RESEARCH ARTICLE

# Stochastic modeling reveals kinetic heterogeneity in post-replication DNA methylation

**Luis Busto-Moner**[1,2], **Julien Morival**[3], **Honglei Ren**[4], **Arjang Fahim**[3], **Zachary Reitz**[3], **Timothy L. Downing**[3,4,5], **Elizabeth L. Read**[2,4,5]*

**1** Institut Químic de Sarrià, Universitat Ramon Llull, Barcelona, Spain, **2** Dept. of Chemical & Biomolecular Engineering, University of California, Irvine, California, United States of America, **3** Department of Biomedical Engineering, University of California, Irvine, Irvine, California, United States of America, **4** NSF-Simons Center for Multiscale Cell Fate Research, University of California, Irvine, Irvine, California, United States of America, **5** Center for Complex Biological Systems, University of California, Irvine, Irvine, California, United States of America

* elread@uci.edu

## Abstract

DNA methylation is a heritable epigenetic modification that plays an essential role in mammalian development. Genomic methylation patterns are dynamically maintained, with DNA methyltransferases mediating inheritance of methyl marks onto nascent DNA over cycles of replication. A recently developed experimental technique employing immunoprecipitation of bromodeoxyuridine labeled nascent DNA followed by bisulfite sequencing (Repli-BS) measures post-replication temporal evolution of cytosine methylation, thus enabling genome-wide monitoring of methylation maintenance. In this work, we combine statistical analysis and stochastic mathematical modeling to analyze Repli-BS data from human embryonic stem cells. We estimate site-specific kinetic rate constants for the restoration of methyl marks on >10 million uniquely mapped cytosines within the CpG (cytosine-phosphate-guanine) dinucleotide context across the genome using Maximum Likelihood Estimation. We find that post-replication remethylation rate constants span approximately two orders of magnitude, with half-lives of per-site recovery of steady-state methylation levels ranging from shorter than ten minutes to five hours and longer. Furthermore, we find that kinetic constants of maintenance methylation are correlated among neighboring CpG sites. Stochastic mathematical modeling provides insight to the biological mechanisms underlying the inference results, suggesting that enzyme processivity and/or collaboration can produce the observed kinetic correlations. Our combined statistical/mathematical modeling approach expands the utility of genomic datasets and disentangles heterogeneity in methylation patterns arising from replication-associated temporal dynamics versus stable cell-to-cell differences.

## Author summary

Cytosine methylation is a chemical modification of DNA that, in concert with other associated epigenetic marks, plays a role in regulating gene expression. When DNA is replicated in the cell in advance of mitotic cell division, not only is the genetic sequence copied, but the patterns of epigenetic marks on DNA are faithfully copied, also. New experimental techniques are capable of measuring the presence or absence of DNA methylation on individual nucleotide sites across the genome on newly-formed DNA shortly after replication. In this study, we apply statistical inference techniques to quantify the rate at which DNA methylation appears on nascent DNA post replication in human embryonic stem cells. We find a broad range of per-site rate constants, ranging from shorter than ten minutes to five hours and longer. We furthermore found that these rate constants are correlated with distance along the genome. By comparison with computer simulation results, we identify enzymatic reaction mechanisms that are consistent with experimental measurements.

## Introduction

DNA methylation is an essential epigenetic modification found in a diversity of organisms, which is broadly associated with silencing of genes [1]. Methylation patterns across the genome encode epigenetic information associated to cellular processes including differentiation [2, 3] and genomic imprinting [4, 5]. These patterns are also conserved in distinct cell types, and clearly distinguish cell types in mammalian tissues [6–8]. Failure in the transmission of such patterns from one generation to the next and the appearance of aberrant methylation patterns have been associated with cancer [9, 10], aging [11], or organismal death [12].

In mammals, DNA methylation is primarily found in the cytosine-phosphate-guanine (CpG) dinucleotide context, which presents a symmetric substrate for inheritance that echoes the Watson-Crick model of genetic inheritance [12, 13]. Methylation patterns are generally transmitted with high fidelity from the parent template strand to nascent DNA over cycles of DNA replication. The classic model of *maintenance* methylation holds that DNA Methyltransferase 1 (DNMT1) is primarily responsible for this inheritance, which it accomplishes by localizing to replication foci [14] and preferentially catalyzing addition of methyl groups onto hemi-methylated CpG substrates (i.e., those CpG substrates with methylation present on only the parent-strand cytosine) [15–17]. In contrast to DNMT1, DNMT3A and 3B are often termed *de novo* methyltransferases because their catalytic activity shows no preference for hemimethylated versus unmethylated DNA and they are essential in the establishment of genome-wide methylation patterns during embryogenesis [18]. However, in recent years it has been pointed out that this classical model is overly-simplistic [19], since, for example, DNMT1 and DNMT3s are both essential for development, both contribute to maintenance methylation [20], and these enzymes work together with methyl-eraser enzymes (Ten-eleven translocation proteins (TETs) [21]) to control methylation across the genome and over time.

Whole genome bisulfite sequencing, which maps the methylation status of individual CpGs, shows generally bimodal patterns comprising fully methylated or fully demethylated regions. That is, the fraction of cells in a population with methylation at a given site tends to be near 1 or 0. However, intermediate methylation (IM), where methylation fraction is between 0 and 1, is also widespread. Despite broad conservation of genomic methylation patterns in distinct cell types, some loci show this type of non-uniformity in methylation across homogeneous cell populations. This heterogeneity appears to be itself conserved, as common IM

regions have been identified across individuals and even species [22]. IM appears to be critical for proper organism development and cell fate determination [22–24], contributes to genomic imprinting [8, 25], and plays a prominent role in tumor cell evolution [26].

The determinants of IM are not fully understood. In some contexts, cell-to-cell heterogeneity within populations has been implicated as the chief contributor to IM [27–29]. However, as methylation levels result from dynamic processes carried out asynchronously in different cells, IM could result not only from stable cell-to-cell differences, but also from temporal heterogeneity. For example, in an unsynchronized population of replicating cells, a subset of cells would be in the process of re-establishing methylation marks post-replication, thus contributing to lowered methylation fractions at the bulk level. A recently-developed experimental technique, Replication-associated Bisulfite Sequencing (Repli-BS), enables time-resolved measurement of genomic methylation patterns, including in newly replicated DNA [29], shedding light on dynamic re-establishment of methylation that must occur after each round of DNA replication. Using this technique, Charlton and Downing, et al. reported a pronounced genome-wide delay of several hours in post-replication nascent strand DNA methylation in human Embryonic Stem Cells (hESCs). These results echoed previous observations of a lag in maintenance methylation following replication in a variety of mammalian cell types [20, 30–32]. Furthermore, Charlton and Downing, et al., reported that the delay in post-replication nascent strand methylation accounts for a significant amount of the IM observed in hESCs in WGBS experiments.

Along with experimental evidence, mathematical modeling has informed understanding of DNA methylation dynamics. Population epigenetic models have explored the interplay between processes including enzyme-mediated *de novo* methylation, maintenance methylation, demethylation, and replication [33–38]. Some models have incorporated various mechanisms of interdependence of CpGs, where, for example, the efficiency of maintenance methylation at a given site depends on the methylation status of its neighbors [39–44]. Biochemical studies have enabled the development of enzyme-kinetic models and parameter quantification for methyltransferase activity [17, 45, 46]. While a number of modeling studies based on *in vivo* data in various cellular contexts have quantified the relative efficiency of maintenance methylation (i.e., the probability that the methylated state is successfully propagated through one cell division cycle), genome-wide quantification of sub-cell-cycle kinetics of maintenance methylation *in vivo* has not been possible.

The expansion in recent years of genomic measurement techniques provides an increasingly fine-grained view of methylation patterns across the genome, across cells, and across time. However, there remains a major gap in our understanding of the molecular sources and regulatory consequences of most of the heterogeneity present within the mammalian methylome. In this work, we combine statistical inference and mathematical modeling to analyze genome-wide post-replication methylation kinetics, making use of published Repli-BS data from hESCs. First, using Maximum Likelihood Estimation (MLE), we infer parameters quantifying remethylation kinetics of nascent DNA post-replication to individual CpG-site resolution, genome wide. Second, we perform stochastic simulation of various candidate enzyme-kinetic models of maintenance methylation in order to identify potential mechanisms consistent with the experimentally-inferred parameter distributions. Our combined statistical/mathematical modeling approach expands the utility of genomic datasets such as those resulting from Repli-BS experiments. The approach enables a basepair-level view of the combined influences of temporal and cell-to-cell heterogeneity across the genome.

## Methods

### Methods overview

The workflow of our approach is summarized as follows. We analyzed published Repli-BS data [29], which tracks re-establishment of genomic methylation patterns in newly replicated DNA over time. A schematic of DNA remethylation process is shown in Fig 1A. We first employed analytical, stochastic models of remethylation kinetics to serve as a framework for analysis of the experimental data. These analytical models with few parameters (two to three) served primarily as a tool to quantify kinetics via statistical inference of post-replication DNA methylation, to single CpG-site resolution, genome-wide (Fig 1B and 1C). We then developed a set of candidate kinetic models of enzyme-mediated maintenance methylation; the aim in studying these more detailed and biologically motivated models was to provide mechanistic insight on maintenance methylation processes in conjunction with the inferred parameters from Repli-BS (Fig 1D). The connection between the two modeling frameworks (i.e., between the small analytical models with inferred parameters, and the more complex, enzyme-kinetic mechanistic models) was achieved as follows. We first present the primary outputs of the statistical inference: namely, (1) the distributions of per-site inferred parameter values across different chromosomes, (2) the correlation of parameter values with genomic distance (GD), and (3) the distribution of inferred parameter values with respect to local CpG density ($CpG_d$). Next, we perform stochastic simulations of candidate enzyme-kinetic models, using parameters derived from previous literature where possible. Finally, we extract *in silico* Repli-BS read-data from the simulations, subject to the same experimental constraints (i.e., measured timepoints, read-depth) as the experimental data. We then compare outputs (1-3) from simulated and experimental read-data in order to assess the ability of different model mechanisms to reproduce features of the experimental outputs.

### Experimental data from Repli-BS

In the Repli-BS experiments [29], human embryonic stem cells (HUES64) were pulsed for one hour with bromodeoxyuridine (BrdU), and bisulfite sequencing measurements were obtained at multiple timepoints between 0 and 16 hr post-pulse. Methylation was measured on BrdU-labeled DNA, thereby selecting only those cells in which DNA replication occurred during the pulse interval ([-1,0] hours). The captured bisulfite read-data measured the presence (1) or absence (0) of methylation at individual CpG sites. Thus, the experimental data is of the form $\{N_{ij}^0, N_{ij}^1\}$, or observed numbers $N$ of unmethylated reads ("0") and methylated reads ("1") on nascent DNA at each timepoint $j$ at site $i$. Each measured site comprised a variable number of acquired reads at each timepoint. For parameter inference, we analyzed four timepoints (0, 1, 4, and 16 hr). In the original Repli-BS dataset, nascent DNA (0 hr) was collected from cells that were sorted according to their stage in S-phase of the cell cycle (S1-S6). In order to obtain a single nascent DNA methylation file, data from the six fractions were merged. Prior to merging the six datasets, methylation data were first filtered according to replication timing so as to capture only actively replicating regions within each file and avoid aggregation of background signal. Replication timing region files for each S-phase fraction were created based on sequencing read enrichment over genomic background. CpG methylation data from each fraction were then intersected with their corresponding replication timing regions and filtered. The remaining data were then merged into a single '0 hr' file. We restricted analysis only to those sites that had a minimum total read-depth of 15, with at least ten at time 0 (i.e., $N_{i,0h}^0 + N_{i,0h}^1 \geq 10$, $N_{i,[1h,4h,16h]}^0 + N_{i,[1h,4h,16h]}^1 \geq 5$). After these restrictions, the dataset contained 10,435,822 analyzed unique CpG sites, which constitute $\approx 40\%$ of the total number of sites in
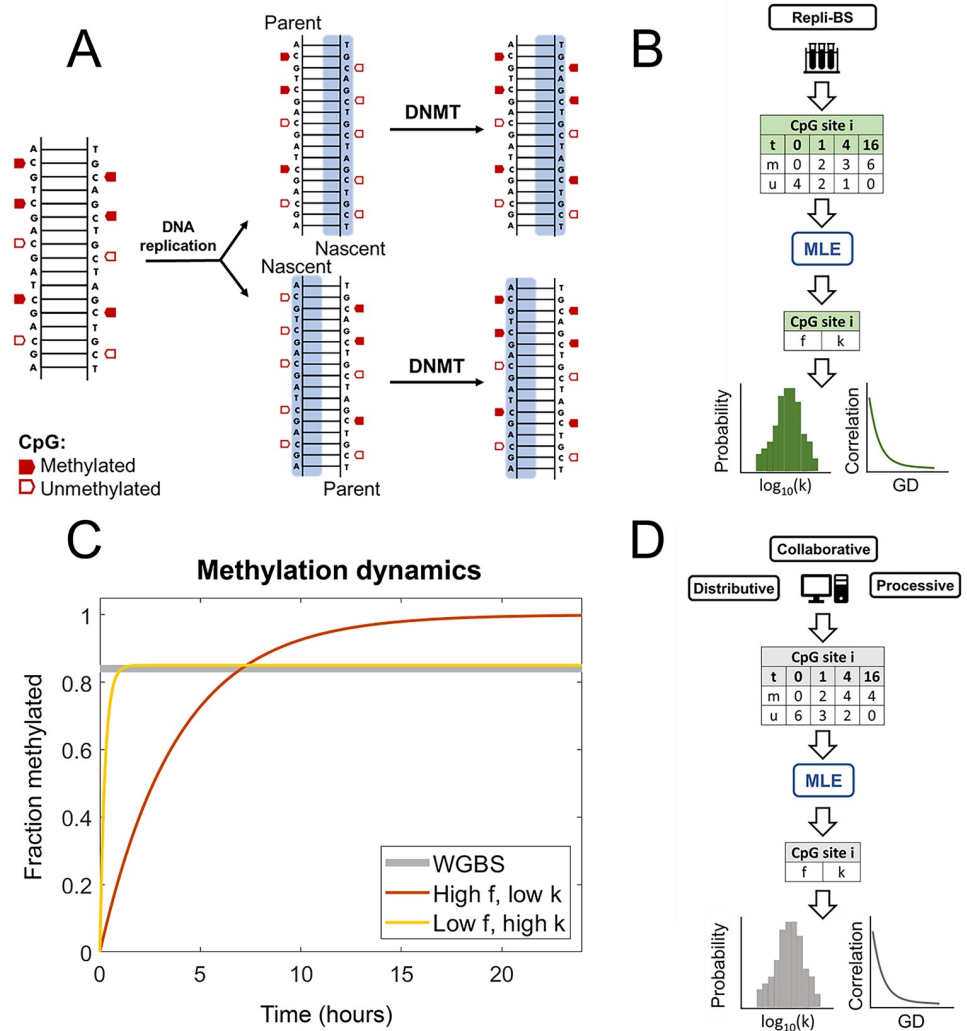
**Fig 1. A**: DNA methylation in the context of replication. Upon replication, complementary unmethylated nascent strands are synthesized for each parent strand, such that fully methylated CpGs become hemimethylated. Classically, full methylation is restored by DNMT1 (though DNMT3s have also been shown to contribute to this maintenance). **B**: Work-flow of the data analysis: Repli-BS data records methylated (m) and unmethylated (u) reads on the nascent strand for each site $i$ genome-wide, with timepoints over 16 hours. MLE allows the inference of stochastic model parameters for each site $i$, giving outputs of parameter distributions and parameter-correlation with genomic distance (GD). **C**: The MLE procedure assesses remethylation rate ($k$) and steady-state remethylation fraction ($f$) for each CpG site, thus distinguishing between sites that are remethylated faster while reaching a lower average methylation level (yellow line) and sites with slower kinetics but higher methylation overall (orange line). In contrast, traditional WGBS would not distinguish these cases, as they have roughly the same time-average (grey line). **D**: Work-flow of the enzyme-kinetic simulations: Stochastic modeling of remethylation kinetics according to either a Distributive, Processive, or Collaborative mechanism generates simulated datasets, which are then analyzed with the same MLE procedure used for the experimental data, shedding light on *in vivo* mechanisms.

https://doi.org/10.1371/journal.pcbi.1007195.g001

the original set. Note that this read-depth-based threshold for keeping sites in the analysis has the drawback that the amount of information gained from a given read/timepoint depends on the true kinetics. To address this issue, we also developed an alternative, confidence-based-threshold, by first analyzing all CpG sites in the dataset and then keeping them for analysis depending on the width of the confidence interval computed by the Profile Likelihood

method. Qualitative results were highly consistent between the analyses with read-depth- or confidence-based thresholds. Details can be found in Figs B and D in S1 File.

## Statistical inference of CpG remethylation kinetics

**Analytical kinetic models.** We employed analytical, stochastic models of remethylation kinetics to serve as a framework for analysis of the experimental Repli-BS data. (Analytical means here that models admit a simple, analytical formula for the likelihood function used for parameter estimation). The basic model assumes that fully methylated CpG sites (i.e., dual-methylated on both strands) become hemi-methylated at the time of DNA replication, followed by subsequent remethylation of the nascent strand over time (Fig 1A). Each CpG site $i$ is characterized by two parameters: $f_i$ (the fraction of cells in the population that are methylated at site $i$ in the steady-state) and $k_i$ (the rate of remethylation at site $i$). Methylation of an individual, hemimethylated site is assumed to be an independent, memoryless, stochastic process. Under these assumptions, a CpG site in an individual cell, which is hemimethylated at the time of replication, has probability to experience remethylation at time $t$ post-replication given by the exponential distribution:

$$p(t; k_i) = k_i \exp(-k_i t). \tag{1}$$

Thus, for a population of cells, the probability of observing a methylated read (denoted '1') on the nascent strand at site $i$ is given by

$$p(1|k_i, f_i, t) = f_i (1 - \exp(-k_i t)). \tag{2}$$

Since each read is 0 or 1, the probability of observing an unmethylated read ('0') is given by

$$p(0|k_i, f_i, t) = 1 - p(1|k_i, f_i, t) = 1 - f_i (1 - \exp(-k_i t)). \tag{3}$$

The half-life to remain unmethylated, or the "maintenance methylation lag-time" of an individual site $i$ is then

$$t_{1/2,i} = \ln 2 / k_i. \tag{4}$$

We emphasize that the primary utility of this simple model is to enable estimation of rate constants (and thus timescales) of remethylation kinetics across the genome from Repli-BS data. The inference results for this simple model with independent CpGs can nevertheless reveal more complex mechanisms of maintenance methylation, as described below (see Results).

The basic, two-parameter ($\{k_i, f_i\}$) model assumes that post-replication methylation is strictly irreversible, i.e., it enables description of only monotonically increasing methylation-fractions at any given site over time, and cannot account for any loss of methylation within one cell cycle. (The model can, however, account for passive demethylation, i.e. if $k$ is too slow for parent-strand methylation levels to be fully re-established within one cell cycle). In light of the demethylating activity of TET enzymes, we also performed inference using a three-parameter, reversible model, with time-dependent probability of methylation given by:

$$p(1|k_{1i}, k_{2i}, t) = \frac{f_i k_{1i}}{k_{2i} - k_{1i}} \left( e^{-k_{1i} t} - e^{-k_{2i} t} \right), \tag{5}$$

where $k_{1i}$ is the rate constant to acquire methylation over time post-replication and $k_{2i}$ is the rate constant of loss of methylation. This model provides the simplest means of fitting non-monotonic, reversible kinetics, potentially representing both methylation and active demethylation processes at a given site.

Due to the simplistic nature of the above models (Eqs 2 and 5), they can be considered to be agnostic to underlying biological mechanisms and to serve merely as tools to fit either irreversible or reversible kinetics. That is, the model fits cannot directly distinguish between various plausible mechanisms that could give rise to monotonic or non-monotonic kinetics of methylation on individual CpG sites. For example, if methylation and active demethylation processes occur continuously at a given site, this could result in apparently monotonic kinetics and $f < 1$.

Both the two- and three-parameter models can be extended to incorporate sources of experimental error. Given a false-positive rate $E_p$ (the probability of a false methylation count) and false-negative rate $E_n$ (the probability of a false non-methylation count), then the probability of experimental observation of a methylation read is:

$$P_{obs}(1|\theta_i, t) = p(1|\theta_i, t)(1 - E_p - E_n) + E_p,\qquad(6)$$

where the parameter vector $\theta_i = \{k_i, f_i\}$ for the two-parameter (irreversible) model and $\theta_i = \{k_{1i}, k_{2i}, f_i\}$ for the three-parameter (reversible) model. The probability of experimental observation of an unmethylated read is $P_{obs}(0|\theta_i, t) = 1 - P_{obs}(1|\theta_i, t)$.

In the above models, the time variable $t$ denotes the time that has elapsed post-replication. More specifically, it can be taken as the instant at which nucleotides (including the thymidine analog BrdU) were added to the nascent DNA strand at a given locus. The experiments have inherent uncertainty related to this timing. Since the BrdU pulse was one hour in duration, replication could have occurred anytime within the hour-long pulse. As such, we convert the experimental "post-pulse" time to the model's "post-replication" time in an unbiased way by adding one half hour. In other words, the experimental timepoint of 0-hour-post-pulse is converted to $t = 0.5$ hours post-replication (and similar for the other experimental timepoints). This conversion does not correct for any additional variability in replication timing that occurs within the pulse window. An alternative method, treating "time-post-replication" as a uniformly distributed random variable over the interval of one hour, was also studied.

**Maxmimum likelihood estimation of remethylation rates.**   In order to estimate the model parameters (i.e., the rate and steady-state fraction-methylated), we ask how likely it is that the model would "produce" the measured experimental data. In general, for a model that describes the probability, $p(\mathbf{x}|\theta)$ of an outcome given parameter(s) $\theta$, the likelihood function for $N$ independent observations is

$$L(\theta) = \prod_{n=1}^{N} p(x_n|\theta)\qquad(7)$$

and the Maximum Likelihood Estimate of the parameters, given the data, is

$$\hat{\theta}_{ML}(\mathbf{x}) = \operatorname{argmax}_{\theta} L(\theta) = \operatorname{argmax}_{\theta} l(\theta) = \operatorname{argmax}_{\theta} \sum_{n=1}^{N} \ln p(x_n|\theta).\qquad(8)$$

where $l(\theta)$ is the log-likelihood. The experimental data is of the form $\{N_{ij}^0, N_{ij}^1\}$, that is, observed numbers $N$ of unmethylated reads ("0") and methylated reads ("1") at timepoint $j$ at site $i$. Applying maximum likelihood estimation to the stochastic model of remethylation with parameters $\theta_i = [k_i, f_i]$ for site $i$, one obtains

$$\hat{\theta}_i = \operatorname{argmax}_{\theta} \sum_{j} [N_{ij}^0 \ln(p(0|\theta_i, t_j)) + N_{ij}^1 \ln(p(1|\theta_i, t_j))]\qquad(9)$$

where $i$ is the site index and $j$ is the timepoint index. (This expression omits the combinations factor, which does not affect maximization).

In order to estimate the parameters, the log-likelihood surface $l(k, f)$ was computed numerically for each set of site-specific read-data as a function of discrete $k$ and $f$ values, with domains $k \in [10^{-2}, 10], f \in [0, 1]$. The limits in $k$ were chosen by performing MLE for simulated data with the same timepoints and average read-depths as the experimental data, and identifying the approximate range of values over which $k$-estimation was possible.

The values of $k$ and $f$ for which the log-likelihood was maximum were taken as the estimated best-fit parameters for a given site. The exception to this was when the $k$ maximum was located on the edge of the $k$-domain precluding unambiguous assignment (this generally only occurred on the upper edge, i.e., for very fast rates). In such cases, a Confidence Interval (CI)-based estimate of the lower bound on $k$ was used (see below). All codes were written in MATLAB and are available on Github, along with the complete set of fitted parameters (https://github.com/Read-Lab-UCI/DNA-methylation-maintenance-kinetics).

**Confidence interval estimation and parameter bounds.**   Confidence Intervals around ML estimates of the parameters for each site were computed using a Profile Likelihood method [47]. The Profile Likelihood corresponding to a specific value of a given model parameter, $\sigma_i \in \theta$ (where $i$ here indexes the set of model parameters) refers to the maximum likelihood obtained when that parameter value is fixed while all remaining model parameters are freely varied. That is,

$$PL(\sigma_i) = max_{j \neq i} L([\sigma_i, \sigma_j]). \tag{10}$$

CIs for the parameter $\sigma_i$ are then estimated by the range of values $[\sigma_i^-, \sigma_i^+]$ for which the Profile Likelihood falls within a defined range of the Maximum Likelihood, $L(\hat{\theta})$. To approximate the 95% CI,

$$\frac{PL([\sigma_i^-, \sigma_i^+])}{L(\hat{\theta})} \leq \exp^{-3.841/2}, \tag{11}$$

where the value of 3.841 derives from the 95th-percentile of a 1-degree-of-freedom $\chi^2$ distribution [48]. The estimation of ML parameters and Profile-Likelihood-based confidence intervals is represented in Fig 2.

**Parameter correlation function.**   Correlation of inferred parameters is calculated as a function of genomic distance (i.e., number of basepairs separating individual CpG sites). As analyzed CpGs are unevenly spaced along the genome, correlation is calculated for binned distances [49]. The correlation function of parameter $\theta$ is given by

$$\hat{C}(\theta, d_n) = \frac{\langle (X_i - \mu_X)(Y_j - \mu_Y) \rangle}{\sigma_X \sigma_Y} \tag{12}$$

where $d_n$ is the $n$th discrete distance bin, and $(\mathbf{X}, \mathbf{Y})$ are the pairs of parameters $(\theta_i, \theta_j) = (k_i, k_j)$ or $(f_i, f_j)$ at sites with genomic positions $i$ and $j$ where $d_{n-1} < |i - j| \leq d_n$. $\mu_X$ and $\sigma_X$ are the mean and standard deviation, respectively, of the parameter values in $\mathbf{X}$ (and similar for $\mathbf{Y}$). This definition is identical to that used in other analyses of correlated methylation fractions [43].

## Single-basepair-level stochastic enzyme-kinetic models

We performed simulations of single-CpG stochastic enzyme-kinetic models according to a set of candidate mechanisms, called the Distributive, Processive, and Collaborative models. The
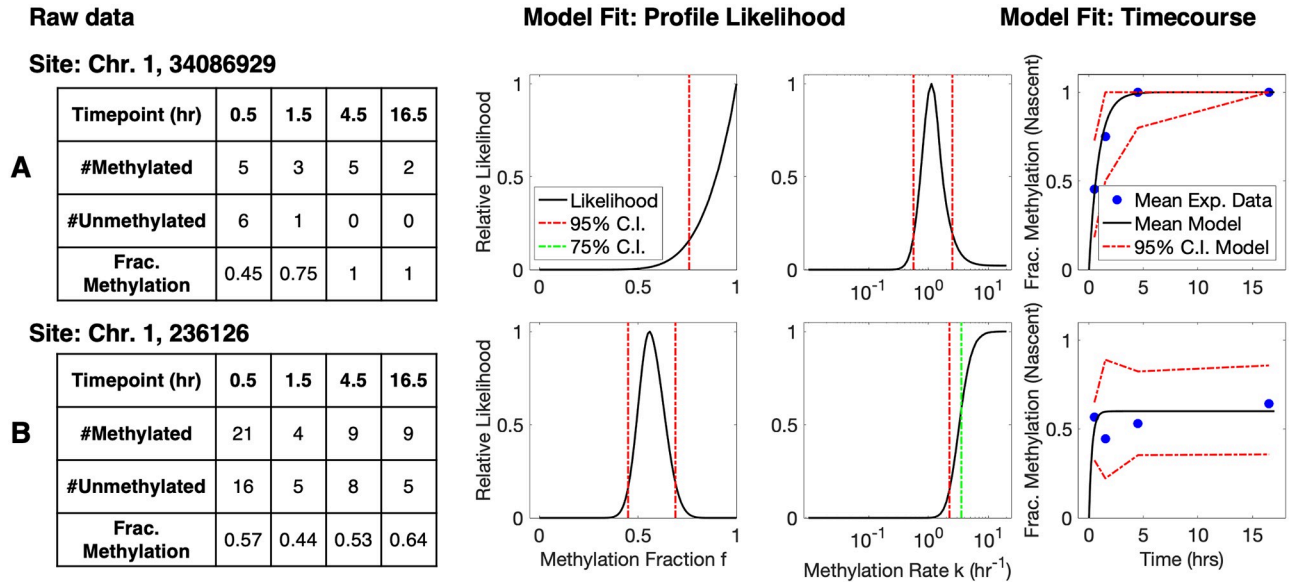
**Fig 2. Representative read-data and model fits for two individual CpG sites on Chromosome 1.** (A: site 34086929; B: site 236126) (left to right: raw experimental data, profile likelihood function for parameter $f$, profile likelihood function for parameter $k$, and model-predicted mean timecourse overlaid with experimental datapoints). **A**: Representative site with a global maximum in $k$, corresponding to parameter values $k = 1.1 \text{ hr}^{-1}$ ([0.56, 2.5] for the 95% CI) and $f = 1$ ([0.76, 1] for the 95% CI). The maximum-likelihood model prediction of mean fraction-methylation versus time (right, black curve) is overlaid with averaged experimental data (right, blue dots). The 95% confidence intervals for the model-predicted timecourse are simulated by Eq 2, while accounting for the variable number of samples (reads) at each timepoint. **B**: A representative site where the remethylation kinetic are too fast to measure, given the time resolution of experiments. In this case, the likelihood function increases to a plateau that extends infinitely in the direction of increased $k$, and only a lower bound on $k$ can be determined unambiguously. In such cases, we take the lower 75% confidence bound as the parameter estimate for subsequent analysis (see Methods). Thus, the estimated parameter values for this site are $k = 3.5 \text{ hr}^{-1}$ ([2.2, $\infty$]) and $f = 0.6$ ([0.45, 0.69]).

models are formulated as stochastic biochemical kinetic reaction models (or, in the case of the Processive mechanism, a stochastic reaction-diffusion model). The model reactions and associated rate parameters are shown graphically in Fig 3 and described in more detail in S1 File. These models focus only on the process of maintenance methylation, i.e., the remethylation process occurring over < 20 hours, and neglecting additional processes such as methyl erasure. In the three models, DNA is treated as a one-dimensional system of CpG sites which can be either unmethylated (u), hemimethylated (h), or methylated (m). Immediately after replication, sites are assumed to be in either the unmethylated or hemimethylated states, with hemimethylated sites being susceptible to remethylation by the enzyme (E, assumed namely to be DNMT1). The reaction also requires the methyl donor substrate, S, which stands for S-adenosylmethionine (SAM), while Q stands for its unmethylated form, S-adenosyl homocysteine (SAH). While sharing a common backbone in terms of E and S binding, as well as the remethylation reaction, the three models differ in the manner in which the enzyme reaches new hemimethylated sites after catalyzing methylation at one site.

**Distributive mechanism.** The backbone Distributive mechanism is based on a Compulsory-Order Ternary-Complex Mechanism (COTCM), by which DNMT1 (*E*) first binds the hemimethylated CpG (*h*) to form the *Eh* complex, and subsequently a SAM molecule (*S*) forming the ternary complex *EhS* (Fig 3A). While *m* stands for the methylated CpG, *Q* stands for SAH, the unmethylated form of SAM. The Distributive mechanism treats individual CpG sites as fully independent. The value of the forward and reverse rate constants for the first two binding reactions 1 and 2 ($k_{1f}$, $k_{1r}$, $k_{2f}$, and $k_{2r}$), as well as the catalytic step 3 ($k_3$) have been derived from experimental values in [17] (See S1 File for more details). (Note that there is no direct
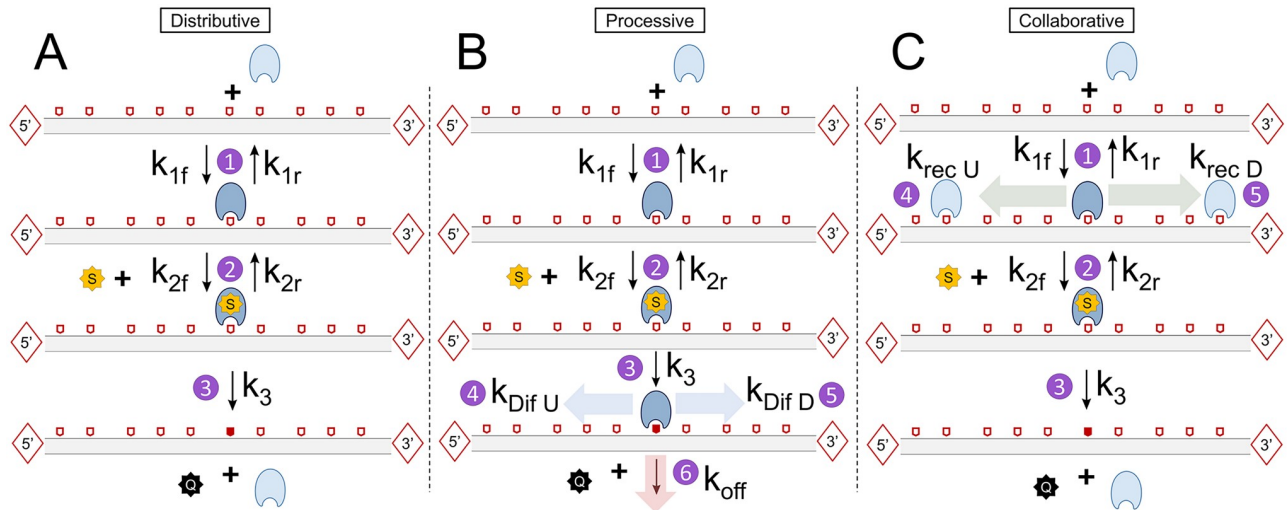
**Fig 3. DNA remethylation Distributive (A), Processive (B) and Collaborative mechanisms (C).** Hemimethylated sites (e.g. sites that can be remethylated) are indicated as empty pentagons, while methylated sites are represented as red-filled pentagons. Unmethylated sites are not represented in the scheme. In the Distributive model DNMT1 binds to a hemimethylated CpG site, incorporates SAM, and catalyzes methylation. Methylation and release of both DNMT1 and SAH occur in a single step. In the Processive model, after methylation DNMT1 can diffuse towards its immediate neighbor sites either upstream (U) or downstream (D). In the Collaborative model, once DNMT1 is bound to a hemimethylated CpG, a second DNMT1 molecule can be recruited onto nearby CpG sites. A distance function adapted from [42] favors recruitment at nearer CpGs.

relationship between the kinetic constants here and in the analytical kinetic models). All parameter values for this and the other models can be found in Table A in S1 File.

**Processive mechanism.** The Processive mechanism assumes that DNMT1 can remain bound to DNA after performing the catalytic step and reach subsequent hemimethylated CpG sites by diffusing along DNA. The first two steps are identical to the Distributive model, with $E$ binding $h$ to form $Eh$, and $Eh$ subsequently bonding $S$ to form $EhS$ (Fig 3B). The same assumptions were made to determine the rate constants associated with these steps.

In the Processive model, however, the catalytic step ($k_3$) does not directly imply DNMT1 to drop off from the DNA chain, returning as a free species into solution. Instead, $E$ remains bound to the DNA molecule onto the recently methylated site, forming the $Em$ complex, while releasing a molecule of $Q$. Then, two different events can take place: on one hand, DNMT1 can move towards its neighbor CpG sites either upstream (towards the 5' end) or downstream (towards the 3' end) through linear diffusion along DNA. A new $Eh$ complex with the destination site will be formed. Alternatively, DNMT1 can drop off the DNA chain and return into solution, with a rate constant $k_{off}$. In both events, a methylated CpG site is left behind. The model is processive in the sense that DNMT1 can with high likelihood perform successive methylation on sufficiently close $h$ sites. However, there is no explicit requirement in the model that DNMT1 move unidirectionally. To incorporate diffusion into the stochastic simulations, we use a First Passage Time Kinetic Monte Carlo algorithm, based on ref. [50]. We assume the enzyme travels with 1D diffusion coefficient $D$ and unbinds with rate $k_{off}$. The algorithm requires computation of the probability that an enzyme will reach each of three "exit-states": the nearest upstream neighbor $h$ at distance $d_U$, the nearest downstream neighbor $h$ at distance $d_D$, or the solution (by unbinding). The algorithm also requires computation of the First Passage Time density function, i.e. the distribution of waiting times at which the enzyme will first reach one of these three exit states, which is performed using Gillespie's Eigenvalue approach [51]. Details of the Processive Model can be found in S1 File.

**Collaborative mechanism.** The Collaborative mechanism shares reactions 1, 2 and 3 (and associated parameters) with the other two models. In this case, the catalytic step $k_3$ implies enzyme drop-off after methylation, just like in the Distributive model. However, here sites are interdependent through a phenomenological mechanism of "collaboration" between enzyme molecules: after DNMT1 binds, a second enzyme can be recruited to any nearby CpG site upstream or downstream the original site (not necessarily the contiguous). The stochastic propensity for each recruitment reaction $k_{Rec}$, notwithstanding, is indeed a function of the distance of a neighboring hemimethylated site to the recruiting site according to:

$$k_{Rec} = \frac{a}{b + d_i} \cdot k_{1f} \tag{13}$$

Where $d_i$ is the distance between the recruiting and the neighbor hemimethylated CpG sites, and $a$ and $b$ are free parameters integrated into a non-dimensional distance-dependent function. Therefore, the recruitment propensity decreases with distance. The phenomenological model and the mathematical form of the distance function was adopted from a previous modeling study [52].

Note that classical views of a collaborative mechanisms are based on the fact that DNMT1 is recruited by agents such as UHRF1 [53]. Our model does not implicitly consider UHRF1, but through the distance-dependent function assumes that DNMT1 recruitment after a first copy is bound to DNA indirectly account for the fact after recruiting the first enzyme copy, UHRF1 can recruit a second copy close to it, hence being a simplification of a more complex biological reality.

**Stochastic simulation.** Stochastic simulation was carried out using the Stochastic Simulation Algorithm [54], except in the case of the Processive model as described above.

Before any reaction could take place, the first step consisted of simulating the substrate, i.e. DNA containing $N_{sites}$ CpG sites, each assigned to be either unmethylated or hemimethylated at time 0 post-replication. Site numbers (and resulting inter-CpG distances) and methylation assignments were mined from an independent experimental dataset (WGBS measurements from Chr1 in arrested HUES64 cells, [29]) in the following manner: A start site was randomly sampled from Chr1, and the following contiguous $N_{sites}$ measured sites from the WGBS dataset were taken as the population-average, steady-state methylation landscape for the simulated substrate (with $N_{sites}$ = 100,000). Each site in a simulated "cell" was assigned a steady-state methylation status of m or u (i.e., methylated or unmethylated) randomly, with probability of methylation matching the population average; these assigned steady-state fractions are denoted $f_a$. If a given site in a cell is assigned to be methylated at steady-state, then it is assumed to be hemimethylated at time 0, and kinetics of re-methylation proceed according to the model mechanisms described above. Unmethylated sites remain as such for the duration of the simulation. Note that data from arrested cells (which are not undergoing DNA replication) are chosen to estimate $f_a$, as measurements in these cells are assumed to reflect a steady-state methylation landscape in the absence of replication-associated temporal dynamics.

Simulations of time-trajectories were performed for multiple cells in order to obtain multiple "reads" of methylation status for each site, in accordance with the experimental read-depth afforded by the Repli-BS data. In this way, simulated datasets were produced for each model *in silico* of the form $\{N_{ij}^0, N_{ij}^1\}$ matching the experimental dataset in timepoints and distributions of read-depth for each site (see S1 File for details).

## Results

### Identification of single-CpG remethylation kinetic parameters

Statistical analysis of Repli-BS data by Maximum Likelihood Estimation enabled per-site inference of the rate of post-replication methylation of the nascent strand ($k$) and the steady-state

fraction of cells methylated on the parent strand ($f$), according to Eq 1. Note that the parameter $f$ here has a slightly different meaning than the methylation fractions obtained traditionally from bulk WGBS data. As illustrated in Fig 1C, WGBS typically averages methylation from asynchronously dividing cells, and thus captures DNA strands in different stages of time after replication.

In contrast, $f$ as inferred here represents the fraction methylation in the steady-state (rather than the time-averaged fraction methylation), that is, after the methylation status of a given site has been returned to the "baseline" delineated by the parent strand before replication.

The parameters were generally "identifiable", that is, a single global maximum was present in the computed bivariate ($k$ vs. $f$) likelihood surface, and the parameter values corresponding to this peak were thus obtained as the Maximum Likelihood Estimates. However, due to the limited time-resolution of the experiments, some sites experienced remethylation too quickly to allow unambiguous assignment of $k$. This occurred when all or nearly all reads were methylated from the earliest timepoint (0 hour post-pulse, estimated to be an average of 0.5 hour post-replication, see Methods). In the statistical analysis, this manifested as a one-sided plateau in the profile likelihood function of $k$ (Fig 2B), enabling only identification of a lower bound on the rate constant $k$. In such cases, we used the value of the lower 75% Confidence Interval as an estimate of $k$, reasoning that this provides a conservative estimate of the remethylation rate given the experimental time resolution. Note that the parameter $f$ is by definition bounded, $f \in [0, 1]$, so the maximum likelihood value of $f$ frequently occurred on the edge of the domain (Fig 2A). In such cases the profile likelihood function increased steeply toward the edge (i.e., toward $f = 1$), and since $f$ is also by definition bounded, we directly used the edge-located maximum in $f$, rather than using CI-determined bounds as with $k$.

The distributions of inferred parameters for Chromosome 1 according to the two-parameter model (Eq 2) are presented in Fig 4. The distribution of $f$ values shows a bimodal pattern, similar to the directly-measured fractions from WGBS ($WGBS_f$), with most sites having high methylation fractions ($f > 0.6$, with the majority showing full methylation, $f = 1$) and a small subset of unmethylated sites ($f = 0$). The distribution of inferred $f$-values is qualitatively consistent with previous WGBS measurements in cell-cycle-arrested cells (see Fig J in S1 File), though it shows a relatively smaller fraction of $f = 0$ sites. Our read-depth restrictions were more likely to exclude these unmethylated CpGs, likely resulting from asymmetric PCR amplification efficiencies of methylated versus unmethylated strands [55]. Overall, the results support that our inferred $f$ fractions here can be considered to be analogous to WGBS-derived fractions, albeit "corrected", i.e., with the influence of replication-associated kinetics removed.

The distribution of remethylation rate $k$ values shows high site-to-site variability in remethylation kinetics. Non-zero $k$ values were estimated over a range from 0.01 to 9.5 $\mathrm{hr}^{-1}$, with 95% of the values lying within 0.3 and 6 $\mathrm{hr}^{-1}$ ([2.5%,97.5%]). That is, 95% of sites were found to have a "half-time to remethylation" between 7 minutes and 2.3 hours. A small fraction of sites had $k$ below 0.1 $\mathrm{hr}^{-1}$, or a half-life of 7 hours or more. The median value of inferred $k$ for Chromosome 1 was 2.2 $\mathrm{hr}^{-1}$. As noted previously, for the fastest sites, the identified constant $k$ can only be taken as a lower bound for the true rate; as such, the true $k$–distribution is likely wider than that presented in Fig 4, and the curtailed shape of the $k$–distribution on the right-hand-side likely reflects the experimental limit in time-resolution, rather than the true kinetic distribution. Since kinetics of remethylation are meaningless for sites that remain unmethylated, only sites with $f > 0$ have an associated estimate for $k$. Estimates of $k$ and $f$ on individual sites appear to be slightly negatively correlated (see Table B in S1 File).

The parameter estimates in Fig 4 are based on a model (Eq 2) that assumes monotonic remethylation kinetics, i.e., in the several hour time window post-replication, methylation on the nascent strand is assumed to increase or the site remains unmethylated, but it cannot
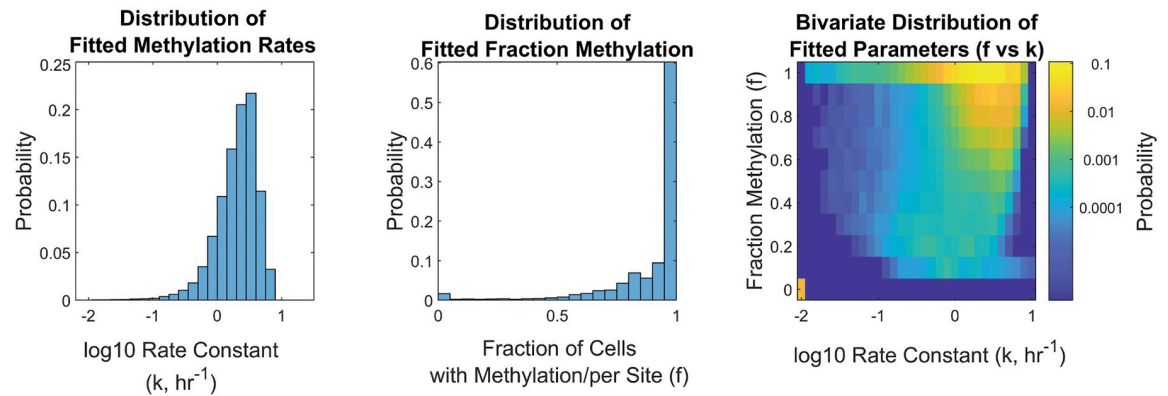
**Fig 4. Histogram of inferred remethylation rate *k* (left), steady-state fraction methylation *f* (middle), and bivariate heatmap of inferred *f* vs. *k* for Chromosome 1 in hESCs (right).** Histograms represent ∼ 0.8 million measured CpGs, and are normalized by probability. Note that *k*-values are only defined for sites with nonzero methylation (i.e., *f* > 0). In the heatmap, these *f* = 0 sites are shown in the lower left corner; the position with respect to the *k*-axis is arbitrary, since they have no defined rate of remethylation.

decrease. To relax this assumption, we also estimated parameters for the 3-parameter "reversible" methylation kinetic model (see Methods, Eq 5). Using a Bayesian Information Criterion-based model selection, we found that <1% of analyzed CpG sites on Chromosome 1 were better fit by the reversible model (Fig A in S1 File), concluding that for the majority of sites, the monotonic 2-parameter model sufficiently captures the kinetics revealed by the Repli-BS measurements.

Error analysis of the parameter estimates was performed in various ways. We generated "ground truth" simulated data with identical timepoints and read-depth distribution as the experimental data, and then tested ability of the MLE approach to recover the correct parameters (See MLE Validation in S1 File for details). In this analysis, the broad features of *in silico*-assigned parameter distributions were recovered accurately (Fig C in S1 File). The error in individual estimates of *k* ranged widely depending on the assigned values of *k* and *f* and the selected read-depth. We estimate an overall average relative error in per-site *k* values of approximately 32%. The average absolute error in inferred *f* values was 0.1. Overall, we concluded from this analysis that individual inferred parameters can be subject to relatively large error, while broader features of the distribution can be accurately inferred. Moreover, individual-site *k*'s can be estimated with high confidence to within an order of magnitude. For example, for assigned $k's$ of 1 hr$^{-1}$ (in the mid-range of the inferred distribution), 50% of inferred values fell within 0.71 and 1.41 hr$^{-1}$, and 95% of estimates fell within 0.50 and 2.51 hr$^{-1}$, which is less than one order of magnitude.

We varied the method of parameter estimation to determine whether the parameter distributions shown in Fig 4 are robust to details of the estimation method. First, we tested the influence of an experimentally-informed Bayesian prior on the estimated parameters. As discussed above, per-site methylation fractions obtained in the same cell line from WGBS measurements in arrested cells are expected to be a reasonable independent measurement of our statistically inferred values of *f*. We therefore used these independent measurements to construct Bayesian priors on $f_i$, and determine the impact on the estimates $k_i$. We found that some individual per-site estimates were affected by this choice of prior, but in general the bulk of estimates were similar between the two approaches and the broad characteristics of the distribution were unchanged. In another approach, we tested whether including explicit treatment of unknown sources of experimental measurement error in the model (Eq 6) would affect the estimates of *k*

and *f*, and found that within standard ranges of error estimates the distributions remained largely the same (Figs E-G in S1 File).

## Inferred parameters reveal high intra-chromosomal variability, but little variation between chromosomes

While the inferred parameters, remethylation rate *k* and fraction methylation *f*, show high variability from site to site, their distributions are highly uniform across different chromosomes. Representative *k* distributions are presented in Fig 5. (see also Figs H and I in S1 File). Distributions of inferred *k* and *f* values were also uniform with respect to mean replication timing (Fig S in S1 File).

## Inferred parameters correlate with local CpG density

Individual-CpG-site inference of *k* and *f* allow the study of how both parameters depend on local CpG density ($CpG_d$). In general, higher density regions show more CpGs with lower methylation fractions (Fig 6). The mean value of *f* for low, medium, and high density regions was 0.91, 0.90, and 0.75, respectively. These averages reflect increasing probability of sites in each density group with *f* = 0. This result is in agreement to what other authors have reported for WGBS methylation fraction: high-$CpG_d$ areas are more likely to be hypomethylated than low $CpG_d$ areas [56]. The distributions of inferred *k* parameters also show dependence on $CpG_d$. In general, the distributions shift rightward with higher density, i.e., faster remethylation is associated with higher $CpG_d$. However, in the highest density group there is also the appearance of an extended tail toward lower rates. These high-CpG-density sites with slower rates were not restricted to low-*f*-sites (Fig L in S1 File). Together these features give rise to a nonmonotonic dependence of average rate on density; mean remethylation rates *k* were 2.4, 3.1, and 2.5 $hr^{-1}$ in the low, medium, and high density groups, respectively.

Similar correlations between *k* and $CpG_d$, as well as between *f* and $CpG_d$, are observed along the genome (see Figs K and L in S1 File, respectively).

## Remethylation parameters are correlated among neighboring sites

Individual-CpG-site estimation of kinetic parameters enables analysis of correlation between parameters of neighboring sites (see Methods, 12). We computed correlation as a function of genomic distance (GD), based on the individual CpG site IDs for analyzed sites (Fig 7).

We found that both parameters were correlated on neighbor sites, albeit with different shapes and lengths of their correlation functions. *f*-correlations are stronger than *k*-correlations for sites in the immediate vicinity: for example, adjacent CpGs on Chromosome 1 have an average *f*-correlation of 0.83 and an average *k*-correlation of 0.64. The *f*-correlation first drops below 0.5 at a distance of approximately 300 bp, while for *k* this dropoff to < 0.5 occurs around only 16 bp. Despite this more rapid initial dropoff in *k*-correlation, *k* values appear to have a weak but consistent correlation that persists out beyond 10 kilobasepairs. As with the distributions and density-dependence, the correlation functions showed remarkable uniformity across different chromosomes (Figs N and O in S1 File). Correlation with genomic distance is robust with respect to uncertainty in numeric *k* estimates: correlation was maintained after binning rates into ordinal categories (Fig P in S1 File). Correlation was also maintained when MLE was performed after combining read-data from CpGs grouped in 200bp tiles (Figs Q and R in S1 File).

Correlation of fitted parameters with genomic distance was found to be broadly consistent, regardless of mean replication timing of CpGs in S-phase (Figs T and U in S1 File) or genomic context (Figs V and W in S1 File). However, we observed some specific dependencies of the
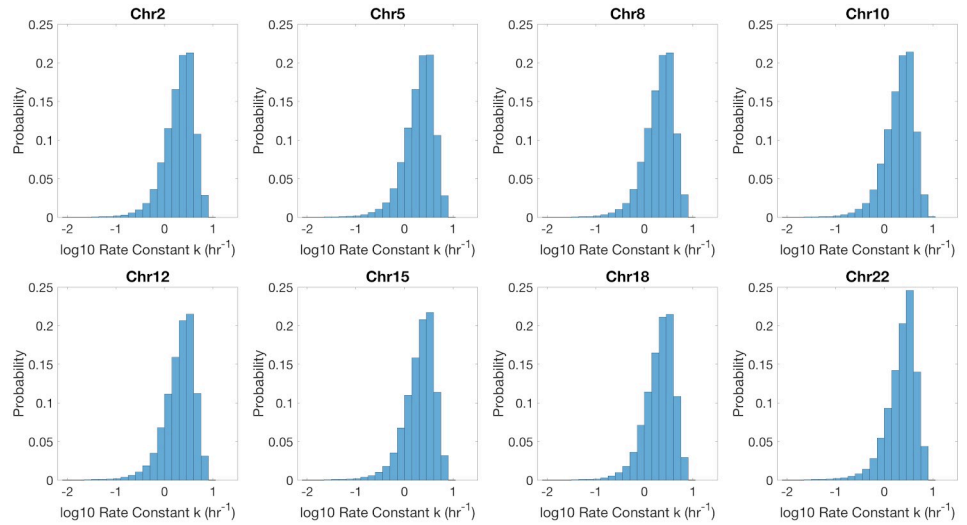
**Fig 5. Histogram of remethylation rates, *k*, for Chromosomes 2, 5, 8, 10, 12, 15, 18, and 22.** Histograms are normalized by probability.

correlation. For example, the correlation with genomic distance of *f* becomes weaker with later replication-timing in S-phase (Fig U in S1 File).

## Different enzyme models produce distinct parameter correlations with CpG distance and density

In order to further understand the inference results from Repli-BS data and gain mechanistic insight into the process of maintenance methylation, we studied three mathematical models encoding different candidate mechanisms for DNMT1-mediated remethylation post-replication (Fig 3). First, we employed a Distributive mechanism in which remethylation at each CpG
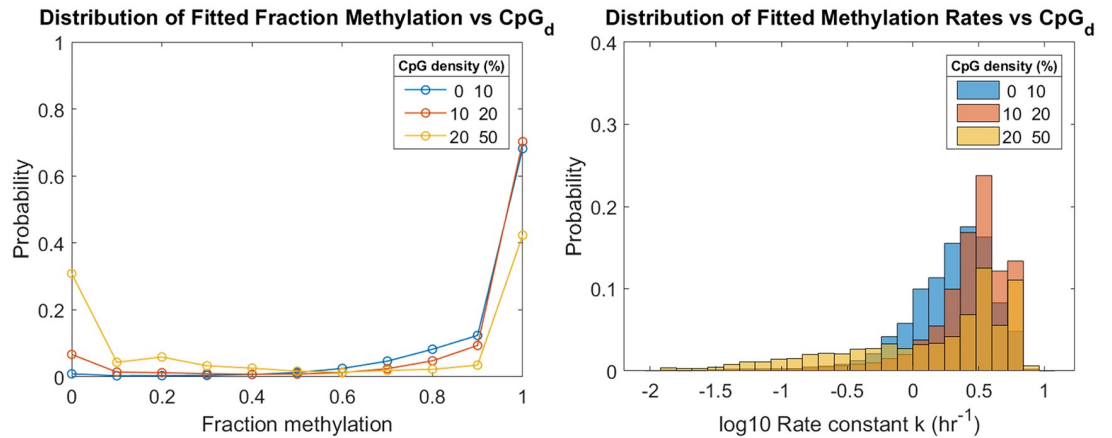


**Fig 6. Remethylation rates and fraction methylation distributions for low, medium, and high-density CpG sites of Chr1.** CpG density ($CpG_d$) of a site *i* is determined as the fraction of *bp* that are part of a CpG dinucleotide within a radius of 50 *bp* upstream and downstream the DNA molecule. Low-density CpG-sites represent 87% of the total sites analyzed in Chr1. Medium-density CpG sites represent a 12%, and high-density sites less than 1%. Low density is defined as [0,10)%. Medium density is defined as [10,20)%. High density is defined as [20, $CpG_{d_{max}}$]%, where $CpG_{d_{max}}$ is the maximum $CpG_d$ found in Chr1 (50%).
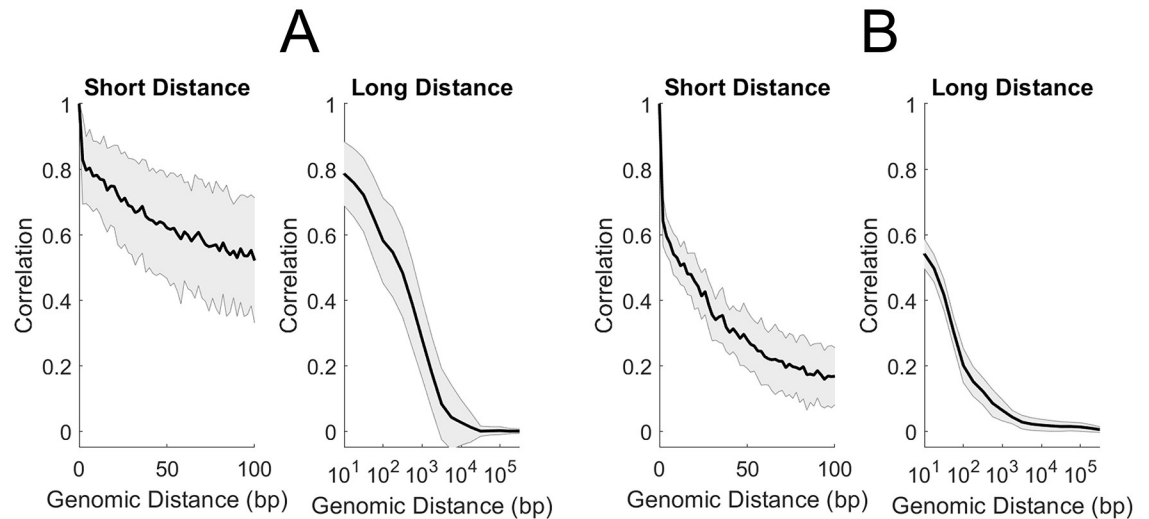
**Fig 7. Correlation of fraction methylation $f$ (A) and remethylation rates $k$ (B) with Genomic Distance (GD).** Correlation over short distances (left) and long distances (right) for Chr1.

https://doi.org/10.1371/journal.pcbi.1007195.g007

site is independent from the surrounding sites. Second, we employed a Processive mechanism in which DNMT1 can linearly diffuse along DNA after methylating one site, potentially accessing contiguous hemimethylated neighbor sites in this manner. Finally, we employed a Collaborative mechanism in which DNMT1 can be recruited onto nearby hemimethylated CpG sites after a first enzyme-copy is bound on a nearby site. These different mechanisms capture aspects of previous mathematical models of maintenance methylation (see Methods).

For each model, we performed stochastic simulations of the remethylation process over a 17 hour period post-replication in order to generate simulated read-data with the same characteristics as the experimental Repli-BS data. Simulations were performed for DNA substrates containing 100, 000 CpG-sites (comprising $\approx$ 4.5 million bp), with steady-state methylation landscapes derived from experimental data (see Methods). We then analyzed the per-site kinetics of the simulated data with the same MLE procedure as used for the experimental data. In this way, we could determine the effect of the more complex enzyme-kinetic mechanisms on the per-site inferred kinetics. We could furthermore determine which model mechanisms generated data features in common with the Repli-BS experiments.

When using the different molecular mechanisms to stochastically simulate remethylation kinetics, using parameter derived from enzyme kinetic studies [17], we observed distributions in per-site $k$ parameters ($k_{model}$) that are somewhat slower overall and narrower than the corresponding experimental distributions (Fig 8, top). Furthermore, the $k_{model}$ distributions are generally similar for the three models.

Major differences appear in terms of $k_{model}$ correlations with GD (Fig 8 middle). Kinetic rates derived from the Distributive model do not correlate with GD to any extent, i.e., the correlation function immediately drops from 1 for $GD = 0$ (a given site is fully correlated with itself) to fluctuate around 0 for all $GD \geq 2$ (the minimum distance between CpG dinucleotides). In contrast, $k_{model}$ values derived from the Processive and the Collaborative mechanisms show distinctive correlation functions that decrease with GD. The precise shapes and persistence of the correlation functions of both Processive and Collaborative mechanisms were found to depend on the models' parameters (Figs X and Y in S1 File), but the existence of correlation is robust. In contrast, the Distributive model cannot produce neighbor correlations
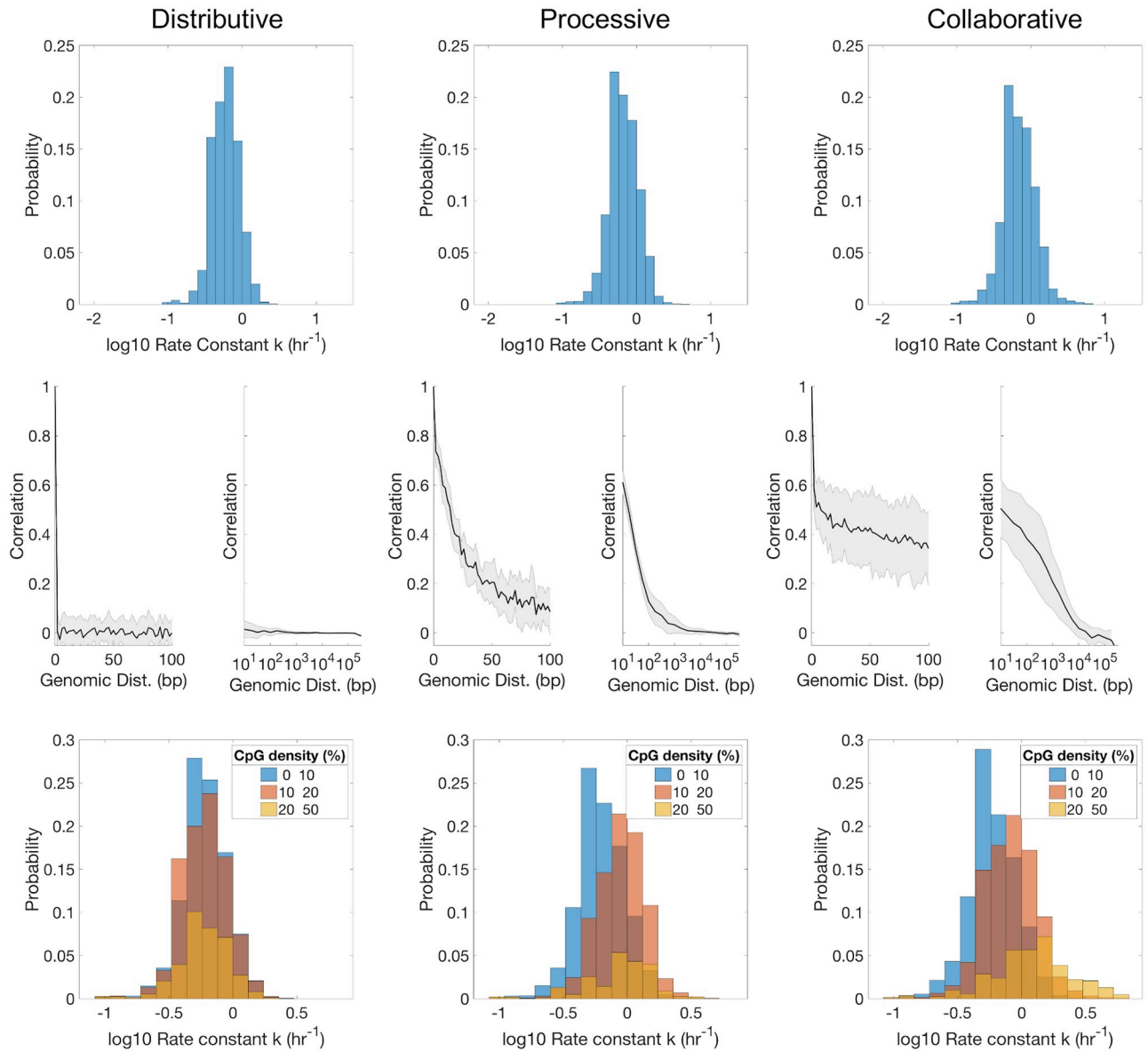
**Fig 8. Simulated remethylation rate histograms (top), *k*-correlation with Genomic Distance (GD) (middle), and *k*-dependence on local CpG density (CpG$_d$) (bottom) when using each of the proposed mechanisms (Distributive, Processive, and Collaborative).** The same *in silico* cluster containing $10^5$ sites was used for all models. Both the position and the $f_a$ for every site were sampled from an independent experimental dataset of WGBS measurements from Chr1 in arrested HUES64 cells [29].

https://doi.org/10.1371/journal.pcbi.1007195.g008

for any choice of model parameters. Overall, these results show that correlation between kinetics on different CpG sites is not imposed by the local features of steady-state methylation patterns, nor by the MLE procedure itself (as these are common to all three models). Rather, the neighbor-correlations result from the DNMT1-mediated interdependence between neighboring sites. Moreover, the results show that neighbor correlation can result from two disparate mechanisms of inter-dependence (single enzyme processivity versus cooperation between multiple enzyme molecules).

Remethylation rates generated from the Processive and the Distributive mechanisms also show dependence on CpG$_d$, with faster $k_{model}$ values inferred for higher-density sites (Fig 8

bottom). This observation is in agreement with the mechanistic aspects of both models, since proximity between sites increase both diffusing and recruiting propensities. Again, the results of both models are in stark contrast with remethylation rates derived from the Distributive mechanism, for which $k_{model}$ distributions remain centered around the same value for the three density groups. The shifted distributions of the Processive and Collaborative models with $CpGd$ are in partial agreement with the experimental results of Fig 6, where the bulk of the distribution is also seen to shift to higher $k$ values with increasing density. However, none of the models capture the extended slow-kinetics tail observed experimentally for the high-density group seen in Fig 6 and Figs K and L in S1 File.

## Discussion

In this work, our approach combining statistical inference and mathematical modeling reveals genome-wide temporal heterogeneity in DNA methylation maintenance kinetics. Inferred kinetic rates of maintenance methylation varied by about two orders of magnitude across individual CpG sites in hESCs. The results further show how kinetics of maintenance methylation at individual CpGs depends on local CpG density and correlates with kinetics on neighboring sites. Stochastic simulations revealed that these correlations could be introduced by enzyme-mediated remethylation through either a Processive or Collaborative mechanism.

Our mathematical modeling-aided analysis approach helps to extend the utility of genomic datasets emerging from techniques such as Repli-BS to shed light on processes of epigenetic regulation. Specifically, the approach implemented here gives a deeper understanding of sources of IM, by disentangling heterogeneity in DNA methylation levels resulting from replication-associated temporal heterogeneity (i.e., due to lag in remethylation) versus stable cell-to-cell differences (Fig 1C). In terms of the inferred parameters, sites with slower remethylation kinetics (lower $k$ values) experience a longer delay in methylation inheritance and thus exhibit this type of temporal heterogeneity. While the biological significance of sites exhibiting this pronounced lag is not clear, Charlton and Downing, et al. suggest that hESCs show a more pronounced genome-wide lag while IM levels were reduced in more specialized cell types. A separate study reported that DNA methylation is relatively stable during replication in primary dermal fibroblasts [28]. Together, these observations suggest the lag may have a potential role in fate specification; for example, a delay in methylation restoration could provide transcription factors with a "window of opportunity" to bind methylation-protected loci [57]. Our results suggest that regulation of such a window is dynamic and temporally heterogeneous within a population of hESCs. Given that various forms of cellular heterogeneity are known to play key roles in stem cell fate decision-making and embryonic development [58], we posit that it will be important to develop a more comprehensive picture of how post-replication methylation timing and variability impact probabilistic differentiation systems.

Our results can potentially aid in the development of more detailed mathematical models of DNA methylation dynamics; for example, the results indicate that remethylation rates vary by two orders of magnitude or more across the genome. The observed broad distribution of kinetic rates may reflect the multiple ways in which DNMT1 can reach hemimethylated CpGs, i.e., directly and independently from solution, from neighboring sites (e.g., as in the processive and collaborative mechanisms), or through additional recruitment mechanisms that were not studied here. For instance, active recruitment of DNMT1 to the replication fork may account for the fastest inferred rates, as discussed previously [29]. The variations in kinetic rates could also be the result of increased competition between other DNA-associated factors and DNMT1 for CpG sites. In this case, variation in kinetic rates would reflect accessibility of hemimethylated DNA based on the local chromatin landscape. These types of recruitment/

competition are not present in any of the models studied here, potentially explaining why the mathematical models (with *in vitro*-derived parameters) showed consistently slower and more narrowly-distributed kinetics as compared to the Repli-BS-inferred parameters.

For parameter inference, we chose simplistic analytical models (i.e., the two- and three-parameter models, Eqs [2](#) and [5](#), respectively). The bulk of the analysis and modeling effort of this paper was then focused on the two-parameter model, since a model selection procedure indicated that the vast majority of sites were better described by this "irreversible" model. The rationale for applying such simplistic models for our initial inference was, first, that they impose minimal mechanistic assumptions on the observed data, and second, that the few-parameter models afforded straightforward application of MLE for parameter estimation. Our initial parameter inference of $k$ and $f$ assumes only irreversibility of post-replication methylation; it makes no assumption of which molecular species control the reactions or by what mechanism. Thus, this model cannot capture the full complexity of methylation dynamics (neglecting, e.g., active demethylation), but we employ it for its expedience in analyzing the Repli-BS results for the majority of measured CpGs.

Previous studies have used statistical inference to quantify per-site parameters governing maintenance methylation dynamics [35, 40]. A key difference between those studies and this one is that parameters in those and other models [33, 34, 36–38] quantified the probability of methylation to be correctly reestablished before the next round of division, whereas our parameters quantify per-hour kinetic rate constants on a sub-cell-cycle timescale. Therefore, the focus and scope of previous *in vivo* inference/modeling efforts has been on the stability of methylation patterns over longer timescales, e.g., over hundreds of generations [33] or over days to weeks in the context of epigenetic reprogramming [38], whereas the scope of our study is the enzyme-kinetic processes occurring within one round of replication. As such, one unique feature of our study is that it more closely connects the enzyme-kinetic literature on DNMTs with statistical analysis of genomic data. The temporal nature of Repli-BS experiments enables this connection.

The difference in timescales between our model and others may also account for the suitability here of an irreversible model that neglects active demethylation: whereas the interplay of TET, DNMT3a/b, and DNMT1-mediated processes has been shown to be necessary to account for overall stability of epigenetic inheritance over many generations [41, 59], we found that the classical model of DNMT1-mediated maintenance methylation described reasonably well the reestablishment of methylation at most CpG sites within one cell cycle. Nevertheless, the ≈1% of sites at which reversible kinetics was apparent could possibly reveal loci of preferential active demethylation and be of future interest.

Interdependence of CpGs in methylation dynamics has also been the subject of previous modeling studies. Multiple mechanisms have been suggested for this interdependence [39–44, 59], from the enzymatic behavior of DNMT1 itself (e.g., through processivity) or in conjunction with other molecular species. This interdependence was found to play an important role in the stability of methylation inheritance [39, 42]. Processivity was suggested by biochemical studies, in which longer oligonucleotides experienced faster methylation kinetics [16, 46]. A linear diffusion model (which we consider a type of processivity because it often results in sequential methylation of neighboring sites) was previously found to be consistent with the enzymatic behavior of DNMT1 [39]. Additional phenomenological models of interdependence were introduced in [41, 42], one of which we adapted for use herein. We found that the presence (versus absence) of neighbor correlations was robust to other details of the models, such as the other kinetic constants and the initial conditions of the DNA methylation landscape. Therefore, the $k_{model}$-GD correlation can be unequivocally attributed to enzyme-kinetic mechanisms of Processivity and/or Collaboration. This idea is reinforced in light of the fact

that the three mechanisms share a common "backbone" in terms of the reactions they feature, with two sets of binding reactions and a catalytic step. In a similar way, $k_{model}$ correlation with $CpG_d$, is again only observed for the Processive and Collaborative models. Overall, these findings support the hypothesis that the rate of remethylation of one site is affected by the state and the position of surrounding ones, and show how independent-site-inferences can nevertheless reveal interdependence and thus reflect more complex mechanisms.

In future studies, it may be possible to use the specific features of the inferred correlation functions to shed light on the enzyme-mediated mechanism of remethylation *in vivo*. From our simulations, the linear diffusion model is more consistent with the rapid fall-off, but low-persistent $k$-correlation inferred from data, and as such appears more in line with experiments than the single Collaborative model studied here. However, we note that the specific features of the simulated correlation functions depend on a number of unknown parameters (see SI), and comprehensive parameter optimization for the enzyme kinetics is outside the scope of the present work. As such, we conclude both the Processive and Collaborative models to be broadly consistent with the Repli-BS data. However, neither of these models perfectly matched the experiment-derived correlation functions, nor did they account for the apparent bimodality of $k$'s in high density CpG regions. Therefore, our inference results may suggest more complex mechanisms of density/neighbor dependence in future studies.

A limitation of the current work is the various sources of uncertainty that contribute to individual site-estimates. A major source of uncertainty in the MLE estimates is the variable, often low, read-depth for individual sites. (We note that the individual site-uncertainties, for example, as quantified by the width of 95% confidence intervals from profile likelihood functions, depend not only on read-depth but also on the observed kinetics and their relationship to the available experimental timepoints.) Consideration of read-depth leads to a necessary tradeoff between the minimum per-site read-depth admitted for analysis and the total number of sites maintained in the analysis (here, 40% of the original set with the chosen restrictions). We sought to balance these factors and found through in silico "ground-truth" testing and alternative analysis methods that, while individual site estimates could be susceptible to significant uncertainty, the shapes of parameter distributions and their correlation, e.g., with GD, were robust. Additional sources of error include the hour-long BrdU pulse window that limits resolution of precise time-of-replication, the number and/or choice of experimental timepoints, and experimental errors in bisulfite conversion. Future experiments on sub-cell-cycle timescales with additional experimental conditions or increased sampling should enable an increasingly detailed understanding of maintenance methylation kinetics and, more broadly, of DNA methylation heterogeneity.

## Supporting information

**S1 File. All supporting information is located in the accompanying supporting information file.** This file contains supporting text, 26 supplemental figures, and two supplemental tables. (PDF)

## Author Contributions

**Conceptualization:** Luis Busto-Moner, Timothy L. Downing, Elizabeth L. Read.

**Data curation:** Julien Morival, Honglei Ren, Arjang Fahim, Zachary Reitz.

**Formal analysis:** Luis Busto-Moner, Julien Morival, Honglei Ren, Elizabeth L. Read.

**Funding acquisition:** Luis Busto-Moner, Timothy L. Downing, Elizabeth L. Read.

**Investigation:** Luis Busto-Moner, Timothy L. Downing, Elizabeth L. Read.

**Methodology:** Luis Busto-Moner, Timothy L. Downing, Elizabeth L. Read.

**Project administration:** Timothy L. Downing, Elizabeth L. Read.

**Software:** Luis Busto-Moner, Honglei Ren.

**Supervision:** Timothy L. Downing, Elizabeth L. Read.

**Validation:** Luis Busto-Moner.

**Writing – original draft:** Luis Busto-Moner.

**Writing – review & editing:** Luis Busto-Moner, Julien Morival, Zachary Reitz, Timothy L. Downing, Elizabeth L. Read.

# References

1. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nature Reviews Genetics. 2010; 11(3):204–220. https://doi.org/10.1038/nrg2719 PMID: 20142834

2. Jones PA, Taylor SM. Cellular differentiation, cytidine analogs and DNA methylation. Cell. 1980; 20 (1):85–93. https://doi.org/10.1016/0092-8674(80)90237-8 PMID: 6156004

3. Takizawa T, Nakashima K, Namihira M, Ochiai W, Uemura A, Yanagisawa M, et al. DNA Methylation Is a Critical Cell-Intrinsic Determinant of Astrocyte Differentiation in the Fetal Brain. Developmental Cell. 2001; 1(6):749–758. https://doi.org/10.1016/s1534-5807(01)00101-0 PMID: 11740937

4. Riggs AD. X inactivation, differentiation, and DNA methylation. Cytogenetic and Genome Research. 1975; 14(1):9–25. https://doi.org/10.1159/000130315

5. Kaneda M, Okano M, Hata K, Sado T, Tsujimoto H, Li E, et al. Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature. 2004; 429(6994):900–903. https://doi.org/10.1038/nature02633 PMID: 15215868

6. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009; 462:315 EP –. https://doi.org/10.1038/nature08514 PMID: 19829295

7. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature. 2008; 454:766 EP –. https://doi.org/10.1038/nature07107 PMID: 18600261

8. Ziller MJ, Gu H, Müller F, Donaghey J, Tsai LTY, Kohlbacher O, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013; 500:477 EP –. https://doi.org/10.1038/nature12433 PMID: 23925113

9. Jones PA, Liang G. Rethinking how DNA methylation patterns are maintained. Nature Reviews Genetics. 2009; 10(11):805–811. https://doi.org/10.1038/nrg2651 PMID: 19789556

10. Saito Y, Kanai Y, Nakagawa T, Sakamoto M, Saito H, Ishii H, et al. Increased protein expression of DNA methyltransferase (DNMT) 1 is significantly correlated with the malignant potential and poor prognosis of human hepatocellular carcinomas. International Journal of Cancer. 2003; 105(4):527–532. https://doi.org/10.1002/ijc.11127 PMID: 12712445

11. Richardson B. Impact of aging on DNA methylation. Ageing Research Reviews. 2003; 2(3):245–261. https://doi.org/10.1016/s1568-1637(03)00010-2 PMID: 12726774

12. Li E, Bestor TH, Jaenisch R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell. 1992; 69(6):915–926. https://doi.org/10.1016/0092-8674(92)90611-f PMID: 1606615

13. Holliday AR, Pugh JE. DNA Modification Mechanisms and Gene Activity during Development. Science. 1975; 187(4173):226–232. PMID: 1111098

14. Prelich G, Stillman B. Coordinated leading and lagging strand synthesis during SV40 DNA replication in vitro requires PCNA. Cell. 1988; 53(1):117–126. https://doi.org/10.1016/0092-8674(88)90493-x PMID: 2894900

15. Bestor T, Laudano A, Mattaliano R, Ingram V. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. Journal of Molecular Biology. 1988; 203(4):971–983. https://doi.org/10.1016/0022-2836(88)90122-2 PMID: 3210246

16. Hermann A, Goyal R, Jeltsch A. The Dnmt1 DNA-(cytosine-C5)-methyltransferase Methylates DNA Processively with High Preference for Hemimethylated Target Sites. Journal of Biological Chemistry. 2004; 279(46):48350–48359. https://doi.org/10.1074/jbc.M403427200 PMID: 15339928

17. Pradhan S, Bacolla A, Wells RD, Roberts RJ. Recombinant Human DNA (Cytosine-5) Methyltransferase. Journal of Biological Chemistry. 1999; 274(46):33002–33010. https://doi.org/10.1074/jbc.274.46.33002 PMID: 10551868

18. Smith ZD, Meissner A. DNA methylation: Roles in mammalian development; 2013.

19. Jeltsch A, Jurkowska RZ. New concepts in DNA methylation. Trends in Biochemical Sciences. 2014; 39 (7):310–318. https://doi.org/10.1016/j.tibs.2014.05.002 PMID: 24947342

20. Liang G, Chan MF, Tomigahara Y, Tsai YC, Gonzales FA, Li E, et al. Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements. Molecular and Cellular Biology. 2002; 22(2):480–491. https://doi.org/10.1128/MCB.22.2.480-491.2002 PMID: 11756544

21. Kohli RM, Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. Nature. 2013; 502 (7472):472–479. https://doi.org/10.1038/nature12750 PMID: 24153300

22. Elliott G, Hong C, Xing X, Zhou X, Li D, Coarfa C, et al. Intermediate DNA methylation is a conserved signature of genome regulation. Nature Communications. 2015; 6(1). https://doi.org/10.1038/ncomms7363

23. Feinberg AP, Irizarry RA. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proceedings of the National Academy of Sciences. 2010; 107(suppl_1):1757–1764. https://doi.org/10.1073/pnas.0906183107

24. Pervjakova N, Kasela S, Morris AP, Kals M, Metspalu A, Lindgren CM, et al. Imprinted genes and imprinting control regions show predominant intermediate methylation in adult somatic tissues. Epigenomics. 2016; 8(6):789–799. https://doi.org/10.2217/epi.16.8 PMID: 27004446

25. Stöger R, Kubicka P, Liu CG, Kafri T, Razin A, Cedar H, et al. Maternal-specific methylation of the imprinted mouse Igf2r locus identifies the expressed locus as carrying the imprinting signal. Cell. 1993; 73(1):61–71. https://doi.org/10.1016/0092-8674(93)90160-r PMID: 8462104

26. Brocks D, Assenov Y, Minner S, Bogatyrova O, Simon R, Koop C, et al. Intratumor DNA Methylation Heterogeneity Reflects Clonal Evolution in Aggressive Prostate Cancer. Cell Reports. 2014; 8(3):798–806. https://doi.org/10.1016/j.celrep.2014.06.053 PMID: 25066126

27. Singer Z, Yong J, Tischler J, Hackett J, Altinok A, Surani M¬, et al. Dynamic Heterogeneity and DNA Methylation in Embryonic Stem Cells. Molecular Cell. 2014; 55(2):319–331. https://doi.org/10.1016/j.molcel.2014.06.029 PMID: 25038413

28. Vandiver AR, Idrizi A, Rizzardi L, Feinberg AP, Hansen KD. DNA methylation is stable during replication and cell cycle arrest. Scientific Reports. 2016; 5(1):17911. https://doi.org/10.1038/srep17911

29. Charlton J, Downing T, Smith Z, Gu H, Clement K, Pop R, et al. Global delay in nascent strand DNA methylation. Nature Structural & Molecular Biology. 2018; 25(4):327–332. https://doi.org/10.1038/s41594-018-0046-4

30. Adams RL. The relationship between synthesis and methylation of DNA in mouse fibroblasts. Biochimica Et Biophysica Acta. 1971; 254(2):205–212. https://doi.org/10.1016/0005-2787(71)90829-x PMID: 5136448

31. Woodcock DM, Simmons DL, Crowther PJ, Cooper IA, Trainor KJ, Morley AA. Delayed DNA methylation is an integral feature of DNA replication in mammalian cells. Experimental Cell Research. 1986; 166(1):103–112. https://doi.org/10.1016/0014-4827(86)90511-2 PMID: 2427346

32. Shirodkar AV, St Bernard R, Gavryushova A, Kop A, Knight BJ, Yan MSC, et al. A mechanistic role for DNA methylation in endothelial cell (EC)-enriched gene expression: relationship with DNA replication timing. Blood. 2013; 121(17):3531–3540. https://doi.org/10.1182/blood-2013-01-479170 PMID: 23449636

33. Otto SP, Walbot V. DNA methylation in eukaryotes: kinetics of demethylation and de novo methylation during the life cycle. Genetics. 1990; 124(2):429–437. PMID: 2307364

34. Pfeifer GP, Steigerwald SD, Hansen RS, Gartler SM, Riggs AD. Polymerase chain reaction-aided genomic sequencing of an X chromosome-linked CpG island: methylation patterns suggest clonal inheritance, CpG site autonomy, and an explanation of activity state stability. Proceedings of the National Academy of Sciences. 1990; 87(21):8252–8256. https://doi.org/10.1073/pnas.87.21.8252

35. Genereux DP, Miner BE, Bergstrom CT, Laird CD. A population-epigenetic model to infer site-specific methylation rates from double-stranded DNA methylation patterns. Proceedings of the National Academy of Sciences. 2005; 102(16):5802–5807. https://doi.org/10.1073/pnas.0502036102

36. Arand J, Spieler D, Karius T, Branco MR, Meilinger D, Meissner A, et al. In Vivo Control of CpG and Non-CpG DNA Methylation by DNA Methyltransferases. PLoS Genetics. 2012; 8(6):e1002750. https://doi.org/10.1371/journal.pgen.1002750 PMID: 22761581

37. McGovern AP, Powell BE, Chevassut TJT. A dynamic multi-compartmental model of DNA methylation with demonstrable predictive value in hematological malignancies. Journal of Theoretical Biology. 2012; 310:14–20. https://doi.org/10.1016/j.jtbi.2012.06.018 PMID: 22728673

38.  von¬†Meyenn F, Iurlaro M, Habibi E, Liu N, Salehzadeh-Yazdi A, Santos F, et al. Impairment of DNA Methylation Maintenance Is the Main Cause of Global Demethylation in Naive Embryonic Stem Cells. Molecular Cell. 2016; 62(6):848–861. https://doi.org/10.1016/j.molcel.2016.04.025

39.  Goyal R. Accuracy of DNA methylation pattern preservation by the Dnmt1 methyltransferase. Nucleic Acids Research. 2006; 34(4):1182–1188. https://doi.org/10.1093/nar/gkl002 PMID: 16500889

40.  Bonello N, Sampson J, Burn J, Wilson IJ, McGrown G, Margison GP, et al. Bayesian inference supports a location and neighbour-dependent model of DNA methylation propagation at the MGMT gene promoter in lung tumours. Journal of Theoretical Biology. 2013; 336:87–95. https://doi.org/10.1016/j.jtbi.2013.07.019 PMID: 23911575

41.  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. https://doi.org/10.1093/nar/gkt1235 PMID: 24288373

42.  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. https://doi.org/10.1093/nar/gkw124 PMID: 26932361

43.  Song Y, Ren H, Lei J. Collaborations between CpG sites in DNA methylation. International Journal of Modern Physics B. 2017; 31(20):1750243. https://doi.org/10.1142/S0217979217502435

44.  Lück A, Giehr P, Walter J, Wolf V. A Stochastic Model for the Formation of Spatial Methylation Patterns. In: Feret J, Koeppl H, editors. Computational Methods in Systems Biology. vol. 10545. Cham: Springer International Publishing; 2017. p. 160–178. Available from: http://link.springer.com/10.1007/978-3-319-67471-1_10.

45.  Yokochi T, Robertson KD. Preferential Methylation of Unmethylated DNA by Mammalian de Novo DNA Methyltransferase Dnmt3a. Journal of Biological Chemistry. 2002; 277(14):11735–11745. https://doi.org/10.1074/jbc.M106590200 PMID: 11821381

46.  Svedruzic ZÃ, Reich NO. Mechanism of allosteric regulation of Dnmt1's processivity. Biochemistry. 2005; 44(45):14977–14988. https://doi.org/10.1021/bi050988f PMID: 16274244

47.  Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics. 2009; 25(15):1923–1929. https://doi.org/10.1093/bioinformatics/btp358 PMID: 19505944

48.  Cole SR, Chu H, Greenland S. Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer. American Journal of Epidemiology. 2014; 179(2):252–260. https://doi.org/10.1093/aje/kwt245 PMID: 24173548

49.  Edelson RA, Krolik JH. The discrete correlation function—A new method for analyzing unevenly sampled variability data. The Astrophysical Journal. 1988; 333:646. https://doi.org/10.1086/166773

50.  Bezzola A, Bales BB, Alkire RC, Petzold LR. An exact and efficient first passage time algorithm for reaction-diffusion processes on a 2D-lattice. Journal of Computational Physics. 2014; 256:183–197. https://doi.org/10.1016/j.jcp.2013.08.053

51.  Gillespie DT. Markov processes: an introduction for physical scientists. Boston: Academic Press; 1992.

52.  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. https://doi.org/10.1093/nar/gkw124 PMID: 26932361

53.  Bostick M, Kim JK, Esteve PO, Clark A, Pradhan S, Jacobsen SE. UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells. Science. 2007; 317(5845):1760–1764. https://doi.org/10.1126/science.1147939 PMID: 17673620

54.  Gillespie DT. Exact stochastic simulation of coupled chemical reactions. The Journal of Physical Chemistry. 1977; 81(25):2340–2361. https://doi.org/10.1021/j100540a008

55.  Laird PW. Principles and challenges of genome-wide DNA methylation analysis. Nature Reviews Genetics. 2010; 11(3):191–203. https://doi.org/10.1038/nrg2732 PMID: 20125086

56.  Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes & Development. 2011; 25 (10):1010–1022. https://doi.org/10.1101/gad.2037511

57.  Mirny L, Slutsky M, Wunderlich Z, Tafvizi A, Leith J, Kosmrlj A. How a protein searches for its site on DNA: the mechanism of facilitated diffusion. Journal of Physics A: Mathematical and Theoretical. 2009; 42(43):434013. https://doi.org/10.1088/1751-8113/42/43/434013

58.  Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. Nature. 2010; 467(7312):167–173. https://doi.org/10.1038/nature09326 PMID: 20829787

59.  Jeltsch A, Jurkowska RZ. New concepts in DNA methylation. Trends in Biochemical Sciences. 2014; 39 (7):310–318. https://doi.org/10.1016/j.tibs.2014.05.002 PMID: 24947342