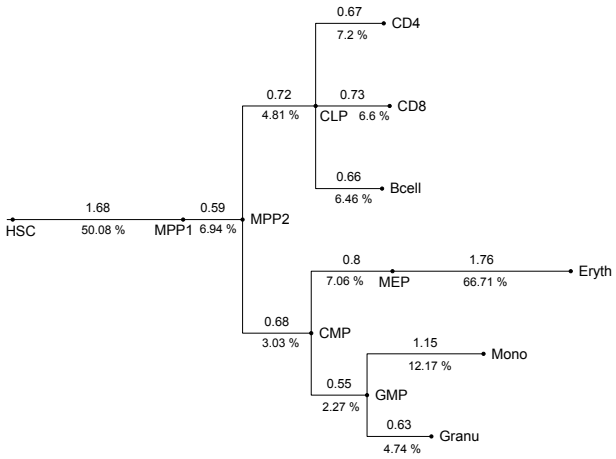

Figure 4: Tree resulting from estimates modeling IWEs. The logarithmic branch lengths (above branches) and percentages of CpG islands affected by IWEs (below branches) are the means of the MCMC samples after a burn in phase of 10^6 steps.

methylation and demethylation 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 analysed the data set of Bock et al. (2012) 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. 4).

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. 4). Regarding the number of IWEs, the formation of the first multi-pluripotent 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.

3.1.1 Evidence that IWE rate vary among branches

In the tree that we inferred with the IWE model (Fig. 4), 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. 5) 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. 2).

3.2 Simulation experiments

To validate the accuracy of our inference we conducted, we simulated 150 data sets with parameters values drawn from the prior distributions (see 2.3 and online appendix 1). Each of the simulated data sets contained 100 islands with sizes varying uniformly between 10 and 400. For each of the simulated data sets we inferred the posterior distribution of the parameters we used to produce the simulated

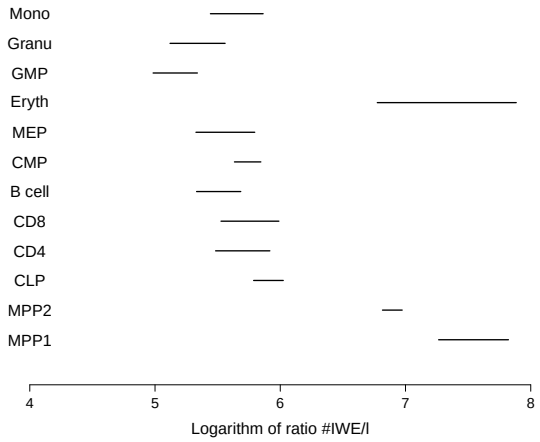

Figure 5: Multiple testing corrected 95% intervals of the ratio of estimated number of events to the estimated branch length.

data. In Fig. 6 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 the 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.

4 Discussion

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 model suggested for certain developmental phases in haematopoiesis that many CpG islands were affected by IWEs, which may indicate enhanced activity in gene regulation.

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 Capra and Kostka (2014), who consider all reversible 3×3 rate matrices for the three states.

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. 5). Further, Figure 2 suggests that overall methylation frequencies vary among the branches. 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 (Huelsenbeck et al., 2000) 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 (Drummond et al., 2006), would be to assume that IWE rates or other parameters are sampled from a prior distribution independently for each branch.

In the application example above with the data of Bock et al. (2012), the tree topology and the methylation states at the internal nodes were given. Our computational approach for the IWE-SSE

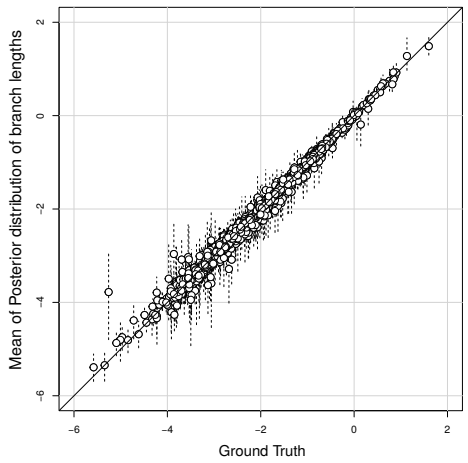

Figure 6: Comparison between estimated logarithms of branch lengths and target values in simulation study. Dotted lines indicate 95% credibility intervals.

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 (Felsenstein, 2004). 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 (Brocks et al., 2014; Sottoriva et al., 2013). 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 known to be among the hallmarks of cancer (Hanahan and Weinberg, 2011). 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 it seems worthwhile to explore possibilities of inferring single cell genealogies from long-read methylation data, which are now becoming available (Simpson et al., 2017).

Acknowledgments

We thank Tobias Altmiks for software testing an for spotting a bug in an early version of our program code.

Funding

This project was funded by the German Science Foundation DFG through the Collaborative Research Consortium SFB 1243.

References

Bock, C., Beerman, I., Lien, W.-H., Smith, Z. D., Gu, H., Boyle, P., Gnirke, A., Fuchs, E., Rossi, D. J., and Meissner, A. (2012). DNA methylation dynamics during in vivo differentiation of blood and skin stem cells. *Molecular cell*, 47(4):633–647.

Brocks, D., Assenov, Y., Minner, S., Bogatyrova, O., Simon, R., Koop, C., Oakes, C., Zucknick, M., Lipka, D. B., Weischenfeldt, J., et al. (2014). Intratumor DNA methylation heterogeneity reflects clonal evolution in aggressive prostate cancer. *Cell Reports*, 8(3):798–806.

Capra, J. A. and Kostka, D. (2014). Modeling DNA methylation dynamics with approaches from phylogenetics. *Bioinformatics*, 30(17):i408–i414.

Deaton, A. M. and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & Development*, 25(10):1010–1022.

Drummond, A. J., Ho, S. Y., Phillips, M. J., and Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol.*, 4(5):e88.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., et al. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, 38(12):1378.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, 25(5):471.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376.

Felsenstein, J. (2004). *Inferring Phylogenies*, volume 2. Sinauer associates Sunderland, MA.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *cell*, 144(5):646–674.

Hastie, D. I. and Green, P. J. (2012). Model choice using reversible jump Markov chain Monte Carlo. *Statistica Neerlandica*, 66(3):309–338.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Huelsenbeck, J. P., Larget, B., and Swofford, D. (2000). A compound Poisson process for relaxing the molecular clock. *Genetics*, 154(4):1879–1892.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res*, 12(6):996–1006.

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution dna methylation analysis. *Nucleic acids research*, 33(18):5868–5877.

Saxonov, S., Berg, P., and Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*, 103(5):1412–1417.

Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods*, 14(4):407–410.

Smith, Z. D. and Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, 14(3):204.

Sorensen, D. and Gianola, D. (2002). *Likelihood, Bayesian, and MCMC methods in quantitative genetics*. Springer-Verlag, New York.

Sottoriva, A., Spiteri, I., Shibata, D., Curtis, C., and Tavaré, S. (2013). Single-molecule genomic data delineate patient-specific tumor profiles and cancer stem cell organization. *Cancer research*, 73(1):41–49.

Xie, W., Schultz, M. D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J. W., Tian, S., Hawkins, R. D., Leung, D., Yang, H., Wang, T., Lee, A. Y., Swanson, S. A., Zhang, J., Zhu, Y., Kim, A., Nery, J. R., Urich, M. A., Kuan, S., Yen, C. A., Klugman, S., Yu, P., Suknuntha, K., Propson, N. E., Chen, H., Edsall, L. E., Wagner, U., Li, Y., Ye, Z., Kulkarni, A., Xuan, Z., Chung, W. Y., Chi, N. C., Antosiewicz-Bourget, J. E., Slukvin, I., Stewart, R., Zhang, M. Q., Wang, W., Thomson, J. A., Ecker, J. R., and Ren, B. (2013). Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, 153(5):1134–48.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39(3):306–314.