Hidden Markov Modelling Reveals Neighborhood Dependence of Dnmt3a and 3b Activity

Alexander Lück, Pascal Giehr, Karl Nordström, Jörn Walter, and Verena Wolf

Abstract—DNA methylation is an epigenetic mark whose important role in development has been widely recognized. This epigenetic modification results in heritable information not encoded by the DNA sequence. The underlying mechanisms controlling DNA methylation are only partly understood. Several mechanistic models of enzyme activities responsible for DNA methylation have been proposed. Here, we extend existing Hidden Markov Models (HMMs) for DNA methylation by describing the occurrence of spatial methylation patterns over time and propose several models with different neighborhood dependences. Furthermore, we investigate correlations between the neighborhood dependence and other genomic information. We perform numerical analysis of the HMMs applied to comprehensive hairpin and non-hairpin bisulfite sequencing measurements and accurately predict wild-type data. We find evidence that the activities of Dnmt3a and Dnmt3b responsible for *de novo* methylation depend on 5' (left) but not on 3' (right) neighboring CpGs in a sequencing string.

Index Terms—DNA methylation, hidden Markov model, spatial stochastic model

1 Introduction

The DNA code of an organism determines its appearance and behavior by encoding protein sequences. In addition, there is a multitude of additional mechanisms to control and regulate the ways in which the DNA is packed and processed in the cell and thus determine the fate of a cell. One of these mechanisms in cells is DNA methylation, which is an epigenetic modification that occurs at cytosine (C) bases of eukaryotic DNA. Cytosines are converted to 5-methylcytosine (5mC) by DNA methyltransferase (Dnmt) enzymes. The neighboring nucleotide of a methylated cytosine is usually guanine (G) and together with the CG-pair on the opposite strand, a common pattern is that two methylated cytosines are located diagonally to each other on opposing DNA strands. DNA methylation at CpG dinucleotides is known to control and mediate gene expression and is therefore essential for cell differentiation and embryonic development. In human somatic cells, approximately 70-80 percent of the cytosine nucleotides in CpG dyads are methylated on both strands and methylation near gene promoters varies considerably depending on the cell type. Methylation of promoters often correlates with low or no transcription [27] and can be used as a predictor of gene expression [14]. Also, significant differences in overall and specific methylation levels exist between different tissue types and between normal cells and cancer cells from the same tissue. However, the exact mechanism which leads to a methylation of a specific CpG and the formation of distinct methylation patterns at certain genomic regions is still not fully understood. Recently proposed measurement techniques based on hairpin bisulfite sequencing (BS-seq) allow to determine the level of 5mC at individual CpGs dyads on both DNA strands [19]. Based on a small hidden Markov model, the probabilities of the different states of a CpG can be accurately estimated (assuming that enough samples per CpG are provided) [1], [9], [16].

Mechanistic models for the activity of the different Dnmts usually distinguish *de novo* activities, i.e., adding methyl groups at cytosines independent of the methylation state of the opposite strand, and maintenance activities, which refers to the copying of methylation from an existing DNA strand to its newly synthesized partner (containing no methylation) after replication [12], [23]. Hence, maintenance methylation is responsible for re-establishment of the same DNA methylation pattern before and after cell replication. A common hypothesis is that the copying of DNA methylation patterns after replication is performed by Dnmt1, an enzyme that shows a preference for hemimethylated CpG sites (only one strand is methylated) as they appear after DNA replication. Moreover, studies have shown that Dnmt1 is highly processive and able to methylate long sequences of hemimethylated CpGs without dissociation from the target DNA strand [12]. However, an exact transmission of the methylation information to the next cellular generation is not guaranteed. The enzymes Dnmt3a and Dnmt3b show equal activities on hemi- and unmethylated DNA and are mainly responsible for *de novo* methylation, i.e., methylation without any specific preference for the current state of the CpG (hemi- or unmethylated) [23]. However, by now evidence exists that the activity of the different enzymes is not that exclusive, i.e., Dnmt1

- *A. Lück and V. Wolf are with the Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany. E-mail: {alexander.lueck, verena.wolf}@uni-saarland.de.*
- *P. Giehr, K. Nordström, and J. Walter are with the Department of Biological Sciences, Saarland University, 66123 Saarbrücken, Germany. E-mail: pascalgiehr@googlemail.com, karl.nordstroem@uni-saarland.de, j.walter@mx.uni-saarland.de.*

Manuscript received 15 June 2018; revised 1 Mar. 2019; accepted 8 Apr. 2019. Date of publication 23 Apr. 2019; date of current version 7 Oct. 2019. (Corresponding author: Alexander Lück.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2019.2910814

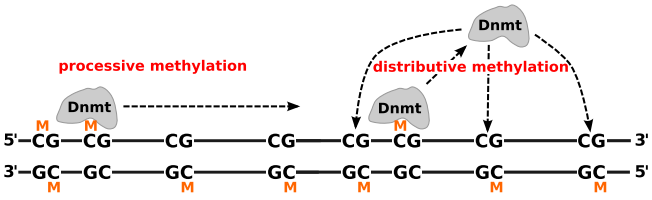

Fig. 1. Dnmts can methylate DNA in a processive way where the enzyme starts at one CpG and slides in 5' to 3' direction over the DNA or in a distributive manner, "jumping" randomly from one CpG to another.

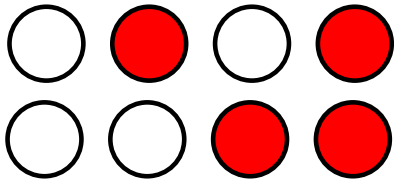

Fig. 2. A lattice of length $L = 4$ containing all possible states 0, 1, 2, and 3, forming the pattern 0123.

shows to a certain degree also *de novo* and Dnmt3a/b maintenance methylation activity [2]. The way how methyltransferases interact with the DNA and introduce CpG methylation was investigated in many *in vitro* studies. Basically, one can distinguish between two mechanisms. A distributive one, where the enzyme periodically binds and dissociates from the DNA, leaping more or less randomly from one CpG to another and a processive one in which the enzyme migrates along the DNA without detachment from the DNA [10], [13], [22], as illustrated in Fig. 1. Note that for Dnmt1, for instance, it is reasonable to assume that it is processive in 5′ to 3′ direction since it is linked to the DNA replication machinery. In particular for the Dnmt3's different hypotheses about the processivity and neighborhood dependence exist [3], [6], but the detailed mechanisms remain elusive.

Several models that describe the dynamics of the formation of methylation patterns have been proposed. In the seminal paper of Otto and Walbot, a dynamical model was proposed that assumed independent methylation events for a single CpG. The main idea was to track the frequencies of fully, hemi- and unmethylated CpGs during several cell generations [24]. Later, refined models allowed to distinguish between maintenance and *de novo* methylation on the parent and daughter strands [8], [26]. More sophisticated extensions of the original model of Otto and Walbot models have been successfully used to predict *in vivo* data still assuming a neighbor-independent methylation process for a single CpG site [2], [9]. However, measurements indicate that methylation events at a single CpG may depend on the methylation state of neighboring CpGs, which is not captured by these models.

Here, we follow the dynamical HMM approach proposed in [2] where knockout data was used to train a model that accurately predicts wild-type methylation levels for BS-seq data of repetitive elements from mouse embryonic stem cells. We extend this model by describing the methylation state of several CpGs instead of a single CpG and use similar dependence parameters as introduced in Bonello et al. [4]. More specifically, we design different models by combining the activities of the two types of Dnmts and test for both, maintenance and *de novo* methylation the hypotheses illustrated in Fig. 1. The models vary according to the order in which the enzymes act, whether they perform methylation in a processive manner or not, and how much their action depends on the left/right CpG neighbor. We use the same BS-seq hairpin data as in [2], i.e., data where Dnmt1 or Dnmt3a/b was knocked out (KO) and learn the parameters of the different models. We also relate the estimated dependence parameters to the distance between the respective adjacent CpGs in order to investigate their possible influence. Then, similar as in [2], we predict the behavior of the measured wild-type (WT), in which both types of enzymes are active, by designing a combined model that describes the activity of both enzymes and compare the results to the WT data. Finally, we apply our model to non-hairpin data.

We found that all proposed models show a similar behavior in terms of prediction quality such that no model can be declared as the best fit. However, our results indicate that Dnmt1 works independently of the methylation state of its neighborhood, which is in accordance to the current hypothesis that Dnmt1 is linked to the replication machinery and copies the methylation state on the opposite strand. On the other hand, Dnmt3a/b shows a dependence to the left but no dependence to the right, which supports hypotheses of processive or cooperative behavior. Furthermore, we find evidence that at least for small distances rather the genetic region than the distance determines the dependence on the neighbors. Applying our model to a genome-wide data set we find three distinct clusters based on the dependence parameters and distances between adjacent CpGs. These clusters also show different methylation levels and reveal that hypomethylated CpGs in promoter regions behave independent of their neighborhood. Finally our results show that our model can also be used for non-hairpin data as long as no information from the opposite strand is needed as for example in Dnmt1KO data.

This paper is organized is as follows: Our model is introduced in Section 2 and the results are presented in Section 3. In Section 4 we discuss the related work. We conclude the paper in Section 5 and give a brief outline on future work.

2 Model

2.1 Notation

Consider a sequence of L neighboring CpG dyads,[1] which is represented as a lattice of length L and width two (for the two strands). Each cytosine in the lattice can either be methylated or not, leading to four possible states at each position l:

- *State 0*: Both cytosines are not methylated.
- *State 1*: The cytosine on the upper strand is methylated, the lower one not.
- *State 2*: The cytosine on the lower strand is methylated, the upper one not.
- *State 3*: Both cytosines are methylated.

A sequence of four CpGs, each of which is in one of the four possible states, is shown in Fig. 2.

For a system of length L there are in total 4^L possibilities to combine the states of individual CpGs. These combinations are called *patterns* in the following. A pattern is denoted by a concatenation of states, e.g., 321, 0123 or 33221.

1. The exact nucleotide distance between two neighboring dyads is not considered here explicitly, but we assume that this distance is small. For the BS-seq data that we consider, the average distance between two CpGs is 14 bps (base pairs) and the maximal distance is 46 bps.

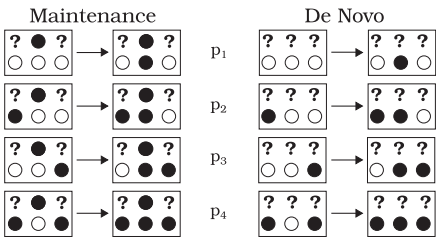

Fig. 3. Possible maintenance and *de novo* transitions depicted for the lower strand, where ○ denotes an unmethylated, ● a methylated site, and **?** a site where the methylation state does not matter. Note that the same transitions can occur on the upper strand.

In order to represent the pattern distribution as a vector it is necessary to uniquely assign a reference number to each pattern. A pattern can be perceived as a number in the tetral system, such that converting to the decimal system leads to a unique reference number. After the conversion an additional 1 is added in order to start the referencing at 1 instead of 0.

Examples for $L = 3$:

$$\begin{aligned} 000 &\longrightarrow 1\ (=0+1)\\ 123 &\longrightarrow 28\ (=27+1)\\ 333 &\longrightarrow 64\ (=63+1). \end{aligned}$$

This reference number then corresponds to the position of the pattern in the respective distribution vector.

We describe the state of a sequence of L CpGs by a discrete-time Markov chain with pattern distribution $\pi(t)$, i.e., the probability of each of the 4^L patterns after t cell divisions. For the initial distribution $\pi(0)$, we use the distribution measured in the wild-type when the cells are in equilibrium. Note, that other initial conditions gave very similar results, i.e., the choice of the initial distribution does not significantly affect the results. The reason is that also the KO data is measured after a relatively high number of cell divisions where the cells are almost in equilibrium. Transitions between patterns are triggered by different processes: First due to *cell division* the methylation on one strand is kept as it is (e.g., the upper strand), whereas the newly synthesized strand (the new lower strand) does not contain any methyl group. Afterwards, methylation is added due to different mechanisms. On the newly synthesized strand a site can be methylated if the cytosine at the opposite strand is already methylated (*maintenance*). It is widely accepted that maintenance in form of Dnmt1 is linked to the replication machinery and thus occurs during/directly after the synthesis of the new strand. Furthermore, CpGs on both strands can be methylated independent of the methylation state of the opposite site (*de novo*). The transition matrix P is defined by composition of matrices for cell division, maintenance and *de novo* methylation of each site.

2.2 Cell Division

Depending on which daughter cell is considered after cell replication, the upper ($s = 1$) or lower ($s = 2$) strand is the parental one after cell division. Then, the new pattern can be obtained by applying the following state replacements:

$$s=1:\begin{cases} 0 \longrightarrow 0\\ 1 \longrightarrow 1\\ 2 \longrightarrow 0\\ 3 \longrightarrow 1 \end{cases} \qquad s=2:\begin{cases} 0 \longrightarrow 0\\ 1 \longrightarrow 0\\ 2 \longrightarrow 2\\ 3 \longrightarrow 2 \end{cases}. \tag{1}$$

Given some initial pattern with reference number i, applying the transformation Eq. (1) to each of the L positions leads to a new pattern with reference number j (notation: $i \overset{(1)}{\rightsquigarrow} j$). The corresponding transition matrix $D_s \in \{0,1\}^{4^L \times 4^L}$ has the form

$$D_s(i,j) = \begin{cases} 1, & \text{if } i \overset{(1)}{\rightsquigarrow} j,\\ 0, & \text{else.} \end{cases}. \tag{2}$$

2.3 Maintenance and De Novo Methylation

For maintenance and *de novo* methylation, the single site transition matrices are built according to the following rules:

Consider at first the (non-boundary) site $l = 2, \ldots, L-1$ and its left and right neighbor $l-1$ and $l+1$ respectively. The remaining sites do not change and do not affect the transition. The probabilities of the different types of transitions in Fig. 3 have the form

$$p_1 = 0.5\cdot(\psi_L + \psi_R)x, \tag{3}$$

$$p_2 = 0.5\cdot(\psi_L + \psi_R)x + 0.5\cdot(1-\psi_L), \tag{4}$$

$$p_3 = 0.5\cdot(\psi_L + \psi_R)x + 0.5\cdot(1-\psi_R), \tag{5}$$

$$p_4 = 1 - 0.5\cdot(\psi_L + \psi_R)(1-x), \tag{6}$$

where we set the probability x to $x = \mu$ in case of maintenance and to $x = \tau$ in case of *de novo* methylation. $\psi_L,\ \psi_R \in [0,1]$ are the dependence parameters for the left and right neighbor.

A dependence value of $\psi_i = 1$ corresponds to a total independence on the neighbor whereas $\psi_i = 0$ leads to a total dependence. Hence, μ and τ can be interpreted as the probability of maintenance and *de novo* methylation of a single cytosine between two cell divisions assuming independence from neighboring CpGs. Moreover, all CpGs that are part of the considered window of the DNA have the same value for the parameters μ, τ, ψ_L, and ψ_R, since in earlier experiments only very small differences have been found between the methylation efficiencies of nearby CpGs [2].

In order to understand the form of the transition probabilities consider at first a case with only one neighbor. The probabilities then have the form ψx if the neighbor is unmethylated and $1 - \psi(1-x)$ if the neighbor is methylated. Note that both forms evaluate to x for $\psi = 1$, meaning that a site is methylated with probability x, independent of its neighbor. For $\psi = 0$ the probabilities become 0 and 1, meaning that if there is no methylated neighbor the site cannot be methylated or will be methylated for sure if there is a methylated neighbor respectively.

The probabilities for two neighbors are obtained by a linear combination of the one neighbor cases, with ψ_L for the left and ψ_R for the right neighbor, and an additional weight of 0.5 to normalize the probability. The same considerations also apply to the boundary sites however there is no way of knowing the methylation states outside the boundaries (denoted by ?). Therefore instead of a concrete methylation state (○ for unmethylated, ● for methylated site) the average methylation density ρ is used to compute the transition probabilities at the boundaries (depicted here for *de novo*):

$$? \circ \circ \rightarrow ? \bullet \circ \qquad \tilde{p}_1 = (1-\rho)\cdot p_1 + \rho\cdot p_2, \tag{7}$$

$$? \circ \bullet \rightarrow ? \bullet \bullet \qquad \tilde{p}_2 = (1-\rho)\cdot p_3 + \rho \cdot p_4, \tag{8}$$

$$\circ \circ ? \rightarrow \circ \bullet ? \qquad \tilde{p}_3 = (1-\rho)\cdot p_1 + \rho \cdot p_3, \tag{9}$$

$$\bullet \circ ? \rightarrow \bullet \bullet ? \qquad \tilde{p}_4 = (1-\rho)\cdot p_2 + \rho \cdot p_4. \tag{10}$$

Note that the same considerations hold for maintenance at the boundaries if the opposite site of the boundary site is already methylated.

For each position l, there are four transition matrices: two for maintenance and two for *de novo*, namely one for the upper and one for the lower strand in each process. In order to construct these matrices consider the three positions $l-1$, l and $l+1$, where the transition happens at position l. Only the transitions depicted in Fig. 3 can occur. Furthermore the transitions are unique, i.e., for a given reference number i the new reference number j is uniquely determined. For patterns not depicted in Fig. 3 no transition can occur, i.e., the reference number does not change.

The matrix describing a maintenance event at position l and strand s has the form

$$M_s^{(l)}(i,j) = \begin{cases} 1, & \text{if } \; i=j \text{ and } \nexists j' : \; i \rightsquigarrow j', \\ 1-p, & \text{if } \; i=j \text{ and } \exists j' : \; i \rightsquigarrow j', \\ p, & \text{if } \; i \neq j \text{ and } \; i \rightsquigarrow j, \\ 0, & \text{else,} \end{cases} \tag{11}$$

where the probability p is given by one of the Eqs. (3), (4), (5), (6), (7), (8), (9), (10) that describes the corresponding case and $x = \mu$. Note that $M_s^{(l)}$ depends on s and l since it describes a single transition from pattern i to pattern j, which occurs on a particular strand and at a particular location with probability p. We define matrices $T_s^{(l)}$ for *de novo* methylation according to the same rules except that $x = \tau$ and the possible transitions are as in Fig. 3, right. All matrices are of size $4^L \times 4^L$.

The advantage of defining the matrices position- and process-wise is that different models can be realized by changing the order of multiplication of these matrices.

It is important to note that 5mC can be further modified by oxidation to 5-hydroxymethyl- (5hmC), 5-formyl- (5fC) and 5-carboxyl cytosine(5caC) by Tet enzymes. These modifications are involved in the removal of 5mC from the DNA and can potentially interfere with methylation events. However, our data does not capture these modifications and therefore we are not able to consider these modifications in our model.

2.4 Combination of Transition Matrices

For all subsequent models it is assumed that first of all cell division happens and maintenance methylation only occurs on the newly synthesized strand given by s, whereas *de novo* methylation happens on both strands. Given the mechanisms in Fig. 1, the two different kinds of methylation events, and the two types of enzymes, there are several possibilities to combine the transition matrices. We consider the following four models, which we found most reasonable based on the current state of research in DNA methylation:

1) first processive maintenance and then processive *de novo* methylation

$$P_s = \prod_{l_1=1}^{L} M_s^{(l_1)} \prod_{l_2=1}^{L} T_1^{(l_2)} \prod_{l_3=1}^{L} T_2^{(l_3)}, \tag{12}$$

2) first processive maintenance and then *de novo* in arbitrary order

$$P_s = \frac{1}{(L!)^2} \prod_{l_1=1}^{L} M_s^{(l_1)} \left(\sum_{\sigma_1 \in S_L} \prod_{l_2=1}^{L} T_1^{(\sigma_1(l_2))} \right) \cdot \left(\sum_{\sigma_2 \in S_L} \prod_{l_3=1}^{L} T_2^{(\sigma_2(l_3))} \right), \tag{13}$$

3) maintenance and *de novo* at one position, processive

$$P_s = \prod_{l=1}^{L} M_s^{(l)} T_1^{(l)} T_2^{(l)}, \tag{14}$$

4) maintenance and *de novo* at one position, arbitrary order

$$P_s = \frac{1}{L!} \sum_{\sigma \in S_L} \prod_{l=1}^{L} M_s^{(\sigma(l))} T_1^{(\sigma(l))} T_2^{(\sigma(l))}, \tag{15}$$

where S_L is the set of all possible permutations for the numbers $1, \ldots, L$.

Note that the *de novo* events on both strands are independent, i.e., the *de novo* events on the upper strand do not influence the *de novo* events on the lower strand and vice versa, such that $[T_1^{(l)}, T_2^{(l')}] = 0$ independent of ψ_i.[2] Obviously it is important whether maintenance or *de novo* happens first, since the transition probabilities and the transitions themselves depend on the actual pattern. Furthermore in the case $\psi_i < 1$ (dependence on right and/or left neighbor) the order of the transitions on a strand matters, i.e., $[M_s^{(l)}, M_s^{(l')}] \neq 0$ and $[T_s^{(l)}, T_s^{(l')}] \neq 0$ for $l \neq l'$. Note that this definition of models in principle allows to consider an arbitrary number of CpGs. However, at least three CpGs are needed to properly include the influence of the left and right neighbor in the transitions. It is also important to note that independent of the number of considered CpGs the window size of the influential CpGs for the transition rates is always kept at size three. However, treating more than three CpGs at once has two major drawbacks: First of all the number of possible patterns grows rapidly (recall 4^L possible patterns for L CpGs) and hence the transition matrices become very large as well ($4^L \times 4^L$). This may lead to memory issues while calculating the distributions, which can however be circumvented by sampling approaches, i.e., stochastic simulation of the underlying Markov chain. Another problem with the large number of possible patterns is that more data is required in order to ensure a good coverage, i.e., the number of measurements should be larger than the number of patterns.

The second main problem is that using the same dependence parameters for all pairs of adjacent CpGs is a rather strong assumption. Note that this assumption becomes more problematic for larger windows, due to e.g., different distances between the CpGs. One solution would be to introduce extra dependence parameters for each pair, however this may lead to difficulties in the parameter identification.

2. $[A, B] = AB - BA$ is the commutator of the matrices A and B.

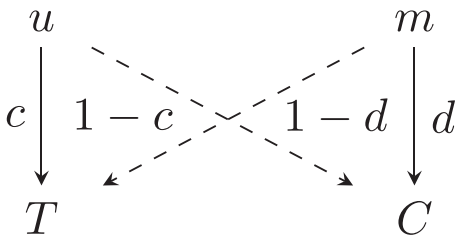

Fig. 4. Conversions of the unobservable states u, m to observable states T, C with respective rates.

The total transition matrix is then given by a combination of the cell division and maintenance/*de novo* matrices. Recall that we consider two different types of Dnmts, i.e., Dnmt1 and Dnmt3a/b. If only one type of Dnmt is active (KO data) the matrix has the form

$$P = 0.5 \cdot (D_1 \cdot P_1 + D_2 \cdot P_2) \tag{16}$$

and if all Dnmts are active (WT data)

$$P = 0.5 \cdot (D_1 \cdot P_1 \cdot \tilde{P}_1 + D_2 \cdot P_2 \cdot \tilde{P}_2), \tag{17}$$

where P_s and $\tilde{P}_s$ have one of the forms Eqs. (12), (13), (14), (15). This leads to four different models for one active enzyme or 16 models for all active enzymes respectively. In the second case P_s represents the transitions caused by Dnmt1 and $\tilde{P}_s$ the transitions caused by Dnmt3a/b. Note that if $\psi_L = \psi_R = 1$ all models are the same within each case since they reduce to the neighborhood independent model from [2]. Furthermore, the cell division, maintenance, and *de novo* transition matrices for a single CpG at a given position are sparse. However, upon combining them to the full transition matrices in Eq. (16) or (17), the final matrices become dense and therefore have higher memory requirements.

2.5 Conversion Errors

The actual methylation state of a C cannot be directly observed. During BS-seq, with high probability every unmethylated C (denoted by u) is converted into thymine (T) and every 5mC (denoted by m) into C. However, conversion errors may occur and we define their probability as $1 - c$ and $1 - d$, respectively, as shown by the dashed arrows in Fig. 4. It is reasonable that these conversion errors occur independently and with approximately identical probability at each site and thus the error matrix for a single CpG takes the form

$$\Delta_1 = \begin{pmatrix} c^2 & c\bar{c} & c\bar{c} & \bar{c}^2 \\ c\bar{d} & cd & \bar{c}\bar{d} & d\bar{c} \\ c\bar{d} & \bar{c}\bar{d} & cd & d\bar{c} \\ \bar{d}^2 & d\bar{d} & d\bar{d} & d^2 \end{pmatrix}, \tag{18}$$

with $\bar{c} = 1 - c$ and $\bar{d} = 1 - d$. Due to the independence of the events this matrix can easily be generalized for systems with $L > 1$ by recursively using the Kronecker-product

$$\Delta_L = \Delta_1 \otimes \Delta_{L-1} \qquad \text{for} \quad L \geq 2. \tag{19}$$

Hence, Δ_L gives the probability of observing a certain sequence of C and T nucleotides for each given unobservable methylation pattern. In order to compute the likelihood $\hat{\pi}$ of the observed BS-seq data, we therefore first compute the transient distribution $\pi(t)$ of the underlying Markov chain at the corresponding time instant[3] t by solving

$$\pi(t) = \pi(0) \cdot P^t \tag{20}$$

and then multiply the distribution of the unobservable patterns with the error matrix.

$$\hat{\pi} = \pi(t) \cdot \Delta_L. \tag{21}$$

Note that this yields a hidden Markov model with emission probabilities Δ_L. In the following the values for c were chosen according to [2]. Since the value for d was not determined in [2], we measured the conversion rate $d = 0.94$ in an independent experiment under comparable conditions [9]. In this study we used hairpin linkers, which contain C, 5mC, as well as 5hmC. After sequencing we determine the conversion state of each particular C from within each read. Note, that we calculated the average conversion rate of all experiments for the present study.

2.6 Maximum Likelihood Estimator

In order to estimate the parameters $\theta = (\mu, \psi_L, \psi_R, \tau) \in [0, 1]^4$, we employ a Maximum (Log)Likelihood Estimator (MLE)

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta), \quad \ell(\theta) = \sum_{j=1}^{4^L} \log\left(\hat{\pi}_j(\theta)\right) \cdot N_j, \tag{22}$$

where $\hat{\pi}$ is the pattern distribution obtained from the numerical solution of Eqs. (20) and (21) for a given time t and N_j is the number of occurrences of pattern j in the measured data. The parameters $\theta = \hat{\theta}$ are chosen in such a way that ℓ is maximized. In order to ensure that the global maximum in $[0, 1]^4$ is found during the optimization, we ran the estimation several times with different random starting points. In all cases the estimation yielded the same results, such that it is very likely that indeed the global optimum was found.

We employ the MLE twice in order to estimate the parameter vector $\hat{\theta}_1$ for Dnmt1 from the 3a/b DKO (double knockout) data and the vector $\hat{\theta}_{3a/b}$ for Dnmt3a/b from the Dnmt1 KO data, where transition matrix Eq. (16) is used. The corresponding time instants are $t = 26$ for the 3a/b DKO data and $t = 41$ for the 1KO data.

We approximate the standard deviations of the estimated parameters $\hat{\theta}$ as follows: Let $\mathcal{I}(\hat{\theta}) = \mathbb{E}[-\mathcal{H}(\hat{\theta})]$ be the expected Fisher information, with the Hessian $\mathcal{H}(\hat{\theta}) = \nabla\nabla\ell(\hat{\theta})$. The inverse of the expected Fisher information is a lower bound for the covariance matrix of the MLE such that we can use the approximation $\sigma(\hat{\theta}) \approx \sqrt{\text{diag}(-\mathcal{H}(\hat{\theta}))}$.

A prediction for the wild-type can be computed by combining the estimated vectors such that in the model both types of enzymes are active. For this, we insert $\hat{\theta}_1$ in P_s and $\hat{\theta}_{3a/b}$ in $\tilde{P}_s$ in Eq. (17) to obtain the transition matrix for the wild-type.

3. The number of cell divisions is estimated from the time of the measurement since these cells divide once every 24 hours.

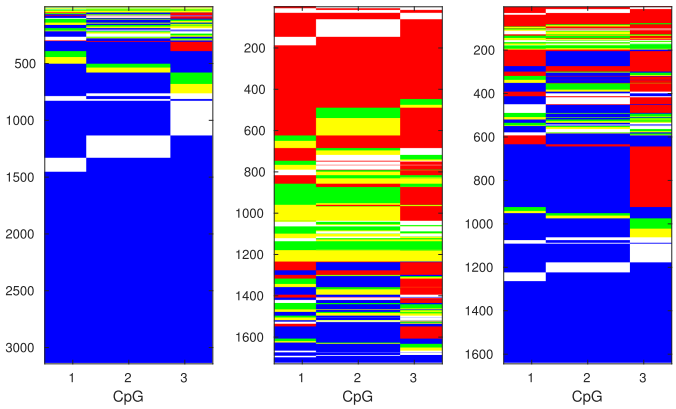

Fig. 5. Representations of WT (left), Dnmt1KO (middle), and Dnmt3a/b DKO (right) data for mSat. On the X axis, the CpGs and on the Y axis the measured cells are shown. The different colors encode the states as follows: Red: 0, green: 1, yellow: 2, blue: 3, and white: "no measurement".

2.7 Data

For our analysis we focused on hairpin data of the single copy genes Afp (5 CpGs) and Tex13 (10 CpGs) as well as the repetitive elements IAP (intracisternal A particle) (6 CpGs), L1 (Long interspersed nuclear elements) (7 CpGs) and mSat (major satellite) (3 CpGs). During the workflow of hairpin bisulfite sequencing, the two DNA strands are linked together covalently, i.e., the methylation status of both strands from an individual chromosome (DNA molecule) is known. Repetitive elements occur in multiple copies and are dispersed over the entire genome. Therefore they allow capturing an averaged, more general behavior of methylation dynamics. Typical data sets are shown in Fig. 5. Note that the WT data is almost always fully methylated, while the Dnmt1KO data is mostly un- or hemimethylated. The Dnmt3ab DKO data is somewhat in between.

3 Results

3.1 Parameter Estimation

In the following we focus on the hairpin data for the single copy genes and repetitive elements as introduced in the previous section. If a locus contains more than three CpGs, the analysis is done for all sets of three adjacent sites independently, in order to keep computation times short and memory requirements low. In the sequel, we mainly focus on the estimated dependence parameters ψ_L and ψ_R and on the prediction quality of the different models.

The estimates for all the available KO data and all suggested models obtained using the transition matrix in Eq. (16) are summarized as histograms in Fig. 6. Because of the different possibilities to combine the four different models in Eqs. (12), (13), (14), (15) and because of the different loci considered, in total there are 84 estimates for each KO data set. We plot the number of occurrences N of ψ_L (left) and ψ_R (right) in different ranges for both sorts of KO data (Dnmt1KO and Dnmt3a/b DKO).

The estimates of ψ_L spread over the whole interval [0, 1] while in the case of ψ_R, nearly all estimates are larger than 0.99 and only in a few cases the dependence parameter is significantly smaller than 1. Hence, in most cases the methylation probabilities are independent of the right neighbor for both Dnmt1KO and Dnmt3a/b DKO. For ψ_L the dependence

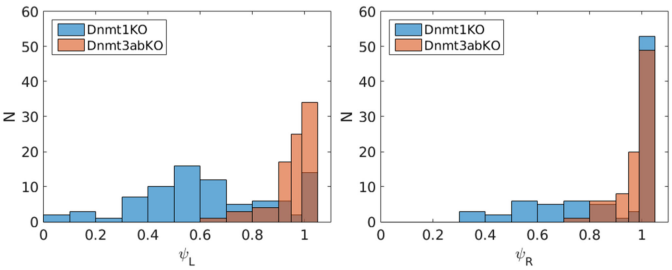

Fig. 6. Histograms for the estimated dependence parameters ψ_L and ψ_R for all sets of three adjacent CpGs in all loci and for all suggested models.

parameter in the Dnmt3a/b DKO case occurs more often close to 1, meaning that the transitions induced by Dnmt1 have little to no dependence on the left neighbor. On the other hand for Dnmt1KO the dependence parameter occurs more often at smaller values giving evidence that there is a dependence on the left neighbor for the activity of Dnmt3a/b. Note that all models show a similar behavior in terms of the dependence parameters for a given locus or position within a locus respectively, i.e., either $\psi_i \approx 1$ or $\psi_i < 1$ for all models. Since the histograms for Dnmt3a/b DKO look very similar for ψ_L and ψ_R, we used a two-sample Kolmogorov-Smirnov test to assess if they differ significantly. The resulting p-value of 1 indicates that there is no significant difference in this case. Note that we also get quite high p-values (0.786 and 0.433) when applying the test to the Dnmt1KO histogram for ψ_R and the two Dnmt3a/b DKO histograms. On the other hand, the p-values are significantly smaller for the Dnmt1KO ψ_L histogram, with a minimum of 0.019, indicating a different behavior for the dependence on the left neighbor for Dnmt3a/b.

Since ψ_R is usually close to 1 a smaller model with only three parameters $\theta = (\mu, \psi, \tau)$ can be proposed, where ψ is a dependence parameter for the left neighbor. This model can either be obtained by fixing $\psi_R = 1$ in the original model and setting $\psi = \psi_L$ or by redefining the transition probabilities to ψx if the left neighbor is unmethylated and $1 - \psi(1 - x)$ if the left neighbor is methylated. In that case ψ and ψ_L are related via $\psi = 0.5(\psi_L + 1)$. Note that both versions yield the same results. In order to check whether there is a significant difference in the original and the smaller model, we performed a Likelihood-ratio test with the null hypothesis that the smaller model is a special case of the original model. Since the original model with more parameters is always as least as good as the smaller model, our goal is to check in which cases the smaller model is sufficient. Indeed, if ψ_R was estimated to be approximately 1 the Likelihood-ratio test indicates that the smaller model is sufficient (p-value ≈ 1). On the other hand, for the few cases where ψ_R differs significantly from 1 the original model has to be used (p-value < 0.01).

3.2 CpG Distances

We now take a closer look at the estimated dependence parameters shown in the histograms in Fig. 6 and link the parameters to their respective loci and distances between adjacent CpGs. The results for the estimation of the left and right dependence parameter for both Dnmt3a/b DKO and Dnmt1KO data, based on the transition matrix in Eq. (12) are shown in Fig. 7. The results based on the other transition matrices yielded similar results and are therefore not

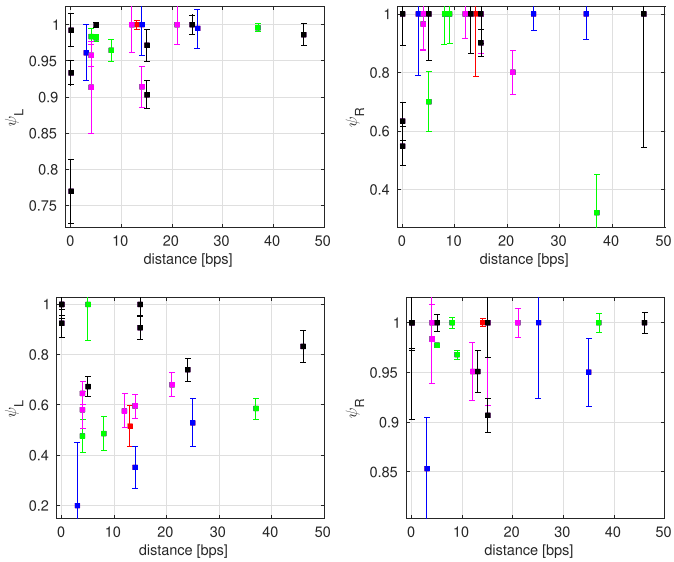

Fig. 7. Dependence parameter versus distance between CpGs measured in bps. The top row shows the results for the Dnmt3a/b DKO data, the bottom row for Dnmt1KO. The left (right) column shows results for the dependence parameter to the left (right). The different colors of the symbols represent the different loci and are explained in the main text. Note the different ranges on the Y axes.

presented here. The coloring of the symbols for the different loci is as follows: mSat (red), Afp (blue), IAP (green), L1 (pink) and Tex13 (black). As already seen before, in all cases, except for the dependence of the activity of Dnmt3a/b on the left neighbor, the dependence parameter is always close to 1, independent of the distance between the CpGs, i.e., the majority of the estimates for the dependence parameters fall into the interval $0.9 < \psi < 1$. Only Dnmt3a/b shows a stronger dependence on the left neighbor, i.e., in most cases $\psi < 0.9$, but no simple relation to the distance is visible. Another observation from Fig. 7c is that the depencency parameters show very similar behaviors within the same locus. However, it is impossible to draw reliable conclusions due to the small sample size within each locus.

3.3 Wild-type Prediction

As a next step we used the estimated parameters from the KO data to predict the WT data. The models from Eqs. (12), (13), (14), (15) are referred to as *Models 1-4*. For the prediction, the notation (x, y) is used to refer to Model x for the Dnmt3a/b DKO (only Dnmt1 active) and Model y for the Dnmt1KO case (only Dnmt3a/b active). One instance of the prediction, for which Model 1 was used for both Dnmt1KO and Dnmt3a/b DKO, i.e., (1,1), are shown in Fig. 8. Note that all wild-type predictions yielded a very similar accuracy. We list the corresponding estimations for the parameters for an example of a single copy gene (Afp) and a repetitive element (L1) in Table 1. While the standard deviation of the estimated parameters for μ is always of the order 10^{-2} and for τ of order 10^{-3}, it is usually of order 10^{-2} for ψ_i. Depending on the model, locus and position, standard deviations up to order 10^{-1} may occur for the dependence parameters in a few cases.

In Fig. 8 the predictions for the pattern distribution together with the WT pattern distribution and a prediction from the neighborhood independent model ($\psi_L = \psi_R = 1$) for all loci are shown in the main plot. As an inset the distributions are shown on a smaller scale to display small deviations. With the exception of patterns 1 and 64 (which corresponds to no methylation/full methylation of all sites) in L1 and pattern 64 in all loci, where the difference between WT and the numerical solution is about 10 percent, the difference is always small ($< 5\%$) as seen in the insets. In order to compare the performance of the neighborhood dependent and

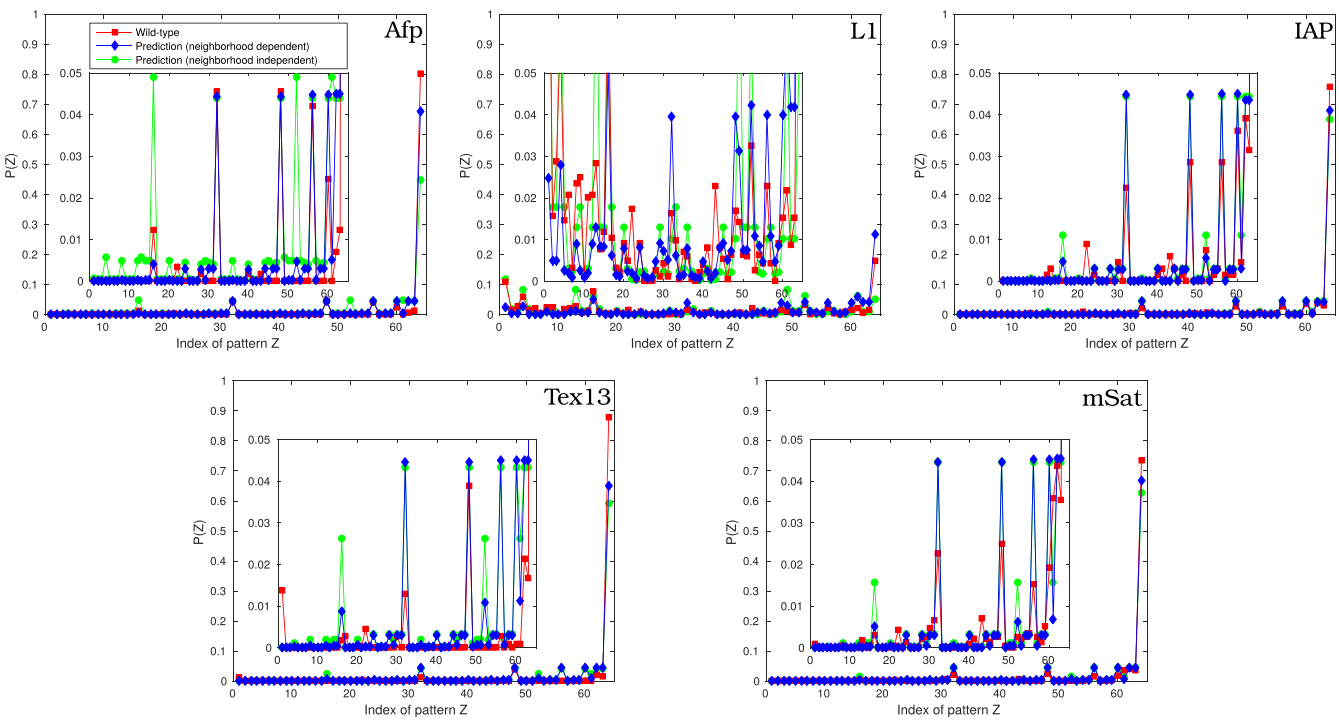

Fig. 8. The figures show an example for the predicted (neighborhood dependent and neighborhood independent) and the measured pattern distribution for each locus. The inset shows a zoomed in version of the distribution.

TABLE 1
Estimated Parameters for the KO Data and Model (1,1) Based on Eq. (12) for the Loci Afp and L1 with Sample Size n

KO	μ	ψ_L	ψ_R	τ	n	Locus
Dnmt1	0.452 ± 0.062	0.383 ± 0.076	1.000 ± 0.094	0.091 ± 0.016	134	Afp
Dnmt3a/b	0.990 ± 0.003	0.984 ± 0.011	1.000 ± 0.006	$10^{-10} \pm 0.011$	186	Afp
Dnmt1	0.334 ± 0.051	0.576 ± 0.067	1.000 ± 0.122	0.038 ± 0.004	1047	L1
Dnmt3a/b	0.789 ± 0.037	1.000 ± 0.038	0.984 ± 0.045	$10^{-10} \pm 0.002$	805	L1

TABLE 2
Kullback-Leibler Divergence KL for the Neighborhood Dependent and Independent Predictions at All Loci

Locus	Afp	L1	IAP	Tex13	mSat
KL_{dep}	0.6820 ± 0.0914	0.5342 ± 0.0638	0.3615 ± 0.0482	1.3364 ± 0.3235	0.1398 ± 0.0134
KL_{ind}	3.3557 ± 0.0979	0.5639 ± 0.0771	0.5390 ± 0.0602	2.0120 ± 0.3637	0.2582 ± 0.0286

neighborhood independent model, we compute the Kullback-Leibler divergence

$$KL = \sum_{j=1}^{4^L} \pi_j(\text{WT}) \log\left(\frac{\pi_j(\text{WT})}{\pi_j(\text{pred})}\right) \tag{23}$$

for both cases and each locus and list the results in Table 2. The mean and standard deviation were obtained via bootstrapping of the wild-type data (10.000 bootstrap samples). The results show that the mean of KL as well as its standard deviation are always smaller for the neighborhood dependent model, i.e., the neighborhood dependent model yields more accurate predictions.

For the 16 proposed models from Eq. (17) we observe a similar performance for all loci and positions in terms of accuracy of the prediction. On the large scale the differences are not visible and even for the smaller scale the differences are small. We therefore only show two examples for mSat in Fig. 9. By comparing KL that we list in Table 3, the similar performance of all 16 models can clearly be seen. The difference in KL between the "best" and the "worst" case is about 0.01. Again, the mean and standard deviation for KL were obtained via bootstrapping of the wild-type data (10.000 bootstrap samples for each model). Since no confidence intervals of the parameters are included, this standard deviation can be regarded as a lower bound. However, even with these lower bounds the intervals of KL overlap for all models, such that no model can be favorized.

3.4 Non-Hairpin Data

So far we restricted the usage of the model to hairpin data, i.e., for one DNA molecule the methylation state of both strands is measured. For non-hairpin data there is only knowledge available for each strand independently. The information which strands stem from the same chromosome is not known. However, it is possible to compute the product of the likelihood of the individual strand patterns, which resembles the likelihood of real hairpin data (assuming independence). Our results show that this approach works well as long as the states of the opposite strand do not determine the transition probabilities, which is the case for Dnmt1KO data, since Dnmt3a/b shows only little maintenance activity. Since Dnmt1's main activity is maintenance, we indeed found that the WT and Dnmt3a/b DKO data does not yield good results (results not shown).

To compare the performance of the model for hairpin and non-hairpin data, we split the original hairpin data in upper and lower strand and computed the product of likelihoods for the patterns using the independence assumption. We then

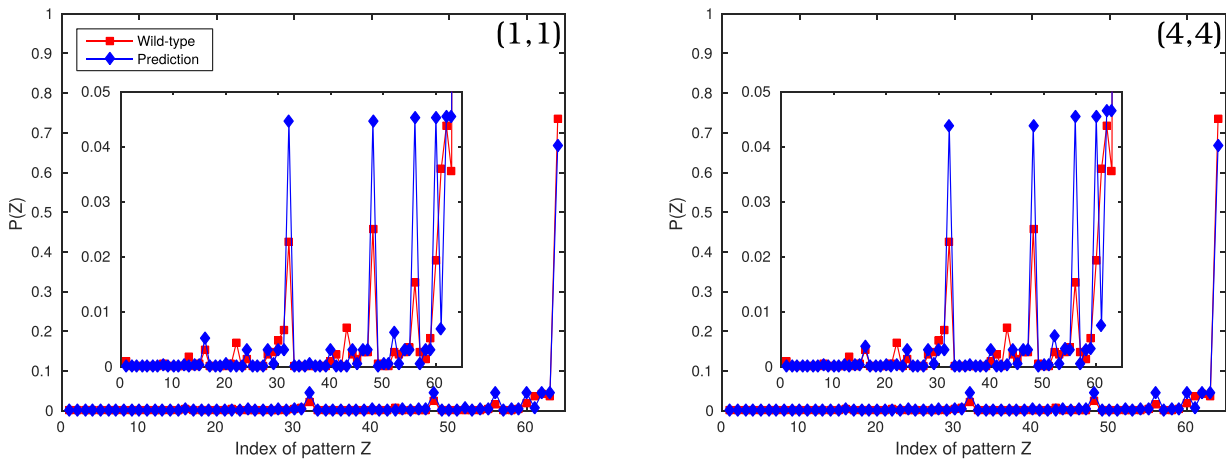

Fig. 9. The figures show the predicted and the measured pattern distribution for two of the 16 models for mSat. The inset shows a zoomed in version of the distribution. The red WT distribution is the same in both plots. Note the slight differences in both predictions for example in pattern 16, 62, and 63.

TABLE 3
Kullback-Leibler Divergence KL for All 16 Models at the Locus mSat

Model	(1,1)	(1,2)	(1,3)	(1,4)
KL	0.1398 ± 0.0134	0.1398 ± 0.0134	0.1398 ± 0.0134	0.1337 ± 0.0127
Model	(2,1)	(2,2)	(2,3)	(2,4)
KL	0.1438 ± 0.0137	0.1439 ± 0.0136	0.1439 ± 0.0137	0.1374 ± 0.0133
Model	(3,1)	(3,2)	(3,3)	(3,4)
KL	0.1399 ± 0.0134	0.1399 ± 0.0134	0.1398 ± 0.0133	0.1337 ± 0.0127
Model	(4,1)	(4,2)	(4,3)	(4,4)
KL	0.1410 ± 0.0137	0.1411 ± 0.0136	0.1409 ± 0.0135	0.1349 ± 0.0130

estimated the parameters via MLE with our model and the computed distributions. We found that for Dnmt3a/b the results are very close to the original hairpin data in terms of dependence parameter ψ_L and ψ_R, since in the model definition these parameters rely only on information on the same strand. No information from the opposite strand influences the dependence parameters. The ratio $R = \mu/\tau$ is usually smaller, i.e., the maintenance is under- and the *de novo* activity overestimated, for the non-hairpin data as shown in Fig. 10. However, this does not lead to contradictory results since maintenance and *de novo* methylation can not be distinguished by the model if the CpG on the opposite strand is methylated.

3.5 Genome-Wide Data

Due to the limited amount of CpGs for the experiments in the previous sections, we also considered genome-wide hairpin data obtained from mouse embryonic stem cells to substantially increase the number of measured CpGs and hence also the number of possible distances between adjacent CpGs. In the genome-wide data the methylation state of the CpGs were recorded in windows of approximately 150 bps for a subset of CpGs, such that there is information available for about 4 million CpGs of the entire genome. The data contains the methylation state of each CpG and the position on the DNA, from which the distance between adjacent CpGs can be derived. For our analysis, we only consider CpGs within the sames read i.e., in the 150 bp window. This last information is of great importance since we want to investigate the neighborhood dependence and have to ensure that the three

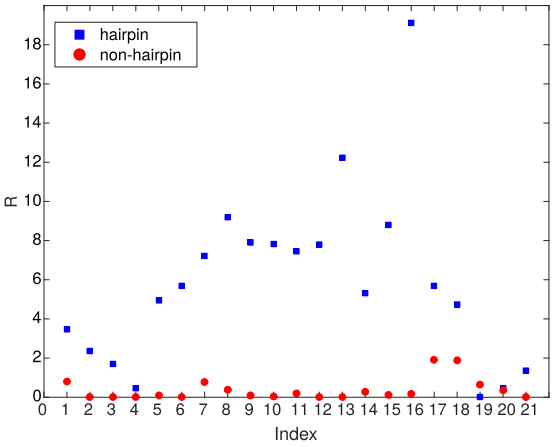

Fig. 10. Ratio $R = \mu/\tau$ between maintenance and *de novo* rate for hairpin (blue) and non-hairpin data (red) for all loci. The loci are mapped to the indices as follows: mSat:1, Afp:2–4, IAP:5–8, L1:9–13, Tex13:14–21.

adjacent CpGs stem from the same DNA molecule. Therefore the data is filtered such that we omit all CpGs which do not form a sequence of at least three consecutive CpGs within one read. Note that we do not consider all cases where either only one or two CpGs were covered in the measurement window or because of missing CpGs the consecutive sequence is split in chunks of two CpGs or smaller. Furthermore we only considered CpG triples for which at least 64 (i.e., the number of possible patterns) measurements were taken. After applying these constrains there are $3{,}489$ CpG triples left.

Since only WT data (and no KO data) was available for the whole genome, we had to use a modified version of the parameter estimation based on Eq. (17), which contains eight parameters (four for each enzyme). In order to reduce the model complexity we use the observations from the previous experiments, namely that only Dnmt3a/b shows a dependence to the left, and we therefore set the remaining dependence parameters $\psi_L^{(1)}$, $\psi_R^{(1)}$ and $\psi_R^{(3a/b)}$ to 1. The conversion errors for the data set are $c = 0.996$ and $d = 0.93$. The conversion rates are derived from short synthetic DNA fragments containing different cytosine forms at definite positions. These oligos become part of the hairpin bisulfite library and therefore undergo the same treatment as the stem cell DNA. Thus, after sequencing, we can determine the conversion rate of C and 5mC independently of our biological sample.

Despite considering only CpG triples with a coverage of at least 64, in general the coverage is pretty low compared to the hairpin data used for the parameter estimation in the previous section. We therefore employ Bayesian inference rather than MLE for the parameter estimation in the genome-wide data. We use a Metropolis Hastings algorithm with the estimations from ML as starting points and a Gaussian proposal distribution with mean 0 and a standard deviation of 0.01 such that on average 40 percent of the 5000 total trials per CpG triplet are accepted for the posterior distribution. Afterwards a variant of the k-means algorithm is applied, which also considers standard deviations of the quantities that should be clustered [15]. Note that in order to avoid a domination by the much larger distances in the clustering, the distance is normalized before the algorithm is applied. The ideal number of clusters is chosen by minimizing the Davies-Bouldin index [5], which is defined as the ratio between cluster separation and similarity within the clusters. The results of the parameter estimation and the clustering is shown in Fig. 11. Note that the clustering is based on dependence parameter and distance only. The methylation state is not an input of the clustering algorithm.

In our results the methylation state of a CpG shows a strong dependence on the methylation state of the left

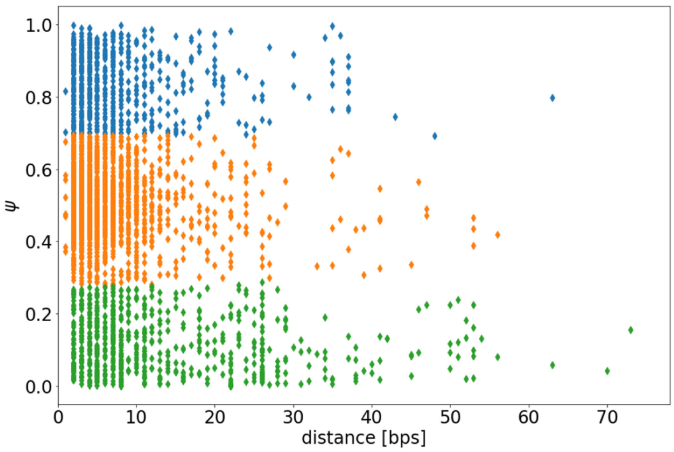

Fig. 11. Dependence parameter versus distance between CpGs for the genome-wide data. The three colors represent three clusters. Cluster 0: blue, cluster 1: orange, and cluster 2: green.

neighbor even for distances up to 70 bps. We therefore conclude that the independence starts at much larger distances. Note that due to the restriction that the three CpGs have to be within the same 150 bps window during the measurement, even for the genome-wide data the distances between the CpGs are rather short. It is therefore not possible with the current data and measurement techniques to check hypotheses such as the independence of neighboring CpGs for large distances. Nevertheless, we see distinct methylation profiles for the three individual clusters as shown in Fig. 12.

The CpGs of these three individual clusters differ also in their genomic localization. Whereas most of the CpGs in cluster 2 are located in introns or intergenic regions, the majority of CpGs in cluster 0 and cluster 1 are found at promoters (Fig. 12a). Fig. 12b shows the genomic localization of CpGs from within (non-)CpG islands (CGIs). We also analyzed the frequencies of the four methylation states of each cluster as displayed in Fig. 12c. CpGs in cluster 0 show low frequencies of fully- or hemimethylated states and in general appear to be unmethylated. Cluster 2 exhibits an inverse behavior compared to cluster 0, meaning that CpGs are more often found in a fully methylated state. Lastly, cluster 1 displays a bimodal distribution of fully- and unmethylated states but similar frequencies in 5mC/C and C/5mC. In other words, unmethylated CpGs seem to show less dependence compared to methylated ones. Fig. 12d shows that CpGs within CGIs tend to behave independently of the left neighbor, however this behavior is not exclusive to CpGs from CGIs since CpGs in non-CGIs can also show an independent behavior. In addition, we conducted an enrichment analysis of transcription factors using the recently developed R package LOLA [25]. We found strong enrichment of cluster 2 CpGs at transcription factor binding sites (TFBS) including Pol2 and Polr2a pointing towards a relation of active transcription (results not shown). Taken together, our findings suggest that hypomethylated CpGs at promoters and TFBS behave more independently. One possible explanation would be the constant setting (most likely by Dnmt3a/b) and removal of CpG methylation at these regions, which would point towards a constant turn over of 5mC. However, a more detailed analysis is needed to address this question.

4 Related Work

In [4] location- and neighbor-dependent models are proposed for single-stranded DNA methylation data in blood and tumor cells. The (de-)methylation rates depend on the position of the CpG relative to the 3′ or 5′ end and/or on the methylation state of the left neighbor only. The dependence is realized by the introduction of an additional parameter. In our proposed models we use double-stranded DNA and can therefore include hemi-methylated sites and even distinguish on which strand the site is methylated. Furthermore we allow dependences on both neighbors by introducing two different dependence parameters. In contrast [7] copes with the neighborhood

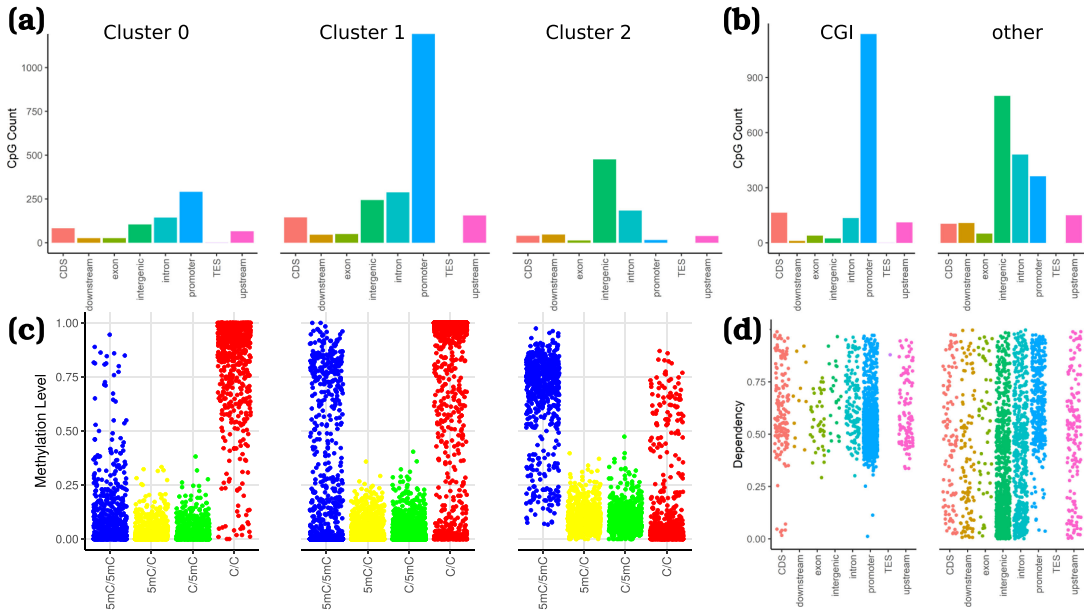

Fig. 12. Biological context of CpG clustering. Counts of annotated genomic features within the individual clusters (a) and within (non-)CGIs (b). Frequency CpG methylation state (c); states are indicated as follows: state 0 = C/C: red, state 1 = 5mC/C: yellow, state 2 = C/5mC: green, and state 3 = 5mC/5mC: blue. Dependence on left neighbor in (non-)CGIs (d).

dependence indirectly by allowing different parameter values for different sites. In order to reduce the dimensionality of the parameter vector, a hierarchical model based on beta distributions is proposed. Another difference to our model is the distinction between *de novo* rates for parent and daughter strand. However, this can easily be included in future work. A density-dependent Markov model was proposed [18]. In this model, the probabilities of (de-)methylation events may depend on the methylation density in the CpG neighborhood. In addition, a neighboring sites model has been developed, in which the probabilities for a given site are directly influenced by the states of neighboring sites to the left and right [18]. When these models were tested on double-stranded methylation patterns from two distinct tandem repeat regions in a collection of ovarian carcinomas, the density-dependent and neighboring sites models were superior to independent models in generating statistically similar samples. Although this model also includes the dependence on the methylation state on the left and right neighbor for double-stranded DNA the approach is different. The transition probabilities of the neighbor-independent model are transformed into a transition probability of a neighbor-dependent model by introducing only one additional parameter. The state of the left and right neighbor are taken into account by exponentiating this parameter by some norm. In addition, this approach does not allow the intuitive interpretation of the dependence parameter. Recently the model from [18] was extended to include the influence of different distances between the CpGs [21]. However this model is still restricted to single-stranded methylation data. In [11] it has been shown that the collaboration between CpG sites is required to obtain stable fractions of methylation states over time in CpG islands. In this model another nearby CpG serves as a mediator such that its state influences the possible reactions. In a more recent version of this model the distance to the mediator CpG is taken into account [20]. However, both models feature active demethylation, have no explicit dependence parameter and do not distinguish between the two different hemimethylated states.

5 Conclusion

We proposed a set of stochastic models for the formation and modification of methylation patterns over time. These models take into account the state of the CpG sites in the spatial neighborhood and allow to describe different hypotheses about the underlying mechanisms of methyltransferases adding methyl groups at CpG sites. We used knockout data from bisulfite sequencing at several loci to learn the efficiencies at which these enzymes perform methylation. By combining these efficiencies, we accurately predicted the probability distribution of the patterns in the wild-type. Moreover, we found that in all cases the models predict values for the dependence parameters ψ_L and ψ_R close to 1 and therefore independence of methylation for the Dnmt3a/b DKO meaning that Dnmt1 methylates CpGs independent of the methylation of neighboring CpGs. For Dnmt3a/b on the other hand we could identify dependences on the neighboring CpGs. Both findings are in accordance with current existing mechanistic models: Dnmt1 reliably copies the methylation from the template strand to maintain the distinct methylation patterns, whereas Dnmt3a/b try to establish and keep a certain amount of CpG methylation at a given loci. Interestingly, our models only suggest dependences of *de novo* methylation activity on the CpGs in the 5′ neighborhood. This indicates that Dnmt3a and Dnmt3b show a preference to methylate CpGs in a 5′ to 3′ direction and could point towards a processive or cooperative behavior of these enzymes like recently described in *in vitro* experiments [6], [13]. Our results indicate that, at least for small distances, rather the genetic region than the distance determines the dependence on the neighbors. Compared to a neighborhood independent model with $\psi_L = \psi_R = 1$, a neighborhood dependent model shows better predictions and furthermore allows to investigate (possible) connections of adjacent CpGs and their methylation states. As long as no information from the opposite strand is needed, i.e., if maintenance activity is not too high, as in the Dnmt1KO data, our model can also be used for non-hairpin data. Applying our model at genome-wide data reveals distinct dependence clusters with individual methylation patterns. We find, that hypomethylated CpGs at promoter and TFBS are more likely to behave independent of their neighborhood compared to hypermethylated CpGs.

As future work, we plan to investigate models in which we distinguish between the actions of Dnmt3a and Dnmt3b and in which we allow a diagonal dependence for *de novo* methylation, i.e., a dependence on the state of neighboring CpGs on the opposite strand. Furthermore, we intend to explicitely include the actual distance of neighboring CpGs in our model by making the dependence parameters distance dependent. This also eases the modelling of more than three CpGs since we then do not longer assume the same dependence parameters for all CpGs and therefore make the model more flexible. To investigate a potential impact of oxidized cytosine forms on the methylation at neighboring CpG sites we further plan to include the CpG states 5hmC, 5fC and 5caC in our model and use a hybrid approach as presented in [17] in order to omit the necessity of specifying the order of certain events a priori.

Acknowledgments

This research has been partially funded by the German Research Council (DFG) as part of the Cluster of Excellence on Multimodal Computing and Interaction and as part of the Collaborative Research Centre (SFB) 1309 'Chemical Biology of Epigenetic Modifications'. KN was supported by the de.NBI grant (031L0101D) from the Federal Ministry of Education and Research in Germany. Conceived and designed the experiments: VW and JW. Performed the experiments: PG. Analyzed the data: AL, PG, and KN. Wrote the paper: AL and PG. Designed/implemented the software used in analysis: AL. Alexander Lück and Pascal Giehr contributed equally to this work.

References

[1] T. Äijö, Y. Huang, H. Mannerström, L. Chavez, A. Tsagaratou, A. Rao, and H. Lähdesmäki, "A probabilistic generative model for quantification of DNA modifications enables analysis of demethylation pathways," *Genome Biol.*, vol. 17, no. 1, 2016, Art. no. 49.

[2] J. Arand, D. Spieler, T. Karius, M. R. Branco, D. Meilinger, A. Meissner, T. Jenuwein, G. Xu, H. Leonhardt, V. Wolf, et al., "In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases," *PLoS Genetics.*, vol. 8, no. 6, 2012, Art. no. e1002750.

[3] T. Baubec, D. F. Colombo, C. Wirbelauer, J. Schmidt, L. Burger, A. R. Krebs, A. Akalin, and D. Schübeler, "Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation," *Nature* vol. 520, no. 7546, pp. 243–247, 2015.
[4] N. Bonello, J. Sampson, J. Burn, I. J. Wilson, G. McGrown, G. P. Margison, M. Thorncroft, P. Crossbie, A. C. Povey, M. Santibanez-Koref, et al., "Bayesian inference supports a location and neighbour-dependent model of DNA methylation propagation at the MGMT gene promoter in lung tumours," *J. Theoretical Biol.*, vol. 336, pp. 87–95, 2013.
[5] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.
[6] M. Emperle, A. Rajavelu, R. Reinhardt, R. Z. Jurkowska, and A. Jeltsch, "Cooperative DNA binding and protein/DNA fiber formation increases the activity of the Dnmt3a DNA methyltransferase," *J. Biol. Chemistry*, vol. 289, no. 43, pp. 29602–29613, 2014.
[7] A. Q. Fu, D. P. Genereux, R. Stöger, C. D. Laird, and M. Stephens, "Statistical inference of transmission fidelity of DNA methylation patterns over somatic cell divisions in mammals," *Ann. Appl. Statist*, vol. 4, no. 2, 2010, Art. no. 871.
[8] D. P. Genereux, B. E. Miner, C. T. Bergstrom, and C. D. Laird, "A population-epigenetic model to infer site-specific methylation rates from double-stranded DNA methylation patterns," *Proc. Nat. Acad. Sci. United States America*, vol. 102, no. 16, pp. 5802–5807, 2005.
[9] P. Giehr, C. Kyriakopoulos, G. Ficz, V. Wolf, and J. Walter, "The influence of hydroxylation on maintaining CpG methylation patterns: A Hidden Markov model approach," *PLoS Comput. Biol.*, vol. 12, no. 5, 2016, Art. no. e1004905.
[10] H. Gowher and A. Jeltsch, "Molecular enzymology of the catalytic domains of the Dnmt3a and Dnmt3b DNA methyltransferases," *J. Biol. Chemistry*, vol. 277, no. 23, pp. 20409–20414, 2002.
[11] J. O. Haerter, C. Lövkvist, I. B. Dodd, and K. Sneppen, "Collaboration between CpG sites is needed for stable somatic inheritance of DNA methylation states," *Nucleic Acids Res.*, vol. 42, no. 4, pp. 2235–2244, 2013.
[12] A. Hermann, R. Goyal, and A. Jeltsch, "The Dnmt1 DNA-(cytosine-c5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites," *J. Biol. Chemistry*, vol. 279, no. 46, pp. 48350–48359, 2004.
[13] C. Holz-Schietinger and N. O. Reich, "The inherent processivity of the human de novo methyltransferase 3A (DNMT3A) is enhanced by DNMT3L," *J. Biol. Chemistry*, vol. 285, no. 38, pp. 29091–29100, 2010.
[14] C. A. Kapourani and G. Sanguinetti, "Higher order methylation features for clustering and prediction in epigenomic studies," *Bioinf.*, vol. 32, no. 17, pp. i405–i412 2016.
[15] M. Kumar and N. R. Patel, "Clustering data with measurement errors," *Comput. Statist. Data Anal.*, vol. 51, no. 12, pp. 6084–6101, 2007.
[16] C. Kyriakopoulos, P. Giehr, and V. Wolf, "H(O)TA: Estimation of DNA methylation and hydroxylation levels and efficiencies from time course data," *Bioinf.*, vol. 33, no. 11, pp. 1733–1734, 2017.
[17] C. Kyriakopoulos, P. Giehr, A. Lück, J. Walter, and V. Wolf, "A hybrid HMM approach for the dynamics of DNA methylation," in *Proc. Int. Workshop Hybrid Syst. Biol.*, 2019.
[18] M. R. Lacey, M. Ehrlich, et al., "Modeling dependence in methylation patterns with application to ovarian carcinomas," *Statistical Appl. Genetics Molecular Biol.*, vol. 8, no. 1, 2009, Art. no. 40.
[19] C. D. Laird, N. D. Pleasant, A. D. Clark, J. L. Sneeden, K. A. Hassan, N. C. Manley, J. C. Vary, T. Morgan, R. S. Hansen, and R. Stöger, "Hairpin-bisulfite PCR: Assessing epigenetic methylation patterns on complementary strands of individual DNA molecules," *Proc. Nat. Acad. Sci. United States America*, vol. 101, no. 1, pp. 204–209, 2004.
[20] C. Lövkvist, I. B. Dodd, K. Sneppen, and J. O. Haerter, "DNA methylation in human epigenomes depends on local topology of CpG sites," *Nucleic Acids Res.*, vol. 44, no. 11, pp. 5123–5132, 2016.
[21] K. N. Meyer and M. Lacey, "Modeling methylation patterns with long read sequencing data," *IEEE ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 4, pp. 1379–1389, Jul./Aug. 2018.
[22] A. B. Norvil, C. J. Petell, L. Alabdi, L. Wu, S. Rossie, and H. Gowher, "Dnmt3b methylates DNA by a noncooperative mechanism, and its activity is unaffected by manipulations at the predicted dimer interface," *Biochemistry*, vol. 57, pp. 4312–4324, 2016.
[23] M. Okano, D. W. Bell, D. A. Haber, and E. Li, "DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development," *Cell*, vol. 99, no. 3, pp. 247–257, 1999.
[24] S. P. Otto and V. Walbot, "DNA methylation in eukaryotes: Kinetics of demethylation and de novo methylation during the life cycle," *Genetics*, vol. 124, no. 2, pp. 429–437, 1990.
[25] N. C. Sheffield and C. Bock, "LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor," *Bioinf.*, vol. 32, no. 4, pp. 587–589, 2015.
[26] L. B. Sontag, M. C. Lorincz, and E. G. Luebeck, "Dynamics, stability and inheritance of somatic DNA methylation imprints," *J. Theoretical Biol*, vol. 242, no. 4, pp. 890–899, 2006.
[27] M. M. Suzuki and A. Bird, "DNA methylation landscapes: Provocative insights from epigenomics," *Nat. Rev. Genetics*, vol. 9, no. 6, pp. 465–476, 2008.

Alexander Lück received the BSc and MSc degrees in physics, both from Saarland University, in 2012 and 2014, respectively. Currently, he is working toward the PhD degree in the Modelling and Simulation Group, Saarland University.

Pascal Giehr received the diploma degree in biology from Saarland University, in 2012. Currently, he is working toward the PhD degree in the Department of Genetics and Epigenetics, Saarland University.

Karl Nordström received the MSc degree in applied physics from the Chalmers University of Technology, in 2006 and the PhD degree from Uppsala University, in 2010. He is currently a postdoc in the de.NBI project (BMBF, grant 031L0101D) at the Department of Genetics and Epigenetics, Saarland University.

Jörn Walter received the diploma degree in biology, in 1986 and the PhD degree from the Free University in Berlin, in 1990. He holds the chair of the Department of Genetics and Epigenetics at Saarland University, focusing on the role of DNA-methylation in development and disease. He contributed to more than 100 research papers in epigenetics and coordinated several large epigenetics and epigenomics research activities such as the German Epigenome Program DEEP (2012-2017). Currently, he is elected spokesman of the reviewing panel "life science 2" of the Deutsche Forschungsgemeinschaft (DFG) and co-chair of the Scientific Steering Board of the International Human Epigenome Consortium IHEC. More information can be found at http://epigenetik.uni-saarland.de/de/epigenetics/.

Verena Wolf received the diploma degree in computer science from the University Bonn, in 2003 and the PhD degree from the University Mannheim, in 2008. She is a full professor with Saarland University since 2012 and leads the Modelling and Simulation Group at the Department of Computer Science. She is currently working on discrete stochastic modelling as well as efficient simulation methods and has been on the program committees of more than 50 international conferences. More information about her can be found at https://mosi.uni-saarland.de.