Model-based whole-genome analysis of DNA methylation fidelity*

Christoph Bock[3,4,5], Luca Bortolussi[1,2], Thilo Krüger[1], Linar Mikeev[1], and Verena Wolf[1]

[1]Modelling and Simulation Group, University of Saarland, Germany
[2]DMG, University of Trieste, Italy
[3]CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria
[4]Department of Laboratory Medicine, Medical University of Vienna, Austria
[5]Max Planck Institute for Informatics, Saarbrücken, Germany
cbock@cemm.oeaw.ac.at
luca@dmi.units.it
{thilo.krueger,linar.mikeev,verena.wolf}@uni-saarland.de

Abstract. We consider the problem of understanding how DNA methylation fidelity, i.e. the preservation of methylated sites in the genome, varies across the genome and across different cell types. Our approach uses a stochastic model of DNA methylation across generations and trains it using data obtained through next generation sequencing. By training the model locally, i.e. learning its parameters based on observations in a specific genomic region, we can compare how DNA methylation fidelity varies genome-wide. In the paper, we focus on the computational challenges to scale parameter estimation to the whole-genome level, and present two methods to achieve this goal, one based on moment-based approximation and one based on simulation. We extensively tested our methods on synthetic data and on a first batch of experimental data.

Keywords: DNA methylation, Epigenomics, Branching processes, Parameter Estimation, Next Generation Sequencing

1 Introduction

Epigenetic marks such as DNA methylation provide a mechanism by which cells can control gene activity in a manner that is heritable between cell generations and adaptive to external stimuli [3]. Biochemically, DNA methylation is a covalent modification of the DNA. In vertebrates, DNA methylation occurs almost exclusively in the context of a cytosine (C) followed by a guanine (G). These so-called CpG sites are palindromic and can carry one DNA methylation group on

* L.B., T.K., L.M., and V.W. are partially funded by the German Research Council (DFG) as part of the Cluster of Excellence on Multimodal Computing and Interaction at Saarland University and the Collaborative Research Center SFB 1027. C.B. was supported by a New Frontiers Group award of the Austrian Academy of Sciences.

each strand. As the result, a single CpG site can be symmetrically unmethylated, asymmetrically methylated on either the forward or the reverse strand of the DNA (hemimethylated), or symmetrically methylated on both strands. In most cases, DNA methylation is symmetrical, which provides redundant information on both strands. When cells divide and the DNA is copied in a semi-conservative manner (i.e., each daughter cell receives one strand of the double-stranded DNA molecule), the DNA methylation on the newly synthesized strand can be reconstructed from the DNA methylation on the conserved DNA strand. The process of copying DNA methylation patterns is called maintenance methylation.

Compared to the very high efficiency with which the DNA sequence is copied and maintained during cell division (typically with error rates in the order of 10^{-8}), the fidelity of DNA methylation maintenance is much lower. Based on single-locus studies, error rates have been estimated to be in the order of 10^{-2} to 10^{-3}. To maintain high DNA methylation levels in specific regions of the genome despite the relatively low fidelity of maintenance DNA methylation, cells utilize a second mechanism called de novo methylation to methylate previously unmethylated cytosines independent of the DNA methylation status of the second cytosine within a CpG site. The rates of de novo methylation during normal cell growth are relatively low, but they appear to be sufficient to compensate for the gradual loss of DNA methylation that would normally result from the limited fidelity of DNA methylation maintenance.

Comprehensive genome-wide assessments of the fidelity of DNA methylation and of the de novo DNA methylation rate and their comparison between different cell types have been lacking, and prior work has focused on small parts of the genome. With genome-wide methods for DNA methylation mapping and analysis [5, 4], even in single cells [7], it is now possible to collect comprehensive datasets to estimate these important biological parameters in a genome-wide manner and to systematically search for differences between cell types. In this study, we address the computational challenges of inferring these parameters in a manner that is sufficiently high-throughput and scalable to support the genome-wide analysis of large numbers of samples.

The assumed experimental design is as follows: A single cell is isolated and left to grow exponentially over n generations, typically $n = 20$. The resulting cell population (approximately $2^{20} \approx 1$ million cells) is subjected to genome-scale bisulfite sequencing, the reads are aligned to the reference genome, and for each CpG site the number of methylated and unmethylated reads are counted for a subsample of the population (typically 10 to 100 measurements per CpG, with several million assayed CpGs). In these experiments, we are interested in the dynamic nature of the methylation process, in particular how it is propagated through the n cell generations. In order to understand this behaviour, we need a mathematical model of the methylation fidelity and of the de novo methylation rate, which must be trained using the experimental data available. In building the model, our main goal was to predict its parameters from data, which are directly connected to de novo methylation and fidelity probabilities. The model we construct is based on those proposed in [2, 12], and it is a discrete-time

Markov chain describing how the methylation progresses through generations at the population level (i.e. counting how many cells have a specific CpG site are unmethylated, hemi-methylated, or fully methylated).

The main challenge with this model is computational, as we need to perform the parameter estimation task genome-wide, for each CpG site and each biological sample. The model cannot be solved analytically, and it is too large for being solved with standard numerical techniques. Hence, we engineered two different strategies for computing the likelihoods required to train the model parameters, one based on an analytical approximation, and the other based on simulation. In this paper, we present and compare the two approaches, both theoretically and experimentally.

To validate the accuracy of the presented methods we simulated test data using our model, and we estimated the parameters for these test data sets. Furthermore, we used our the methods to estimate the parameters in real, experimentally derived, data sets.

The paper is organised as follows: in Section 2, we discuss the mathematical model, and in Section 3 we present the parameter estimation techniques. Preliminary results are shown and explained in detail in Section 4, and conclusions are drawn in Section 5.

2 Stochastic Model of DNA Methylation

We propose a stochastic model for the dynamics of DNA methylation of a cell population over a certain number of cell divisions. This model is an extension of previous models that have described the average state of a single CpG site within a certain cell population [2, 12].

Single cell model. To describe the DNA methylation dynamics of a single CpG site, we consider three possible site states: unmethylated on both DNA-strands (*unmethylated, U*), methylated on both strands (*fully methylated, F*) or methylated on one out of the two strands (*hemimethylated, H*). This naturally leads to a (discrete-time) Markov model description where one time step corresponds to one cell cycle or the time between two cell divisions (in cultured cells, this time is often on the order of 24h). Over the course of one cell cycle, the DNA methylation state of the CpG site changes in three phases. In phase one, the two strands of DNA are separated such that each daughter cell receives one strand, and a complementary stand is synthesised. This complementary strand is always unmethylated, such that this step dilutes the DNA methylation levels compared to the parent cell. Thus, the transitions for this phase are from U to U and from F to H with probability one, respectively, as well as from H to H or to U with probability 0.5, respectively (Fig. 1, left). In the second phase, which occurs during and after the synthesis of the new strand, a special class of enzymes, called DNA methyltransferases (DNMTs), try to maintain the pattern of the mother strand by methylating hemimethylated CpG sites. Maintenance methylation is a stochastic process [2], such that the state of a site changes

from H to F with probability f_m and from U to U with probability f_u. Successful maintenance typically occurs with a relatively high probability. However, in both cases maintenance might fail with probability $1 - f_m$ (transition from H to H) and with probability $1 - f_u$ (transition from U to H). The third phase, which lasts from the end of a cell division to the beginning of the next, allows for de novo methylation, where methyl groups are transferred by DNMTs to sites that are in state U or H. Here, the assumption is that de novo methylation occurs at a given site and strand independently of the DNA methylation state of the CpG on the other strand [2]. Thus, with probability μ the state changes either from H to F or from U to H (Fig. 1, left). Note that we neglect the extremely rare transition from U to F through de novo methylation, in order to keep the model simple. Simulations of the model show no significant differences if the transition from U to F due to de novo methylation is added (results not shown).

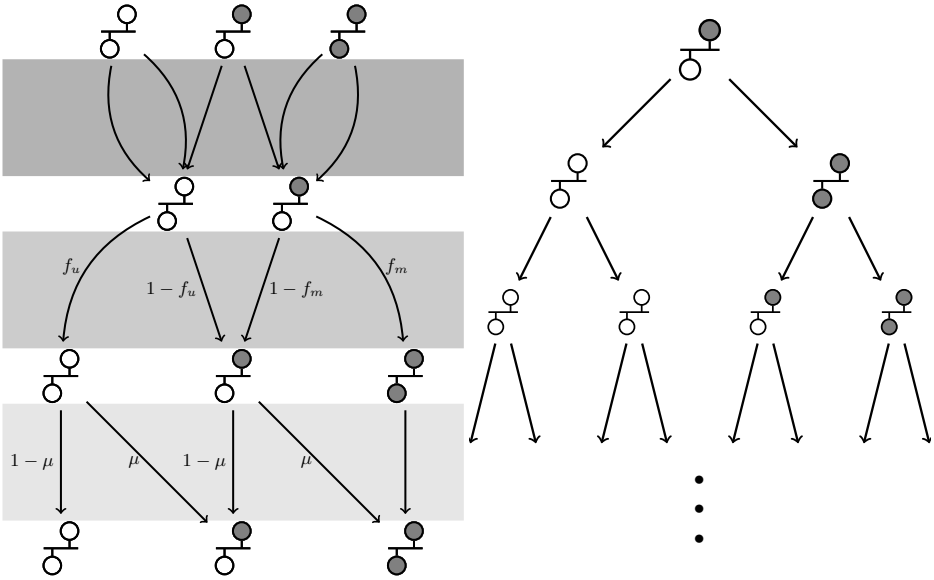


Fig. 1: The three phases of the single cell model (left) with cell-division (dark grey, the division of single cells into two new cells is represented by two arrows.), maintenance methylation (grey), and de novo methylation (light grey) as well as the population model (right) over 20 generations

Population model. In the proposed population model we consider a fixed CpG site in a single allele, thus modelling independently CpG sites in different alleles, and start initially with a single cell. Then, after the next time step we consider all daughter cells (two at time $t = 1$, four at time $t = 2$, etc.). Thus,

the resulting process is a branching process and after 20 generations we have to consider $2^{20} \approx 10^6$ cells. In order to describe the state of the CpG site in the cell population at time t, we compute the probabilities that, when a site in the parent cell is in state X, the two states of the daughter cells are Y and Z, where $X, Y, Z \in \{U, H, F\}$, according to the following matrix M

$$M = \begin{array}{c} \\ U \\ H \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} UU & UH & UF & HH & HF & FF \end{array} \\ \begin{pmatrix} t_{UU}^2 & 2 \cdot t_{UU}t_{UH} & 2 \cdot t_{UU}t_{UF} & t_{UH}^2 & 2 \cdot t_{UH}t_{UF} & t_{UF}^2 \\ 0 & t_{UU}t_{HH} & t_{UU}t_{HF} & t_{UH}t_{HH} & t_{UF}t_{HH} + t_{UH}t_{HF} & t_{UF}t_{HF} \\ 0 & 0 & 0 & t_{HH}^2 & 2 \cdot t_{HH}t_{HF} & t_{HF}^2 \end{pmatrix} \end{array},$$

with $(t_{UU}, t_{UH}, t_{UF}, t_{HH}, t_{HF}) = (f_u \cdot (1-\mu), (1-f_u)(1-\mu) + f_u\mu, (1-f_u)\mu, (1-f_m)(1-\mu), f_m + (1-f_m)\mu)$. Note that the entries for column UH considers the two symmetric cases that either the site in one daughter cell is in state U and the other one in state H or vice versa. The same holds for the columns UF and HF. The entries of M are computed by considering the corresponding paths in the diagram in Fig. 1, left. Consider for example the entry $M_{H,HF} = t_{UF}t_{HH} + t_{UH}t_{HF}$. During cell division a site in state H is divided into a site in state U and a site in state H (upper grey block in Fig. 1, left). Next we have to consider the paths from the nodes in the second line to those in the last line, i.e. from U to F ($t_{UF} = (1 - f_u) \cdot \mu$) and from H to H ($t_{HH} = (1 - f_m)(1 - \mu)$). Analogously, we consider the paths from U to H and from H to F which yields the second term $t_{UH}t_{HF}$.

In Fig. 1, right, we illustrate a trace of the population model over time. Assuming the state of the CpG site is H in the initial cell, then the two daughter cells of the next generation could be in state U and F, while in the following generation we could have states U,U,H and F, etc. After $n = 20$ generations, in the final population there are 2^{20} cells, in each of them, the state of the tracked CpG site will be in one of the three states. Note that we apply matrix M to all cells of the current generation to determine the possible daughter cells. In addition, we assume that DNA methylation in a cell and on a given allele occurs independently of other cells and alleles.

The model we are considering belongs to the well-known family of multi-type Galton-Watson branching process [8], giving us the possibility of exploiting the vast theory developed for them.

Our wet-lab data contain only information about how many methylated ($m = m(n)$) and unmethylated ($u = u(n)$) single strand sites are present in a subset of the final population (after 20 generations). However, this means that for the unknown full state we have the relationships

$$m(n) = Y_H(n) + 2 \cdot Y_F(n) \tag{1}$$

$$u(n) = Y_H(n) + 2 \cdot Y_U(n) \tag{2}$$

if $Y_U(n), Y_H(n)$, and $Y_F(n)$ are the numbers of unmethylated, hemimethylated, and fully methylated CpG sites in the final population. As $n = 20$ will be fixed, we will often omit this index it in the following.

3 Parameter Estimation

Since our ultimate goal is to do whole-genome studies and apply our model to different cell types, we are interested in parameter estimation procedures that are computationally efficient. Our model is parametric in $\Theta = (f_m, f_u, \mu)$ as well as in the state of the initial cell. In order to estimate these parameters we use a maximum likelihood approach and compute the likelihood of observing m and u (single strand observations). Computing this exactly is computationally very expensive due to the large size of the state space and the stiffness of the model (f_m and f_u being close to one). In the following we present two methods to approximate the maximum likelihood and estimate such parameters. The first one is based on stochastic simulations of the model and a statistical estimation of the likelihood of the observed data. The other approach uses a moment-based numerical method to approximate the likelihood.

Description of data. The wet-lab DNA methylation data comprise lists of integer pairs $\lambda = (u_e, m_e)$. Each pair describes the DNA methylation measurements for a given CpG site e, where one of the strands was observed u_e times unmethylated and m_e times methylated (the experimental setup does not allow to distinguish between upper and lower strand). Since it is known that certain groups of CpG sites behave similarly we will also use our model to describe the average behaviour of a CpG site within such a group and collect all observation pairs for these sites in a set Λ. If we only consider a single CpG site e, then $\Lambda = \{(u_e, m_e)\}$.

Likelihood for single data pairs. Consider a possible state of the model $\boldsymbol{Y} = \boldsymbol{Y}(n) = (Y_U(n), Y_H(n), Y_F(n))$ after $n = 20$ generations, whose entries sum up to $Y_U + Y_H + Y_F = 2^{20}$. From this vector it is possible to compute the numbers m and u of methylated and unmethylated strands (see equations 1 and 2). In the following, we use m_X and u_X to denote the number of methylated and unmethylated strands conditional on the site of the initial cell being in state $X \in \{U, H, F\}$ and we define the two relative frequencies $p_{uX} = \frac{u_X}{u_X + m_X}$ and $p_{mX} = \frac{m_X}{u_X + m_X} = 1 - p_{uX}$.

We now want to compute the likelihood that a certain data pair $\lambda = (u_e, m_e)$ is observed given the parameter set Θ. We assume that the measured cells are randomly chosen and therefore we can reduce the computation of the emission probabilities to the urn problem (drawing with replacement). Hence, the emission probability for observation λ, if the final state is $\boldsymbol{y} = (y_U, y_H, y_F)$, is given by

$$P_{X,\Theta}(\lambda \mid \boldsymbol{y}) = \binom{m_e + u_e}{m_e} (p_{uX})^{u_e} \cdot (p_{mX})^{m_e}. \tag{3}$$

Thus, the likelihood of the observation λ, conditional on a given initial state $X \in \{U, H, F\}$ and the parameters Θ, is

$$P(\lambda \mid X, \Theta) = \sum_{\boldsymbol{y}} P(\boldsymbol{Y}(n) = \boldsymbol{y}) \cdot P_{X,\Theta}(\lambda \mid \boldsymbol{y}) = \mathbb{E}_{\boldsymbol{Y}}[P(\lambda \mid X, \boldsymbol{Y})]. \tag{4}$$

The exact computation of the expectation $\mathbb{E}_{\boldsymbol{Y}}[P(\lambda \mid X, \boldsymbol{Y})]$ requires the knowledge of the probability of each possible final stage $\boldsymbol{y}$, which is computationally expensive. Therefore, we approximate such an expectation in two possible ways, either statistically relying on simulations of the model, or by stochastic approximation.

Simulation-based approach. An estimator for the likelihood $\mathbb{E}_{\boldsymbol{Y}}[P(\lambda \mid X, \boldsymbol{Y})]$ can be obtained by taking the sample mean of the emission probabilities of all trajectories.[1] We compute the sample mean that approximates the likelihood by generating 10000 trajectories using the method explained below. Note that during the optimisation process, we vary this number for performance reasons.

To generate a trajectory of the process, we use a standard simulation algorithm for discrete time Markov chains. The simulation is initialized by computing the matrix M (see Section 2) as well as setting the initial state $\boldsymbol{Y}(0)$, i.e. the state of one site of the single initial cell of the zeroth generation. Then in each step we determine the state of the site of the next generation according to the distributions in M. Instead of repeatedly generating the two daughter cells for all parent cells, we draw samples from a multinomial distribution according to the number of sites in state $X \in \{U, M, F\}$ and the probabilities in the corresponding line of matrix M. For instance, if the state of the initial site is H (see Fig. 1, right) we set the initial counting vector $\boldsymbol{Y}(0) = (Y_U(0), Y_H(0), Y_F(0)) = (0, 1, 0)$. In the next step a new counting vector, say $\boldsymbol{Y}(1) = (1, 0, 1)$ as in Fig. 1, right, is determined according to the multinomial distribution as described above. We iterate this process until we reach generation $n = 20$, thus obtaining a sample of the final state $\boldsymbol{Y}(n)$.

Moment-based approach. An alternative to simulation is to try to approximate the likelihood $\mathbb{E}_{\boldsymbol{Y}}[P(\lambda \mid X, \boldsymbol{Y})]$ by resorting to ideas of stochastic approximation. Our approach is conceptually simple: first, we compute the first two moments of the distribution of $\boldsymbol{Y} = \boldsymbol{Y}(n)$, conditional on the initial site state being $X \in \{U, H, F\}$, namely its mean $\mathbf{e}^X = \mathrm{E}[\boldsymbol{Y}(n)]$, and the covariance matrix $\mathbf{C}^X = (C_{ij})$, $C_{ij} = \mathrm{Cov}[Y_i(n), Y_j(n)]$, $i, j \in \{U, H, F\}$. Then, we assume $\boldsymbol{Y}$ takes continuous values rather than integer ones, and invoke the maximum entropy principle [1] to approximate it by a 2-dimensional normal distribution with mean $\mathbf{e}^X$ and covariance matrix $\mathbf{C}^X$ (we can get rid of one dimension exploiting the fact that the population at generation n equals 2^n). By letting $f_{\mathbf{e}^X, \mathbf{C}^X}$ be the corresponding normal density, we then have

$$P(\lambda \mid X, \Theta) \approx \int_{u,h} \binom{m_e + u_e}{m_e} \left(\frac{y_U + 0.5 y_H}{2^{20}}\right)^{u_e} \left(\frac{1 - y_U - 0.5 y_H}{2^{20}}\right)^{m_e} f_{\mathbf{e}^X, \mathbf{C}^X}(\boldsymbol{y})\, d\boldsymbol{y}.$$

This integral is then numerically approximated by using the two-dimensional Simpson's rule [6].

[1] Note that the emission probabilities are dependent on the relative frequencies p_{uX} and p_{mX}, which are random variables as they depend on the random quantities u_X and m_X.

In order to compute mean and covariance of $\boldsymbol{Y}(n) = (Y_U(n), Y_H(n), Y_F(n))$, $n = 0 \dots 20$, we exploit the fact that $\boldsymbol{Y}$ is a multi-type Galton-Watson branching process [8]. Following [11], we define the expectation matrix $\mathbf{M}$ with elements

$$M_{ij} = \mathrm{E}[Y_j(1) \mid Y(0) = \mathbf{b}^i], \tag{5}$$

where $i, j \in \{U, H, F\}$ and $\mathbf{b}^U = (1, 0, 0)$, $\mathbf{b}^H = (0, 1, 0)$, $\mathbf{b}^F = (0, 0, 1)$. We also define the covariance matrices $\mathbf{V}^k, k \in \{U, H, F\}$ such that

$$V^k_{ij} = \mathrm{Cov}[V_i(1), V_j(1) \mid \boldsymbol{Y}(0) = \mathbf{b}^k]. \tag{6}$$

Then, the following recurrence holds [11]:

$$[\mathbf{e}(n+1)\, \mathcal{C}(n+1)] = [\mathbf{e}(n)\, \mathcal{C}(n)] \left[\begin{array}{c|c} \mathbf{M} & \begin{array}{c} \mathcal{V}^U \\ \mathcal{V}^H \\ \mathcal{V}^F \end{array} \\ \hline \mathbf{0} & \mathbf{M} \times \mathbf{M} \end{array}\right] = [\mathbf{e}(n)\, \mathcal{C}(n)]\mathbf{T},$$

where $\mathbf{M} \times \mathbf{M}$ is the Kronecker product, $\mathcal{C}(n) = (C_{UU}(n), C_{UH}(n), C_{UF}(n), C_{HU}(n), C_{HH}(n), C_{HF}(n), C_{FU}(n), C_{FH}(n), C_{FF}(n))$ and $\mathcal{V}^i = (V^i_{UU}, V^i_{UH}, V^i_{UF}, V^i_{HU}, V^i_{HH}, V^i_{HF}, V^i_{FU}, V^i_{FH}, V^i_{FF})$. For each initial state $k \in \{U, H, F\}$ we also compute

$$[\mathbf{e}^k\, \mathcal{C}^k] = [\mathbf{b}^k\, \mathbf{0}]\mathbf{T}^{20}.$$

Estimating the initial state. The previously discussed approach allows us to compute the likelihood for a single pair λ conditional on the initial state $X \in \{U, H, F\}$. In order to estimate such an initial configuration, we consider the estimated likelihoods $P(\lambda \mid X, \Theta)$ in a Bayesian context. We start by assuming a prior distribution $P(X \mid \Theta)$ over the initial states, and then compute the posterior distribution $P(X \mid \lambda, \Theta)$ according to Bayes theorem as

$$P(X \mid \lambda, \Theta) = \frac{P(\lambda \mid X, \Theta) \cdot P(X \mid \Theta)}{\sum_{X \in \{U,H,F\}} P(\lambda \mid X, \Theta) \cdot P(X \mid \Theta)}.$$

In order to fix the prior, we need to take into account that it is unlikely that the original cell has a hemimethylated site (which is very uncommon for living cells), so the prior should give it a small probability for $X = H$. Our solution is to consider as prior probability the state of the model after one generation, starting from the distribution $(U\ H\ F) = (0.5\ 0\ 0.5)$. For instance, we have $P(H \mid \Theta) = t_{UH}(t_{UU} + t_{UH} + t_{UF}) + t_{HH}(t_{HH} + t_{HF})$ (see also Section 2). Then, we can compute the model likelihood, for a given λ, independent of initial conditions, as $P(\lambda \mid \Theta) = \sum_{X \in \{U,H,F\}} P(\lambda \mid X, \Theta) \cdot P(X \mid \Theta)$.

Likelihood optimisation. The model likelihood for all data pairs $\Lambda = \{\lambda_1, \lambda_2, \dots\}$ is finally obtained by taking the product of the likelihood of all individual pairs. By taking the logarithm, the model log-likelihood then is

$$\log(P(\Lambda \mid \Theta)) = \sum_{\lambda \in \Lambda} \log(P(\lambda \mid \Theta)).$$

To estimate the parameters we used a simple maximum likelihood approach. We computed the likelihoods $-\log(P(\Lambda \mid \Theta))$ for varying Θ and converged to a minimum using simple optimisation procedures. In the final version, we use the *Nelder-Mead* procedure which is a derivative-free optimisation that performed best in our tests [10].

4 Results

In order to validate the proposed estimation algorithms, we ran detailed tests with simulated data (Section 4.1). We also present preliminary results of the whole genome analysis based on real experimental data (Section 4.2).

4.1 Results for simulated data

Generation of synthetic data. In order to simulate realistic experimental data with our model, we need two additional parameters governing the behaviour of the experiment: *coverage*, which is the average number of measurements per CpG site, and *length*, which is the number of CpG sites in the simulated dataset Λ_{sim}. Given such information, synthetic experimental data is generated according to Algorithm 1. In order to vary the coverage and keep the variance of the coverage as realistic as possible, we determine for a fixed coverage the number of measurements per CpG site in such a way that it resembles this number in the truly measured data.[2]

1: prepare a list L_{real} of numbers of measurements per site as follows: choose randomly a sequence of measurement numbers from the real data with *length* entries and compute the average $\overline{C_{real}}$ over all entries of this list
2: set $\Lambda_{sim} = \emptyset$
3: **for** i := 1 **to** *length* **do**
4: draw probabilistically the initial state X (as described in Section 3)
5: run 20 generations from X
6: compute $p(methylated) = (\#methylated\ sites)/(\#sites)$
7: get C_{real} as the ith entry of L_{real}
8: compute $C = Round((C_{real} \cdot (coverage - 0.5))/(\overline{C_{real}}) + 0.5)$
9: draw a random number m from a binomial with $p = p(methylated)$ and C
10: add $\lambda = (m, C - m)$ to Λ_{sim}
11: **end for**

Algorithm 1: Generation of synthetic data.

[2] We avoid the number of measurements C to be set to zero by subtracting 0.5 from the reduced coverage and add 0.5 to the quantity to round (Algorithm 1, line 8).

Scanning the parameter space of simulated data sets. We first examine the likelihood landscape by deep sampling of the parameter space, fixing the coverage to 5 and the length to 1000, as these values are typical for some of the real data sets considered in the following section. We use the moment-based method described in Section 3 to approximate the likelihood.

We consider synthetic data obtained from the model with the arbitrarily chosen values of $\Theta_{sim} = (f_u, f_m, \mu)$ from Table 7. For each parameter set we generated a data set using Algorithm 1. To get an impression of the likelihood landscapes, we computed the likelihood for a fine grid of the parameter space Θ with the proposed approximative approach for parameter sets 1-3. In Fig. 2 we show the results. For better visualisation purposes, we report 2-dimensional plots. We represent for each pair of parameters the negative log-likelihood as a grey value and restrict to the maximum log-likelihood for each pair of parameters in the plot.

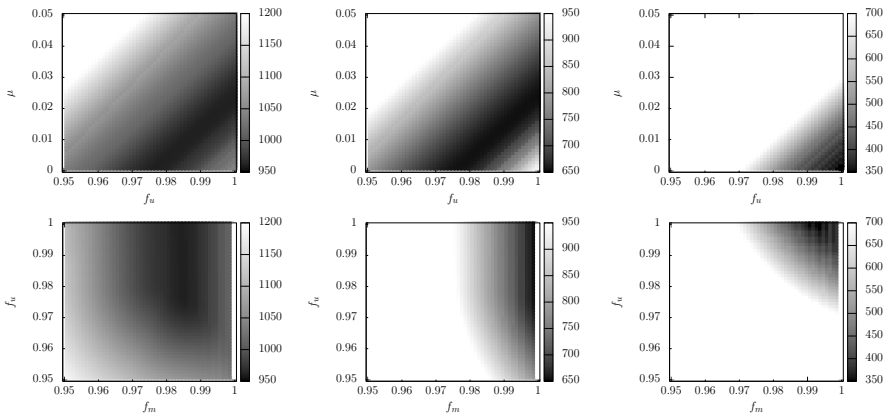
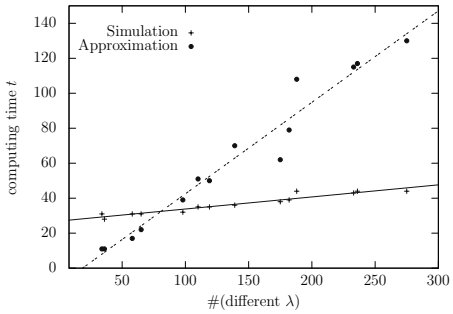


Fig. 2: Likelihood landscape for low coverage data. Parameter set 1 (left), 2 (middle), and 3 (right) from Eq. (7).

In all three cases the estimation resulted in a very flat likelihood landscape since a low coverage of 5 was used. For high coverage synthetic data (as we have it in the real data) the estimation was much more accurate (see below). In particular the sum $\alpha := \mu + (1 - f_m) + (1 - f_u)$, which reflects the probability of copy mistakes in the methylation pattern, is very close to the true value. Nevertheless, in the low coverage case for parameter set 1 we find high likelihood values if the sum $\mu + (1 - f_u) < 0.031$, in rough agreement with the value of the parameter set that was used to simulate the data. Also there is a tendency for f_m to be at a value > 0.97. In all three of the upper plots it can be seen that there is a dependency between f_u and μ. Increasing f_u seems to have the same effect on the likelihood as decreasing μ, which makes sense because in both cases

the U state is copied more often and it is less often the case that the daughter cell has state H.

The parameters of the second and third set resulted in observations that are similar to real data as maintenance typically occurs with high probability. In the case of parameter set 3 the likelihood increases significantly when crossing the line $\mu = f_u - 0.97$ and becomes maximal at $f_u = 1$ and $\mu = 0.001$, while f_m is estimated to be smaller than 0.999.



Θ_{sim}	f_u	f_m	μ	α
1	0.999	0.97	0.03	0.061
2	0.999	0.999	0.03	0.032
3	0.999	0.999	0.001	0.003
4	0.97	0.999	0.03	0.061
5	0.97	0.999	0.001	0.032
6	0.999	0.97	0.001	0.032

(7)

Fig. 3: Comparison of the performances of the simulation and the approximation approach (left). The parameter sets of the simulated data (right). Note that we also list $\alpha := \mu + (1 - f_m) + (1 - f_u)$ in this table, which is used as an indicator number later.

Comparison of simulation and approximation approaches Next we compare the two approaches for approximating the likelihood, namely the simulation-based and the moment-based methods explained in Section 3 . We used the same three parameter sets 1-3 from the previous section (see Tab. (7)) but instead of optimising the third parameter we fix $f_u = 0.999$ and plot the computed likelihood depending on μ and f_m. The results are shown in Fig. 4. It can be seen that the likelihoods obtained by the moment-based method (lower plots) are much smoother, being free from the random effects of the simulation approach (upper plots). Nevertheless, the results of both methods are very similar and the maximum likelihood points in the parameter spaces are very close. For example for parameter set 1, 87% of the log-likelihoods differ by not more than 10% from each other. Furthermore, for the plots in the middle of Fig. 4, both methods find as optimal f_m the true value of 0.999, while μ is optimal at 0.023 for the moment-based approach and 0.026 for the simulation approach (true value is 0.03). Note that the optimal parameters that are recovered with the different methods differ more for the plots in Fig 4, left and less for the plots in Fig. 4, right, due to the different kinds of likelihood landscapes. For performance reasons, we restrict ourselves to the simulative approach for the remainder of the paper. In fact, as soon as there are more than approximatively 80 different data pairs λ in one data set, the numerical method becomes slower than the simulation. For 80 different data pairs both methods need approximately 30 sec, while for a data set with 200 different data pairs, the simulation needs 40 sec and the numerical method

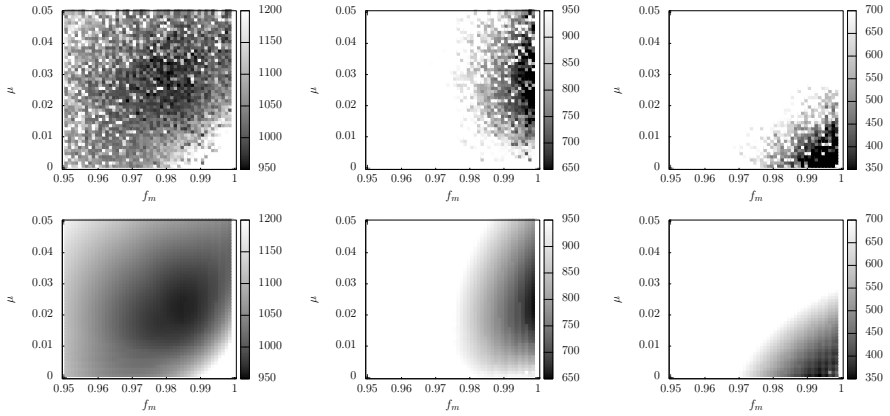


Fig. 4: Comparison between numerical (lower row of plots) and simulative (upper row of plots) approximation of the likelihood: Parameter set 1 (left), 2 (middle), and 3 (right) from Tab. (7).

needs 100 sec. (see Fig. 3 for a complete comparison). For huge data sets, which are common in real experimental data, the simulation would clearly outperform the numerical approach.

Parameter estimation for simulated data To explore the quality of our estimation procedure, data for all parameter sets Θ_{sim} listed in Tab. (7) with different coverages and lengths were generated. Then, we estimated ten times such parameters with the simulation approach (see Section 3). Since we have seen that the model cannot distinguish well between parameter sets with similar $\mu+(1-f_u)$, we concentrate in the following on the sum $\alpha = \mu+(1-f_m)+(1-f_u)$, which reflects the probability of copy mistakes in the methylation pattern. If α is zero, then each site in state U or F will be in state U or F in all daughter cells again. The higher α becomes, the more errors happen during one generation. The results for the (average of the ten) differences $\Delta\alpha$ between the estimated α and the true α with which the simulated data was generated is plotted as a function of the chosen lengths and coverages in Fig. 5.
It can be seen that the coverage of a certain data set plays a crucial role when estimating the parameters of the proposed model. For coverages between 16 and 64, a constant value for $\Delta\alpha$ is reached, which becomes greater again for coverages greater than 100. In contrast, raising the length of a data set is not leading to less accurate parameter estimations. From length 500 on the distances are converging. Since parameter estimation on real data is typically done for large genome regions with many CpG sites and each site produces one data pair of methylation data, it is very common to have a data set of at least 500 entries.
The estimated α when only a single observation λ is given, is obviously rather inaccurate. In Fig. 5 we see that the estimation based on ten observations gives

much better results. Since we only have one observation for each CpG site in the data of the real system and since CpG sites that belong to the same region typically show similar DNA methylation dynamics, we estimate in the next section the parameters of the average behaviour of a group of several CpG sites whose observations are collected by the set Λ.

4.2 Parameter Estimation for Real Data

The main motivation behind our work is the availability of huge datasets of DNA methylation data that we will use to investigate methylation fidelities, i.e. learn the parameters μ, f_m, and f_u. In the following, we use two data sets with human blood samples and solid tumour samples. We ran the simulation-based parameter estimation procedure for both cell types and examined the differences between the estimated fidelities in blood and in tumour cells. In order to estimate parameters for sets Λ of observations of CpG sites in close proximity, we grouped CpG sites in consecutive ranges of 5000 base pairs into genomic regions and used our model to describe the average behaviour of a site in each region. For the analysis, first a region was identified, the methylation data of this region were extracted from the data of the blood and tumour samples, and if there was information about at least 100 different CpG sites, the parameters were estimated for the extracted data. Fig. 6 shows estimated parameters of different regions of chromosome 7. Both plots look broadly similar, but there is a certain number of regions where f_m is reduced in tumour samples. In these regions μ tends to be increased (the points are brighter). To visually investigate this observation, Fig. 7 shows for each region f_x(cancer) as a function of f_x(blood), with $x = m, u$. While f_u is distributed more or less equally over the whole pictured region, f_m is more clustered and in average reduced in cancer-cells. The corresponding plot for μ looks similar to the plot of f_u and is not reported. To validate this visual impression, we tested the null hypothesis $H_{0,1} : f_m(blood) \leq f_m(cancer)$ with a Mann-Whitney test. The resulting p-value was $p < 2^{-32}$,

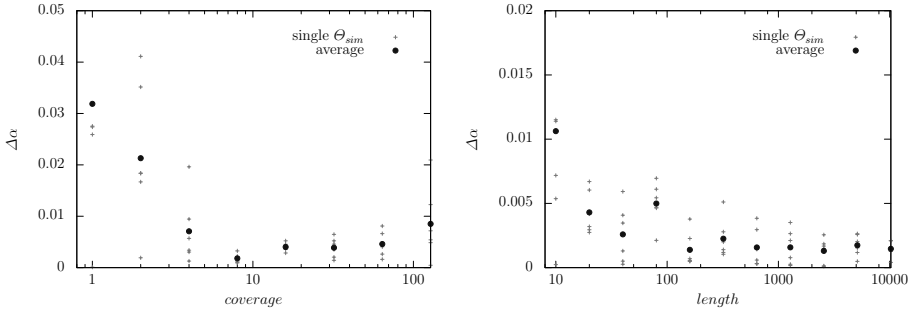


Fig. 5: Distances between estimated α and true α for a fixed length of 1000 (left). Distances for all six parameter sets for a fixed coverage of 6(right).

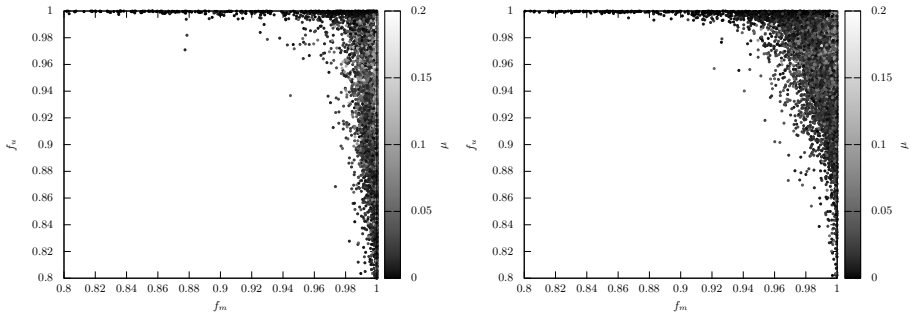


Fig. 6: Estimated parameter sets for groups of 5000 base pairs of chromosome 7. Left: data from 306 blood samples. Right: data from 101 solid tumor samples.

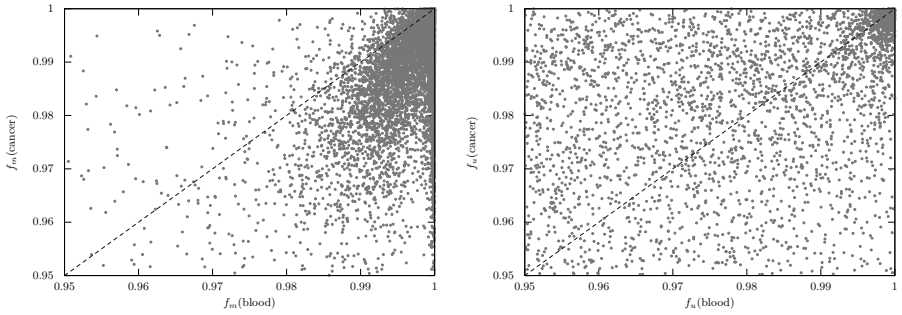


Fig. 7: Comparison of estimated single parameters. Left: f_m. Right: f_u

so we can safely reject this hypothesis. Although the plots for f_u and μ look very similar according to the Mann-Whitney test, we also have to reject $H_{0,2}$: $f_u(blood) \geq f_u(cancer)$ and $H_{0,3}$: $\mu(blood) \leq \mu(cancer)$. Hence, the Mann-Whitney statistically hints at the fact that blood and tumour samples behave differently in terms of DNA methylation fidelity, which is relevant for cancer biology, although the test fails to suggest a clear criterion which can be used as an indicator for detecting abnormal cells. Visual inspection of Fig. 7 points to f_m as a potential candidate. This issue will be further investigated when running the analysis on additional data sets.

5 Conclusions

In this paper we introduce a model of DNA methylation fidelity taking into account the behaviour of individual cells over generations, which can be trained by experimental data obtained from next generation sequencing technology. We carefully crafted efficient parameter estimation techniques to scale the analysis

to the whole-genome level. Currently, parameter estimation for a single group of sites takes less than one minute on a single core. Given that one chromosome contains approximately 10^9 CpG sites, with the grouping of CpG sites considered, we will need approximately 14 days for the complete analysis on a single core machine. However, as this code is fully and straightforwardly parallelizable, the whole-genome analysis is feasible on a high-performance cluster.
In this paper we also present preliminary tests on simulated data and on a real whole-genome dataset, trying to detect differences in methylation fidelities between human blood and tumour samples. Preliminary statistical analysis of data appears to support that there is indeed a systematic difference in the DNA methylation dynamics of the two samples. Deeper investigations are currently carried out on the whole-genome scale to better understand the statistical nature and biological significance of these differences. We will also investigate if and how differences in methylation fidelity reflect on differences in the shape of methylation profiles, comparing with state-of-the-art statistical tests [9].

References

1. Abramov, R.: A practical computational framework for the multidimensional moment-constrained maximum entropy principle. Journal of Computational Physics 211(1), 198–209 (2006)
2. Arand, J., Spieler, D., Karius, T., Branco, M.R., Meilinger, D., Meissner, A., Jenuwein, T., Xu, G., Leonhardt, H., Wolf, V., et al.: In vivo control of CpG and non-CpG DNA methylation by DNA methyltransferases. PLOS Genetics 8(6), e1002750 (2012)
3. Bird, A.: DNA methylation patterns and epigenetic memory. Genes & development 16(1), 6–21 (2002)
4. Bock, C.: Analysing and interpreting DNA methylation data. Nature Reviews Genetics 13(10), 705–719 (2012)
5. Bock, C., Tomazou, E.M., Brinkman, A.B., Müller, F., Simmer, F., Gu, H., Jäger, N., Gnirke, A., Stunnenberg, H.G., Meissner, A.: Quantitative comparison of genome-wide DNA methylation mapping technologies. Nature Biotechnology 28(10), 1106–1114 (2010)
6. Burden, R. L .and Faires, J.D.: Numerical analysis. Brooks/Cole, Cengage Learning (2011)
7. Farlik, M., Sheffield, N.C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., Bock, C.: Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell Reports 10(8), 1386–1397 (2015)
8. Harris, T.E.: The theory of branching processes. Courier Corporation (2002)
9. Mayo, T.R., Schweikert, G., Sanguinetti, G.: M^{3}D: a kernel-based test for spatially correlated changes in methylation profiles. Bioinformatics 31(6), 809–816 (2015)
10. Nelder, J.A., Mead, R.: A simplex method for function minimization. The Computer Journal 7(4), 308–313 (1965)
11. Quine, M.: A note on the moment structure of the multitype Galton-Watson process. Biometrika 57(1), 219–222 (1970)
12. Sontag, L.B., Lorincz, M.C., Luebeck, E.G.: Dynamics, stability and inheritance of somatic DNA methylation imprints. Journal of Theoretical Biology 242(4), 890–899 (2006)