BIOINFORMATICS

Vol. 30 ECCB 2014, pages i408–i414
doi:10.1093/bioinformatics/btu445

Modeling DNA methylation dynamics with approaches from phylogenetics

John A. Capra[1,2,*] and Dennis Kostka[3,4,*]

[1]Center for Human Genetics Research, [2]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN 37232, [3]Departments of Developmental Biology and [4]Computational & Systems Biology, University of Pittsburgh, Pittsburgh, PA 15201, USA

ABSTRACT

Motivation: Methylation of CpG dinucleotides is a prevalent epigenetic modification that is required for proper development in vertebrates. Genome-wide DNA methylation assays have become increasingly common, and this has enabled characterization of DNA methylation in distinct stages across differentiating cellular lineages. Changes in CpG methylation are essential to cellular differentiation; however, current methods for modeling methylation dynamics do not account for the dependency structure between precursor and dependent cell types.

Results: We developed a continuous-time Markov chain approach, based on the observation that changes in methylation state over tissue differentiation can be modeled similarly to DNA nucleotide changes over evolutionary time. This model explicitly takes precursor to descendant relationships into account and enables inference of CpG methylation dynamics. To illustrate our method, we analyzed a high-resolution methylation map of the differentiation of mouse stem cells into several blood cell types. Our model can successfully infer unobserved CpG methylation states from observations at the same sites in related cell types (90% correct), and this approach more accurately reconstructs missing data than imputation based on neighboring CpGs (84% correct). Additionally, the single CpG resolution of our methylation dynamics estimates enabled us to show that DNA sequence context of CpG sites is informative about methylation dynamics across tissue differentiation. Finally, we identified genomic regions with clusters of highly dynamic CpGs and present a likely functional example. Our work establishes a framework for inference and modeling that is well suited to DNA methylation data, and our success suggests that other methods for analyzing DNA nucleotide substitutions will also translate to the modeling of epigenetic phenomena.

Availability and implementation: Source code is available at www.kostkalab.net/software.

Contact: tony.capra@vanderbilt.edu or kostka@pitt.edu

1 INTRODUCTION

DNA methylation is a common epigenetic modification essential to organism development (Smith and Meissner, 2013). In vertebrates, DNA is most commonly methylated at the fifth carbon position on cytosine nucleotides (5mC) that are followed by a guanine, so-called CpG sites. A family of three DNA methyltransferase enzymes (DNMT1, DNMT3A, DNMT3B) is responsible for the establishment and maintenance of methylation state at the millions of CpG sites in most mammalian genomes (Smith and Meissner, 2013). Recently, the ability to perform genome-wide assays of the methylation state of individual CpGs has become a reality because of advances in microarray and DNA sequencing technology. Several approaches that vary in their accuracy, biases, coverage and cost are commonly used; see Laird (2010) for a detailed review of current methods.

Systematic screening of DNA methylation across tissue differentiation and development has improved our knowledge of its role in these processes (Bock *et al.*, 2012; Xie *et al.*, 2013). The methylation profile of the mammalian genome is largely stable, but the methylation of specific genomic regions changes dynamically across development, and different cellular lineages have unique methylation profiles (Ziller *et al.*, 2013). Additionally, the DNA methylation state nearby a gene's transcription start site (TSS) correlates with gene expression (Xie *et al.*, 2013), and the correct orchestration of methylation changes is essential for proper cellular differentiation. Aberrant methylation changes may lead to tumorigenesis and other diseases (Bergman and Cedar, 2013; Hansen *et al.*, 2011; Portela and Esteller, 2010; Tost, 2010).

Studies assaying DNA methylation often focus on the comparison of two types of conditions, like tumor versus normal tissue (Nordlund *et al.*, 2013), or stem cells versus lineage-committed cells (Xie *et al.*, 2013). However, the natural process of cellular differentiation and development has an essentially tree-like topology, in which precursor cell types are connected to their descendants by edges, thereby forming a so-called *lineage tree* (Frumkin *et al.*, 2005). For example, Figure 1 depicts a lineage tree for blood cell differentiation, where DNA methylation has been assayed in cell types represented by nodes (Bock *et al.*, 2012). Independent pairwise comparisons cannot accommodate this structure.

To address this issue, we introduce an approach to model methylation state changes between cell types that explicitly takes dependencies induced by the lineage tree into account. In this setup, modeling methylation changes over developmental time is in many ways reminiscent of describing DNA nucleotide changes over evolutionary time (Fig. 1). As a result, we adapt established continuous-time Markov models of sequence evolution to fit this task.

In addition to accommodating cell lineage relations during development, our approach has the benefit that it works at single CpG dinucleotide resolution and does not require the spatial aggregation of methylation measurements across the genome. Finally, the analogy with models for DNA sequence evolution provides intuitive means to handle missing data, which are common in many DNA methylation datasets. During parameter

*To whom correspondence should be addressed.

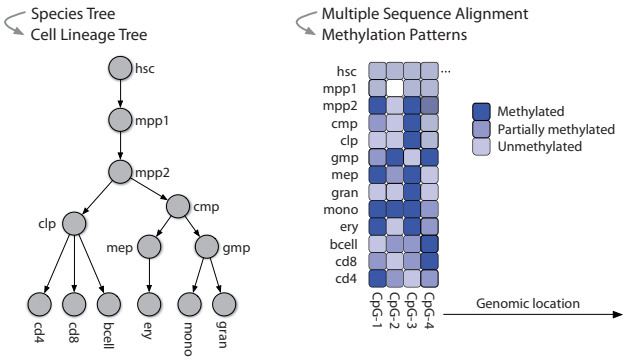

Fig. 1. CpG methylation dynamics can be modeled with an approach inspired by phylogenetic analysis of nucleotide substitutions. Left: A lineage tree of sampled blood cell types during hematopoietic differentiation with stem cells on top and terminally differentiated cells on the bottom (see Section 2.4 for details). The lineage tree takes the role played by the species tree in the phylogenetic context. Right: Examples of methylation patterns across differentiation (columns) for CpG sites at different genomic locations. Each row corresponds to a cell type in the linage tree on the left. The white block represents missing data. The discretized methylation states are analogous to DNA sequence data

estimation, missing data can be marginalized over, and the equivalent of joint ancestry reconstruction (Pupko *et al.*, 2000) allows efficient inference of the most likely methylation states for unobserved data in context of the lineage tree.

As an illustration of our approach, we analyzed methylation data collected across the cell lineage tree in Figure 1 from Bock *et al.* (2012). On this dataset, our method enabled the accurate reconstruction of missing methylation states. The single CpG resolution of our analysis allowed us to discover that the identity of neighboring dinucleotides is strongly correlated with CpG methylation dynamics at many sites in the mouse genome. Finally, using our predictions of CpG methylation variability, we identified a cluster of highly dynamic CpG sites that show evidence of enhancer activity in blood cells.

2 METHODS

2.1 Modeling methylation changes across tissue differentiation

We model the dynamics of DNA methylation across cellular differentiation using an approach motivated by phylogenetic models. In the phylogenetic context, a continuous time Markov chain is used to quantify DNA sequence changes between species over a known species tree. Intuitively, we adapt this approach and replace the species tree with a cell lineage tree and the four-state alphabet of DNA with an alphabet based on methylation status. In our model, the cell lineage tree consists of nodes that correspond to cell types and edges indicate precursor–descendant relationships. For example, the lineage tree shown in Figure 1 traces the differentiation of adult hematopoietic stem cells (HSCs) through several intermediate states into different terminally differentiated blood cell types. To describe methylation patterns, we define three discrete methylation states $\{\mathtt{u}, \mathtt{p}, \mathtt{m}\}$, corresponding to unmethylated, partially methylated and methylated CpGs dinucleotides, respectively.

To model transitions between CpG methylation states along edges of the cell lineage tree, we associate each node with a discrete random variable X_i ($1 \le i \le N$, assuming N nodes), which is dependent on its parents (i.e. its direct precursor cell types) in the lineage tree. As in DNA sequence evolution models, we use a continuous time Markov chain to describe this dependency structure. Specifically, if nodes i and j are connected in the lineage tree by an edge $i \to j$ of length t, then the probability of the methylation state at node j being l conditional on node i being in methylation state k is given by

$$P(X_j = l | X_i = k) = [\mathrm{expm}(Qt)]_{kl}$$

for $k, l \in \{\mathtt{u}, \mathtt{p}, \mathtt{m}\}$ (Guttorp and Minin, 1995). Q is a 3×3 rate matrix (or generator), and expm denotes the matrix exponential. We assume a time-reversible Markov chain with equilibrium frequency π, which implies Q is fully parameterized by three non-negative rate parameters, $\{a_i\}$, and π. (The number of expected transitions along an edge is $(-1)\sum_i \pi_i Q_{ii} t$, and therefore, we will enforce $(-1)\sum_i \pi_i Q_{ii} = 1$ and report t in units of expected methylation state transitions.) In summary, our model is parameterized by ϑ, which consists of the topology of the lineage tree (which we assume is fixed, known and consists of N nodes and E edges), the branch lengths $\{t_i\}_{i=1}^{E}$, the equilibrium frequencies $\{\pi_i\}_{i=1}^{3}$ and the rate parameters $\{a_i\}_{i=1}^{3}$. The likelihood of an observed methylation pattern $\mathbf{x} = \{x_i\}_{i=1}^{N}$ is then

$$P(\mathbf{x}|\vartheta) = \prod_{i=1}^{N} P_\vartheta(X_i = x_i | X_{\mathrm{pa}(i)} = x_{\mathrm{pa}(i)})$$

where $\mathrm{pa}(i)$ is the parent of node i in the lineage tree. For the root node, we have $P(X = i) = \pi_i$, and assuming independence between methylation patterns at different CpG sites, we have for the likelihood of all observed patterns $D = \{\mathbf{x}_i\}_{i=1}^{L}$ (assuming there are L CpG sites):

$$L(\vartheta) = P(D|\vartheta) = \prod_{i}^{L} P(\mathbf{x}_i|\vartheta). \quad (1)$$

In contrast to most applications dealing with DNA sequence changes, non-leaf nodes can be observed in our setting. We handle missing data by marginalization, i.e. summation over all possible configurations of unobserved nodes in the lineage tree, which can be done efficiently (linear in the number of tree nodes) via the elimination algorithm (Siepel and Haussler, 2005). Maximum likelihood parameter estimates are then obtained by maximizing Equation (1) over branch lengths, equilibrium frequencies and rate parameters. In summary, we have adapted a well-known class of models that is typically used in the context of DNA sequence evolution to model the dynamics of methylation changes during tissue differentiation.

2.2 Integrating rate heterogeneity

2.2.1 Modeling rate heterogeneity The approach described so far models methylation dynamics using the same process at all CpG sites in the genome, and thereby assumes homogeneity of methylation dynamics. This assumption is not always reasonable. For instance, CpGs located in CpG islands have a propensity to be unmethylated (compared with other CpG sites), and a disposition to stay in that state (Jones, 2012).

To address this issue, we incorporate rate heterogeneity into our model using a mixture modeling approach similar to phylogenetic models for DNA changes under heterogeneous substitution rates. First, we assume a certain fraction (β) of CpG sites to be *invariant*, i.e. they do not change their methylation state during tissue differentiation. For the $(1-\beta)$ fraction of *variable* CpG sites, we assume M different equiprobable rate categories $\{r_m\}_{m=1}^{M}$ such that the probability of methylation pattern $\mathbf{x}_i$ is now

$$P(\mathbf{x}_i|\tilde{\vartheta}) = \beta P(\mathbf{x}_i|r_0, \vartheta) + (1-\beta)\frac{1}{M}\sum_{m=1}^{M} P(\mathbf{x}_i|r_m, \vartheta),$$

where we have used the 'rate category' r_0 to denote invariance and $\tilde{\vartheta}$ to denote the new parameter set. For the invariant term on right side above, $P(\mathbf{x}_i|r_0, \vartheta) = p_k$ if all methylation states in $\mathbf{x}_i$ are k (for $k \in \{\mathrm{u}, \mathrm{p}, \mathrm{m}\}$) and zero otherwise. For the variable part, we have $P(\mathbf{x}_i|r_m, \vartheta) = P(\mathbf{x}_i|\overline{\vartheta}(m))$, where we use Equation (1), but with all branch lengths in ϑ scaled by the factor r_m. The scale factors $\{r_m\}$ are determined by a Gamma distribution with shape parameter α and scale parameter $1/\alpha$ (setting the scale parameter to $1/\alpha$ ensures that the Gamma distribution has a mean of one). Next, the probability density function of the Gamma distribution is discretized by splitting its domain into M equal-mass bins and setting r_m equal to the mean conditional on bin m. Thus, a single positive parameter α determines all M rates. The additional parameters to account for rate variation between CpG sites are the fraction of invariant sites β, the frequencies of invariant states $\{p_i\}_{i=1}^{3}$ and the shape parameter α of the Gamma distribution. Maximum likelihood estimates are again obtained considering CpG sites as independent. In summary, we use the $\Gamma + I$ model (Gu *et al.*, 1995) to account for rate heterogeneity across different CpG sites.

2.2.2 Assigning CpG sites to rate categories To assign CpG sites to rate categories we use an empirical Bayes approach (Galtier *et al.*, 2005). Let $\hat{\vartheta}$ denote the maximum likelihood estimates for $\tilde{\vartheta}$. We assign methylation pattern $\mathbf{x}_i$ to rate category $\hat{m} = \mathrm{argmax}_m P(\mathbf{x}_i|\hat{r}_m, \hat{\vartheta})P(m)/P(\mathbf{x}_i|\hat{\vartheta})$, where $P(m) = \hat{\beta}$ for $m = 0$ and $P(m) = (1 - \hat{\beta})/M$ for $1 \leq m \leq M$.

2.3 Reconstructing missing data

Our model of DNA methylation dynamics can reconstruct missing or unobserved methylation states in a cell type. Intuitively, for a given CpG site, nearby cell types in the lineage tree carry information about its likely methylation state. We quantify this relationship using joint maximum likelihood ancestry reconstruction (Pupko *et al.*, 2000). In essence, assume methylation pattern $\mathbf{x}_i$ contains one or more missing values (i.e. unobserved methylation states). Further assume the empirical Bayes procedure discussed above assigns pattern $\mathbf{x}_i$ to rate category m. Note that during this procedure missing values in $\mathbf{x}_i$ had been 'marginalized out'. Then, we assign the missing values in $\mathbf{x}_i$ to the methylation state configuration that maximizes the likelihood $P(\mathbf{x}_i|\hat{r}_m, \hat{\vartheta})$. The algorithm of Pupko *et al.* (2000) is linear in the number of tree nodes, enabling efficient reconstruction of missing methylation states.

This reconstruction strategy shares methylation state information for a CpG site 'vertically' across the lineage tree and is complementary to approaches leveraging 'horizontal' correlations between different but nearby CpG sites across the genome.

2.4 Data sources and processing

Our algorithm requires two inputs: (i) the topology of the lineage tree and (ii) discrete methylation state data for the stages in the lineage tree at specific positions along a genome. Missing methylation states are allowed (see above).

We analyzed DNA methylation maps from 13 cell populations from stages of a differentiation of adult mouse HSCs to different blood lineages (Bock *et al.*, 2012). The purified cell types were obtained at progressive levels of differentiation, starting with HSCs, followed by multipotent progenitor cells (MPP1 and MPP2) and progenitor cells of the lymphoid (CLP) and myeloid (CMP) lineages. For the lymphoid progenitors, further differentiated cells included T helper cells (CD4), T cells (CD8) and B cells (BCELL). For myeloid progenitor cells, the next stages were granulocyte-monocyte progenitors and megakaryocyte–erythroid progenitors (MEP); the former was followed by monocytes (MONO) and granulocytes (GRAN), whereas the latter was followed by erythrocytes (ERY). The relationships between cell types are summarized in the lineage tree in Figure 1. Bock *et al.* (2012) generated a methylation map for each cell type using reduced representation bisulfite sequencing (RRBS).

We downloaded counts of methylated and unmethylated reads at each sequenced CpG dinucleotide for the two replicates performed in each cell type from the Supplementary Materials Web site (http://info.medical-epigenomics.org/papers/broad_mirror/invivomethylation/) for Bock *et al.* (2012). Values for each CpG were averaged over the two replicates for each cell type (and strand where applicable). Then we discretized the methylation status into methylated (>0.8), partially methylated (between 0.1 and 0.8) and unmethylated (<0.1) categories based on the fraction of methylated reads for the site. Histograms of these values showed clear peaks at the ends of the spectrum and were similar between replicates. We defined CpG islands using the `cpgIslandExt` table for the `mm9` build of the mouse gene from the UCSC genome browser (Kent *et al.*, 2002).

We implemented our algorithms in the R language (R Core Team, 2014). To estimate the parameters of the model described in Section 2.2.1, we used the observed frequencies (excluding CpGs with missing values) for $\{p_i\}_{i=1}^{3}$, and used a box-constaint enabled version of the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS-B) algorithm to obtain maximum likelihood estimates for all other parameters. Because we treat CpG positions independently (see above), the computational complexity of estimating rate categories and reconstructing missing data over a set of CpG sites scales linearly with the number of assayed CpGs. Optimizing the likelihood over all CpGs on Chromosome 1 in the Bock *et al.* (2012) data (about 6% of the dataset) took 10 min and 15 s on a single core of an Intel(R) Xeon(R) X5690 CPU with 3.47 GHz clock speed.

3 RESULTS

We applied our methylation dynamics model to an RRBS dataset tracing the differentiation of adult HSCs (Bock *et al.*, 2012). This study queried a set of over 2 million CpG sites at different stages during blood lineage differentiation, and the assayed cell types with their relationships are summarized in the lineage tree in Figure 1.

We fit a model with four rate categories (three variable and one invariant) to the discretized methylation status of CpG sites along each chromosome (Section 2). The maximum likelihood estimates of the model parameters qualitatively agree across chromosomes, and the resulting model is consistent with several previous findings. Invariant CpG sites are more prevalent than variable sites (Bock *et al.*, 2012; Ziller *et al.*, 2013); 61% of CpGs are invariant in our analysis, and 13, 12 and 14% fall into the slow, medium and fast rate categories, respectively. As expected, the equilibrium distribution for variable states favors methylated CpGs, i.e. $\hat{\pi}_\mathrm{u} < \hat{\pi}_\mathrm{p} < \hat{\pi}_\mathrm{m}$ for all chromosomes. Contrasting the dynamics at variable and invariant CpG sites, we find that invariant sites are most likely to remain unmethylated during differentiation (61%), whereas variable sites are most likely to be in methylated states (57%). The branch lengths obtained in our fitted models reflect the number of expected methylation state transitions between cell types, and we see the longest branch lengths between MEP and ERY cells and between HSC and MPP1 cells ($t = 4.65$ and $t = 0.97$ averaged over chromosomes, respectively). These numbers correspond to an expected fraction of CpG sites with *observed* methylation changes of 24% for MEP $\rightarrow$ERY and of 17% and for HSC $\rightarrow$MPP1, taking into account the prevalence of invariant sites and the different rate categories for variable sites. This is in qualitative agreement with previous

results, and it provides a quantitative underpinning of the known 'methylation divergences' between these contexts.

In the next three sections, we present examples of how our modeling approach enables analysis of methylation dynamics across cellular differentiations.

3.1 Reconstructing unobserved methylation states

Over the 2 079 144 CpG sites assayed in 13 cellular contexts in the blood differentiation dataset, 5 940 467 of 27 028 872 (22%) methylation states are missing because of a range of technical issues (Bock *et al.*, 2012). Given the prevalence of missing data, we assessed the ability of our model to reconstruct unobserved values using information from the observed methylation status of the same sites at other nodes in the lineage tree. Conceptually, this is akin to the problem of ancestral sequence reconstruction for DNA substitution models (Pupko *et al.*, 2000), but in the context of methylation, we also have observed data on internal nodes of the tree.

For each cell type, we masked 10 000 CpG sites with measured methylation state, re-estimated model parameters on data missing the masked methylation states, reconstructed the masked values as described in the Section 2 and then compared the reconstructed methylation values with the actual values.

The reconstructed methylation states are generally accurate (90% correct overall), with some differences in performance between different cell types (Fig. 2). As expected, the length of the edges connecting cell types is correlated with the accuracy of the reconstruction of missing values; the nodes with the longest incident edges in the lineage tree (HSC and ERY) are the most difficult to reconstruct.

To compare our model's reconstruction with a baseline method, we also reconstructed the methylation state of each masked CpG based on the methylation state of its nearest neighbor (in terms of genomic location) in the same cell type. This type of reconstruction assumes a 'horizontal' (i.e. location-wise) correlation between methylation states of neighboring CpG sites, whereas our approach can be viewed as assuming a 'vertical' (i.e. progenitor to descendant) correlation between the same CpG site in neighboring cell types.

Overall, lineage tree-based reconstruction performs significantly better than location-based reconstruction (Fig. 2; 90% correct versus 84% correct; $P \approx 0$, binomial test). However, we note that more sophisticated 'horizontal' methods have achieved higher performance in some contexts; see Section 4.

Stratifying reconstructed methylation states by our inferred rate categories revealed that, unsurprisingly, lineage tree-based reconstruction is hardest for 'fast' CpG sites, i.e. those in the fastest rate category according to the empirical Bayes procedure. Therefore, the lineage tree-based reconstruction approach not only performs better than a location-based method but also contributes valuable information about the confidence in the reconstruction result. We anticipate that combining these largely independent approaches could improve reconstruction further.

3.2 DNA sequence context is correlated with CpG methylation dynamics

Having established that our approach can successfully reconstruct unobserved methylation states, we now describe several analyses that use our model's estimates of CpG methylation dynamics at single CpG resolution. Specifically, we show that outside CpG islands, and particularly for CpGs in promoters, the immediate DNA sequence content around CpG dinucleotides is correlated with the variability of their methylation state.

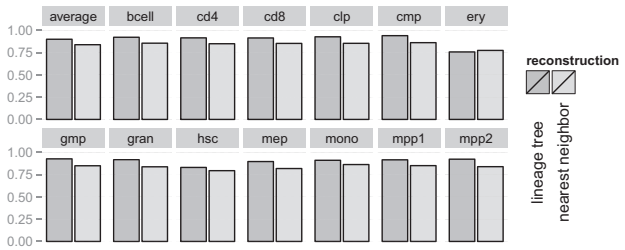

Fig. 2. Lineage tree-based reconstruction of methylation state is more accurate than nearest-neighbor-based reconstruction. We masked the methylation status for 10 000 CpG sites in each cell type and reconstructed these values using 'vertical' information from our lineage tree model and 'horizontal' information from neighboring CpG sites. The lineage tree approach proved significantly more accurate overall (90% versus 84%; $P \approx 0$, binomial test) and within every cell type except erythrocytes ($P = 0.99$), which have the longest branch length ($p < 4E\text{-}17$ for all other cell types)

3.2.1 Local CpG sequence context correlates with methylation dynamics outside CpG islands CpG islands are genomic regions with high CpG dinucleotide frequency, in which methylation status has been reported to influence gene expression (Illingworth and Bird, 2009; Jones, 2012). Here we study methylation dynamics by analyzing the rate category that our model assigns to each assayed CpG dinucleotide.

As expected, we find that invariant CpG dinucleotides are strongly enriched in CpG islands, and that this enrichment falls off with increasing distance from the island (Fig. 3a). To obtain this aggregate view, we split the roughly 16 000 annotated CpG islands into an equal number of bins and discretized flanking genomic regions into equal-sized non-overlapping tiles. The averages for corresponding locations across CpG island loci are shown as points. The strong invariance of CpG islands is in agreement with the notion that CpG islands tend to retain their methylated state (Jones, 2012).

Taking advantage of the single CpG resolution of our approach, we explored whether the enrichment of invariant CpG dinucleotides is exclusive to CpG islands. We hypothesized that local CpG content could be important, so we stratified each CpG dicnucleotide by (i) whether its two neighboring dinucleotides contain none, one or two CpGs and (ii) whether it is located inside a CpG island. For CpGs without neighboring CpG sites, those in CpG islands are strongly enriched for invariance over those outside of CpG islands (factor 1.67), but this enrichment decreases for CpGs with one or two neighboring CpGs (Fig. 3b). In other words, outside of CpG islands, local CpG sequence context is strongly correlated with the absence of methylation state changes across hematopoietic differentiation.

Next, given the importance of invariant CpGs, we assessed whether there is a preference for methylated (or partially methylated) states compared with unmethylated states among invariant CpG sites. We find that invariant CpG sites in CpG islands are almost always unmethylated (96%), which is expected from

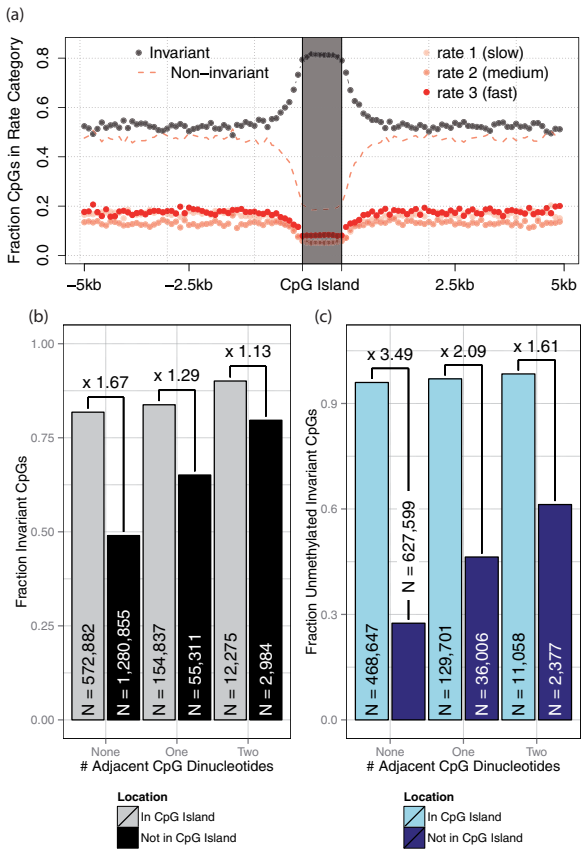

Fig. 3. Methylation dynamics are influenced by DNA sequence context outside of CpG islands. (**a**) Invariant CpG sites (gray circles) are the most common category in our analysis, and they are strongly enriched in CpG islands. The other rate categories (red circles) are roughly equally likely in and around CpG islands. Each dot represents an average over evenly sized bins centered on the 16 024 CpG Islands. The red dashed line gives the sum of the variable categories. (**b**) The presence of adjacent CpG sites is strongly correlated with CpG methylation dynamics outside of CpG islands. The effect of DNA sequence context is much weaker within CpG islands. (**c**) Nearly all invariant CpG sites within CpG Islands are unmethylated. Outside of CpG islands, the adjacent CpG count for a site is strongly correlated with its methylation state

previous analyses (Jones, 2012). However, invariant CpG sites outside of CpG islands are significantly more likely to be methylated (65% methylated and 28% unmethylated; $P \approx 0$, binomial test). We again hypothesized that adjacent sequence context could influence methylation at these sites. Figure 3c shows that invariant CpGs with neighboring CpG dinucleotides outside CpG islands are more unmethylated compared with their counterparts without neighboring CpG dinucleotides.

3.2.2 Low CpG content promoters are enriched for variable CpG sites The previous subsection shows that local CpG context is associated with the dynamics of individual CpG sites in certain settings. Promoter CpGs are known to be functionally important and influenced by CpG density, so next we analyzed single CpG dynamics in promoters with respect to their overall CpG density.

The methylation state of CpGs in gene promoters is associated with transcription levels (Jones, 2012). In several cell types,

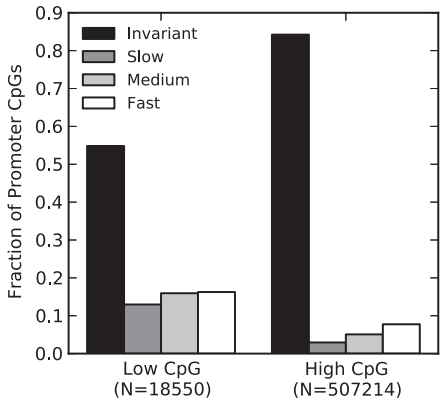

Fig. 4. Low CpG content promoters are enriched for CpGs with variable methylation state. We stratified mouse gene promoters into low and high CpG content groups and then compared the inferred dynamics of CpG sites in these groups. Low CpG content promoters were significantly less likely to be in the invariant rate category ($P \approx 0$; chi-squared test). This pattern remained when CpG islands were not considered

promoter methylation is negatively correlated with gene expression, and the effect is strongest in promoters with low CpG density (Xie *et al.*, 2013). The rate category assignments from our model enabled us to test whether promoter CpG content is also correlated with methylation dynamics across hematopoietic differentiation. Following Xie *et al.* (2013), we defined 'promoters' as regions 500 bp upstream and downstream of TSSs, and we analyzed CpG sites within this window for 19 244 mouse genes. We stratified promoters into low and high CpG density groups. As seen in human data, the CpG density distribution surrounding the mouse TSSs has two peaks; one at low (<0.034 CpG/bp) and one at high CpG density (≥0.034 CpG/bp).

The low CpG content promoters have a significantly higher fraction of variable CpG sites compared with the high CpG promoters (Fig. 4; $P \approx 0$, chi-squared test). Nearly all (84%) of the high CpG promoter CpG sites were invariant across the differentiation, whereas only 55% of the low CpG sites were invariant. This pattern could be driven by the invariance of CpG Islands (Fig. 3) and their prevalence in high CpG content promoters, but the effect remained when CpG island sites were removed (76% versus 54%; $P \approx 0$). These results are consistent with the previous observation that the correlation between methylation state and gene expression is strongest in low CpG promoters (Xie *et al.*, 2013).

3.3 Identification of genomic regions with variable methylation state

The CpG site rate category predictions from our model enable us to identify genomic regions with frequent methylation state changes across hematopoietic differentiation. To do this, we discarded CpG dinucleotides in repeat masked regions (`rmsk` track for `mm9` from UCSC genome browser) and then identified maximal subsets in the RRBS data, for which each site is no >20 bp from the nearest other CpG in the subset. We further filtered out short regions (<50 bp) and focused on regions without evidence for accelerated substitution rates (average `phyloP` score from UCSC genome browser >0); this approach leaves 61 980 regions

with a high density of assayed CpGs that are between 50 and 758 bp long (mean: 95.1 bp). The density of fast (in terms of their annotated rate category) CpG sites in these regions ranges from ~10 to 100%, with 1766 sites exceeding 50%. Figure 5 shows the longest of the 215 regions with all constituent CpGs in the fast rate category. This 129 bp region with 16 assayed CpG sites overlaps a CpG island, has strong evolutionary sequence conservation across placental mammals, and displays histone modifications correlated with transcriptional enhancer activity in various blood-related cell types (ENCODE Project Consortiumm; Bernstein *et al.*, 2012). These attributes suggest a gene regulatory role for this locus. This type of simple candidate approach based on examination of the extremes of the rate category distribution may shed light on regions whose methylation state influences transcriptional regulation during hematopoietic differentiation.

4 DISCUSSION

In this article, we adapt phylogenetic Markov models, which are prevalent in comparative genomics and statistical genetics, to accurately and efficiently analyze DNA methylation dynamics across lineage specification. As a proof of concept, we model RRBS methylation data collected from 13 related stages of blood cell development (Bock *et al.*, 2012). Using our model, we illustrate that (i) CpG site methylation status can be accurately reconstructed using data from related cell types at the same site, (ii) the single CpG site resolution of our methylation dynamics estimates enable the discovery of attributes, such as DNA sequence context, that correlate with CpG methylation dynamics and (iii) our models facilitate the identification of genomic regions with highly variable CpG methylation states that are likely functional.

There are many additional methodologies that could be mapped from the rich reservoir of statistical genomics to the application of modeling methylation dynamics. It will be exciting to see which will prove most useful as genome-wide methylation data continue to be collected to elucidate tissue differentiation and development. For instance, our analyses confirm that methylation dynamics are different between CpG sites located in CpG islands and those elsewhere in the genome. Thus, one way to extend our current approach would be to use different parameterizations based on such 'external' annotations, as is commonly done when modeling coding versus non-coding sequence in comparative genomics. Further on, such models could also be integrated in a hidden Markov model (HMM) framework (as in phylo-HMMs (Siepel and Haussler, 2005)), which could lead to genome segmentations that account for methylation dynamics. An HMM framework has already proven useful in modeling the density of CpG sites across a genome and defining CpG islands (Hsieh *et al.*, 2009; Wu *et al.*, 2010).

In Section 3.1, we demonstrated that progenitor–descendant relationships in the lineage tree can be used to accurately reconstruct the methylation status of CpG sites in different cellular contexts (average of 90% accuracy). These 'vertical' relationships enabled more accurate reconstruction on the RRBS dataset analyzed here than using nearest genomic neighbors to predict missing values. However, we note that there are many methods for reconstructing missing CpG methylation status using genomic information. These methods have largely focused on CpGs in CpG islands, but a recent approach (Zhang *et al.*, 2013) used a random forest classifier to accurately (91–94%) predict CpG

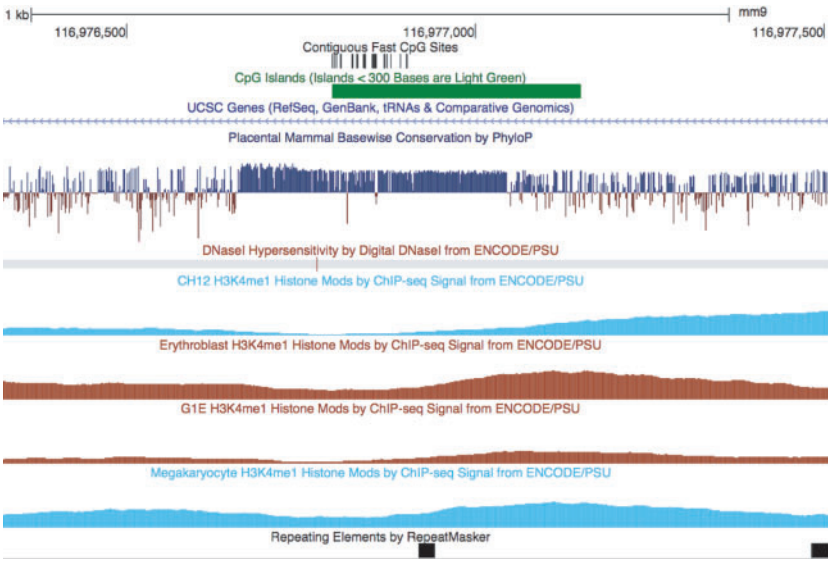

Fig. 5. A block of highly variable CpG sites has evidence of gene regulatory enhancer activity in several blood cells. This region on Chromosome 4 (mm9.chr4:116 976 784–116 976 912) contains 16 CpG sites that our model places in the fast rate category within 129 bp. The region is located within a CpG island in an intron of the gene *Rnf220*, a ubiquitin ligase. The DNA sequence at this locus is strongly conserved across placental mammals; this suggests that it is likely functionally important. In addition, functional genomics data collected by the ENCODE project (ENCODE Project Consortiumm; Bernstein *et al.*, 2012) suggest that this locus is a regulatory enhancer in several blood cell types. It overlaps a DNaseI hypersensitive site in an erythroid progenitor (G1E), and it has peaks of the H3K4me1 enhancer-associated histone modification in B-cell lymphoma cells (CH12), erythroblasts, G1E cells and megakaryocytes

methylation from a methylation array based on a suite of features including neighboring CpG methylation status and overlapping genomic elements. However, the CpG coverage provided by methylation arrays and RRBS assays is different (Laird, 2010), so it is difficult to directly compare the results of these methods on different datasets. Nonetheless, our approach is complementary to existing methods for reconstructing DNA methylation that use neighboring CpG sites along the genome. Context-dependent models that integrate such 'vertical' (i.e. progenitor to descendant) and 'horizontal' (i.e. distance on chromosome) relationships have the potential to further improve methods for reconstructing unobserved states and highlighting functionally relevant shifts in methylation.

Most existing methods for analyzing DNA methylation dynamics are based on pairwise comparisons of cellular contexts (Xie *et al*., 2013; Ziller *et al*., 2013). For example, Ziller *et al*. (2013) identify CpGs with large differences in their estimated methylation between pairs of tissues or cell lines, and then they cluster these to highlight differentially methylated regions. This type of pairwise comparison is appropriate for much of the methylation data currently available, but existing methods are challenging to generalize to analysis of more densely sampled sets of dependent cell types. To address this, our approach explicitly models the existence of statistical dependencies between methylation states from multiple related cell types. Thus, it enables multivariate analyses of methylation patterns in differentiating cell lineages (like those from Bock *et al*., 2012), but it may not provide much improvement when analyzing essentially independent samples (like distantly related terminally differentiated cell types). In addition, our strategy is subject to the discretization of methylation status in a population of cells into discrete methylation states. The three states and the thresholds we use are supported by the distribution of methylation values in our data, but including direct modeling of counts of methylated versus unmethylated instances of a CpG site (Ziller *et al*., 2013) into our approach is a promising future direction.

DNA methylation is just one of several dynamic epigenetic biochemical modifications regulating precise spatiotemporal gene expression patterns that are essential for proper development. The approach we have demonstrated here provides an integrative, multivariate framework for modeling any epigenetic changes across multiple cell types and lineages. As phylogenetic models proved essential in the identification and interpretation of functional DNA sequence regions, we believe that lineage tree-aware Markov models of epigenetic dynamics can play a similar role in developing a deeper understanding of epigenetic phenomena and their roles in tissue differentiation and vertebrate development.

Funding: JAC was supported by institutional funds from Vanderbilt University, DK was supported by institutional funds from the University of Pittsburgh School of Medicine.

Conflict of interest: none declared.

REFERENCES

Bergman,Y. and Cedar,H. (2013) DNA methylation dynamics in health and disease. *Nat. Struct. Mol. Biol.*, **20**, 274–281.

Bock,C. *et al.* (2012) DNA methylation dynamics during *in vivo* differentiation of blood and skin stem cells. *Mol. Cell*, **47**, 633–647.

ENCODE Project Consortiumm, Bernstein,B.E. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Frumkin,D. *et al.* (2005) Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput. Biol.*, **1**, e50.

Galtier,N. *et al.* (2005) Markov models in molecular evolution. In: Nielsen,R. (ed.) *Statistical Methods in Molecular Evolution*. Springer, New York, NY, pp. 3–24.

Gu,X. *et al.* (1995) Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.*, **12**, 546–557.

Guttorp,P. and Minin,V.N. (1995) *Stochastic Modeling of Scientific Data*. Chapman and Hall/CRC, Taylor and Francis Group, Boca Raton, FL, USA.

Hansen,K.D. *et al.* (2011) Increased methylation variation in epigenetic domains across cancer types. *Nat. Genet.*, **43**, 768–775.

Hsieh,F. *et al.* (2009) A nearly exhaustive search for CpG islands on whole chromosomes. *Int. J. Biostat.*, **5**, 14.

Illingworth,R.S. and Bird,A.P. (2009) CpG islands –'A rough guide'. *FEBS Lett.*, **583**, 1713–1720.

Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, **13**, 484–492.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Laird,P.W. (2010) Principles and challenges of genome-wide DNA methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.

Nordlund,J. *et al.* (2013) Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biol.*, **14**, r105.

Portela,A. and Esteller,M. (2010) Epigenetic modifications and human disease. *Nat. Biotech.*, **28**, 1057–1068.

Pupko,T. *et al.* (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol. Biol. Evol.*, **17**, 890–896.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Siepel,A. and Haussler,D. (2005) Phylogenetic hidden Markov models. In: Nielsen,R. (ed.) *Statistical Methods in Molecular Evolution*. Springer, New York, pp. 325–351.

Smith,Z.D. and Meissner,A. (2013) DNA methylation: roles in mammalian development. *Nat. Rev. Genet.*, **14**, 204–220.

Tost,J. (2010) DNA methylation: an introduction to the biology and the disease-associated changes of a promising biomarker. *Mol. Biotechnol.*, **44**, 71–81.

Wu,H. *et al.* (2010) Redefining CpG islands using hidden Markov models. *Biostatistics*, **11**, 499–514.

Xie,W. *et al.* (2013) Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell*, **153**, 1134–1148.

Zhang,W. *et al.* (2013) Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *ArXiv e-prints*, http://arxiv.org/abs/1308.2134 (29 July 2014, date last accessed).

Ziller,M.J. *et al.* (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature*, **500**, 477–481.