bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cluster mean-field theory accurately predicts statistical properties of large-scale DNA methylation patterns

Lyndsay Kerr*[1], Duncan Sproul†[2], and Ramon Grima‡[3]

[1]MRC Institute of Genetics and Cancer, University of Edinburgh, UK
[2]MRC Human Genetics Unit and CRUK Edinburgh Centre, Institute of Genetics and Cancer, University of Edinburgh, UK
[3]School of Biological Sciences, University of Edinburgh, UK

Abstract

The accurate establishment and maintenance of DNA methylation patterns is vital for mammalian development and disruption to these processes causes human disease. Our understanding of DNA methylation mechanisms has been facilitated by mathematical modelling, particularly stochastic simulations. Mega-base scale variation in DNA methylation patterns is observed in development, cancer and ageing and the mechanisms generating these patterns are little understood. However, the computational cost of stochastic simulations prevents them from modelling such large genomic regions. Here we test the utility of three different mean-field models to predict large-scale DNA methylation patterns. By comparison to stochastic simulations, we show that a cluster mean-field model accurately predicts the statistical properties of steady-state DNA methylation patterns, including the mean and variance of methylation levels calculated across a system of CpG sites, as well as the covariance and correlation of methylation levels between neighbouring sites. We also demonstrate that a cluster mean-field model can be used within an approximate Bayesian computation framework to accurately infer model parameters from data. As mean-field models can be solved numerically in a few seconds, our work demonstrates their utility for understanding the processes underpinning large-scale DNA methylation patterns.

1 Introduction

DNA methylation is a repressive epigenetic mark [1] which is primarily found on the cytosines of CpG dinucleotides in mammals. Double-stranded CpG dyads can be unmethylated or methylated on both strands (u and m respectively) or methylated on only one strand (hemimethylated, h). DNA methylation is largely erased from the genome during early mammalian development [2]. It is then re-established by the *de novo* DNA methyltransferases DNMT3A and DNMT3B [3] resulting in a landscape where 70-80% of CpGs are methylated in most human cells [4]. Regulatory elements such as promoters and enhancers

*lyndsay.kerr@ed.ac.uk
†d.sproul@ed.ac.uk
‡ramon.grima@ed.ac.uk

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

often remain methylation free [1]. During DNA replication, the nascent strand is synthesised with unmethylated cytosines and methylation patterns are copied by the maintenance methyltransferase, DNMT1 [5]. Failure to maintain DNA methylation at a locus results in passive DNA demethylation. Methylation can also be removed actively through transient modification by Ten Eleven Translocation (TET) enzymes and subsequent DNA repair [6].

Waves of demethylation and remethylation take place during early development and the generation of germline cells [2]. Changes in DNA methylation patterns also occur during development and cellular differentiation, resulting in cell type specific methylation patterns [7]. The correct establishment of DNA methylation patterns is vital for normal development. Mutations in DNMTs cause Mendelian disorders in humans [8, 9, 10] and mice knockouts die before or shortly after birth [3, 5]. Widespread alterations in DNA methylation patterns occur in cancer and ageing [11, 12], but the significance of these changes is unclear. It has been observed that globally hypomethylated mice expressing a single hypomorphic DNMT1 allele develop cancer, suggesting that altered DNA methylation can cause cancer [13]. However, the mechanisms underpinning DNA methylation changes remain unclear preventing the robust delineation of their role in development and disease.

Mathematical models are powerful tools for understanding complex biological processes, including DNA methylation. The importance of interactions between CpGs in maintaining DNA methylation patterns was first postulated through modelling [14]. Specifically, the authors modelled collaborative interactions where CpGs within a region of the genome can influence the state of other CpGs, e.g. through enzyme recruitment. Models including such collaborativity were subsequently found to explain experimental measurements of methylation maintenance *in vitro* and *in vivo* more closely than those that did not include it [15, 16]. A recent study also suggests that collaborativity mediated by neighbour-guided error correction through DNMT1 is important for maintaining DNA methylation [17]. Deterministic models, non-spatial stochastic models and spatial stochastic models have all been used to describe DNA methylation [18]. Deterministic models are based on rate equations while stochastic models are based on Fokker-Planck equations or chemical master equations (CMEs). CMEs are ideal because they take into account the inherent discreteness of molecular fluctuations [19] which is well known to play an important role in cellular dynamics [20]. The CME of simple non-spatial stochastic models can be solved exactly in closed-form [21], but this is often not possible for spatial stochastic models. Rather in this case, stochastic simulations are used to model the individual reaction processes described by the CME. Various types of stochastic models have been used to describe collaborative methylation systems (see for example [22, 23, 24, 25, 26]). To date, such mathematical models have been applied to understand methylation patterns on a kilo-base scale. However, mega-base scale alterations to DNA methylation patterns occur in development, cancer and ageing [27]. Existing models rely on simulations that are too computationally expensive to run for such large genomic regions.

Here we test the idea that large-scale steady-state methylation patterns can be modelled in a tractable manner using mean-field (MF) models. By comparison to synthetic data generated from stochastic simulations, we demonstrate that a type of cluster mean-field model can predict the statistical properties of large-scale methylation patterns. In Section 2 we introduce a nearest-neighbour collaborative model for DNA methylation and describe the process used to simulate data from this model. We describe the three MF models

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

we test in Section 3. In Section 4 we compare the ability of each MF model to predict statistics associated with methylation patterns resulting from the simulations. We find that a type of cluster mean-field model provides excellent predictions and demonstrate that this model can be used within an Approximate Bayesian Computation (ABC) framework to infer parameters underpinning large-scale methylation systems. Finally, in Section 5, we discuss the implications of our findings.

2 Nearest-neighbour collaborative model

2.1 Model set-up

We consider the reaction system in Fig. 1, where some reactions are non-collaborative (involving only the "target" CpG, whose methylation state changes during the reaction), while others are collaborative (involving both a target CpG and a "mediator" CpG). The role of the mediator is to encourage the reaction to occur, e.g. via the recruitment of methylase or demethylase enzymes. This system, and reduced versions, have previously been used to examine small-scale methylation patterns [14, 28, 29].

Non-collaborative reactions:

$$u \xrightarrow{k_1} h, \qquad h \xrightarrow{k_2} m, \qquad m \xrightarrow{k_3} h, \qquad h \xrightarrow{k_4} u.$$

Collaborative reactions:

$$u + h \xrightarrow{k_5} h + h, \qquad u + m \xrightarrow{k_6} h + m, \qquad h + h \xrightarrow{k_7} m + h, \qquad h + m \xrightarrow{k_8} m + m,$$
$$m + h \xrightarrow{k_9} h + h, \qquad m + u \xrightarrow{k_{10}} h + u, \qquad h + h \xrightarrow{k_{11}} u + h, \qquad h + u \xrightarrow{k_{12}} u + u.$$

Figure 1: System of reactions under consideration. Here u, h and m represent unmethylated, hemimethylated and methylated CpGs, respectively. Non-collaborative reactions involve only one CpG, while collaborative reactions involve two CpGs. For each collaborative reaction, the second reactant (the mediator—see text) recruits an enzyme that changes the methylation state of the first reactant (the target). For example, the reaction $u + h \to h + h$ involves a hemimethylated CpG at one site recruiting an enzyme which changes the state of a CpG at another site from unmethylated to hemimethylated. Reaction rates are k_i, $i = \{1, \ldots, 12\}$.

We make the following assumptions:

(i) *CpGs can only influence the methylation state of their nearest neighbours; see Fig. 2.* While both experimental and modelling studies have demonstrated the importance of CpGs being influenced by surrounding CpGs [14, 15], the extent and range of such influence is unknown. We therefore consider only interactions between nearest neighbours.

(ii) *There are no direct transitions between the unmethylated and methylated states.* We justify this with the observation that methylase and demethylase enzymes act on single DNA strands [6, 30]. This implies that hemimethylation is a necessary transition state between unmethylated and fully methylated CpGs.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(iii) *The system has reached a steady state.* Here, we assume that there are no long-term effects of DNA replication on methylation patterns. This assumption is supported by the observation that the DNA methylation patterns of cycling and arrested cells are similar [31].

We also assume that the rates k_i, $i = \{1, 2, \ldots, 12\}$ are of the form given in Table 1, where x measures the strength of collaborativity between CpGs ($x < 1$ indicates that non-collaborative reactions dominate, while $x > 1$ indicates that collaborative reactions dominate). The parameter y measures the strength of methylation vs. demethylation ($y < 1$ corresponds to demethylation dominating and $y > 1$ corresponds to methylation dominating). The parameter a scales the reaction rates.

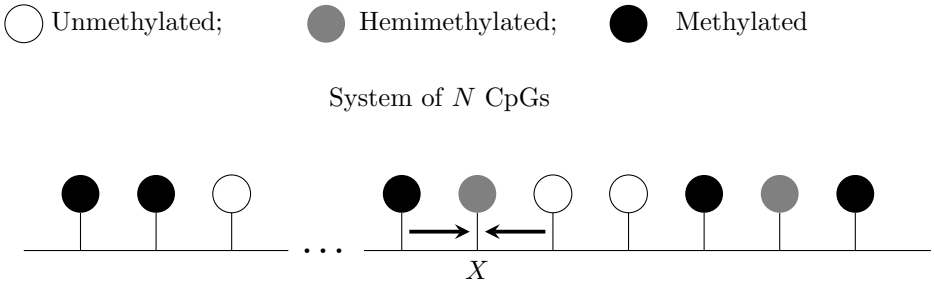

Figure 2: Collaborative interactions that can influence a target X under the nearest-neighbour collaborative model. Individual CpGs are represented by "lollipops", with their colour indicating their methylation status (white: unmethylated, grey: hemimethylated and black: methylated). Collaborative methylation and demethylation reactions can only occur between neighbouring CpGs (potential influences on CpG X are indicated by arrows). There is no upper bound on $N \in \mathbb{N}$, allowing large-scale methylation patterns to be considered.

	Demethylation	**Methylation**
Non-Collaborative	$k_3 = k_4 = a$	$k_1 = k_2 = ay$
Collaborative	$k_9 = k_{10} = k_{11} = k_{12} = ax$	$k_5 = k_6 = k_7 = k_8 = axy$

Table 1: Reaction rates for the model in Fig. 1.

2.2 Simulations of nearest-neighbour collaborative system

In the simple case of two CpGs, the system can be in six possible states: mm, uu, hh, um, hm and uh. For such a small system, all possible transitions between states (via reactions in Fig. 1) can be identified and the evolution of the system can be described exactly by six mathematical equations, one for each state. However, for large systems, it is infeasible to identify all possible states and transitions between states meaning that equations describing the exact evolution of the system can not be formulated. While stochastic simulations are computationally expensive, they are the ground truth of the nearest-neighbour collaborative system to which we compare our MF models and so here we describe the process underlying these simulations.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We focus on the steady-state case so that u, h and m levels fluctuate around some fixed steady-state values. For a system of N CpGs, we simulate the nearest-neighbour collaborative system using Gillespie's algorithm [32]—see Fig. 3 for an illustration of how this algorithm simulates a 4-CpG system. A sample taken at any time point will contain N methylation states. For each parameter set, we therefore sample the system at a total of $T = 10^6/N$ time points after steady state has been reached to obtain a dataset of 10^6 steady-state methylation states.

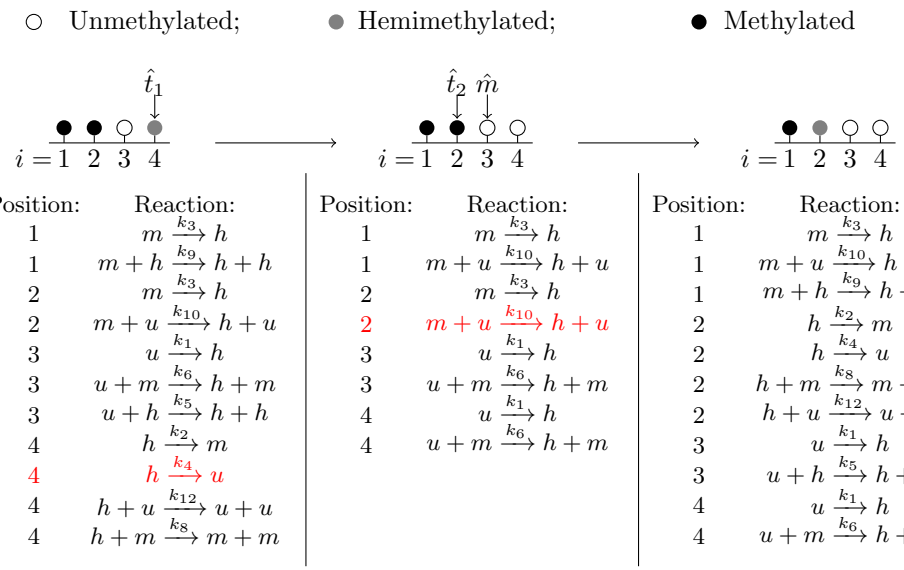

Figure 3: Stochastic simulations of the nearest-neighbour collaborative system. For simplicity, a 4-CpG system is considered here. We fix a, x and y and impose periodic boundary conditions so that the first and final CpG can interact. All potential reactions are first identified, and the Gillespie algorithm is used to choose one of these reactions, and a time for it to occur. The system is updated to account for the reaction occurring and the process is repeated to generate dynamical behaviour. In the example shown, we start with the system on the left. All possible reactions are listed and a $h \xrightarrow{k_4} u$ reaction is chosen to occur at target position $\hat{t}_1$. The system and the list of potential reactions are then updated (middle). Subsequently, another reaction, $m + u \xrightarrow{k_{10}} h + u$, is chosen to occur at target position, $\hat{t}_2$, with CpG $\hat{m}$ acting as mediator. The system and the list of possible reactions is again updated (right). This process is repeated until the system reaches steady state.

2.3 Analysis of simulated data

To facilitate our analysis, we define a sequence u^t, $t \in \{1, \ldots, T\}$, where

$$u_i^t = \begin{cases} 1, & \text{if CpG } i \text{ is in the } u \text{ state at timepoint } t, \\ 0, & \text{otherwise.} \end{cases}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We define h^t and m^t similarly; see Fig.4. We also define z^t via

$$z^t = u^t + 2h^t + 3m^t. \tag{1}$$

The mean u, h and m levels, μ_{us}, μ_{hs}, μ_{ms}, are obtained via

$$\mu_{us} = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} u_i^t, \qquad \mu_{hs} = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} h_i^t, \qquad \mu_{ms} = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} m_i^t. \tag{2}$$

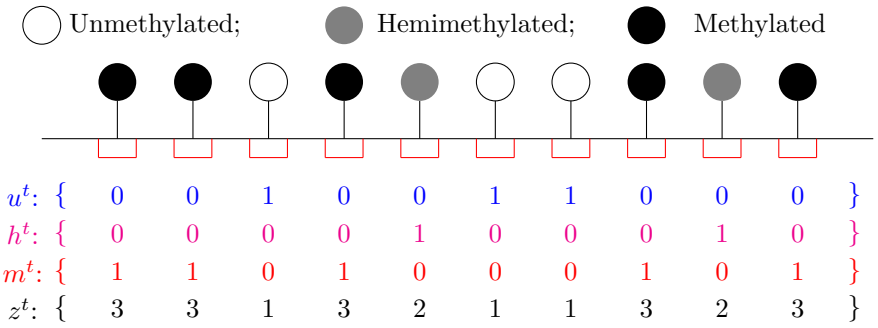

Figure 4: Construction of the sequences u^t, h^t, m^t and z^t. For simplicity, only ten CpGs are shown for a single time point, t. A vector u^t is created, where $u_i^t = 1$ if CpG i is unmethylated at time t and $u_i^t = 0$ otherwise. Vectors h^t and m^t are constructed similarly. Finally, a vector z^t is created via $z^t = u^t + 2h^t + 3m^t$.

For each t, we also calculate the mean and variance over z^t (μ_z^t, $\sigma^2(z^t)$, respectively) and average over all $t \in \{1, \ldots, T\}$ to obtain an overall mean and variance, μ_z and $\sigma^2(z)$, which are given by

$$\mu_z = \frac{1}{T}\sum_{t=1}^{T} \mu_z^t = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} z_i^t, \tag{3}$$

$$\sigma^2(z) = \frac{1}{T}\sum_{t=1}^{T} \sigma^2(z^t) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N-1} \frac{(z_i^t - \mu_z^t)^2}{N-2}. \tag{4}$$

We then define sequences v^t and w^t by

$$v^t = \{z_1^t, z_2^t, \ldots, z_{N-1}^t, z_N^t\}, \qquad w^t = \{z_2^t, z_3^t, \ldots, z_N^t, z_1^t\}, \qquad t \in \{1, \ldots, T\}. \tag{5}$$

For each $t \in \{1, \ldots, T\}$, the sequence v^t is identical to z^t, while w^t is a shifted version of z^t (i.e. $w_i^t = z_{i+1}^t$ for $i = \{1, \ldots N-1\}$ and $w_N^t = z_1^t$). For each $i = \{1, \ldots, N\}$, comparing v_i^t and w_i^t provides information regarding the methylation state of two neighbouring CpGs at time $t \in \{1, \ldots, T\}$. We calculate the covariance between v^t and w^t for each t, Covar(v^t, w^t), averaging over $t \in \{1, \ldots, T\}$ to obtain an overall covariance between neighbouring sites

$$\text{Covar}(z) \coloneqq \text{Covar}(v, w) = \frac{1}{T}\sum_{t=1}^{T} \text{Covar}(v^t, w^t) = \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} (v_i^t - \mu_v^t)(w_i^t - \mu_w^t), \tag{6}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where $\mu_v^t = \mu_w^t = \mu_z^t$. Finally, the correlation between neighbouring sites, $\rho(z) \coloneqq \rho(v, w)$, is calculated via

$$\rho(z) \coloneqq \rho(v, w) = \frac{\text{Covar}(v, w)}{\sqrt{\sigma^2(v)\sigma^2(w)}} = \frac{\text{Covar}(v, w)}{\sigma^2(z)}. \tag{7}$$

3 Mean-field models for DNA methylation maintenance

For large CpG systems, the CME describing the nearest-neighbour collaborative model cannot be easily solved and stochastic simulations are computationally expensive. In contrast, it is often the case that models simplified using the MF approximation can be computationally solved in a time-efficient manner and we aim to test whether they accurately approximate the nearest-neighbour collaborative model for DNA methylation. To this end, we construct three MF models (see Sections 3.1, 3.2, 3.3). These models consider an infinite system of CpGs and so, by design, their ability to accurately describe a genomic region increases with the size of the region. In these models, nearest-neighbour interactions are approximated by considering the mean state of the system. In the first model, nearest-neighbour interactions are entirely approximated by considering the probability that two states are adjacent (one-site MF model). The second model describes distinct pairs of CpGs (distinct pairs MF model). Interactions occurring within a pair are directly accounted for and other nearest-neighbour interactions are approximated by considering the probability that two paired states are adjacent. In the third model we consider overlapping pairs of CpGs (overlapping pairs MF model). Interactions occurring within a pair are again directly accounted for, but now other nearest-neighbour interactions are approximated by considering the probability that two paired states overlap. The remainder of this section is devoted to mathematical descriptions of these models.

3.1 One-site mean-field model

We define the proportion of sites in the u, h, m states to be the mean u, h, m levels, μ_u, μ_h, μ_m, respectively. Here we construct a one-site MF model, where changes in the system are influenced by μ_u, μ_h, μ_m, rather than nearest-neighbour interactions; see Fig. 5.

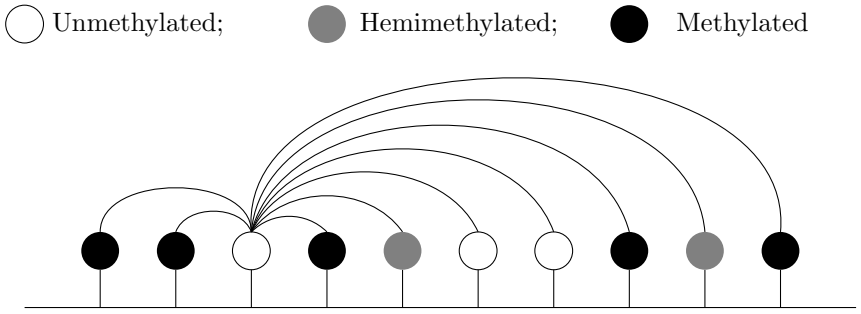

Figure 5: Schematic of the one-site MF model. CpGs are influenced by the mean of the system rather than nearest-neighbour interactions.

Consider the reaction $u + h \xrightarrow{k_5} h + h$ in Fig. 1. Since the h mediator is unchanged by the reaction, we can write this as an effective first-order reaction $u \xrightarrow{2k_5\mu_h} h$, where the h mediator

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

is absorbed into the effective reaction rate by making it proportional to μ_h. The factor of two accounts for the h mediator being on either side of the u target. Similarly, $u + m \xrightarrow{k_6} h + m$ can be written as $u \xrightarrow{2k_6\mu_m} h$. Thus the $u \xrightarrow{k_1} h$, $u + h \xrightarrow{k_5} h + h$, $u + m \xrightarrow{k_6} h + m$ reactions in Fig. 1 can be written as a single effective first-order reaction $u \xrightarrow{r} h$, with $r = k_1 + 2k_5\mu_h + 2k_6\mu_m$. We thus write the system in Fig. 1 as the effective first-order system

$$m \xrightarrow{a_1} h, \qquad h \xrightarrow{a_2} m, \qquad u \xrightarrow{a_3} h, \qquad h \xrightarrow{a_4} u, \tag{8}$$

where

$$\begin{aligned} a_1 &\coloneqq a_1(\mu_u, \mu_h) = k_3 + 2k_9\mu_h + 2k_{10}\mu_u, \\ a_2 &\coloneqq a_2(\mu_h, \mu_m) = k_2 + 4k_7\mu_h + 2k_8\mu_m, \\ a_3 &\coloneqq a_3(\mu_h, \mu_m) = k_1 + 2k_5\mu_h + 2k_6\mu_m, \\ a_4 &\coloneqq a_4(\mu_u, \mu_h) = k_4 + 4k_{11}\mu_h + 2k_{12}\mu_u, \end{aligned}$$

and $\mu_u + \mu_h + \mu_m = 1$. The k_7 and k_{11} terms have an additional factor of two since their associated reactions involve two h reactants, and either of these can change state during the reaction.

Let $L_u,\ L_h,\ L_m$ be the "level" (proportion) of $u,\ h,\ m$ at a single CpG, respectively. A CpG can only be in one state at any time and so

$$(L_u, L_h, L_m) = (1, 0, 0) \quad \text{or} \quad (L_u, L_h, L_m) = (0, 1, 0) \quad \text{or} \quad (L_u, L_h, L_m) = (0, 0, 1), \tag{9}$$

for each CpG. Using (8) we construct a CME describing the probability that a site is in the $u,\ h$ or m state and from this we can obtain moment equations (see Appendix C of Ref [33]), for the statistics of $L_u,\ L_h,\ L_m$. The mean values are $\mu_u = \langle L_u \rangle,\ \mu_h = \langle L_h \rangle,\ \mu_m = \langle L_m \rangle$, where the angled brackets denote the expected value. Due to the conservation law $L_m = 1 - L_u - L_h$, we need only consider equations for u and h. The means are described by the equations,

$$\frac{d\mu_u}{dt} = -a_3\mu_u + a_4\mu_h, \qquad \frac{d\mu_h}{dt} = a_1(1 - \mu_u - \mu_h) - a_2\mu_h + a_3\mu_u - a_4\mu_h. \tag{10}$$

Setting

$$\frac{d\mu_u}{dt} = \frac{d\mu_h}{dt} = 0$$

leads to implicit equations for the steady-state means,

$$\mu_{us} = \frac{a_{1s}a_{4s}}{a_{1s}a_{3s} + a_{2s}a_{3s} + a_{1s}a_{4s}}, \qquad \mu_{hs} = \frac{a_{1s}a_{3s}}{a_{1s}a_{3s} + a_{2s}a_{3s} + a_{1s}a_{4s}}, \tag{11}$$

where $a_{1s} = a_1(\mu_{us}, \mu_{hs}),\ a_{2s} = a_2(\mu_{hs}, 1 - \mu_{us} - \mu_{hs}),\ a_{3s} = a_3(\mu_{hs}, 1 - \mu_{us} - \mu_{hs}),\ a_{4s} = a_4(\mu_{us}, \mu_{hs})$. Since Eq. (11) is independent of a, the means depend only on x and y. For fixed parameters, we can solve Eq. (11) numerically to obtain values for μ_{us} and μ_{hs}.

The second moment equations are given by

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\begin{aligned}
\frac{d\langle L_uL_u\rangle}{dt} &= -2a_3\langle L_uL_u\rangle + 2a_4\langle L_uL_h\rangle + a_4\mu_h + a_3\mu_u,\\
\frac{d\langle L_uL_h\rangle}{dt} &= -a_1\Big(\langle L_uL_u\rangle + \langle L_uL_h\rangle\Big) - a_2\langle L_uL_h\rangle + a_3\Big(\langle L_uL_u\rangle - \langle L_uL_h\rangle\Big)\\
&\qquad - a_4\Big(\langle L_uL_h\rangle - \langle L_hL_h\rangle\Big) + a_1\mu_u - a_3\mu_u - a_4\mu_h,\\
\frac{d\langle L_hL_h\rangle}{dt} &= -2a_1\Big(\langle L_uL_h\rangle + \langle L_hL_h\rangle\Big) - 2a_2\langle L_hL_h\rangle + 2a_3\langle L_uL_h\rangle - 2a_4\langle L_hL_h\rangle\\
&\qquad + a_1\big(1-\mu_u+\mu_h\big) + a_2\mu_h + a_3\mu_u + a_4\mu_h.
\end{aligned} \tag{12}$$

From Eq. (9) we expect $L_uL_u = L_u,\ L_hL_h = L_h$ and $L_uL_h = 0$ at any CpG. Solving Eq. (12) in steady state leads to

$$\langle L_uL_u\rangle_s = \langle L_u\rangle_s = \mu_{us}, \qquad \langle L_uL_h\rangle_s = 0, \qquad \langle L_hL_h\rangle_s = \langle L_h\rangle_s = \mu_{hs}. \tag{13}$$

Variances, $\sigma^2(L_{us})$ and $\sigma^2(L_{hs})$, can then be obtained via

$$\begin{aligned}
\sigma^2(L_{us}) &= \langle L_uL_u\rangle_s - \mu_{us}^2 = \mu_{us} - \mu_{us}^2,\\
\sigma^2(L_{hs}) &= \langle L_hL_h\rangle_s - \mu_{hs}^2 = \mu_{hs} - \mu_{hs}^2,
\end{aligned} \tag{14}$$

along with the covariance $\mathrm{Covar}(L_{us}, L_{hs})$, which is given by

$$\mathrm{Covar}(L_{us}, L_{hs}) = \langle L_uL_h\rangle_s - \mu_{us}\mu_{hs} = -\mu_{us}\mu_{hs}. \tag{15}$$

The mean, variance and covariances associated with the m state can now be obtained using

$$\begin{aligned}
\mu_{ms} &= 1 - \mu_{us} - \mu_{hs},\\
\sigma^2(L_{ms}) &= \sigma^2(1 - L_{us} - L_{hs}) = \sigma^2(L_{us}) + \sigma^2(L_{hs}) + 2\mathrm{Covar}(L_{us}, L_{hs}),\\
\mathrm{Covar}(L_{us}, L_{ms}) &= -\sigma^2(L_{us}) - \mathrm{Covar}(L_{us}, L_{hs}),\\
\mathrm{Covar}(L_{hs}, L_{ms}) &= -\mathrm{Covar}(L_{us}, L_{hs}) - \sigma^2(L_{hs}).
\end{aligned} \tag{16}$$

We calculate the steady-state mean and variance, μ_z and $\sigma^2(z)$, associated with the variable $z = L_u + 2L_h + 3L_m$ via

$$\begin{aligned}
\mu_z &= \mu_{us} + 2\mu_{hs} + 3\mu_{ms},\\
\sigma^2(z) &= \sigma^2(L_{us}) + 4\sigma^2(L_{hs}) + 9\sigma^2(L_{ms})\\
&\qquad + 2\Big(2\mathrm{Covar}(L_{us}, L_{hs}) + 3\mathrm{Covar}(L_{us}, L_{ms}) + 6\mathrm{Covar}(L_{hs}, L_{ms})\Big).
\end{aligned}$$

Note that the superscript t was only used in Section 2.2 to differentiate between samples at different time points. Here we simply have a single z. Since no spatial information is obtained from the one-site MF model, the covariance and correlation between neighbouring sites cannot be extracted.

3.2 Distinct pairs mean-field model

We next construct a two-site MF model, where we consider "clusters" of two adjacent CpGs. Such cluster MF models have been successfully used to study vehicular traffic and driven-diffusive gas models [34, 35]. We define the mean level (proportion) of pairs in the six possible

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

states,

$$mm, \qquad uu, \qquad hh, \qquad um \ (:= mu), \qquad hm \ (:= mh), \qquad uh \ (:= hu), \tag{17}$$

to be $\mu_1,\ \mu_2,\ \mu_3,\ \mu_4,\ \mu_5,\ \mu_6$, respectively. Here μ_4 is the proportion of pairs containing u and m, irrespective of order. Similarly, μ_5 is the proportion of pairs containing h and m, and μ_6 is the proportion of pairs containing u and h, irrespective of order.

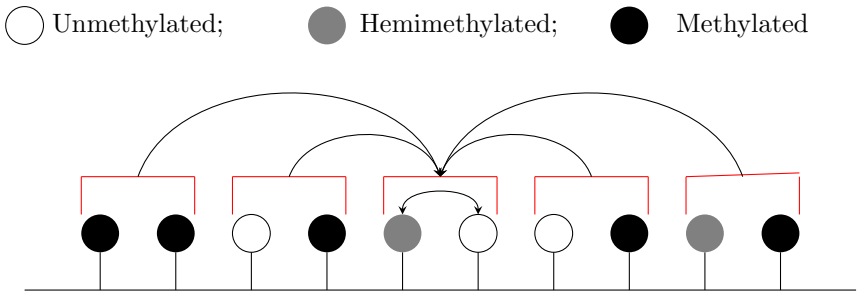

Figure 6: Schematic of the distinct pairs MF model. The two CpGs within a pair can interact directly with each other and the pair is also influenced by the mean state of pairs in the system. In the figure, the uh pair can change state due to interactions between the u and h within the pair, and due to the mean state of pairs in the system.

In the distinct pairs MF model (DPMF model), CpGs within a pair are allowed to interact directly, preserving some nearest-neighbour interactions. The influence of the nearest-neighbour CpGs flanking the pair is then approximated by considering the probabilities that an adjacent pair is in each of the six possible states; see Fig. 6. Here, each CpG belongs to only one pair and each pair of sites is a single reactant. As with the one-site model, we consider an effective first-order reaction system, given by

$$\begin{aligned} &mm \xrightarrow{\hat{a}_1} hm, \quad uu \xrightarrow{\hat{a}_2} uh, \quad hh \xrightarrow{\hat{a}_3} uh, \quad hh \xrightarrow{\hat{a}_4} hm, \quad um \xrightarrow{\hat{a}_5} hm, \quad um \xrightarrow{\hat{a}_6} uh, \\ &hm \xrightarrow{\hat{a}_7} um, \quad hm \xrightarrow{\hat{a}_8} mm, \quad hm \xrightarrow{\hat{a}_9} hh, \quad uh \xrightarrow{\hat{a}_{10}} hh, \quad uh \xrightarrow{\hat{a}_{11}} uu, \quad uh \xrightarrow{\hat{a}_{12}} um, \end{aligned} \tag{18}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

where the effective rates are given by

$$
\begin{aligned}
\hat{a}_1 &= 2k_3 + k_9(2\mu_3 + \mu_5 + \mu_6) + k_{10}(2\mu_2 + \mu_4 + \mu_6),\\
\hat{a}_2 &= 2k_1 + k_5(2\mu_3 + \mu_5 + \mu_6) + k_6(2\mu_1 + \mu_4 + \mu_5),\\
\hat{a}_3 &= 2k_4 + 2k_{11} + k_{11}(4\mu_3 + \mu_5 + \mu_6) + k_{12}(2\mu_2 + \mu_4 + \mu_6),\\
\hat{a}_4 &= 2k_2 + 2k_7 + k_7(4\mu_3 + \mu_5 + \mu_6) + k_8(2\mu_1 + \mu_4 + \mu_5),\\
\hat{a}_5 &= k_1 + k_6 + k_5\Big(\mu_3 + \frac{\mu_5}{2} + \frac{\mu_6}{2}\Big) + k_6\Big(\mu_1 + \frac{\mu_4}{2} + \frac{\mu_5}{2}\Big),\\
\hat{a}_6 &= k_3 + k_{10} + k_9\Big(\mu_3 + \frac{\mu_5}{2} + \frac{\mu_6}{2}) + k_{10}(\mu_2 + \frac{\mu_4}{2} + \frac{\mu_6}{2}\Big),\\
\hat{a}_7 &= k_4 + k_{11}\Big(\mu_3 + 2\frac{\mu_5}{2} + \frac{\mu_6}{2}\Big) + k_{12}\Big(\mu_2 + \frac{\mu_4}{2} + \frac{\mu_6}{2}\Big),\\
\hat{a}_8 &= k_2 + k_8 + k_7\Big(\mu_3 + 2\frac{\mu_5}{2} + \frac{\mu_6}{2}\Big) + k_8\Big(\mu_1 + \frac{\mu_4}{2} + \frac{\mu_5}{2}\Big),\\
\hat{a}_9 &= k_3 + k_9 + k_9\Big(\mu_3 + \frac{\mu_5}{2} + \frac{\mu_6}{2}\Big) + k_{10}\Big(\mu_2 + \frac{\mu_4}{2} + \frac{\mu_6}{2}\Big),\\
\hat{a}_{10} &= k_1 + k_5 + k_5\Big(\mu_3 + \frac{\mu_5}{2} + \frac{\mu_6}{2}\Big) + k_6\Big(\mu_1 + \frac{\mu_4}{2} + \frac{\mu_5}{2}\Big),\\
\hat{a}_{11} &= k_4 + k_{12} + k_{11}\Big(\mu_3 + \frac{\mu_5}{2} + 2\frac{\mu_6}{2}\Big) + k_{12}\Big(\mu_2 + \frac{\mu_4}{2} + \frac{\mu_6}{2}\Big),\\
\hat{a}_{12} &= k_2 + k_7\Big(\mu_3 + \frac{\mu_5}{2} + 2\frac{\mu_6}{2}\Big) + k_8\Big(\mu_1 + \frac{\mu_4}{2} + \frac{\mu_5}{2}\Big),
\end{aligned}
$$

and $\sum_{i=1}^{6} \mu_i = 1$. We describe the construction of $\hat{a}_1$ in Appendix A. Essentially, $\hat{a}_1$—$\hat{a}_{12}$ are constructed by considering all possible ways that each reaction in (18) can occur via a reaction from Fig. 1 taking place. Such reactions can occur within the reactant pair or can take place between a site within the pair and a site from an adjacent pair. Terms associated with interactions between two hh, two hm or two uh pairs have an additional factor of two since either pair can change state during the reaction. While we can, in principle, calculate the distribution of pairs in (18) [36], we restrict our attention to obtaining moments of the system.

Let $L_1,\ L_2,\ L_3,\ L_4,\ L_5,\ L_6$ be the level (proportion) of each of the paired states at a single pair of CpGs. At any time, a pair can be in only one state and so at a single pair we have

$$L_i = 1, \quad \text{for one } i \quad \text{and} \quad L_j = 0 \quad \text{for all } j \neq i. \tag{19}$$

The $u,\ h,\ m$ levels within a pair of CpGs, $\hat{L}_u,\ \hat{L}_h,\ \hat{L}_m$, are then given by

$$\hat{L}_u = L_2 + \frac{L_4}{2} + \frac{L_6}{2}, \qquad \hat{L}_h = L_3 + \frac{L_5}{2} + \frac{L_6}{2}, \qquad \hat{L}_m = L_1 + \frac{L_4}{2} + \frac{L_5}{2}. \tag{20}$$

Using the CME for the system (18) we construct first and second moment equations for L_i, $i = \{1, \ldots, 6\}$; see Appendix C. The first moment equations describe $\mu_i = \langle L_i \rangle$, the mean values of L_i for $i = \{1, \ldots, 6\}$. For fixed parameters, solving these equations numerically in steady state gives the steady-state means, μ_{is}, $i = \{1, \ldots, 6\}$. Note that these means are independent of a.

From the second moment equations, we obtain the steady-state expected values of L_iL_j, $\langle L_iL_j \rangle_s$, for $i, j \in \{1, 2, \ldots, 6\}$. As expected from Eq. (19), $\langle L_iL_i \rangle_s = \langle L_i \rangle_s = \mu_{is}$, $\langle L_iL_j \rangle_s =$

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

0 for all $i \neq j,\ i,j \in \{1,2,\ldots,6\}$. Variances and covariances are then obtained using

$$\begin{aligned}\sigma^2(L_{is}) &= \langle L_i L_i \rangle_s - \mu_{is}^2 = \mu_{is} - \mu_{is}^2, \qquad && i = 1,2,\ldots,6,\\ \mathrm{Covar}(L_{is}, L_{js}) &= \langle L_i L_j \rangle_s - \mu_{is}\mu_{js} = -\mu_{is}\mu_{js}, && i \neq j,\ i,j = 1,2,\ldots,6.\end{aligned}$$

Once again, these are independent of the parameter a.

We now have statistics for the paired states in (17). The steady-state means for the $u,\ h,\ m$ levels in a pair are then given by

$$\mu_{us} = \mu_{2s} + \frac{\mu_{4s}}{2} + \frac{\mu_{6s}}{2}, \qquad \mu_{hs} = \mu_{3s} + \frac{\mu_{5s}}{2} + \frac{\mu_{6s}}{2}, \qquad \mu_{ms} = \mu_{1s} + \frac{\mu_{4s}}{2} + \frac{\mu_{5s}}{2}. \tag{21}$$

The pair-to-pair variance in u level is given by

$$\begin{aligned}\sigma^2(\hat{L}_{us}) = \sigma^2(L_{2s}) &+ \frac{1}{4}\sigma^2(L_{4s}) + \frac{1}{4}\sigma^2(L_{6s})\\ &+ 2\left(\frac{1}{2}\mathrm{Covar}(L_{2s}, L_{4s}) + \frac{1}{2}\mathrm{Covar}(L_{2s}, L_{6s}) + \frac{1}{4}\mathrm{Covar}(L_{4s}, L_{6s})\right),\end{aligned}$$

and similarly for the variances associated with h and m, $\sigma^2(\hat{L}_{hs})$ and $\sigma^2(\hat{L}_{ms})$. Covariances are given by

$$\begin{aligned}\mathrm{Covar}(\hat{L}_{us}, \hat{L}_{ms}) = \mathrm{Covar}(L_{1s}, L_{2s}) &+ \frac{1}{2}\mathrm{Covar}(L_{2s}, L_{4s}) + \frac{1}{2}\mathrm{Covar}(L_{2s}, L_{5s})\\ &+ \frac{1}{2}\mathrm{Covar}(L_{1s}, L_{4s}) + \frac{1}{4}\sigma^2(L_{4s}) + \frac{1}{4}\mathrm{Covar}(L_{4s}, L_{5s})\\ &+ \frac{1}{2}\mathrm{Covar}(L_{1s}, L_{6s}) + \frac{1}{4}\mathrm{Covar}(L_{4s}, L_{6s}) + \frac{1}{4}\mathrm{Covar}(L_{5s}, L_{6s})\end{aligned}$$

and analogously for $\mathrm{Covar}(\hat{L}_{hs}, \hat{L}_{ms}),\ \mathrm{Covar}(\hat{L}_{us}, \hat{L}_{hs})$.

Note that the statistics obtained so far relate to $u,\ h,\ m$ levels within a pair of CpGs. The mean level of a state within a pair is the same as the mean level of the state at each site. However, this is not the case for higher moments. For example, $\sigma^2(\hat{L}_{us}),\ \sigma^2(\hat{L}_{hs}),\ \sigma^2(\hat{L}_{ms})$ are pair-to-pair variances, rather than site-to-site variances.

We aim to obtain statistics relating to $z = L_u + 2L_h + 3L_m$, where $L_u,\ L_h,\ L_m$ are the single-site $u,\ h,\ m$ levels. We define $\hat{z} = \hat{L}_u + 2\hat{L}_h + 3\hat{L}_m$, noting that $\hat{z}$ contains information regarding pairs of CpGs. Essentially,

$$\hat{z} = \frac{v + w}{2}, \tag{22}$$

where v and w are as in Eq. (5) and we again do not require the superscript t.

The means of z, v, w and $\hat{z}$ coincide and

$$\mu_z = \mu_{\hat{z}} = \mu_v = \mu_w = \mu_{us} + 2\mu_{hs} + 3\mu_{ms}.$$

Also, $vw = L_2 + 2L_6 + 3L_4 + 4L_3 + 6L_5 + 9L_1$ and so

$$\langle vw \rangle = \mu_{2s} + 2\mu_{6s} + 3\mu_{4s} + 4\mu_{3s} + 6\mu_{5s} + 9\mu_{1s}.$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

From this we calculate the covariance between neighbouring sites as

$$\text{Covar}(z) = \text{Covar}(v, w) = \langle vw \rangle - \langle v \rangle \langle w \rangle = \langle vw \rangle - \mu_z^2.$$

The variance associated with $\hat{z}$ can be calculated via,

$$\begin{aligned}\sigma^2(\hat{z}) &= \sigma^2(\hat{L}_{us} + 2\hat{L}_{hs} + 3\hat{L}_{ms}) \\ &= \sigma^2(\hat{L}_{us}) + 4\sigma^2(\hat{L}_{hs}) + 9\sigma^2(\hat{L}_{ms}) \\ &\quad + 2\Big(2\text{Covar}(\hat{L}_{us}, \hat{L}_{hs}) + 3\text{Covar}(\hat{L}_{us}, \hat{L}_{ms}) + 6\text{Covar}(\hat{L}_{hs}, \hat{L}_{ms})\Big).\end{aligned}$$

However, $\sigma^2(\hat{z})$ is the pair-to-pair variance. Using $\sigma^2(z) = \sigma^2(v) = \sigma^2(w)$ we obtain

$$\sigma^2(\hat{z}) = \sigma^2\left(\frac{v+w}{2}\right) = \frac{1}{4}\big(\sigma^2(v) + \sigma^2(w) + 2\text{Covar}(v, w)\big) = \frac{1}{2}\big(\sigma^2(z) + \text{Covar}(v, w)\big),$$

leading to the site-to-site variance,

$$\sigma^2(z) = 2\sigma^2(\hat{z}) - \text{Covar}(v, w).$$

The correlation between neighbouring pairs is obtained via

$$\rho(z) = \rho(v, w) = \frac{\text{Covar}(v, w)}{\sqrt{\sigma^2(v)\sigma^2(w)}} = \frac{\text{Covar}(v, w)}{\sigma^2(z)}.$$

To summarise, the statistics of primary interest from our calculations are: the means μ_{us}, μ_{hs}, μ_{ms}, μ_{zs}, the variance $\sigma^2(z)$, the covariance $\text{Covar}(z)$ and the correlation $\rho(z)$.

3.3 Overlapping pairs mean-field model

Similarly to the DPMF model, there are also six possible states for a pair of CpGs in the overlappling pairs MF model (OPMF model), see (17). The OPMF model also incorporates direct interactions within a pair. However, each CpG now belongs to two pairs, one with its left-hand neighbour and one with its right-hand neighbour, leading to a system of overlapping pairs. The influence of CpGs flanking a pair is now approximated by considering the conditional probability that the pair overlaps with another pair of a certain state; see Fig. 7.

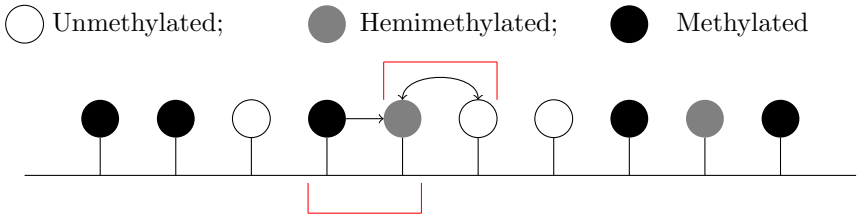

Figure 7: Schematic of the overlapping pairs MF model. A pair of CpGs interact directly with each other and the effect of CpGs flanking the pair is approximated by considering the conditional probabilities that a flanking site is in the u, h or m state, given the state of its neighbour within the pair.

As before, we consider an effective first-order reaction system, given by

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$\begin{gathered} mm \xrightarrow{\tilde{a}_1} hm, \quad uu \xrightarrow{\tilde{a}_2} uh, \quad hh \xrightarrow{\tilde{a}_3} uh, \quad hh \xrightarrow{\tilde{a}_4} hm, \quad um \xrightarrow{\tilde{a}_5} hm, \quad um \xrightarrow{\tilde{a}_6} uh, \\ hm \xrightarrow{\tilde{a}_7} um, \quad hm \xrightarrow{\tilde{a}_8} mm, \quad hm \xrightarrow{\tilde{a}_9} hh, \quad uh \xrightarrow{\tilde{a}_{10}} hh, \quad uh \xrightarrow{\tilde{a}_{11}} uu, \quad uh \xrightarrow{\tilde{a}_{12}} um, \end{gathered} \tag{23}$$

where the effective rates are given by

$$\begin{aligned} \tilde{a}_1 &= 2k_3 + 2k_9\left(\frac{\mu_5/2}{\mu_1+\mu_4/2+\mu_5/2}\right) + 2k_{10}\left(\frac{\mu_4/2}{\mu_1+\mu_4/2+\mu_5/2}\right), \\ \tilde{a}_2 &= 2k_1 + 2k_5\left(\frac{\mu_6/2}{\mu_2+\mu_4/2+\mu_6/2}\right) + 2k_6\left(\frac{\mu_4/2}{\mu_2+\mu_4/2+\mu_6/2}\right), \\ \tilde{a}_3 &= 2k_4 + 2k_{11} + 2k_{11}\left(\frac{\mu_3}{\mu_3+\mu_5/2+\mu_6/2}\right) + 2k_{12}\left(\frac{\mu_6/2}{\mu_3+\mu_5/2+\mu_6/2}\right), \\ \tilde{a}_4 &= 2k_2 + 2k_7 + 2k_7\left(\frac{\mu_3}{\mu_3+\mu_5/2+\mu_6/2}\right) + 2k_8\left(\frac{\mu_5/2}{\mu_3+\mu_5/2+\mu_6/2}\right), \\ \tilde{a}_5 &= k_1 + k_6 + k_5\left(\frac{\mu_6/2}{\mu_2+\mu_4/2+\mu_6/2}\right) + k_6\left(\frac{\mu_4/2}{\mu_2+\mu_4/2+\mu_6/2}\right), \\ \tilde{a}_6 &= k_3 + k_{10} + k_9\left(\frac{\mu_5/2}{\mu_1+\mu_4/2+\mu_5/2}\right) + k_{10}\left(\frac{\mu_4/2}{\mu_1+\mu_4/2+\mu_5/2}\right), \\ \tilde{a}_7 &= k_4 + k_{11}\left(\frac{\mu_3}{\mu_3+\mu_5/2+\mu_6/2}\right) + k_{12}\left(\frac{\mu_6/2}{\mu_3+\mu_5/2+\mu_6/2}\right), \\ \tilde{a}_8 &= k_2 + k_8 + k_7\left(\frac{\mu_3}{\mu_3+\mu_5/2+\mu_6/2}\right) + k_8\left(\frac{\mu_5/2}{\mu_3+\mu_5/2+\mu_6/2}\right), \\ \tilde{a}_9 &= k_3 + k_9 + k_9\left(\frac{\mu_5/2}{\mu_1+\mu_4/2+\mu_5/2}\right) + k_{10}\left(\frac{\mu_4/2}{\mu_1+\mu_4/2+\mu_5/2}\right), \\ \tilde{a}_{10} &= k_1 + k_5 + k_5\left(\frac{\mu_6/2}{\mu_2+\mu_4/2+\mu_6/2}\right) + k_6\left(\frac{\mu_4/2}{\mu_2+\mu_4/2+\mu_6/2}\right), \\ \tilde{a}_{11} &= k_4 + k_{12} + k_{11}\left(\frac{\mu_3}{\mu_3+\mu_5/2+\mu_6/2}\right) + k_{12}\left(\frac{\mu_6/2}{\mu_3+\mu_5/2+\mu_6/2}\right), \\ \tilde{a}_{12} &= k_2 + k_7\left(\frac{\mu_3}{\mu_3+\mu_5/2+\mu_6/2}\right) + k_8\left(\frac{\mu_5/2}{\mu_3+\mu_5/2+\mu_6/2}\right). \end{aligned}$$

We detail the construction of $\hat{a}_1$ in Appendix B.

An identical approach as in Section 3.2 leads to first and second moment equations for the system, see Appendix C, and from these we obtain means, variances and covariances associated with the paired states and with the pair-to-pair u, h, m levels. Again, these depend only on x and y. The statistics associated with z are obtained as for the DPMF model, see Section 3.2.

4 Model comparison and parameter inference

To test whether MF models are capable of modelling large-scale methylation patterns, we now compare model predictions to synthetic data generated using nearest-neighbour collaborative simulations. The statistical properties obtained from our models are independent of

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a and so we fix $a = 0.2$. Since demethylation dominates when $y < 1$ and methylation dominates when $y > 1$, we hypothesise that a sharp change in the behaviour of the system may occur at $y = 1$. To capture this potential transition for different collaborativity strengths, we consider $x = \{0.1,\ 1,\ 5,\ 50\}$, $y = \{0.1,\ 0.2,\ \ldots,\ 2\}$. For every parameter set, we run $n = 10$ stochastic simulations, obtaining ten datasets of 10^6 CpGs, from which we calculate the statistics of interest. We then calculate the means and standard errors over the ten datasets to obtain overall summary statistics.

In this study, we approximate the nearest-neighbour collaborative system by MF models which consider an infinite system of CpGs. As x increases, finite-size effects cause discrepancies between the simulations and models, which can be counteracted by increasing the number of simulated sites. We therefore simulate systems of $N = 200$ sites when $x = \{0.1,\ 1,\ 5\}$ and systems of $N = 500$ sites when $x = 50$.

4.1 Mean-field models capture steady-state methylation levels

We first compare the mean u, h, m levels (μ_{us}, μ_{hs}, μ_{ms}) from the MF models to those from the simulations (Fig. 8). Considering the simulated data first, we observe that u and m dominate when $y < 1$ or $y > 1$ respectively. h is an intermediate state between u and m and peaks at $y = 1$, where there is also a sharp transition between u and m-dominant states.

While all models capture the qualitative behaviour of the means as x and y are varied, we observe that predictions of methylation levels from the OPMF model are closest to those observed in the simulated data (Fig. 8). All three models predict the mean u and m levels reasonably well, however only the OPMF model accurately predicts the mean h level for large x and for y close to one. All predictions from the OPMF model are within the error of the simulated data and it successfully captures the transition observed at $y = 1$ for all x considered. Conversely, the predictions of the other models deviate from the simulations at the transition point when x is large. The one-site MF model deviates to the greatest extent for 87% of the parameter sets. The poorer performance of the one-site model can be seen most clearly in the mean h levels when $x = 1$. This suggests that the one-site model has the worst predictive power and we exclude it from further analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

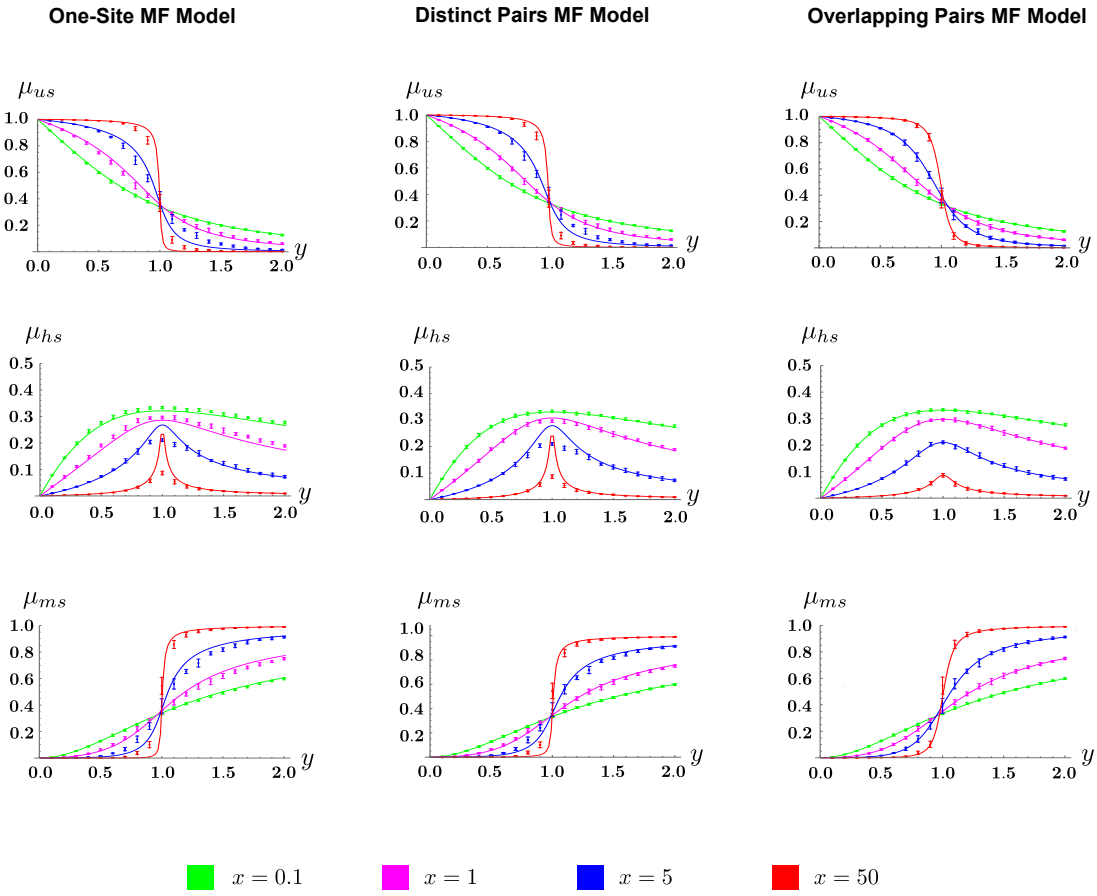

Figure 8: The OPMF model accurately predicts the average behaviour of large-scale methylation levels. The mean u, h and m levels are plotted against the methylation strength y for various values of the collaborativity strength x and the different mean-field models (left: one-site MF model, middle: DPMF model and right: OPMF model). Solid lines denote model predictions and points show the mean of $n = 10$ simulations. Error bars indicate standard error for the simulated data.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4.2 The OPMF model accurately predicts associations between neighbouring sites

To test whether MF models can predict associations between neighbouring CpGs we consider $z = \{z_1, z_2, \ldots, z_N\}$, where $z_i = \{1,\ 2,\ 3\}$ if CpG i is in the $u,\ h,\ m$ state, respectively. From z we calculate the mean and variance associated with the methylation state, and the covariance and correlation in methylation state between neighbouring sites. In the simulated data (Fig. 9), we again observe a transition in these statistics when the methylation and demethylation strengths are equal ($y = 1$). Our results counterintuitively suggest that neighbouring sites are most correlated here (the peak $\rho(z)$ occurs when $y = 1$).

To gain insight into this observation, we examine the patterns that evolve for $x = 50$ (Fig. 10). When y is small, large u clusters form and we intuitively expect neighbouring sites to be highly correlated. However, these large clusters are interspersed with infrequent, isolated occurrences of h and m which have low correlations with their neighbours. Moreover, ten m (or h) sites appearing in isolation will result in smaller u clusters than the ten sites appearing as a single cluster. The overall effect is to reduce the correlation when y is small. A similar rationale explains the low correlation when y is large. u and m cluster sizes are most similar when methylation and demethylation are equally strong, resulting in u and m sites correlating equally with their neighbours and the overall correlation peaking.

We next consider predictions of these statistics from the MF models. Predictions from the OPMF model lie within the error observed in the simulated data for all parameters considered (Fig. 9). Conversely, predicted statistics from the DPMF model show large deviations from the corresponding simulated statistics when x is large and y is close to one demonstrating that it has lower predictive power.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

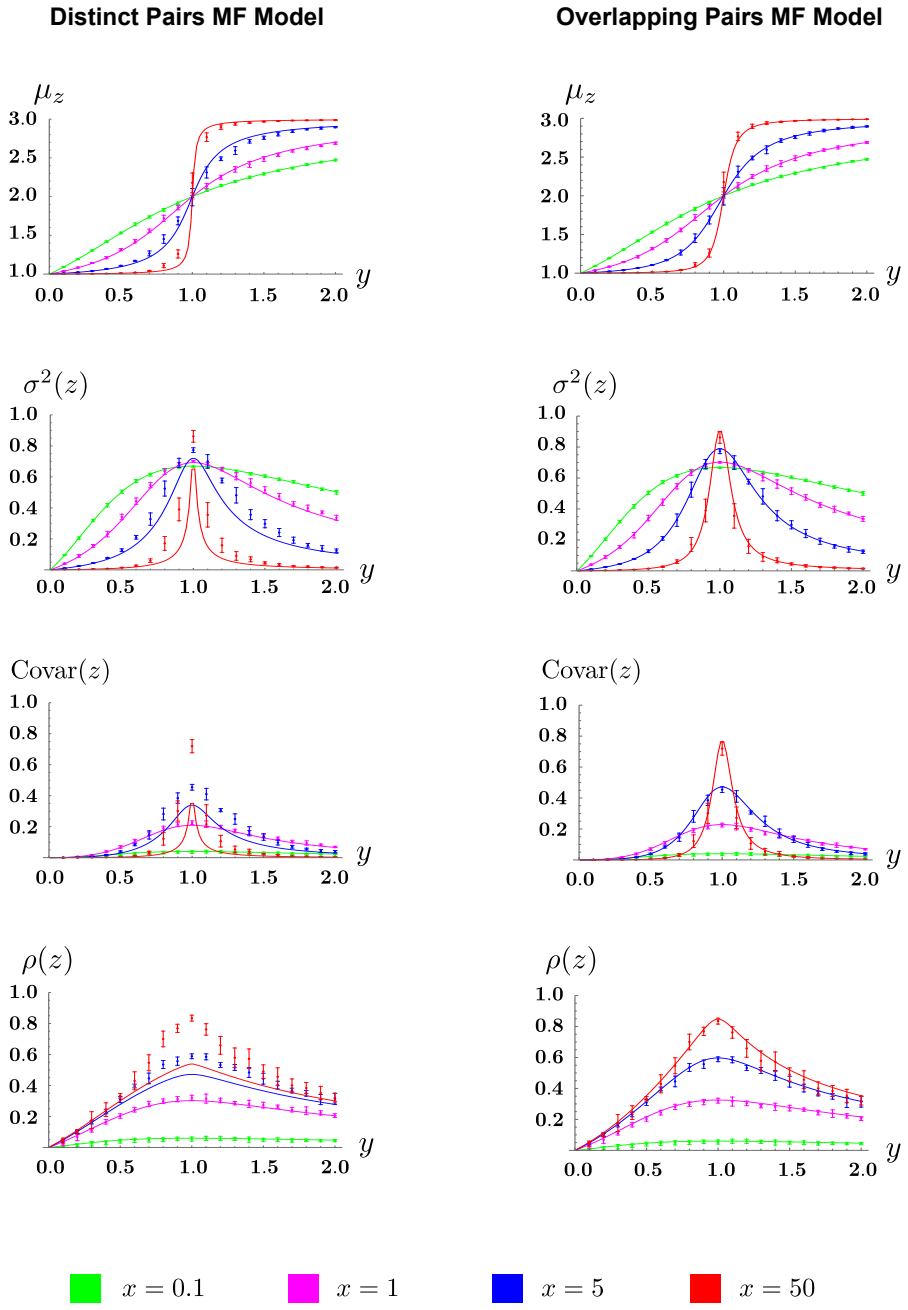

Figure 9: The OPMF model accurately predicts associations between neighbouring CpGs. Predictions of the means (μ_z), variances ($\sigma^2(z)$), covariances ($Covar(z)$) and correlations ($\rho(z)$) are plotted against y for different values of x and for the DPMF model (left-hand panels) and the OPMF model (right-and panels). Solid lines denote model predictions and points show the mean statistics calculated from the simulated data ($n = 10$ in each case). Error bars indicate standard error for the simulated data. Note that, since z is determined by x and y, the statistics plotted here are implicit functions of x and y.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

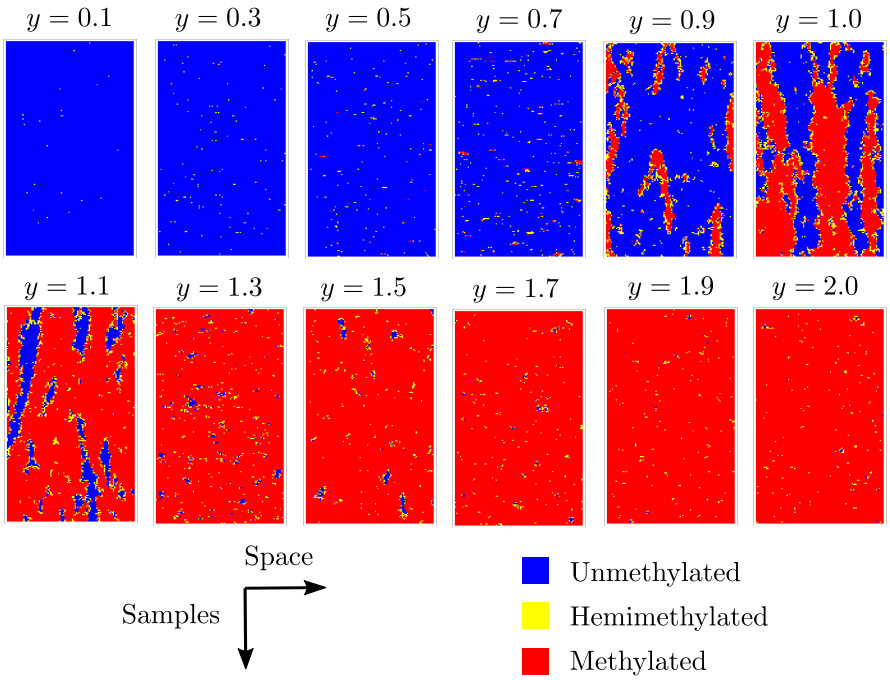

Figure 10: Size of unmethylated (u) clusters and methylated (m) clusters are most similar when $y = 1$. Simulated methylation patterns are shown for 100 CpGs when $x = 50$. For each y, 200 steady-state patterns are displayed, with each row showing the 100 sites at a different time point.

4.3 Overlapping pairs MF model can infer the parameters underpinning large-scale methylation patterns

To test whether the OPMF model could, in principle, be used to infer collaborativity and methylation strengths from data, we generate synthetic data for selected model parameters, see Section 2.2. We then infer these parameters back using the OPMF model.

Methylomes are typically assayed by whole genome bisulfite sequencing [37]. A variant of this, hairpin-bisulfite sequencing, can be used to assay both strands of each DNA molecule [38]. In both cases, the resulting data is composed of short reads. Each read assays few CpGs and we do not know if reads originate from the same cell or DNA molecule. Simulated datasets from previous sections do not provide a good reflection of bisulfite sequencing since all of the CpGs in the simulated system were sampled at the same time points, the equivalent of the CpGs originating from the same molecule. To obtain short-read data, we instead simulate data for $N = 1000$ CpGs. After steady state is reached, we sample the system at $10,000$ different time points. This is equivalent to sampling $10,000$ molecules in steady state at a single time point. For each CpG, we take the methylation state at 30 time points, randomly chosen from the original $10,000$. The time points chosen for each CpG site are independent of those chosen for other CpGs. This emulates hairpin-bisulfite sequencing data with coverage of 30 reads per CpG. We combine the sample states for all CpGs into a single

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

dataset, X, and consider $z = \{z_1, z_2, \ldots, z_{1000}\}$ where $z_i = \{1,\ 2,\ 3\}$ if X_i corresponds to a $u,\ h,\ m$ state, respectively. The mean and variance of z are calculated and used for inference.

There are numerous well-established methods for conducting inference. For example, inference can be conducted using maximum likelihood estimation [39], which provides point estimates for model parameters. Here we use a Bayesian inference approach, which has the advantage of providing us with a distribution of the estimate value from which we can calculate confidence intervals associated with our inferred parameter values [40, 41, 42, 43]. In particular, we use the Approximate Bayesian Computation Sequential Monte Carlo algorithm (ABC SMC), a likelihood-free inference method (see [44] for a comprehensive overview). We use uniform priors, $U(0, 100)$ and $U(0, 2)$, for x and y, respectively. We also define the distance, d, between the simulations and model prediction to be the sum of the absolute relative errors of the mean and variance, i.e.

$$d = \left| \frac{\mu_{model} - \mu_{data}}{\mu_{data}} \right| + \left| \frac{\sigma^2_{model} - \sigma^2_{data}}{\sigma^2_{data}} \right|,$$

where $\mu_{model},\ \mu_{data}$ are the means of z from the model and data, respectively, and $\sigma^2_{model},\ \sigma^2_{data}$ are the variances associated with the model and data, respectively. To rapidly select appropriate tolerances, we calculate the true distances between the simulated data and model predictions at the parameter values of interest.

As in previous sections, we examine $x = \{0.1,\ 1,\ 5,\ 50\}$. For each x, we infer for $y = \{0.3,\ 1,\ 1.7\}$ using the GpABC Julia package [45]. Accepted $x,\ y$ values from the final ABC SMC population make up the posterior distributions for x and y, with the means taken to be the inferred parameter values. 95% confidence intervals were calculated by removing the lowest 2.5% and highest 2.5% from the posteriors.

We find that the inferred parameter values are always of the same order of magnitude as the true values (Table 2), with the most successfully inferred parameters being inferred within 1% of the true values (Fig. 11a, b). There are only two cases where the true parameter values lie outwith the inferred 95% confidence intervals (e.g. Fig. 11c). However, we obtain wide posteriors for large x (e.g. Fig 11d), indicating more uncertainty in the inference.

	$y = 0.3$	$y = 1$	$y = 1.7$
$x = 0.1$	$(x, y) = (0.197, 0.329)$	$(x, y) = (0.066, 0.997)$	$(x, y) = (0.101, 1.701)$
$x = 1$	$(x, y) = (0.676, 0.251)$	$(x, y) = (1.011, 0.990)$	$(x, y) = (0.923, 1.730)$
$x = 5$	$(x, y) = (6.302, 0.338)$	$(x, y) = (5.287, 1.000)$	$(x, y) = (6.432, 1.613)$
$x = 50$	$(x, y) = (59.571, 0.329)$	$(x, y) = (50.963, 0.999)$	$(x, y) = (65.347, 1.599)$

Table 2: Inferred parameters for the nearest-neighbour collaborative model using the OPMF model and the ABC SMC algorithm.

The accuracy of inference is highly dependent on the model sensitivity to parameters, with ease of inference increasing as sensitivity to the parameter increases. To test the sensitivity of our model to model parameters, we calculate the relative sensitivity [46] of the OPMF model to the parameters x and y, for $x = \{0.1,\ 0.2, \ldots 99.9,\ 100\}$ and $y = \{0.1,\ 0.2, \ldots 1.9,\ 2\}$. For all parameters considered, the x-sensitivity divided by the y-sensitivity is strictly less than one, indicating that the model shows more sensitivity to y than x. This means that y will be inferred more accurately than x.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

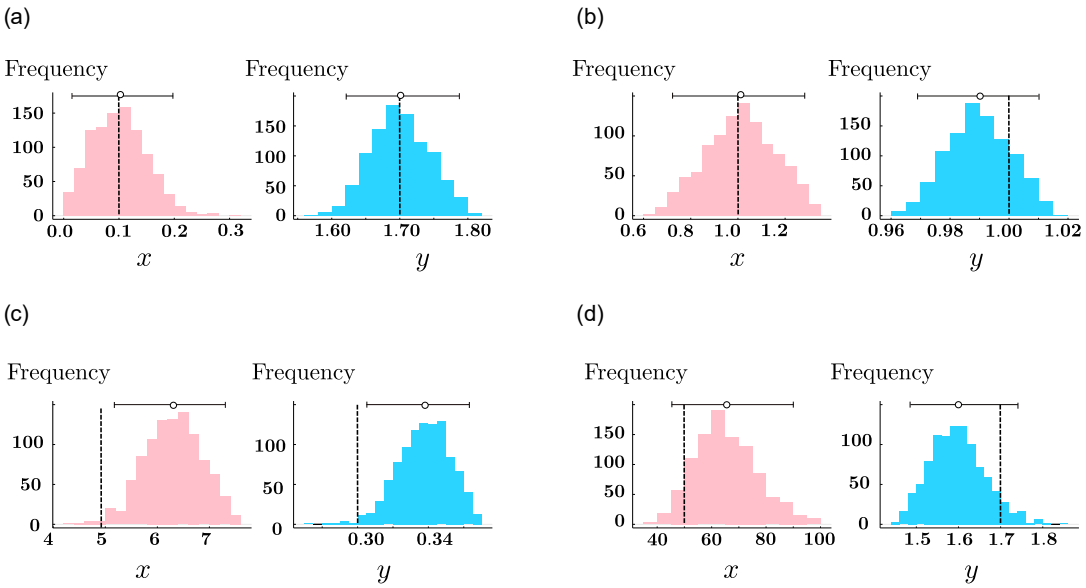

Figure 11: The OPMF model can be used to infer collaborativity and methylation strengths. Example posteriors from inference are shown, with (a), (b), (c) and (d) each corresponding to a different parameter set. True parameter values are denoted by dashed lines, inferred parameter values are shown as dots, with 95% confidence intervals shown as horizontal bars.

5 Discussion

Genomic DNA methylation patterns vary between cell types, across differentiation and in disease. The mechanisms underpinning this variation remain unclear but can be better understood using mathematical models. Current approaches are limited by their inability to feasibly model large systems of CpGs and thus understand known large-scale features of methylomes. Here we show that a cluster MF model, based around overlapping pairs of CpGs, can predict DNA methylation patterns generated under a nearest-neighbour collaborative model. This suggests that MF models are a valuable tool for understanding large-scale DNA methylation features.

Previous studies have used mathematical modelling to gain insight into the mechanisms regulating the establishment and maintenance of DNA methylation patterns. In particular, the requirement of collaborativity between CpGs to maintain DNA methylation patterns was postulated through modelling [14] before being observed experimentally [15, 17]. Previous models of DNA methylation rely on stochastic simulations. However, their computational expense limits their use to the study of promoter-scale DNA methylation (regions around 1Kb in size). The stochastic simulations we run here on 200 or 500 sites can take hours to run. In contrast, the OPMF model can be applied to arbitrarily large systems of CpGs and solved numerically in seconds to give accurate predictions for statistics of interest. Our model is based upon the same, or similar, reaction systems used in previous stochastic modelling studies [14, 28, 29], but can be used to study larger systems of CpGs than previously considered. This makes our MF model far better suited to understanding the mechanisms underpinning megabase-sized variations in DNA methylation observed in development, age-

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ing and cancer, which occur at a scale three orders of magnitude larger than promoters [27]. To our knowledge, the largest system previously examined mathematically contained 10^5 CpGs [15]. However, here simulations were conducted for only a single model parameter set. Running large-scale simulations of this type for many parameter sets will result in computational bottlenecks, meaning that such simulations cannot be used for inference. A previous study has proposed a method, based on the generalized method of moments, for rapid inference using DNA methylation patterns [47]. However, the largest system tackled with this approach contains 10 CpGs. Here we show that our OPMF model can, in principle, be used for accurate, time-efficient inference when modelling arbitrarily large genomic regions.

When used for inference, the OPMF model has a higher sensitivity to methylation strength (y) than collaborativity strength (x) explaining why the former is generally better inferred than the latter. However, some posteriors obtained in Section 4.3 are very wide and/or the true parameter values lie outside the inferred 95% CIs, indicating that there is scope for inference to be improved. Since our model shows impressive performance in forward prediction, discrepancies between true and inferred parameters are likely due to insufficient data or the inference technique used. However, our analysis demonstrates that the OPMF model can in principle be used for inference on DNA methylation data. It's accuracy may be improved in future comprehensive studies by experimenting with different inference techniques and sample sizes. In addition, we have inferred parameters using only the mean and variance in methylation state, which can be estimated from standard short-read data. The recent application of long-read technologies to assay DNA methylation patterns [48] could enable the computation of higher order statistics from experimental data, such as the correlation in methylation state between neighbouring sites. Using these additional summary statistics for inference could improve results.

Here we assume that the processes governing the creation of methylation patterns *in vivo* are described well by our nearest-neighbour collaborative model. It is possible that collaborativity *in vivo* can occur between non-nearest-neighbours, something which is not explicitly accounted for in our nearest-neighbour collaborative model. Collaborative methylation interactions are likely determined by the properties of the DNA methylation machinery. DNMT1 and DNMT3B both methylate processively along DNA strands whereas DNMT3A methylates in a distributive manner but can form multimers along the DNA fibre [30]. However, the range and strengths across which these interactions occur is currently unclear so we focus on nearest-neighbour interactions. We note that our OPMF model does capture interactions beyond nearest neighbours because the mean state of pairs in the system influences the change in state of CpGs. Previous modelling studies have also considered different forms of collaborative interactions between CpGs in a system (Fig. 12a,b). In [14] collaboration between any CpGs in the system and nearest-neighbour collaborative methylation alongside distance-dependent collaborative demethylation were both demonstrated to produce stable CpG clusters that were either methylated or unmethylated. Stable clusters were also observed under a distance-dependent collaborative model, where collaborative demethylation dominates over short ranges and collaborative methylation dominates over long ranges [28].

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

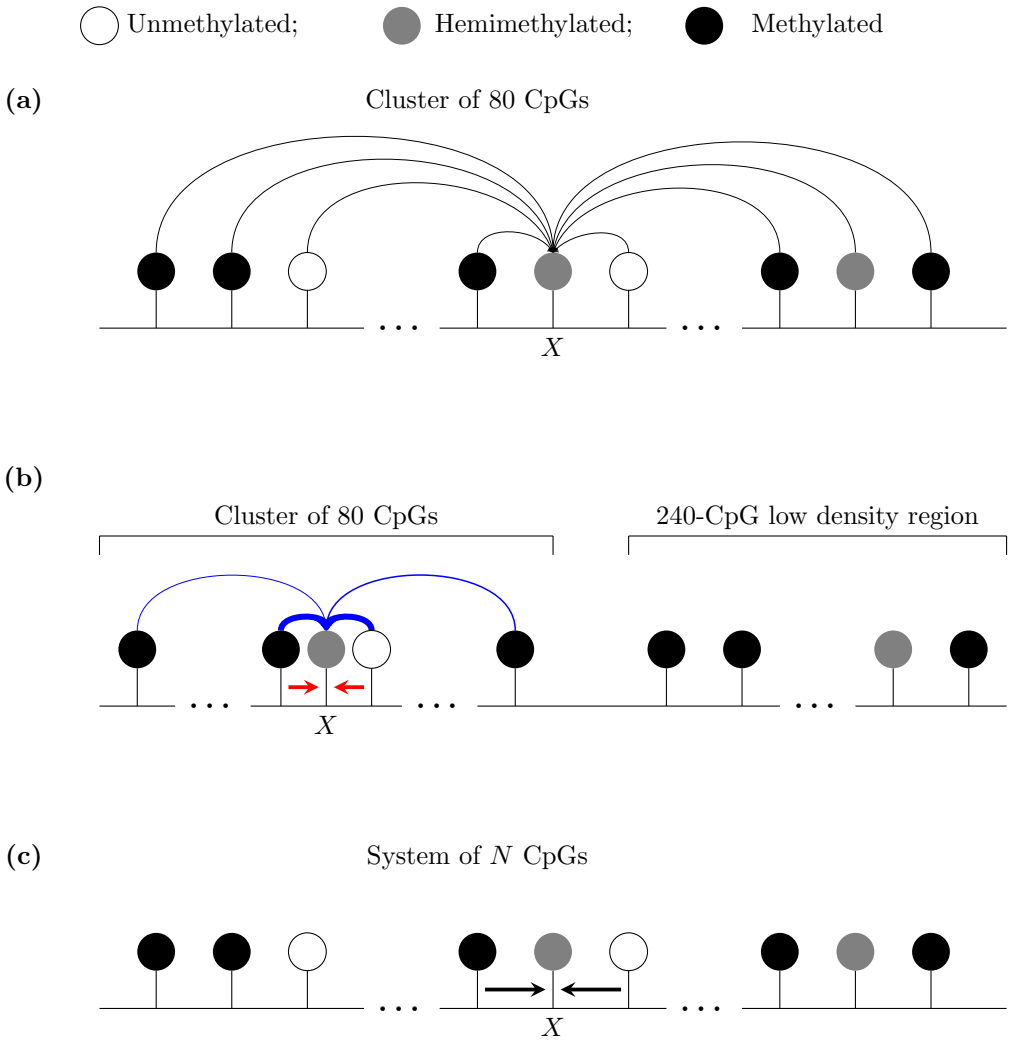

Figure 12: Potential collaborative interactions that can influence a target, X, under the models in [14] and the model considered here. (a) A cluster of 80 CpGs is first considered in [14], where a CpG can collaborate with any other CpG in the system with equal probability. (b) A high-density cluster (of 80 CpGs) adjacent to a highly methylated low-density region (of 240 CpGs) is then considered in [14], where there is nearest-neighbour collaborative methylation (red arrows) and the probability of collaborative demethylation occurring due to interaction between two sites decays as the distance between them increases (blue arrows; decay in reaction probability shown by narrowing width of arrows). Note that collaborative demethylation is restricted to the 80-CpG cluster. (c) In this paper collaborative reactions only occur between neighbouring CpGs and there is no upper bound on the system size, allowing large-scale patterns to be considered.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Here we have also assumed that the system of CpGs we model reaches a steady state. This means that we consider either non-dividing cells, or dividing cells which settle down to steady state between replication events such that DNA replication has no long-term effects on DNA methylation patterns. Whether or not a cell satisfies the latter case is dependent on the real magnitudes of k_i, $i = \{1, \ldots, 12\}$, and the time between replication events. Experimental studies show that arrested cells have similar DNA methylation patterns to those that are cycling, supporting the assumption that DNA replication has no long-term effect on DNA methylation [31]. Furthermore, an analysis of DNA methylation patterns on newly synthesised DNA suggests that re-methylation occurs within 20 minutes of replication [49]. However, another analysis of DNA methylation following replication suggests re-methylation is often delayed [50]. At present, it is unclear whether this delay is sufficient to have an effect on methylation patterns during the following cell cycle.

Our assumption that k_i, $i = \{1, \ldots, 12\}$ take the form in Table 1 could be violated in reality. For example, DNMT1 shows a strong preference for h sites over u sites [30], meaning that methylation reactions with an h target may have higher reaction rates than those with a u target. Future work could relax rate assumptions to account for such factors. Preliminary investigations confirm that the OPMF model provides a good approximation to the nearest-neighbour collaborative system when k_i, $i = \{1, \ldots, 12\}$ are considered as twelve independent parameters (data not shown). However, the difficulty of parameter inference increases with the number of parameters, meaning that relaxing parameter assumptions will likely decrease inference quality. Nonetheless, our model suggests that it is the parameters x and y that determine steady-state methylation patterns, rather than individual reaction rates. This is supported by a study where modelling of experimental data suggested that the ratio between methylation and demethylation rates determines steady-state methylation levels at single CpGs [51].

Here, we demonstrate that MF models can accurately predict the behaviour of large CpG systems subjected to nearest-neighbour collaboration. Our study presents the first mathematical modelling approach that can be applied to arbitrarily large systems of CpGs. The future application of this approach will facilitate the delineation of the methylation dynamics that underpin the formation of large-scale methylation patterns in developmental and disease contexts.

6 Acknowledgements

We thank Chris Ponting, Diego Oyarzun and members of the Grima and Sproul lab for helpful discussions about the manuscript. LK is a cross-disciplinary post-doctoral fellow supported by funding from the University of Edinburgh and Medical Research Council (MC_UU_00009/2). DS is a Cancer Research UK Career Development fellow (reference C47648/A20837), and work in his laboratory is also supported by an MRC university grant to the MRC Human Genetics Unit. RG is supported by Leverhulme Trust research awards (RPG-2018-423 and RPG-2020-327).

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

7 Contributions

LK conducted the modelling and analyses presented in the manuscript. DS and RG planned and supervised the study. LK, DS and RG wrote the manuscript.

8 Competing interests

The authors declare no competing interests.

Appendix

A Derivation of effective reaction rates for the DPMF model

Here we illustrate how $\hat{a}_1$, the reaction rate associated with $mm \to hm$ in (18), is constructed. We first consider all reactions in Fig. 1 that can occur *within* the mm pair, resulting in conversion to an hm pair. Clearly, $\hat{a}_1$ must contain a $2k_3$ term since $mm \to hm$ occurs if either m undergoes $m \xrightarrow{k_3} h$. Note that $mm \to hm$ cannot occur due to a collaborative interaction between the two m sites in mm.

Next, we consider all $mm \to hm$ reactions that can occur due to an m within the mm interacting with a site from an adjacent pair. For example, an m could collaborate with a u from an adjacent pair via $m + u \xrightarrow{k_{10}} h + u$, i.e. we can have

$$mm + uu \to hm + uu, \qquad mm + um \to hm + um, \qquad mm + uh \to hm + uh. \tag{24}$$

The first reaction can be written as an effective first-order reaction, $mm \xrightarrow{2k_{10}\mu_2} hm$, where the uu is absorbed into the reaction rate by making it proportional to μ_2, and the factor of two allows for either m within the mm to undergo this reaction. The second reaction in (24) can be written as $mm \xrightarrow{k_{10}\mu_4} hm$, where the um is absorbed into the reaction rate and either m can undergo the reaction, giving us a factor of two. However, this reaction requires the u within the um to be directly adjacent to the mm. The probability of having um in this particular order is $\frac{1}{2}$, giving an effective reaction rate of $2k_{10}\frac{1}{2}\mu_4 = k_{10}\mu_4$. Similarly, the third reaction in (24) can be written as $mm \xrightarrow{k_{10}\mu_6} hm$.

Finally, an m within the mm can interact with an h from an adjacent pair, via $m+h \xrightarrow{k_9} h+h$, i.e. we can have

$$mm + hh \to hm + hh, \qquad mm + hm \to hm + hm, \qquad mm + uh \to hm + uh. \tag{25}$$

Using similar arguments as above, we can write each of these three reactions as the effective first-order reaction $mm \to hm$, with rates $2k_9\mu_3$, $k_9\mu_5$ and $k_9\mu_6$, respectively.

Hence,

$$\hat{a}_1 = 2k_3 + k_9(2\mu_3 + \mu_5 + \mu_6) + k_{10}(2\mu_2 + \mu_4 + \mu_6). \tag{26}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

B Derivation of effective reaction rates for the OPMF model

We now construct $\tilde{a}_1$, the reaction rate associated with $mm \to hm$ in (23). As with the DPMF model, reactions occurring within the mm pair contribute a $2k_3$ term to $\tilde{a}_1$.

Now, $mm \to hm$ can occur due to a $m + u \xrightarrow{k_{10}} h + u$ reaction if one of the m sites in the mm pair is also in a pair with a u site, i.e. we can have

$$mm + um \to hm + um,$$

where now the um and mm reactants share a common m. This can be written as the effective first-order reaction $mm \to hm$ with rate

$$2k_{10}\left(\frac{\mu_4/2}{\mu_1 + \mu_4/2 + \mu_5/2}\right), \tag{27}$$

where $\frac{\mu_4/2}{\mu_1+\mu_4/2+\mu_5/2}$ is the probability that an m within the mm also forms a um with its other neighbour, i.e. it is the conditional probability that a pair is in the um state given that we know a particular site in the pair is m. Recall that μ_4 gives the proportion of um and mu pairs, while μ_5 gives the proportion of hm and mh pairs. The factors of $1/2$ in (27) account for the fact that the u and h in the um and hm must be on a particular side of the m (since an m is on its other side). The factor of two at the front of (27) allows for either m in the mm to undergo the reaction.

Similarly, $mm \to hm$ can occur due to a $m + h \xrightarrow{k_9} h + h$ reaction if one of the m sites in the mm is also in a pair with a h site, i.e. we can have

$$mm + hm \to hm + hm,$$

where the reactant hm and mm share a common m. This reaction can again be written as $mm \to hm$, where the rate,

$$2k_9\left(\frac{\mu_5/2}{\mu_1 + \mu_4/2 + \mu_5/2}\right),$$

is derived in a similar way as above.

Hence, we have

$$\tilde{a}_1 = 2k_3 + 2k_9\left(\frac{\mu_5/2}{\mu_1 + \mu_4/2 + \mu_5/2}\right) + 2k_{10}\left(\frac{\mu_4/2}{\mu_1 + \mu_4/2 + \mu_5/2}\right). \tag{28}$$

C Moment equations for the cluster MF models

Following the approach in Ref [33] (see Appendix C) we obtain first moment equations for the DPMF and OPMF models. Here μ_1—μ_6 are the mean levels of each paired state. Note that, for $i = 1, \ldots, 12$, $a_i = \hat{a}_i$ in the DPMF model and $a_i = \tilde{a}_i$ in the OPMF model,

$$\frac{d\mu_1}{dt} = -a_1\mu_1 + a_8\mu_5,$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$
\begin{aligned}
\frac{d\mu_2}{dt} &= -a_2\mu_2 + a_{11}(1-\mu_1-\mu_2-\mu_3-\mu_4-\mu_5),\\
\frac{d\mu_3}{dt} &= -(a_3+a_4)\mu_3 + a_9\mu_5 + a_{10}(1-\mu_1-\mu_2-\mu_3-\mu_4-\mu_5),\\
\frac{d\mu_4}{dt} &= -(a_5+a_6)\mu_4 + a_7\mu_5 + a_{12}(1-\mu_1-\mu_2-\mu_3-\mu_4-\mu_5),\\
\frac{d\mu_5}{dt} &= a_1\mu_1 + a_4\mu_3 + a_5\mu_4 - (a_7+a_8+a_9)\mu_5.
\end{aligned}
$$

Note again that we can obtain μ_6 via $\mu_6 = 1-\mu_1-\mu_2-\mu_3-\mu_4-\mu_5$. The second moment equations, describing the evolution of $\langle L_iL_j\rangle$ for $1 \le i,j \le 5$, are given by

$$
\begin{aligned}
\frac{d\langle L_1L_1\rangle}{dt} &= a_1\mu_1 + a_8\mu_5 + 2\Big(a1\langle L_1L_1\rangle + a_8\langle L_1L_5\rangle\Big),\\
\frac{d\langle L_2L_2\rangle}{dt} &= a_2\mu_2 + a_{11}(1-\mu_1-\mu_2-\mu_3-\mu_4-\mu_5)\\
&\quad + 2\Big(-a_2\langle L_2L_2\rangle + a_{11}(\mu_2 - \langle L_1L_2\rangle - \langle L_2L_2\rangle - \langle L_2L_3\rangle - \langle L_2L_4\rangle - \langle L_2L_5\rangle)\Big),\\
\frac{d\langle L_3L_3\rangle}{dt} &= (a_3+a_4)\mu_3 + a_9\mu_5 + a_{10}(1-\mu_1-\mu_2-\mu_3-\mu_4-\mu_5)\\
&\quad + 2\Big(-(a_3+a_4)\langle L_3L_3\rangle + a_9\langle L_3L_5\rangle + a_{10}\big(\mu_3 - \langle L_1L_3\rangle - \langle L_2L_3\rangle - \langle L_3L_3\rangle - \langle L_3L_4\rangle - \langle L_3L_5\rangle\big)\Big),\\
\frac{d\langle L_4L_4\rangle}{dt} &= (a_5+a_6)\mu_4 + a_7\mu_5 + a_{12}(1-\mu_1-\mu_2-\mu_3-\mu_4-\mu_5)\\
&\quad + 2\Big(-(a_5+a_6)\langle L_4L_4\rangle + a_7\langle L_4L_5\rangle + a_{12}\big(\mu_4 - \langle L_1L_4\rangle - \langle L_2L_4\rangle - \langle L_3L_4\rangle - \langle L_4L_4\rangle - \langle L_4L_5\rangle\big)\Big),\\
\frac{d\langle L_5L_5\rangle}{dt} &= a_1\mu_1 + a_4\mu_3 + a_5\mu_4 + (a_7+a_8+a_9)\mu_5\\
&\quad + 2\Big(a_1\langle L_1L_5\rangle + a_4\langle L_3L_5\rangle + a_5\langle L_4L_5\rangle - (a_7+a_8+a_9)\langle L_5L_5\rangle\Big),\\
\frac{d\langle L_1L_2\rangle}{dt} &= -(a_1+a_2)\langle L_1L_2\rangle + a_8\langle L_2L_5\rangle + a_{11}\big(\mu_1 - \langle L_1L_1\rangle - \langle L_1L_2\rangle - \langle L_1L_3\rangle - \langle L_1L_4\rangle - \langle L_1L_5\rangle\big),\\
\frac{d\langle L_1L_3\rangle}{dt} &= -(a_1+a_3+a_4)\langle L_1L_3\rangle + a_8\langle L_3L_5\rangle + a_9\langle L_1L_5\rangle\\
&\quad + a_{10}\big(\mu_1 - \langle L_1L_1\rangle - \langle L_1L_2\rangle - \langle L_1L_3\rangle - \langle L_1L_4\rangle - \langle L_1L_5\rangle\big),\\
\frac{d\langle L_1L_4\rangle}{dt} &= -(a_1+a_5+a_6)\langle L_1L_4\rangle + a_7\langle L_1L_5\rangle + a_8\langle L_4L_5\rangle\\
&\quad + a_{12}\big(\mu_1 - \langle L_1L_1\rangle - \langle L_1L_2\rangle - \langle L_1L_3\rangle - \langle L_1L_4\rangle - \langle L_1L_5\rangle\big),\\
\frac{d\langle L_1L_5\rangle}{dt} &= -a_1\mu_1 - a_8\mu_5 + a_1\langle L_1L_1\rangle + a_4\langle L_1L_3\rangle + a_5\langle L_1L_4\rangle + a_8\langle L_5L_5\rangle\\
&\quad - (a_1+a_7+a_8+a_9)\langle L_1L_5\rangle,\\
\frac{d\langle L_2L_3\rangle}{dt} &= -(a_2+a_3+a_4)\langle L_2L_3\rangle + a_9\langle L_2L_5\rangle\\
&\quad + a_{10}\big(\mu_2 - \langle L_1L_2\rangle - \langle L_2L_2\rangle - \langle L_2L_3\rangle - \langle L_2L_4\rangle - \langle L_2L_5\rangle\big)\\
&\quad\quad + a_{11}\big(\mu_3 - \langle L_1L_3\rangle - \langle L_2L_3\rangle - \langle L_3L_3\rangle - \langle L_3L_4\rangle - \langle L_3L_5\rangle\big),\\
\frac{d\langle L_2L_4\rangle}{dt} &= -(a_2+a_5+a_6)\langle L_2L_4\rangle + a_7\langle L_2L_5\rangle\\
&\quad + a_{11}\big(\mu_4 - \langle L_1L_4\rangle - \langle L_2L_4\rangle - \langle L_3L_4\rangle - \langle L_4L_4\rangle - \langle L_4L_5\rangle\big)\\
&\quad\quad + a_{12}\big(\mu_2 - \langle L_1L_2\rangle - \langle L_2L_2\rangle - \langle L_2L_3\rangle - \langle L_2L_4\rangle - \langle L_2L_5\rangle\big),\\
\frac{d\langle L_2L_5\rangle}{dt} &= a_1\langle L_1L_2\rangle - (a_2+a_7+a_8+a_9)\langle L_2L_5\rangle + a_4\langle L_2L_3\rangle + a_5\langle L_2L_4\rangle
\end{aligned}
$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$
\begin{aligned}
&+ a_{11}\big(\mu_5 - \langle L_1L_5\rangle - \langle L_2L_5\rangle - \langle L_3L_5\rangle - \langle L_4L_5\rangle - \langle L_5L_5\rangle\big),\\
\frac{d\langle L_3L_4\rangle}{dt} &= -(a_3+a_4+a_5+a_6)\langle L_3L_4\rangle + a_7\langle L_3L_5\rangle + a_9\langle L_4L_5\rangle\\
&\quad + a_{10}\big(\mu_4 - \langle L_1L_4\rangle - \langle L_2L_4\rangle - \langle L_3L_4\rangle - \langle L_4L_4\rangle - \langle L_4L_5\rangle\big)\\
&\quad a_{12}\big(\mu_3 - \langle L_1L_3\rangle - \langle L_2L_3\rangle - \langle L_3L_3\rangle - \langle L_3L_4\rangle - \langle L_3L_5\rangle\big),\\
\frac{d\langle L_3L_5\rangle}{dt} &= -a_4\mu_3 + a_9\mu_5 + a_1\langle L_1L_3\rangle + a_4\langle L_3L_3\rangle + a_5\langle L_3L_4\rangle - (a_3+a_4+a_7+a_8+a_9)\langle L_3L_5\rangle\\
&\quad + a_9\langle L_5L_5\rangle + a_{10}\big(\mu_5 - \langle L_1L_5\rangle - \langle L_2L_5\rangle - \langle L_3L_5\rangle - \langle L_4L_5\rangle - \langle L_5L_5\rangle\big),\\
\frac{d\langle L_4L_5\rangle}{dt} &= -a_5\mu_4 - a_7\mu_5 + a_1\langle L_1L_4\rangle + a_4\langle L_3L_4\rangle + a_5\langle L_4L_4\rangle - (a_5+a_6+a_7+a_8+a_9)\langle L_4L_5\rangle\\
&\quad a_7\langle L_5L_5\rangle + a_{12}\big(\mu_5 - \langle L_1L_5\rangle - \langle L_2L_5\rangle - \langle L_3L_5\rangle - \langle L_4L_5\rangle - \langle L_5L_5\rangle\big).
\end{aligned}
$$

Using $L_6 = 1 - L_1 - L_2 - L_3 - L_4 - L_5$, we obtain

$$
\begin{aligned}
\langle L_1L_6\rangle &= \mu_1 - \langle L_1L_1\rangle - \langle L_1L_2\rangle - \langle L_1L_3\rangle + \langle L_1L_4\rangle + \langle L_1L_5\rangle,\\
\langle L_2L_6\rangle &= \mu_2 - \langle L_1L_2\rangle - \langle L_2L_2\rangle - \langle L_2L_3\rangle + \langle L_2L_4\rangle + \langle L_2L_5\rangle,\\
\langle L_3L_6\rangle &= \mu_3 - \langle L_1L_3\rangle - \langle L_2L_3\rangle - \langle L_3L_3\rangle + \langle L_3L_4\rangle + \langle L_3L_5\rangle,\\
\langle L_4L_6\rangle &= \mu_4 - \langle L_1L_4\rangle - \langle L_2L_4\rangle - \langle L_3L_4\rangle + \langle L_4L_4\rangle + \langle L_4L_5\rangle,\\
\langle L_5L_6\rangle &= \mu_5 - \langle L_1L_5\rangle - \langle L_2L_5\rangle - \langle L_3L_5\rangle + \langle L_4L_5\rangle + \langle L_5L_5\rangle,\\
\langle L_6L_6\rangle &= 1 - 2(\mu_1+\mu_2+\mu_3+\mu_4+\mu_5) + \langle L_1L_1\rangle + \langle L_2L_2\rangle + \langle L_3L_3\rangle + \langle L_4L_4\rangle + \langle L_5L_5\rangle\\
&\quad + 2\Big(\langle L_1L_2\rangle + \langle L_1L_3\rangle + \langle L_1L_4\rangle + \langle L_1L_5\rangle + \langle L_2L_3\rangle\\
&\qquad + \langle L_2L_4\rangle + \langle L_2L_5\rangle + \langle L_3L_4\rangle + \langle L_3L_5\rangle + \langle L_4L_5\rangle\Big).
\end{aligned}
$$

References

[1] Schübeler D. Function and information content of DNA methylation. Nature. 2015;517(7534):321–326.

[2] Smallwood SA, Kelsey G. De novo DNA methylation: a germ cell perspective. Trends in Genetics. 2012;28(1):33–42.

[3] Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell. 1999;99(3):247–257.

[4] Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nature reviews genetics. 2008;9(6):465–476.

[5] Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. 1992;69(6):915–926.

[6] Wu X, Zhang Y. TET-mediated active DNA demethylation: mechanism, function and beyond. Nature Reviews Genetics. 2017;18(9):517–534.

[7] Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nature Reviews Genetics. 2013;14(3):204–220.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[8] Heyn P, Logan CV, Fluteau A, Challis RC, Auchynnikava T, Martin CA, et al. Gain-of-function DNMT3A mutations cause microcephalic dwarfism and hypermethylation of Polycomb-regulated regions. Nature genetics. 2019;51(1):96–105.

[9] Xu GL, Bestor TH, Bourc'his D, Hsieh CL, Tommerup N, Bugge M, et al. Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. Nature. 1999;402(6758):187–191.

[10] Winkelmann J, Lin L, Schormair B, Kornum BR, Faraco J, Plazzi G, et al. Mutations in DNMT1 cause autosomal dominant cerebellar ataxia, deafness and narcolepsy. Human molecular genetics. 2012;21(10):2205–2210.

[11] Sproul D, Meehan RR. Genomic insights into cancer-associated aberrant CpG island hypermethylation. Briefings in functional genomics. 2013;12(3):174–190.

[12] Jones MJ, Goodman SJ, Kobor MS. DNA methylation and healthy human aging. Aging cell. 2015;14(6):924–932.

[13] Gaudet F, Hodgson JG, Eden A, Jackson-Grusby L, Dausman J, Gray JW, et al. Induction of tumors in mice by genomic hypomethylation. Science. 2003;300(5618):489–492.

[14] Haerter JO, Lövkvist C, Dodd IB, Sneppen K. Collaboration between CpG sites is needed for stable somatic inheritance of DNA methylation states. Nucleic acids research. 2014;42(4):2235–2244.

[15] Busto-Moner L, Morival J, Ren H, Fahim A, Reitz Z, Downing TL, et al. Stochastic modeling reveals kinetic heterogeneity in post-replication DNA methylation. PLoS computational biology. 2020;16(4):e1007195.

[16] Adam S, Anteneh H, Hornisch M, Wagner V, Lu J, Radde NE, et al. DNA sequence-dependent activity and base flipping mechanisms of DNMT1 regulate genome-wide DNA methylation. Nature communications. 2020;11(1):1–15.

[17] Wang Q, Yu G, Ming X, Xia W, Xu X, Zhang Y, et al. Imprecise DNMT1 activity coupled with neighbor-guided correction enables robust yet flexible epigenetic inheritance. Nature Genetics. 2020;52(8):828–839.

[18] Mc Auley MT, Mooney KM, Salcedo-Sora JE. Computational modelling folate metabolism and DNA methylation: implications for understanding health and ageing. Briefings in bioinformatics. 2018;19(2):303–317.

[19] Gillespie DT. Stochastic simulation of chemical kinetics. Annu Rev Phys Chem. 2007;58:35–55.

[20] Johnston I. The chaos within: exploring noise in cellular biology. Significance. 2012;9(4):17–21.

[21] Schnoerr D, Sanguinetti G, Grima R. Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. Journal of Physics A: Mathematical and Theoretical. 2017;50(9):093001.

[22] Jenkinson G, Abante J, Feinberg AP, Goutsias J. An information-theoretic approach to the modeling and analysis of whole-genome bisulfite sequencing data. BMC bioinformatics. 2018;19(1):1–23.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[23] Jenkinson G, Pujadas E, Goutsias J, Feinberg AP. Potential energy landscapes identify the information-theoretic nature of the epigenome. Nature genetics. 2017;49(5):719–729.

[24] Zhang Y, Wang S, Wang X. Data-Driven-Based Approach to Identifying Differentially Methylated Regions Using Modified 1D Ising Model. BioMed research international. 2018;2018.

[25] Lück A, Wolf V. A Stochastic Automata Network Description for Spatial DNA-Methylation Models. In: International Conference on Measurement, Modelling and Evaluation of Computing Systems. Springer; 2020. p. 54–64.

[26] Lück A, Giehr P, Walter J, Wolf V. A stochastic model for the formation of spatial methylation patterns. In: International Conference on Computational Methods in Systems Biology. Springer; 2017. p. 160–178.

[27] Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nature genetics. 2018;50(4):591–602.

[28] Lövkvist C, Dodd IB, Sneppen K, Haerter JO. DNA methylation in human epigenomes depends on local topology of CpG sites. Nucleic acids research. 2016;44(11):5123–5132.

[29] Utsey K, Keener JP. A mathematical model for inheritance of DNA methylation patterns in somatic cells. Bulletin of Mathematical Biology. 2020;82(7):1–23.

[30] Jeltsch A, Jurkowska RZ. Allosteric control of mammalian DNA methyltransferases–a new regulatory paradigm. Nucleic acids research. 2016;44(18):8556–8575.

[31] Vandiver AR, Idrizi A, Rizzardi L, Feinberg AP, Hansen KD. DNA methylation is stable during replication and cell cycle arrest. Scientific reports. 2015;5(1):1–8.

[32] Gillespie DT. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of computational physics. 1976;22(4):403–434.

[33] Grima R. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. The Journal of chemical physics. 2012;136(15):04B616.

[34] Schreckenberg M, Schadschneider A, Nagel K, Ito N. Discrete stochastic models for traffic flow. Physical Review E. 1995;51(4):2939.

[35] Chowdhury D, Wang JS. Flow properties of driven-diffusive lattice gases: Theory and computer simulation. Physical Review E. 2002;65(4):046126.

[36] Jahnke T, Huisinga W. Solving the chemical master equation for monomolecular reaction systems analytically. Journal of mathematical biology. 2007;54(1):1–26.

[37] Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nature protocols. 2015;10(3):475–483.

[38] Zhao L, Sun Ma, Li Z, Bai X, Yu M, Wang M, et al. The dynamics of DNA methylation fidelity during mouse embryonic stem cell self-renewal and differentiation. Genome research. 2014;24(8):1296–1307.

bioRxiv preprint doi: https://doi.org/10.1101/2021.09.03.458837; this version posted September 3, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[39] Rossi RJ. Mathematical statistics: an introduction to likelihood based inference. John Wiley & Sons; 2018.

[40] Cao Z, Grima R. Accuracy of parameter estimation for auto-regulatory transcriptional feedback loops from noisy data. Journal of The Royal Society Interface. 2019;16(153):20180967.

[41] Golightly A, Wilkinson DJ. Bayesian inference for stochastic kinetic models using a diffusion approximation. Biometrics. 2005;61(3):781–788.

[42] Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002;162(4):2025–2035.

[43] Williams MJ, Werner B, Heide T, Curtis C, Barnes CP, Sottoriva A, et al. Quantification of subclonal selection in cancer from bulk sequencing data. Nature genetics. 2018;50(6):895–903.

[44] Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MP. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface. 2009;6(31):187–202.

[45] Tankhilevich E, Ish-Horowicz J, Hameed T, Roesch E, Kleijn I, Stumpf MP, et al. GpABC: a Julia package for approximate Bayesian computation with Gaussian process emulation. Bioinformatics. 2020;36(10):3286–3287.

[46] Wolkenhauer O, Wellstead P, Cho KH, Ingalls B. Sensitivity analysis: from model parameters to system behaviour. Essays in biochemistry. 2008;45:177–194.

[47] Lück A, Wolf V. Generalized Method of Moments Estimation for Stochastic Models of DNA Methylation Patterns. arXiv preprint arXiv:191101174. 2019.

[48] Gouil Q, Keniry A. Latest techniques to study DNA methylation. Essays in biochemistry. 2019;63(6):639–648.

[49] Xu C, Corces VG. Nascent DNA methylome mapping reveals inheritance of hemimethylation at CTCF/cohesin sites. Science. 2018;359(6380):1166–1170.

[50] Charlton J, Downing TL, Smith ZD, Gu H, Clement K, Pop R, et al. Global delay in nascent strand DNA methylation. Nature structural & molecular biology. 2018;25(4):327–332.

[51] Ginno PA, Gaidatzis D, Feldmann A, Hoerner L, Imanci D, Burger L, et al. A genome-scale map of DNA methylation turnover identifies site-specific dependencies of DNMT and TET activity. Nature communications. 2020;11(1):1–16.