Locally-correlated kinetics of post-replication DNA methylation reveals processivity and region-specificity in DNA methylation maintenance

Honglei Ren[1, 2], **Robert B. Taylor**[2, 3], **Timothy L. Downing**[1, 2, 4, 5, ✉], **and Elizabeth L. Read**[1, 2, 6, ✉]

[1]NSF-Simons Center for Multiscale Cell Fate, University of California, Irvine, CA, USA
[2]Center for Complex Biological Systems, University of California, Irvine, California, USA
[3]Dept. of Physics, University of California, Irvine, California, USA
[4]Dept. of Biomedical Engineering, University of California, Irvine, California, USA
[5]Dept. of Microbiology & Molecular Genetics, University of California, Irvine, California, USA
[6]Dept. of Chemical & Biomolecular Engineering, University of California, Irvine, CA, USA

DNA methylation occurs predominantly on cytosine-phosphate-guanine (CpG) dinucleotides in the mammalian genome, and the methylation landscape is maintained over mitotic cell division. It has been posited that coupling of methylation activity among neighboring CpGs is critical to collective stability over cellular generations, however the mechanism of this coupling is unclear. We used mathematical models and stochastic simulation to analyze data from experiments that probe genome-wide methylation of nascent DNA post-replication. We find that DNA methylation maintenance shows genomic-region-specific kinetics, indicating influence of local CpG density and chromatin accessibility on methyltransferase activity and inter-CpG coupling. We uncover evidence of processive methylation kinetics in post-replication DNA, which manifests as exponential decay of methylation rate correlation on neighboring CpGs. Our results indicate that processivity is a component of inter-CpG-coupling that occurs globally throughout the genome, but other mechanisms of coupling dominate for inter-CpG distances past $\sim$ 100 basepairs. By decomposing local methylation correlations into processive and non-processive components, we estimate that an individual methyltransferase methylates neighbor CpGs in sequence if they are 36 basepairs apart, on average. Our study demonstrates that detailed information on epigenomic dynamics can be gleaned from replication-associated, cell-based genome-wide measurements, by combining data-driven statistical analyses with hypothesis-driven mathematical modeling.

DNA methylation| processivity| stochastic modeling
Correspondence: ***tldownin @uci.edu, elread @uci.edu***

Introduction

DNA methylation is an important epigenetic modification that plays a critical role in development, aging, and cancer, and it is well conserved among most plants, animals and fungi (1, 2). In mammals, DNA methylation occurs predominantly in the cytosine-phosphate-guanine (CpG) dinucleotide context. Across most of the mammalian genome, CpGs occur with low frequency, except for regions called CpG islands (CGIs), which are often associated with promoters (3). In adult mammalian tissues, CpGs within CGIs are mostly unmethylated, whereas CpGs outside of islands are largely methylated. However, some CGIs do show higher levels of methylation, depending on cell type (4). Since methylated promoters are associated with transcriptional repression, this points to a role for DNA methylation as a stable and heritable chromatin mark to program alternative gene expression states (5, 6).

The inheritance and maintenance of methylation patterns across cell cycles is important in development and throughout organismal lifespan. Methylation patterns encode information related to gene expression (7, 8), differentiation (9), and genomic imprinting (10). Failure in maintenance and transmission of such patterns can lead to aberrant gene expression, and diseases including cancer (11), developmental abnormalities, and even death (12). The classical model of DNA methylation maintenance introduced the idea that the symmetrical nature of the CpG dinucleotide provides a biomolecular structure whereby DNA methylation could be inherited across a single CpG site by the activity of a, then posited, "maintenance" methyltranferase enzyme (13). The subsequent discovery of the mammalian DNA methylatransferase enzymes (particularly DNMT1) allowed researchers to measure the biochemical properties of these enzymes, from which emerged a wealth of experimental evidence suggesting that the efficiency and specificity of methylation enzymes alone are far from enough to support the observed high fidelity of inheritance within the classical model (14).

Mechanisms of interdependence, or coupling, of maintenance methylation activities imparted on CpGs located within close proximity has provided some reconciliation between the known biochemistry of methylation reactions and the observed stability of the genomic methylation landscape. In the context of maintenance methylation, this interdependence is thought to occur via multiple molecular mechanisms, including DNMT1 processivity (15–17)(in which an enzyme can methylate multiple neighboring CpGs on nascent DNA sequentially) and cooperative interactions, e.g., with UHRF1, which localizes, and in turn helps recruit DNMT1 molecules, to hemi-methylated CpGs (18, 19). Mathematical modeling has further demonstrated the importance of CpG

interdependence, also called collaboration, both in maintenance and de novo methylation, for long-term collective stability of methylated and unmethylated genomic regions (20–25).

It is not yet known to what extent processivity, versus other mechanisms of CpG interdependence, influences dynamics of maintenance methylation. Nor is it yet well-understood how local genomic context influences these mechanisms *in vivo*. In this study, we elucidate CpG-coupled-dynamics in maintenance methylation by use of statistical inference, bioinformatics, and stochastic modeling. We leverage experiments that measured methylation status of nascent-strand CpGs across post-replication timescales, genome-wide (26). From these data, we infer how the rates with which individual CpGs acquire methylation, post-replication, are correlated on neighboring sites. Using stochastic models, we demonstrate that the rate correlation as a function of genomic distance provides a mechanistic fingerprint for post-replication enzymatic processes. Comparing simulations to data allows us to extract quantitative insights from data, including the relative strengths of processive versus non-processive coupling mechanisms in different genomic regions, and the length of processive steps.

MATERIALS AND METHODS

Methods overview. The methodology of this paper can be summarized as follows. We reanalyzed data from Whole Genome Bisulfite Sequencing (WGBS) (27) and Replication-association Bisulfite sequencing (Repli-BS) (28) in human Embryonic Stem Cells (hESCs) using a combination of data-driven statistical inference and hypothesis-driven stochastic modeling. First, Maximum Likelihood Estimation was used to infer per-CpG post-replication remethylation rates from Repli-BS experiments, following our previously developed method (29). We analyzed the correlation of these data-inferred rates on nearby CpGs in different genomic contexts, such as Enhancer, Promoter, etc., to study regional differences in maintenance kinetics. Next, we studied the association of the strength of nearby-CpG remethylation-rate correlations with other local genomic/epigenomic features. To aid interpretation of the experiment-derived correlation functions and their regional differences and associations, we developed stochastic models of post-replication DNA methylation maintenance kinetics. Using these models, we generated simulated bisulfite sequencing datasets under different mechanistic hypotheses, and we compared the resultant *in silico*-generated rate correlation functions to experimental data.

Site-specific post-replication methylation kinetics inference. The post-replication methylation data (Repli-BS data) of Human Embryonic Stem Cells (HUES 64) were downloaded from GSE82045. In the Repli-BS experiment, cells were pulsed for one hour with bromodeoxyuridine (BrdU). Then, bisulfite sequencing of BrdU-labeled DNA captured CpG methylation reads from DNA that was replicated during the pulse interval. The pulse-chase experiment captured methylation level of CpGs at timepoints 0, 1, 4, and 16 hours post-pulse, thereby giving a genome-wide temporal readout of CpG-methylation over sixteen hours post-replication.

The Maximum Likelihood Estimation (MLE) procedure for inferring per-CpG post-replication methylation rates from Repli-BS data is described in detail in (29). Briefly, the temporally distributed binary read-data (methylated-1 or unmethylated-0) at each CpG site is fit by a Poisson process, with each site i characterized by two inferred constants, k_i and f_i, which represent the rate at which methylation accumulates at the site over the course of the experiment, and the steady-state (or long-time) fraction of cells in the measurement set that exhibit methylation at site i, respectively. We hereon refer to the inferred parameters k_i as the "post-replication remethylation rates" or simply the "remethylation rates". Details of the inference approach can be found in the Supplement (Extended Methods). Note that, while the per-site inferences are obtained based on an analytical, independent Poisson process model, the inferred rate parameters can nevertheless be used to investigate more complex types of dynamics and inter-site dependencies through correlations that are observed among inferred parameters on nearby CpGs.

The ability to infer a remethylation rate for a given CpG site, and the uncertainty associated with that inferred rate, depends on the read-depth of the experimental data, which varies across sites and across timepoints. Details of uncertainty quantification can be found in the Supplement and in our previous study (29). We estimated on average 30% error in any given estimate of rates k_i. Despite these relatively high error-rates, we validated our method by ensuring that ground-truth rate correlations (obtained from simulated data, see below), could be accurately recovered by the MLE inference pipeline.

Annotations of genomic regions. The GRCh37/hg19 genome was used as the reference genome in this paper. The region annotations for genes, promoters, exons, introns, 3'UTRs and 5'UTRs are downloaded from the UCSC Genes track in UCSC Table Browser, whereas the LINEs(Long Interspersed Nuclear Elements), SINEs(Short Interspersed Nuclear Elements), LTRs(Long Terminal Repeats) were extracted from RepeatMasker track. CGIs(CpG Islands), Enhancers were downloaded from CpG Island track and GeneHancer track, respectively. The promoters in this paper were defined as regions 2000bp upstream and 200bp downstream of transcription start sites (TSS). Local CpG density for a site was defined as the number of neighboring CpGs in a 500bp window centered at that site.

The chromatin accessibility data was retrieved from ENCODE/OpenChrom(Duke University) H1 cell line. The histone modification peaks were downloaded from ENCODE/Broad. The regions of transcription factor(TF) peaks or TFBR denoted in this paper were acquired from ENCODE ChIP-seq clusters for 161 TFs in H1 cells.

Stochastic simulation. Region-specific simulations of single-CpG stochastic enzyme-kinetic models were carried out using two candidate mechanisms, the distributive model

and the processive model (Figure 4). The model reactions and associated rate parameters are graphically depicted in Figures 4b and 4c and are described in greater detail in the supplemental material. The Distributive model was simulated using the Gillespie Stochastic Simulation Algorithm (30). To incorporate 1D diffusion into the Processive model, we used a First Passage Time Kinetic Monte Carlo algorithm inspired by (31).

Simulations are performed for stretches of N CpGs, ($N = 25000 - 75000$). Simulations mimic two types of experimental data modalities: WGBS and Repli-BS. To simulate WGBS experiments, for a region of CpGs, the simulation is initialized at post-replicaton time=0, and then read out at randomly sampled timepoints between zero and 24 hours later, to reflect the variable post-replication timings of bulk cells in WGBS experiments. Ten simulation replicates are combined to generate estimates of average per-site methylation levels. To mimic the Repli-BS experiment, the methylation status of CpGs in the simulation were read out at timepoints sampled from intervals matching pulse-chase experiments (28), including uncertainty with respect to true post-replication timing. That is, given the finite BrdU pulse-length, the 0-hour experimental timepoint is assumed to correspond to $t \in [0-1]$ hours post-replication, the 1-hour experimental timepoint corresponds to $t \in [1-2]$ hours post-replication, etc.. Therefore, timepoints of simulation readout were sampled from $t_{chase} + r \in [0,1]$ hour, where experimental timepoints t_{chase} were $0, 1, 4, 16$ hours and r is a uniformly distributed random number. The number of simulations and read-outs at each timepoint were chosen to mimic the distribution of experimental read-depths. The synthetic Repli-BS data were then processed with the same MLE procedure as the experimental data, to infer per-CpG methylation rates.

Simulations were used to validate our statistical inference procedure. We tested that ground-truth correlation functions produced by the two models could be recapitulated by the inference procedure. Further details of the simulated models, algorithms, and parameter values can be found in the Supplement (Extended Methods). MATLAB codes are available in a github repository: Read-Lab-UCI/DNA-methylation-kinetics-correlation.

RESULTS

Correlation of DNA methylation *rate* and *state* on neighboring CpGs. An analysis of cytosine methylation within newly replicated DNA over time (via Repli-BS) revealed that some genomic loci exhibit a pronounced lag in methylation maintenance (26). In our previous work, we developed a statistical inference procedure to obtain post-replication methylation ("remethylation") rates from Repli-BS data (29), which further supported the variability of remethylation rates across the genome. The inferred parameters quantify the rate of accumulation of methylation at each CpG site across cells in the measurement set over the experimental timecourse of 0-16 hours post-replication (see Methods). We first observed that these remethylation rates are correlated on nearby CpGs in our previous study (29).

DNA methylation levels (also hereon denoted methylation "states", or the fraction of cells exhibiting methylation at an individual CpG) on neighboring CpGs are also correlated (24, 33). Computing the correlation as a function of genomic distance in basepairs reflects information such as the size of persistently methylated (or demethylated) domains. We hypothesized that correlation of post-replication methylation rates on individual CpGs ("rate correlation") could give insight to local control of post-replication methylation dynamics, and thereby yield additional information beyond that contained in methylation state-correlation. To this end, we first computed correlation from the two different data modalities (remethylation rate derived from Repli-BS and methylation state derived from WGBS) to relatively short ($<$ 1 kilobase-pair) distances, and compared them across different genomic contexts.

Post-replication remethylation rates are correlated on neighboring CpGs, and correlation varies with genomic regional context. We computed the Pearson correlation coefficient of remethylation rate and state as a function of inter-CpG distance in basepairs. We further analyzed the resultant correlation functions after filtering CpG pairs by genomic context in addition to their inter-CpG distance, from 2 (the minumum) to 1000 bp (Figure 1a). Here, the genomic regional contexts were based on genomic annotations acquired from the UCSC genome table browser and included features such as 3'UTRs, enhancers, etc.(see Methods).

We observe both common and distinct features in the correlation functions, when comparing different genomic regional contexts for both data modalities (i.e., rate-correlations and state-correlations). Common to all regions, and to the genomic average with no region-filtering, remethylation rate-correlation decreases rapidly with distance until approximately 100 bp (red curves in Figure 1a). After 100 bp, a slower decay is evident in all regions. However, this steep decrease in correlation and rapid change in decay is not observed in correlation of methylation state (navy curves), which generally shows a slow (persisting to $>$1 kb), smooth decay, albeit with wide variation in decay-lengths between regions. We quantify the average magnitude of short-range correlation by the mean correlation to 1 kbp. These magnitudes are highly variable across genomic contexts and across modalities. Genome-wide, the magnitude of rate-correlation is low (0.06) compared to that of state-correlation (0.82). Across regions, CGIs(CpG Islands) show the highest magnitude of state-correlation (0.85) while transcription factor binding regions (TFBR) shows the highest magnitude of rate-correlation (0.15). By contrast, the lowest average magnitude of state-correlation and of rate-correlation are both found within SINEs(Short Interspersed Nuclear Elements) regions with 0.18 and 0.035 correlation, respectively. We hereafter refer to these differences in remethylation rate- and state-correlations according to genomic context as "genomic region specificity".

In general, the results for methylation state indicates that methylation on CpGs within 1kb are highly correlated genome-wide. Exceptions are within SINEs and

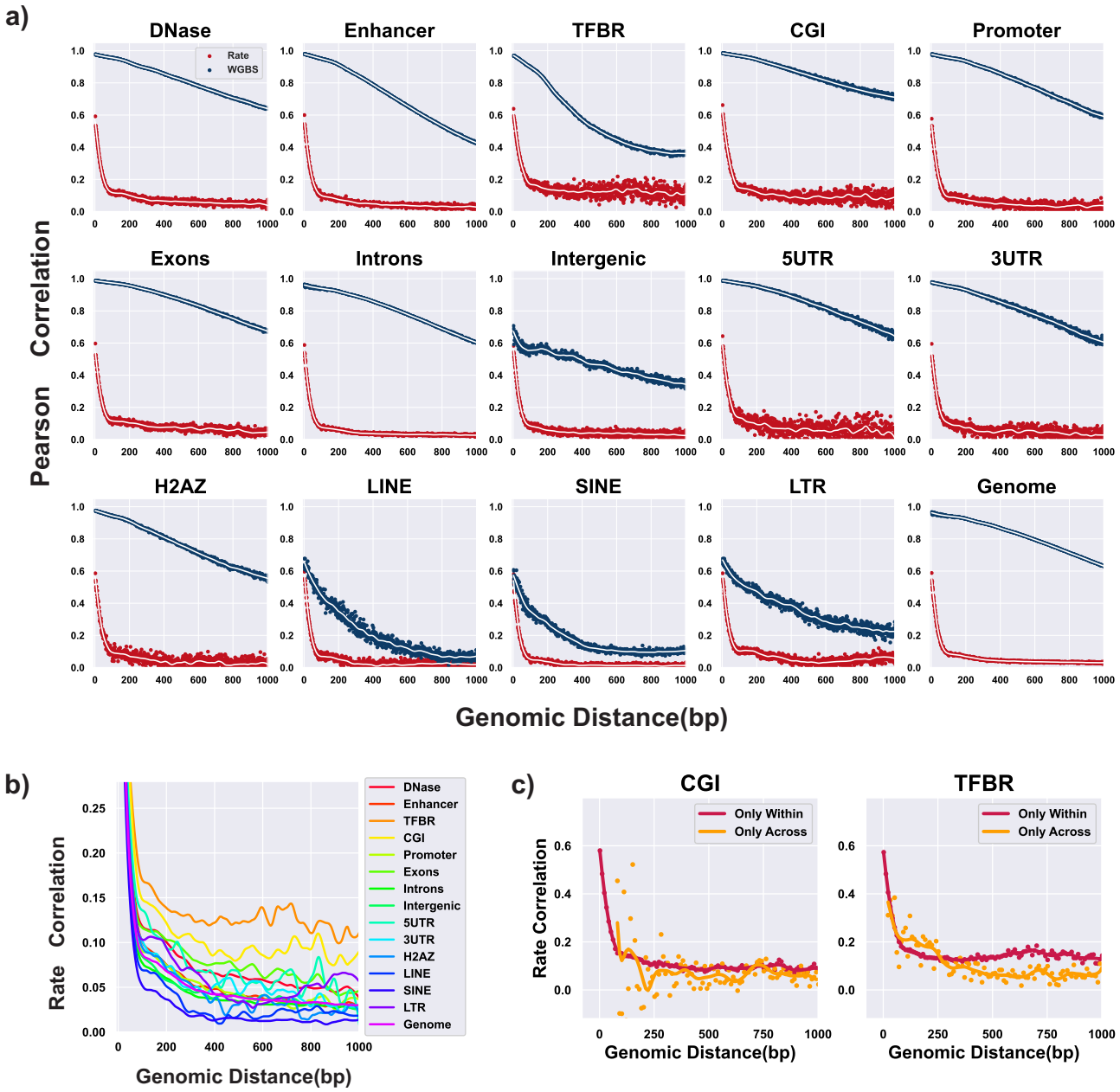

Fig. 1. (a) Correlation of remethylation rates (red, y-axis) and correlation of percent methylations (i.e., methylation "state", blue, y-axis) of nearby CpGs at given genomic distances (x-axis), separated in panels by genomic regional context. The remethylation rates for individual CpG sites were inferred from experimental Repli-BS-seq data, and the percent methylation was obtained from separate WGBS data with the same cell line (HUES64) as Repli-BS-seq. Dots: raw rate correlation in base pair resolution, lines: smoothed correlation curve by LOESS(LOcally Estimated Scatterplot Smoothing,(32)), smooth function in Matlab was used with 'loess' as method and 100 as span. (b) Correlation of remethylation rates with curves from different genomic regional contexts overlapped and y-axis zoomed. Only the LOESS smoothed curves are shown for clarity. (c) Correlation functions computed retaining only pairs of sites identified either as within the same localized region (red) or not within the same localized region (orange).

LINEs(Long Interspersed Nuclear Elements), where state-correlation has significantly decayed by 1 kbp.

Although the methylation rate correlations show an apparently uniform sub-100-bp decay across regions, differences are visible in their longer-lengthscale (>100 bp) decay profiles (Figure 1b). In particular, CGIs and TFBRs (Transcription Factor Binding Regions) show the strongest long-distance correlation, persisting near or above 0.1 past 1 kbp. We further separated the analyzed CpGs from CGIs and TFBRs into "within" and "across" region pairs (1c). "Within" correlation is computed for CpGs that are within one contiguous region (as defined by the filtering protocol), whereas "across" correlation only retains pairs of sites that are assigned to non-contiguous regions. The resultant rate correlation functions indicate generally stronger contribution of within- versus across-region correlation (note that the number of datapoints for "across" correlations is lower for short distances, since nearby CpGs are more likely to be in contiguous regions). In all, these results indicate different magnitudes and decay-lengths of inter-CpG coupling in maintenance methylation. In particular, they point to more pronounced CpG collectivity of maintenance kinetics in contigu-

ous CGIs and TFBRs. That is, neighboring CpGs in CGIs and TFBRs tend to have more correlated remethylation rates and thus more similar methylation across post-replication time, as compared to other regions. However, all regions showed correlated rates to some degree.

We note several additional features in the correlation functions, including apparent periodicity (e.g., in state-correlation, Intergenic region, Figure 1a) and shoulders or local peaks (e.g., in several of the rate-correlation curves in Figure 1b). Some of these features appear consistent with nucleosome positioning, which could indicate coupling between maintenance methylation processes and/or the methylation landscape with nucleosomes. A detailed investigation of all of these features is outside the scope of this work; we focus instead on the broad features of the methylation rate correlation, namely the sub-100 bp rapid decay and region-specific, $>$ 100bp slower decay.

Regional correlation of DNA methylation maintenance kinetics is increased with CpG density and chromatin accessibility, and decreased with higher bulk methylation levels. To investigate the factors associated with the observed region specificity, the association between remethylation rate correlation and other local genomic characteristics were examined. For each individual CpG in the dataset, a measure of the local chromatin accessibility is collected from DNaseI hypersensitivity data in ENCODE/OpenChrom (see Materials and Methods). The local CpG density surrounding a given site is calculated by the number of neighboring CpGs within a 500bp window. "Bulk" methylation refers to measurement of DNA methylation in human embryonic stem cells(HUES64) measured by Whole Genome Bisulfite Sequencing(WGBS). We plotted the magnitude of the remethylation rate correlation (0-1000 bp) for each region versus average chromatin accessibility (e.g., DNAse hypersensitivity enrichment) (Figure 2a), local CpG density (Figure 2b), and WGBS ("bulk") methylation percentage (Figure 2c). In all cases, a strong linear correlation was found between the magnitude of remethylation rate correlation and these three factors; rate correlation is positively correlated with DNase level and CpG density, while it is inversely correlated with bulk methylation level. These results suggest that the genomic regional differences observed in rate correlation in Figure 1 may be driven globally by variation in chromatin accessibility, CpG density, and background methylation landscape.

Local methylation correlation varies across post-replication time and shows persistent region specificity. In order to investigate the origins and regional differences of methylation correlation, we focus on three representative genomic regions characterized by high (CGI), medium (Enhancer) and low (SINE) methylation state correlation. In figure 3, we plot three types of correlation functions for these three regions: in addition to bulk WGBS correlation and rate correlation, we also plot correlation of methylation state in "nascent" DNA (which contain a subset of measurements from the full temporal Repli-BS dataset, corresponding to the 0-hour timepoint of the original pulse-chase experiment(26). Note that, in contrast the Figure 1, the curves in each panel of Figure 3 are site matched. In practice, remethylation rates are available for only a subset of CpGs, as compared to WGBS measurements. This is because the rate constant is undefined where no methylation is measured, and is often unidentifiable when methylation is very low. Thus, there is significant overlap between the sites for which rates are not available and sites for which WGBS percentages are 0 or near 0. We reasoned that much of the methylation state correlation in Figure 1 could arise from the bimodal nature of methylated and demethylated regions. For more direct comparison to rate correlation on a site-specific basis, we thus filtered to a common set of sites, which in practice retains mostly sites with intermediate to high methylation in WGBS.

We observe that local methylation correlation varies across post-replication time (Figure 3). This can be ascertained from lower methylation correlations observed within nascent DNA as compared to bulk DNA in all three regions. We attribute this difference to a general increase in correlation that tracks with post-replication time, since bulk WGBS measurements reflect temporal variability from cells in various stages of the cell cycle and from differences in replication-timing across the genome (meaning it largely contains matured DNA strands), whereas the nascent data in principle captures reads from genome within one hour post-replication (26). We also observe that rate correlations are lower than correlations in both nascent and bulk DNA within all three regions: it is initially lower as genomic distance approaches two (i.e., for directly adjacent CpGs) and falls off more rapidly with distance. In all studied regions, the nascent methylation correlation is intermediate between rate and bulk correlation (with the exception of very short distances in SINEs).

The trend in genomic region specificity of methylation correlation is persistent across the three types of correlation functions. This in turn suggests that the region specificity is persistent across post-replication time. For example, CGIs consistently show the highest correlation compared to the other regions in rate, bulk methylation, and nascent methylation. Conversely, SINE consistently shows the lowest correlation. These results suggest that the processes that govern coupling (or interdependence) of methylation among neighboring CpGs differ depending on the genomic regional context, and that these processes are region-specific already at early post-replication timepoints. Additionally, the distinct shapes and magnitudes of the CpG-neighbor correlations across time suggest that somewhat different processes may control CpG-neighbor-interactions at early versus late post-replication times.

We hypothesize that the three correlation modalities can be interpreted as follows: rate correlation reflects the dynamic mechanisms of maintenance methylation, thus shedding light on early post-replication-time processes. In contrast, bulk methylation state correlation largely reflects the steady-state methylation landscape (where the term *steady-state* is used loosely), i.e. reflecting the balance among methyl-reading/writing/erasing processes operating across post-replication time to regulate the methylation landscape,

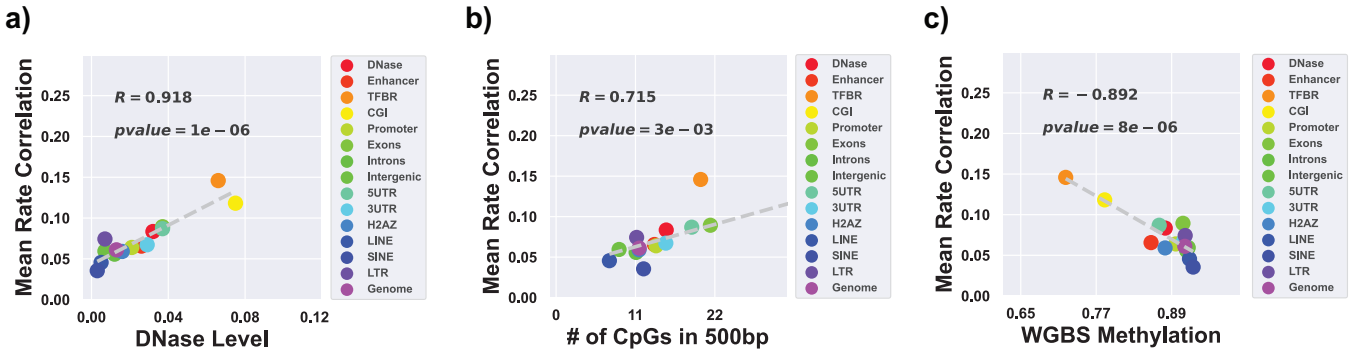

Fig. 2. Mean remethylation rate correlation by region, plotted versus other quantified, localized genomic measurements from independent measurements: (a) Magnitude of mean remethylation rate correlation (equal to the average over rate correlation for all inter-CpG distance < 1 kbp (i.e., integers from 2- 1000) in the given region) versus mean regional chromatin accessibility, as quantified by DNase level; (b) Versus mean local CpG density (mean number of neighboring CpGs within a 500bp window); (c) Versus mean WGBS ("bulk") methylation level. Note that the datasets are site-matched, so the analysis is restricted to sites that tend to have intermediate to high methylation, since these are the sites for which remethylation rates are available.

but largely reflecting the stable methylation landscape of a given cell type. Nascent methylation state reflects a mixture of the two, as the experiment "captures" CpGs in transit between their state immediately post-replication and steady state. In the following sections, we test this hypothesis by use of computer simulations and model-guided data analysis.

Region-specific stochastic simulations of post-replication methylation dynamics. In order to gain further biological insight from the experiment-derived methylation correlation functions, we perform region-specific stochastic simulations of maintenance methylation, and use these simulations to generate synthetic data analogous to the various experimental bisulfite sequencing data modalities. From these synthetic data, we compute regional correlation functions and compare to those derived from experiments. Briefly (see Methods), the simulations track nascent-strand methylation status of stretches of sequentially positioned CpGs, numbering on the order of tens of thousands. Each stochastic simulation tracks the binary (methylated or not methylated) status of the "nascent-strand" CpGs. At the start of the simulation (representing exactly time 0 with respect to DNA replication at that site-i.e., the time of nucleotide addition), all nascent CpGs are assumed to be unmethylated. The presence or absence of methylation on cytosine bases on the opposing parental-strand at time 0 is determined probabilistically from a data-derived regional methylation landscape that acts as the simulation input. If the parental cytosine is methylated at time 0, then the nascent cytosine is assumed to be a target for DNMT1-catalyzed methylation, and it will acquire methylation stochastically at some post-replication timepoint, according to the chemical reaction kinetics encoded in the model. If the parental cytosine is unmethylated at time 0, then DNMT1 does not target the nascent cytosine for methylation, and the site will remain fully unmethylated. In this way, the model tracks only unidirectional maintenance methylation, and does not include active demethylation reactions. It also does not account for any de novo methylation activity. The simulation tracks post-replication timescales (following experiments, to approximately 16 hours), up to but not including subsequent replication events.

Region-specificity is encoded at the start of the simulation in two ways: (1) the CpG positions and (2) the local methylation landscape, meaning: the probability of the nascent CpG to be a target for DNTM1-catalyzed methyl-addition due to the presence of methylation on the parental strand CpG dinucleotide. Both of these quantities are derived from experimental WGBS data with a regional filter to retain only CpGs in the desired region. For example, in simulating a CGI region, the ith simulated CpG ($i \in [1..N]$) has a genomic position x_i and a probability f_i to be targeted for methylation. We obtain both x_i and f_i from WGBS data from hESCs, where x_i is the integer site-ID for the cytosine (which is identified as being located within a CGI) and $f_i \in [0, 1]$ is taken to be equal to the measured methylation fraction at that site. For example, if a given CpG in the dataset has a WGBS-measured methylation fraction of 0.8, then the model assigns the parental cytosine to be methylated at the start of the simulation with a probability equal to 0.8. Strands are simulated in replicate. With sufficient replicates, the simulation eventually recapitulates the experiment-derived input methylation landscape, if it is run for a period of time that is sufficiently long enough to reach steady state. Note that, as discussed previously, the WGBS-data-derived landscape likely reflects some degree of replication-associated temporal variability (26), rather than a true steady-state. Nevertheless, use of the WGBS-background-methylation landscape as the simulation input allows us to encode realistic region-specific differences in CpG densities and qualitative differences in bulk methylation levels.

Simulations probe mechanistic origins of regional methylation correlation. We use simulations to generate synthetic data mimicking the various bisulfite sequencing datatypes (see Methods). We then compute the correlation functions for the synthetic data. Figure 5 shows simulation-derived correlation functions for three representative genomic regions from Chromosome 1 for two mechanistic models termed Distributive and Processive. Briefly, DNMT1 binds to nascent CpGs and catalyzes the addition of a methyl group. In the Distributive model (Figure 4b), the enzyme unbinds after the catalytic step, and must independently rebind

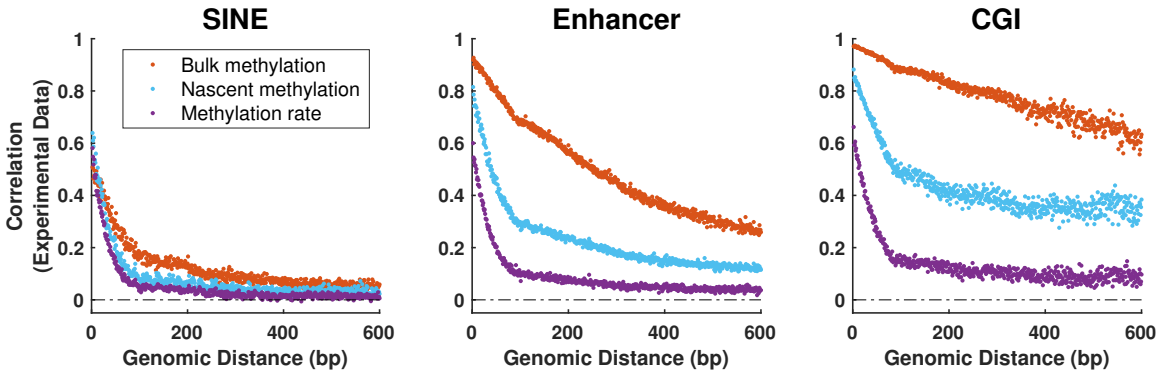

Fig. 3. DNA methylation correlation varies over post-replication time; regional differences in correlation persist across different data modalities and are present when comparing identical sets of CpGs. Data from three representative regions and three methods of extracting methylation correlation; all three curves in each panel are derived from the same set of CpGs. Red: correlation of methylation state of neighboring CpGs from Whole Genome Bisulfite Sequencing Experiments (WGBS, "Bulk"). Blue: Correlation of methylation state of neighboring CpGs from WGBS on nascent DNA from Repli-BS-seq experiments (i.e., < 1 hour post-replication) ("Nascent"). Purple: Correlation of remethylation rates inferred from Repli-BS-seq experiments. Data for rate-correlation (purple) is identical to that of Figure 1a. Data for bulk methylation (WGBS) is also derived from the same dataset as Figure 1a, but filtered to retain only those CpG sites for which a rate was available (determined by per-site read-depths, see Methods). Note that the nascent (blue) and rate (purple) curves are not from independent datasets, since the nascent data is a subset of the total Repli-BS-seq dataset from which the rates were inferred. Also note that the red and blue Pearson correlation functions are computed from raw data, where the purple correlation function is computed on inferred rate parameters from the MLE pipeline.

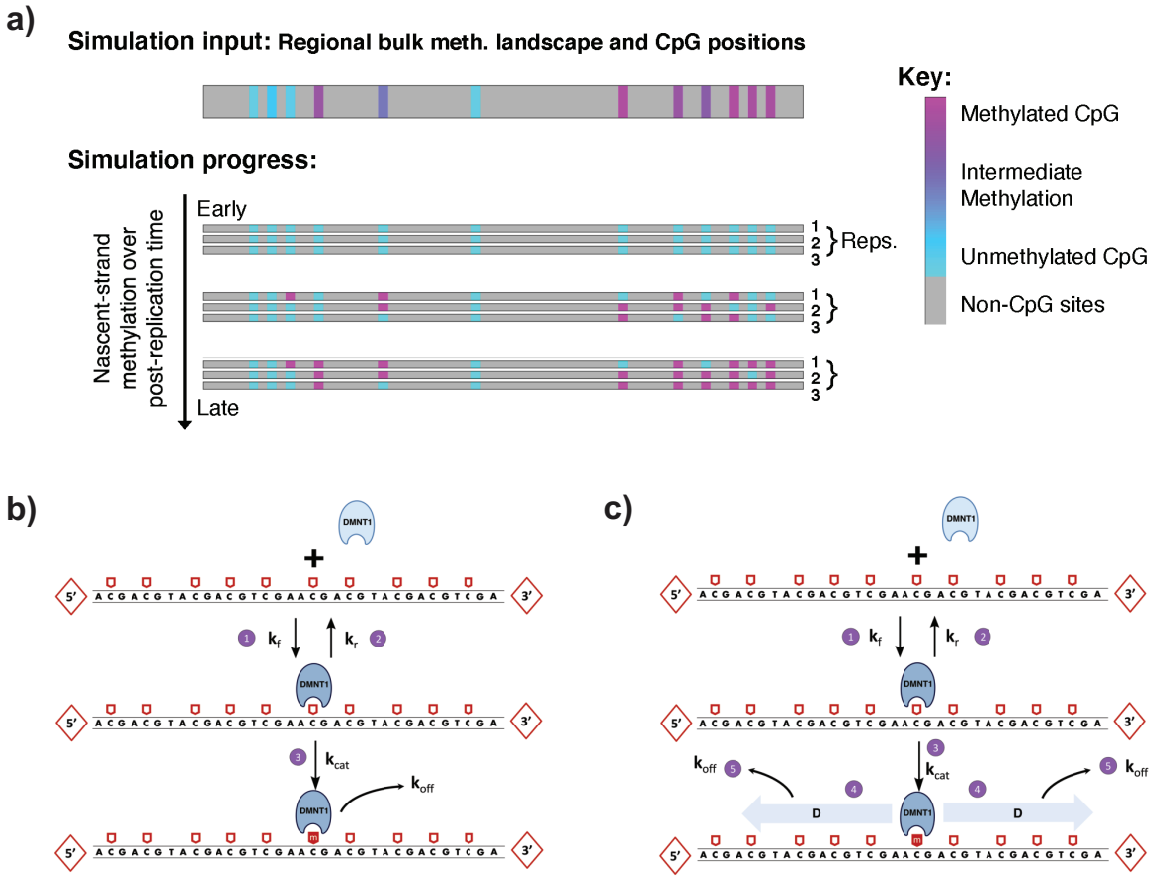

Fig. 4. Schematic of stochastic simulation of DNA methylation maintenance process in different genomic regions. (a) Each region-specific simulation models a strand of N CpGs ($N = 25000$ to 75000) and takes a regional, data-derived background methylation landscape as input. The landscape comprises CpG positions (which vary in local density and inter-CpG distances in different regions) and bulk methylation state as measured in WGBS experiments. Multiple strand replicates ("Reps.") are simulated over time. Immediate-post-replication DNA is assumed to be unmethylated on the nascent strand, and methylation status (0 or 1) of the parent strand is sampled probabilistically from the input methylation landscape. During the simulations, DNMT1 targets hemimethylated CpGs for methylation, according to either a distributive (b) or processive (c) mechanism. The distributive mechanism assumes that the enzyme binds to each hemi-methylated CpG independently. The processive mechanism assumes that, after catalyzing methylation at a CpG, DNMT1 can remain bound to DNA and reach nearby hemimethylated CpGs by linear diffusion along DNA.

to an available hemi-methylated CpG in order to catalyze a subsequent methyl addition. In the Processive model (Figure 4c), the enzyme can remain bound to DNA after methylating a CpG, and can travel along DNA (via a one dimensional diffusive random walk) to reach neighboring hemi-methylated CpGs and again bind and catalyze methylation. The kinetic

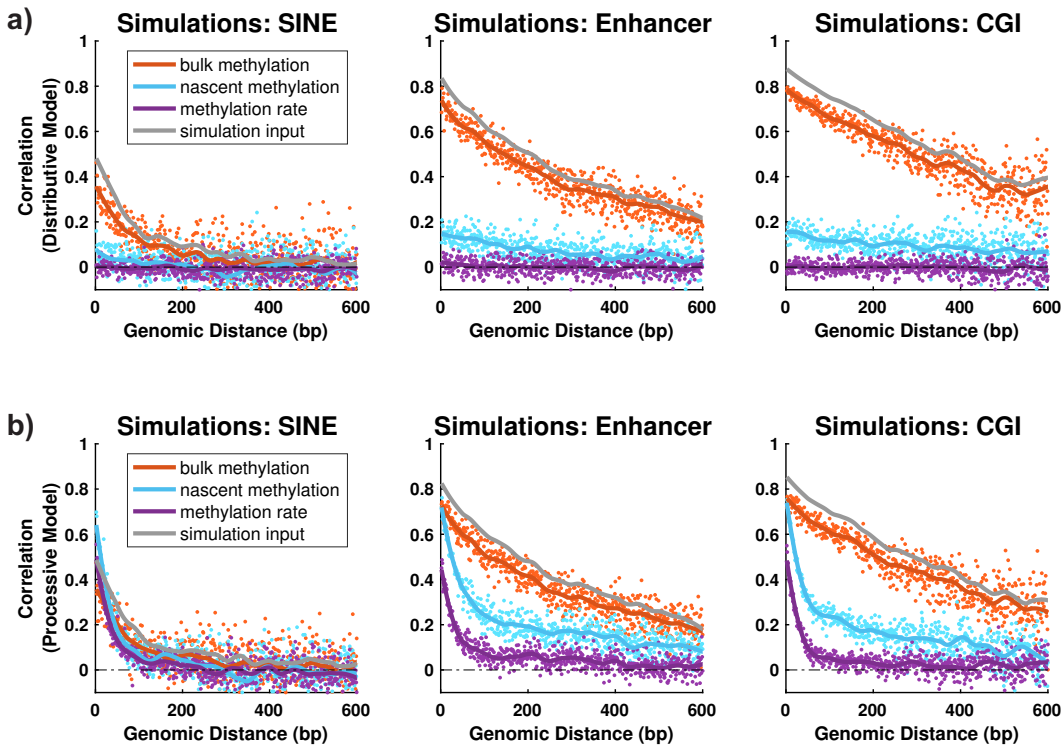

Fig. 5. Simulated correlation functions for three representative regions (SINE, Ehnhancer, CGI) and using the distributive (a) or processive (b) mechanism. The mechanistic simulations are performed using experimental regional methylation landscapes for SINE, Enhancer, and CGI as initial condition inputs (grey curves). The simulations provide synthetic data corresponding to each of the three experimental data modalities, as in Figure 3. Synthetic data are processed in the same way as the experimental data to compute the inferred remethylation rates (purple) and Pearson correlation functions. Both models recapitulate temporal trends in rate correlation seen in experiments, Figure 3 (rate correlation $<$ nascent $<$ bulk). Only the processive model captures nonzero rate correlation sub-100-bp; neither model captures low, persistent correlation $>$ 100 bp observed in Figure 3.

parameters are described in the Supplement.

In both Distributive and Processive mechanisms, the simulated bulk methylation correlation approaches the input methylation landscape in the three regions. This is expected, because the input landscape dictates parental strand methylation in the model, and the model assumes that as time progresses, the nascent strand methylation will ultimately match the parental strand input. In contrast, the shapes of remethylation rate and nascent correlations are distinct from those of the input methylation landscape, and are dependent on the model mechanism. The Distributive model produces no correlation in remethylation rate, in agreement with our previous results (29). In contrast, the Processive model produces a nonzero rate correlation that appears to qualitatively reproduce the experimentally observed rapid decay of correlation with genomic distance up to approximately 100 bp. Neither model reproduces the low but persistent correlation visible at distances greater than 100 bp in CGI and Enhancer in Figure 3.

We first present analysis of the temporal variation in simulated methylation correlation, focusing first on the Distributive mechanism. In the Distributive model, the nonzero correlation that appears in nascent and bulk methylation simulations results entirely from the background methylation landscape that serves as the simulation input, i.e. correlation is not introduced by the model mechanism but rather is encoded in the initial conditions of the simulation. Thus, the nascent methylation correlation function is lower than the bulk due to the fact that, in the simulation, some but not all CpGs have reached their steady-state methylation level before read-out, thereby only partially reflecting the steady-state landscape. This is because the simulation mimics the nascent methylation experiment, in which there is uncertainty with respect to true post-replication timing, but nascent reads are assumed to be captured within up to 1 hour post-replication (see Methods). According to the model, the degree to which the nascent correlation underestimates the bulk can be understood by comparing the timing of the read-out in the simulation (i.e., the sampled post-replication time) with the timing of sites acquiring methylation.

In the model, the timing of individual CpG methylation events is stochastic, but determined on average by the choice of model mechanism and kinetic parameters. Using the Distributive mechanism for illustration purposes, the median time among N CpGs to be methylated in the simulations, ($t_{50}^{stochastic}$) is found to be close to the deterministic prediction of Michaelis-Menten kinetics to reach 50% max of product formation, i.e.,

$$t_{50}^{stochastic} \approx t_{50}^{M.M.} = \frac{K_M + S_0}{2k_{\text{cat}}E_0}. \quad \textbf{(1)}$$

For example, the model parameters of Figure 5 predict $t_{50}^{M.M} = 0.61$ h, while the Distributive simulations in SINE give $t_{50}^{stochastic} = 0.62$ h (note that these numbers are slower than the experiment-derived estimate of 0.34 h). Us-

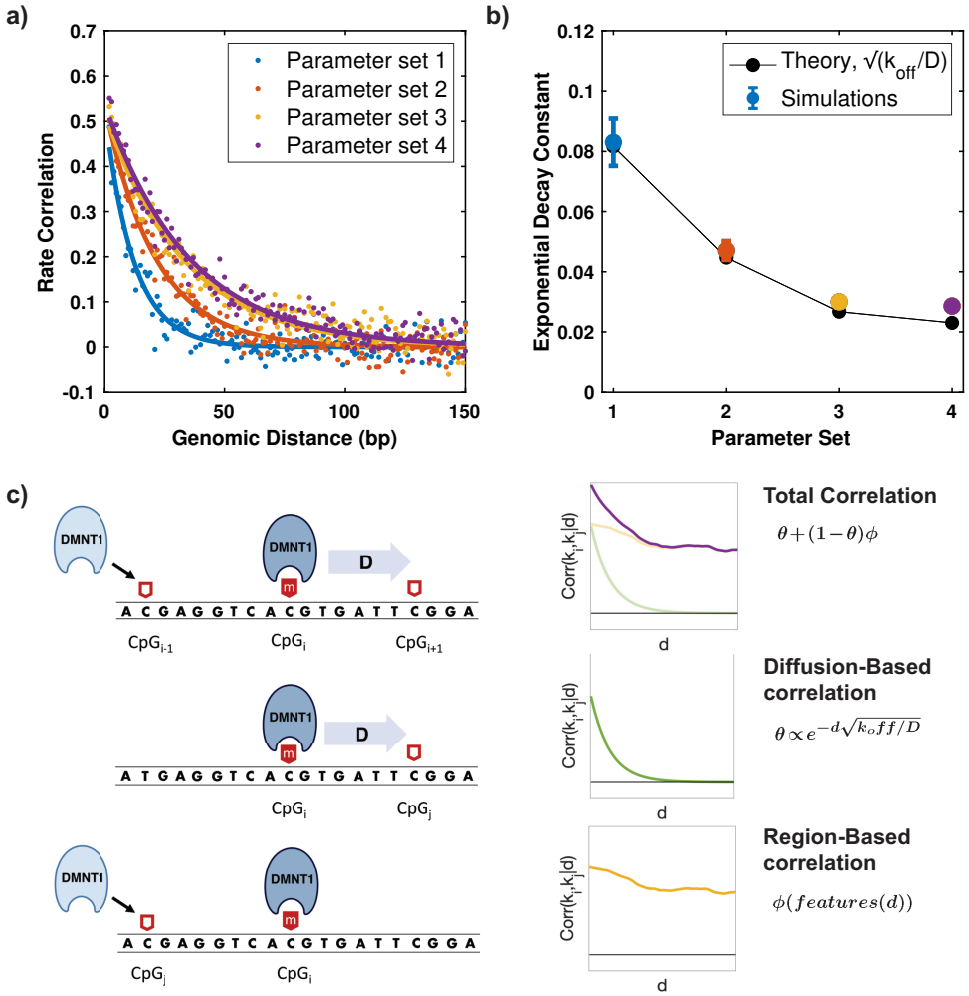

Fig. 6. Using analytical theory and stochastic simulations to analyze mechanistic contributions to DNA methylation rate correlation. (a) Exponential decay with genomic distance of rememthylation rate-correlation, according to the processive mechanism. Processive simulations were performed choosing four different values for the parameter ratio D/k_{off}. After inferring remethylation rates from the synthetic data by the MLE pipeline, ensuing rate correlations (colored dots) were fitted to a single exponential decay function (colored lines) in an unbiased manner. (b) Fitted decay constants from simulated data overlaying the theoretically predicted decay constant from the model parameters ($\sqrt{(k_{off}/D)}$), showing close agreement of theory and simulation, despite the increased complexity of the methylation dynamics simulation, as compared to the 1D analytical model. (c) Mathematical model of mechanistic contributions to rate correlation, termed diffusion-based and region-based. The total rate-correlation (purple curve) is proposed to arise from both types of processes, and can be mathematically decomposed into components. The diffusion-based component (green curve) is due to an individual enzyme at CpG$_i$ reaching nearby target CpG$_j$ by diffusion, and shows exponential decay with genomic distance d. The region-based component (yellow curve) is attributed to any other sources of correlation in remethylation rates, e.g., when two different enzyme molecules reach their targets (CpG$_i$ and CpG$_j$) with correlated arrival times, and does not have an analytic expression. Experimental results from Figures 1-3 indicate that the region-based correlation component decays more slowly and is associated to other features of the genomic/epigenomic landscape that vary by region, including bulk methylation level, CpG density, and chromatin accessibility. These features are themselves correlated as a function of distance d.

ing the simulated data, we estimate that $\sim 30\%$ of CpGs have reached their steady-state methylation level by post-replication $t = 0.5$ (the assumed average timing of readout in the nascent experiment). This explains why the nascent correlation in Distributive model is low but non-zero, as it partially recapitulates the steady-state landscape.

In the Processive model, the nascent correlation reflects a mixture of both the rapid short-distance decay encoded by the model mechanism (i.e., observable in the Processive rate correlation) and the longer-distance correlation that is encoded in the input methylation landscape. This is in contrast to the Distributive mechanism, which introduces no correlation between CpGs, beyond that encoded in the input landscape.

By qualitatively reproducing the experimentally-observed gain of methylation correlation over time and the intermediate nature of the nascent correlation, the simulations lend support to the above-described interpretation of the three correlation functions. In particular, simulations support that mechanistic insights on maintenance methylation can be derived using the rate correlation, since it effectively separates correlations introduced by early post-replication-time processes from those operating at longer timescales (and which dominate the bulk WGBS correlation). That is, the rate correlation can be used to distinguish between hypothesized mechanisms. The simulations also demonstrate that, while the Processive mechanism partially recapitulates experiment-derived rate correlations, it cannot explain the longer-distance, slow correlation decay in CGI.

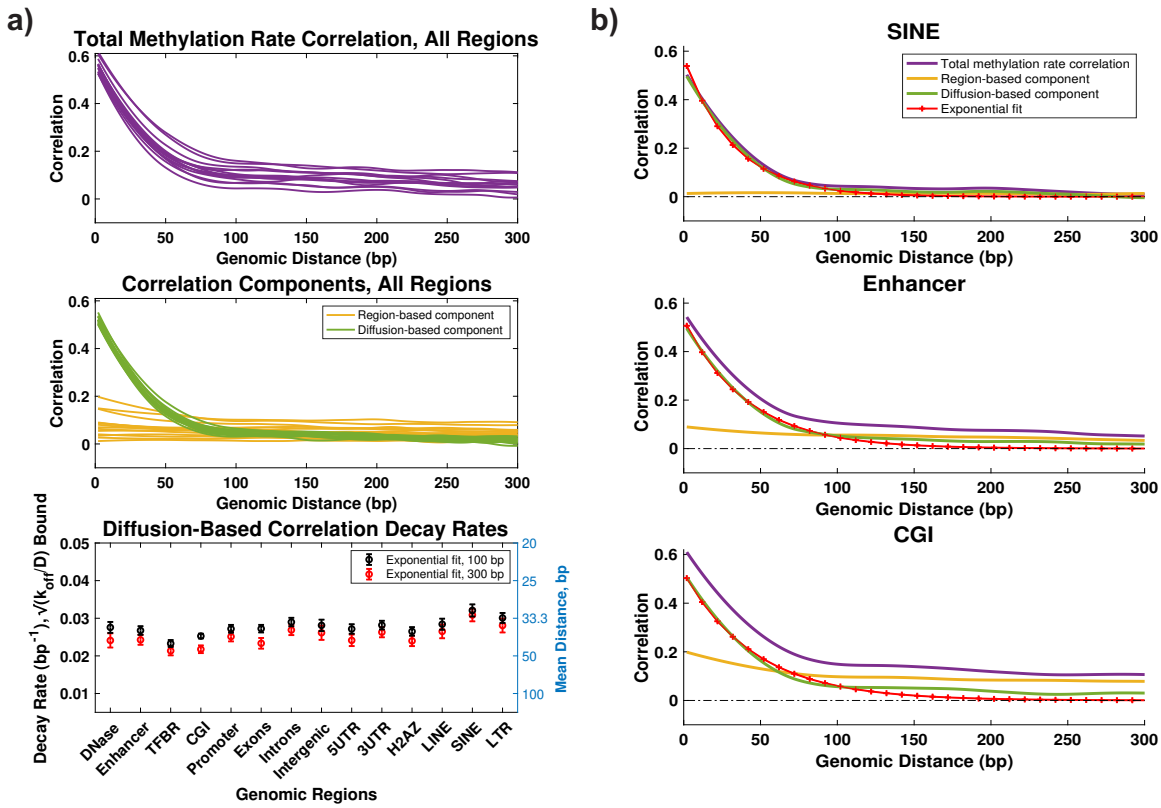

Fig. 7. Decomposition of experiment-derived rate correlation functions by region into region-based and diffusion-based components. (a) 14 purple curves are the remethylation rate correlation functions for each of the 14 studied regions (curves are the same as those shown in separate panels in Figure 1a, except the whole-genome curve, which is not shown here). From each total rate-correlation curve (purple), the region-based correlation component (ϕ, yellow curves) was extracted by a resampling method that retained alignment of remethylation rates to local site-specific features, while deleting the explicit distance-dependence by shuffling CpG positions (see text). The remaining correlation components (θ, green curves), obtained from $Total = \theta + (1-\theta)\phi$ (see Figure 6), collapse onto a common decay profile, after removal of ϕ. Bottom: θ is well-fit in each region by a single exponential decay function, supporting attribution of this correlation component to the processive mechanism (i.e., to diffusion). The decay constants are relatively uniform across regions, supporting common diffusive dynamics across the genome. The average decay constant, $0.028bp^{-1}$, corresponds to an average distance traveled by DNMT1 to a nearby target CpG of 36 basepairs. For the unbiased exponential fits, curves were fitted from 2 up to either 100 or 300 bp; averages were taken from the 100 bp fit, reasoning that these better estimate the rate of $<$ 100 bp decay. The difference between the two values is likely due to residual correlation in θ after 100 bp, due to imperfect estimation of ϕ by our resampling procedure. (b) Simultaneous plotting of correlation components and exponential fits for the three representative regions (same data as left column, but providing a more detailed view). SINE shows a nearly-zero region-based component (yellow), indicating that the total rate-correlation is already reasonably well-fit by an exponential decay (however, imperfect fit is visible in all regions where θ remains slighly above 0 after 100 bp). In contrast, Enhancer and CGI both show small but persistent region-based correlation that is the dominant contributor to the total correlation beyond 100 bp (Enhancer) and beyond 50 bp (CGI). In CGI, it is apparent that the 300-bp exponential fit slightly underestimates the initial decay because of the residual correlation after 100 bp.

Rate correlation in Processive model decays exponentially. We used simulations and theoretical modeling to understand how rate correlations arising from the Processive mechanism depend on model parameters. In an idealized theoretical model, we find that the rate correlations are related to the diffusion constant D and the unbinding rate, k_{off}, as:

$$Corr(X,Y|d) \propto \mathrm{e}^{-d\sqrt{k_{off}/D}} \qquad \textbf{(2)}$$

where X, Y are pairs of remethylation rates on sites that are distance d apart (along the DNA strand). Equation 2 can be understood as follows:

Consider two target sites labeled i and j, located on a 1D strand of DNA, that are located some distance d away. Let τ_i be the time (post-replication) at which site i acquires methylation (similarly, τ_j). Assume that in a particular realization of the stochastic process, two possible events can occur:

1. Each of the sites is independently methylated, in which case τ_i and τ_j are independent samples from some common distribution, denoted $\mathcal{T}_{\mathrm{methyl}}$. $\mathcal{T}_{\mathrm{methyl}}$ reflects the distribution of times required for all biochemical steps in the site acquiring methylation, e.g., for DNMT1 to bind DNA from the nucleosol, form a complex with the CpG site, and successfully catalyze methylation.

2. Alternatively, one of the sites could be methylated by an enzyme that has reached it by diffusion after first acting on the other site. In this case, assuming (without loss of generality) that site i is the first, then τ_i is sampled from $\mathcal{T}_{\mathrm{methyl}}$ and $\tau_j = \tau_i + \tau_D$, where τ_D is sampled from random variable $\mathcal{T}_{\mathrm{diff}}$. $\mathcal{T}_{\mathrm{diff}}$ reflects the distribution of times for DNMT1 to diffuse the distance d to reach the neighbor site, form a complex, and successfully catalyze methylation.

The likelihood of event 1 or 2 above taking place depends in part on the probability that the enzyme can remain bound to DNA and diffuse a distance d before unbinding. In the limit

of $k_{off} \gg D$, then event 1 always occurs, and

$$Corr(X,Y|d) = 0 \quad \textbf{(3)}$$

where X,Y now denotes pairs of methylation times at sites with intervening distance d. The correlation is zero because τ_i and τ_j are assumed to be i.i.d. random variables. Conversely, in the limit of $k_{off} \ll D$, then event 2 always occurs, and

$$Corr(X,Y|d) = 1, \quad \textbf{(4)}$$

(in the limit of $\tau_D \ll \tau_i$). Assuming that the enzyme need only *reach* the next target site by diffusion, in which case the catalytic step occurs subsequently with probability 1, then the correlation is equal to the probability of event 2 occuring, i.e.

$$Corr(X,Y|d) = Prob(\text{enzyme reaches neighbor}) = \mathrm{e}^{-d\sqrt{k_{off}/D}} \quad \textbf{(5)}$$

where the exponential dependence on linear distance, with decay rate $\sqrt{k_{\text{off}}/D}$, is obtained from a 1D analytical diffusion model with parameters k_{off}, D, on a semi-infinite domain with an absorbing state at the target site at distance d from the starting point. (Such models have been used previously for theory of transcription factor searching on DNA (34)).

Finally, we note the distinction between the correlation of methylation times at individual sites (on which Equation 5 is based) and methylation rates, which we extract from experiments and simulations. Exact methylation times at individual sites are not currently measureable in genome-wide experiments, but methylation rates k_i are estimated from Repli-BS-Seq experiments, where, for a given CpG sites i, the experiment-inferred methylation rate constant k_i is by definition taken to be equal to $1/<\tau_i>$, i.e., the inverse of the average over methylation times at site i across different cells. The inverse relationship between rates and times could in principle affect the linear Pearson correlation, however, from simulations we found that the Pearson correlation function was not sensitive to the distinction.

We find a good match between the analytical theory and the stochastic simulations, as presented in Figure 6a. We performed regional simulations for varying values of D/k_{off}, and then processed the simulated data through the MLE inference pipeline and computed the correlation function. We then fit the correlation functions to a single exponential decay, and observe a good match between the fitted decay constant and the theoretically predicted value of $\sqrt{k_{off}/D}$. These results demonstrate that the predicted exponential correlation holds, despite potentially complicating factors in the mechanistic model (e.g., the finite time required for diffusion, many CpG sites and multiple enzymes acting simultaneously on one strand, etc.), and uncertainty introduced by the MLE-fitting pipeline. All in all, these results support the idea that processivity is consistent with exponential decay of rate correlation, and that the decay constant can be interpreted as $\sqrt{k_{off}/D}$.

Separation of remethylation rate correlation functions into diffusion-dependent and region-dependent components supports uniform processivity of DNMT1 across the genome. The Processive model can explain exponential decay of rate correlation, but cannot by itself explain the significant correlation observed past 100 bp in strongly correlated regions such as CGI. Nor can it explain the association of this correlation to other regional genomic characteristics (bulk methylation levels, CpG density, chromatin accessibility). Based on the observations in Fig 2, we reasoned that this additional rate correlation can be understood to result from correlated methylation times on neighboring sites due to a variety of regional features, that we describe as

$$Corr_{\text{region}}(X,Y|d) = \phi(\text{features}(d)) \quad \textbf{(6)}$$

and refer to as "region-based correlation". That is, this component of the correlation function can be expressed as a function ϕ, which depends somehow on local features, e.g., related to the chromatin landscape. These features themselves held some distance-dependent correlation. Consideration of additional correlating factors (beyond diffusion) equates to a revision of the idealized theoretical model above. Now, in event 1, if the neighboring site is not reached by diffusion, but rather the two sites are methylated in separate events, their times τ_i and τ_j are nevertheless correlated in a way that depends not only on their distance apart, but also on various other features of their location within the genome. We do not propose a mechanistic model for this additional correlation, but label it as ϕ, and estimate it for each region based on the data. Given these two contributions, which we term diffusion-dependent (denoted θ) and region-dependent (ϕ), the total correlation from both contributions is given by:

$$TotalCorr(X,Y|d) = \theta + (1-\theta)\phi. \quad \textbf{(7)}$$

where the above equation can be understood by probabilistic arguments, i.e., the probability that a neighbor is reached by diffusion is θ; additional correlation ϕ is only present when the neighbor is *not* reached by diffusion, with probability $1-\theta$. This mathematical decomposition of the rate correlation is shown schematically in Figure 6b.

We developed an approach to separate the experiment-derived total correlation functions into these two components, as follows. The total experiment-derived correlation functions are computed from a list of cytosine positions and their inferred remethylation rates. Denote x_n^m as the position of site n, identified as belonging to region m (where m is one of the fourteen regions of Figure 1, and $n \in [1,N]$ given that data is available for N sites in a region of interest). Denote the remethylation rate as k_n^m, giving a list of pairs $\{x_n, k_n\}_m$ for each region, from which the correlation is computed. We also have additional genomic feature data: let q_n^m, r_n^m, s_n^m denote the measured bulk methylation level, the local CpG density, and the chromatin accessibility acquired from independent datasets (see methods). We then perform unsupervised k-means clustering on the features $\{q,r,s\}_m$ to

obtain clusters of sites that are (i) previously assigned as belonging to the same type of genomic region m, and (ii) share more fine-grained similarity in terms of their bulk methylation level, CpG density, and chromatin accessibility. We then randomly shuffle the nucleotide positions within each subcluster, and denote these shuffled positions as $\tilde{x}_n$. We now have a new list, $\{\tilde{x}_n, k_n\}_m$, where the true nucleotide positions have been randomized, but their reassigned position is still similar to the true position in terms of the features $\{q, r, s\}_m$.

We recalculate the correlation functions for each of the regions with the new lists $\{\tilde{x}_n, k_n\}_m$. We label this new correlation function as the region-dependent component ϕ, reasoning that it captures rate correlation that can be attributed to the regional features, according to the model of Equation 6. We then extract θ from the total correlation using equation 7. The results of this decomposition are plotted in Figure 7. We find that ϕ is more variable between different genomic regions, as compared to the component θ, which appears nearly uniform across genomic regions and in all regions shows rapid decay within about 100 bp. We find that θ is well fit by a single-exponential decay, and the fitted decay constants are similar across genomic regions, with a mean decay length of 36 basepairs. These findings support the idea that θ reflects the diffusion-based contribution to the correlation function, because (i) it is the explicitly distance-dependent part that remains after removing correlation attributed to the other features $\{q, r, s\}_m$, (ii) it is consistent with exponential decay, (iii) it is uniform across the genome, consistent with processivity being an inherent property of DNMT1's mode of action, and thus uniform across genomic regions. If θ is interpreted as reflecting processivity of the enzyme according to a 1D diffusion model, then diffusion parameters can be obtained from the experiment-derived correlation function. We thus estimate D/k_{off} to be 1300 bp^2 and the length of a processive step of DNMT1 to be 36 basepairs, on average, across the genome. From these parameters, it is further possible to compute the length of processive runs of the enzyme. Assuming uniformly spaced CpGs in a given locus at highest possible density (i.e. inter-CpG distance=2 bp), then the average length of processive runs is 34.7 bp, while the 95th percentile of run-lengths is 106 bp. If the inter-CpG distance is increased to 50 bp, then the average run is 16.4 bp (indeed shorter than the inter-CpG distance) while the 95th percentile is 100 basepairs (see Supplement for calculations).

DISCUSSION

In this work, we have analyzed correlations among CpG sites in the genome obtained from estimated kinetics of post-replication DNA methylation and from WGBS in hESCs. We find that post-replication remethylation rates on nearby CpGs are correlated, and the nature of this correlation can shed light on molecular mechanisms of maintenance methylation, when analyzed in conjunction with stochastic simulation and mathematical models. We summarize our key results as follows: (i) Methylation rate correlation contains information distinct from that contained in methylation state correlation. In particular, the rate correlation reveals mechanistic information on early-post-replication processes. (ii) Some, but not all, of the rate correlation observed on nearby CpGs decays exponentially with genomic distance, and is consistent with processive activity of DNMT1. Our analysis indicates an average distance of 36 bp between nearby CpGs that are methylated processively after DNA replication. (iii) In addition to evidence of processivity, we also discovered additional correlation, not consistent with a processive model, that is dependent on the genomic regional context. Among studied regions, CGIs and TFBRs showed the most significant contribution of this non-processive (termed "region-based") correlation, which persisted past 1 kbp. In contrast, SINE and LINE showed the least contribution, suggesting that CpG interaction in these regions results nearly entirely from linear diffusion (processivity) of the enzyme. (iv) Further analysis of the region-based rate correlation indicates that much of the regional variation can be attributed to variation in three genomic/epigenomic features: chromatin accessibility, CpG density, and bulk methylation.

More generally, we show how combining a "top-down" approach (i.e., data-driven, using statistical inference) with a "bottom-up" approach (i.e., using hypothesis-driven, or mechanistic models) can be used to glean kinetic insights from a measurement technique that affords genome-wide readout of the post-replication methylome over time (26). A number of related experimental techniques have been developed, employing nucleoside-analog labeling of replicating DNA, followed by isolation/immunoprecipitation and sequencing (35–44). Our general approach could be adapted in the future to gain further insights into post-replication epigenome dynamics based on these types of measurements.

Processivity of DNMT1. While a number of studies have discovered evidence of DNMT1 processivity previously in *in vitro* systems (15–17, 21, 45, 46), our study is the first to our knowledge to uncover a quantifiable signature of methyltransferase processivity in genome-wide, mammalian cell-based measurements. Thus our findings shed light on enzymatic processivity within the context of maintenance methylation *in vivo*. Our findings are consistent with a picture wherein DNMT1 performs processive catalysis regardless of genomic context and in a quantitatively consistent manner (i.e., with relatively uniform lengths of processive steps between CpG targets, genome-wide).

Previous estimates of the length of processive runs of DNMT1 varied widely. For example, Vilkaitis, et al reported processive runs as long as 520 bp (16) while Goyal et al. reported processive runs of over 6000 bp (21). Our estimates based on the Repli-BS dataset are much shorter than these, with the average length of a processive run being about 35 basepairs for stretches of adjacent CpGs, and still shorter for loci with larger inter-CpG distances. One possible explanation for the discrepancy is that our estimate is from experiments in cells, where the *in vivo* chromatin environment, replication machinery, and full complement of DNA-binding proteins are present. These could limit free diffusion of DNMT1 along DNA. Transcription factors are also thought

to search for targets partly by 1D diffusion along DNA, and the effect of crowding has been considered (34). While *in vitro* estimates for 1D sliding lengths of various DNA binding proteins are as high as 20 kb (47), *in vivo* sliding length of the lac repressor was found to be short, at 45 bp (48). How the chromatin environment affects motion of eukaryotic DNA binding proteins is still poorly understood (49). In particular, the question of how methyltransferase processivity is affected by the crowded environment of the cell warrants further study.

Nonprocessive CpG coupling. While our results are consistent with DNMT1 processivity, additional mechanisms affect post-replication remethylation rate correlation where the local genomic/chromatin landscape allows it. We find this non-processive (i.e. non-exponential) source of rate correlation to be most prevalent where chromatin is locally open, CpGs are dense, and the background methylation is relatively low. Although our study does not attempt to define the mechanistic basis of the non-processive component of the rate correlation, various mechanisms can be proposed based on our observations and on previous literature. For example, any mechanism whereby DNMT1 reaches its CpG target through cooperative interactions with other molecules could be speculated to have kinetics dependent on the local genomic and epigenomic context. Neighbor rate correlations could then be sensitive to local context and have lengthscales determined by the the cooperative molecular interactions, rather than being solely dependent on linear inter-CpG genomic distances, in contrast with the processive mechanism. For example, recruitment of DNMT1 to replicating DNA by UHRF1 (18, 19) likely results in context-dependent kinetics, since UHRF1 targeting is dependent on both histone state and the presence of hemi-methylated CpGs (50). A recent finding of monoubiquinated histone H3 helping recruit DNMT1 to DNA stretches with multiple, but not one, hemimethylated CpGs (51) supports the idea that UHRF1 helps direct DNMT1 to CpG-dense regions, and is consistent with our observation of higher rate correlation in CpG-dense regions. Our findings may also be consistent with a nucleation model, in which the initial binding of DNMT1 to replicating DNA occurs on nucleosomes, directed by UHRF1, after which DNMT1 reaches nearby CpGs processively (52). For example, if the initial binding events of separate DNMT1s on nearby nucleosomes are correlated, such correlation would contribute on a lengthscale on the order of hundreds of basepairs, while shorter-lengthscale correlation would be introduced through processivity.

In addition to UHRF1-mediated mechanisms, additional factors are likely to play a role in the non-exponential correlation we observe. First, DNA is not one-dimensional; DNMT1 could reach nearby CpGs by facilitated diffusion (combining 1D diffusion along DNA with 3D diffusion in the nucleosol to nearby sites), similar to transcription factors (34). Our processive model assumes a 1D substrate, but our results hint at sensitivity of maintenance kinetics to 3D DNA structure in the weak appearance of peaks consistent with nucleosome spacing (Figure 1), though we do not attempt to determine whether such features derive from, e.g., interaction with UHRF1 or from facilitated diffusion inherent to DNMT1. Finally, transient binding of post-replication DNA by transcription factors could introduce correlation into maintenance methylation kinetics, as transcription factors could transiently block access to CpGs by DNMT1 and thus delay remethylation. Such a mechanism could explain why we observe the most pronounced region-based correlation in TFBRs.

A number of mathematical modeling studies have provided support for the presence of interactions (also called “collaboration") among CpGs in dynamic methylation processes, including in maintenance, de novo methylation, and demethylation reactions (20, 22, 24, 25). These models, in which a CpG is in some way affected by the state of nearby CpGs, built upon the earlier, so-called “standard" model, wherein each CpG was considered to be independently targeted by methyltransferases (53). Crucially, as the cited studies showed, these interactions provide the necessary nonlinearity to enable bistability in the dynamic system. That is, they enable the same family of “reader" and “writer" enzymes to simultaneously maintain distinct states of hyper- and hypo-methylation on groups of CpGs in different parts of the genome, thus mimicking observed methylation patterns. While these models tend to be phenomenological in nature (i.e., capturing dynamic phenomena without necessarily encoding detailed molecular mechanisms), processivity can be considered to be one type of molecular mechanism that contributes to inter-CpG interactions. Indeed, mathematical modeling also supports the idea that diffusive processivity enhances multi-generational stability of methylation pattterns (21), just as does other mechanisms of CpG interaction (22). Our study contributes insights in this area by (1) providing data-driven estimates of inter-CpG coupling, (2) providing a means of separating and quantifying the contribution of processive versus other collaborative mechanisms, and (3) comparing across genomic regions.

Relationship between WGBS and rate correlation. Our study reveals a difference between the local correlation of methylation state versus rate. We interpret this result as being due to the different information content of the two: WGBS experiments capture reads that are likely to largely reflect the stable methylation landscape (though still partially influenced by replication-associated temporal variability (26), while the Repli-BS-derived methylation rates reflect the transient dynamic processes that help give rise to the stable landscape. Of note, there appears to be some relationship between the two, as we observed similar trends in region specificity in methylation state and rate (specifically in the region-specific, or non-processive, contribution to correlation). We furthermore note that, when we restrict the analysis to common sites (thus retaining CpGs with intermediate to high methylation levels, as rates are only inferrable for these sites), the similarity between rate- and state-correlation is increased (i.e., as in Figure 3 versus Figure 1, where, for example, the decay of bulk state-correlation in SINEs is nearly as rapid as the decay of rate-correlation). This suggests cross-talk between

the transient post-replication methylation events (including processivity) and the stable methylation landscape. Indeed, it was previously reported that CpG co-methylation decays within tens to hundreds of bp, with enzyme processivity proposed as its mechanistic origin (33).

Limitations of our study. Our study shows that novel experimental techniques that probe replication-associated dynamics genome-wide can yield surprisingly detailed dynamic insights, despite the limited time resolution. However, the statistical inference approach is nevertheless limited. First, each site-specific inferred rate constant is the result of a fit of the data to a stochastic Poisson process; undoubtedly this is a simplistic model for temporal dynamics that could potentially be temporally complex in reality, e.g., non-exponential or even non-monotonic. Thus, individual estimates can be error-prone for a number of reasons, from the inability of the simplistic model to capture complex dynamics, to the limited time-resolution or sampling depth for a given site, or all of the above. Because of the inherent difficulty in characterizing dynamics in detail at any individual CpG genome-wide, we focused here on general features of the rate correlation functions that appear robust across a given genomic region. In this way, the whole-genome nature of the data partially compensates for the sparse temporal resolution. Going forward, it may be possible to yield more detailed dynamic insights with deeper sampling and more fine-grained time resolution, which could enable further investigation into detailed, local features in correlation.

While we focus our models around the enzyme kinetics of DNMT1 (for simplicity and because it is the dominant methyltransferase responsible for carrying out maintenance activity), we acknowledge that our rate correlations are likely impacted by the presence of other (*de novo*) methylatransferases, which have been known to associate with highly methylated CpG-dense location (i.e., CGIs) (54). Although our mechanistic models rigorously incorporate 1D diffusion, they lack the dynamic interplay of *de novo*, maintenance, and demethylation reactions that has been studied in mathematical models previously (22, 55). Our approach could be applied to different cell lines in the future, e.g., with specific methyltransferases disrupted, to further disentangle the molecular basis of kinetic correlation.

Reference

1. Suhua Feng, Steven E Jacobsen, and Wolf Reik. Epigenetic reprogramming in plant and animal development. *Science*, 330(6004):622–627, 2010.
2. Zachary D Smith and Alexander Meissner. Dna methylation: roles in mammalian development. *Nature Reviews Genetics*, 14(3):204–220, 2013.
3. Adrian Bird. Dna methylation patterns and epigenetic memory. *Genes & development*, 16 (1):6–21, 2002.
4. Aimée M Deaton, Shaun Webb, Alastair RW Kerr, Robert S Illingworth, Jacky Guy, Robert Andrews, and Adrian Bird. Cell type–specific dna methylation at intragenic cpg islands in the immune system. *Genome research*, 21(7):1074–1086, 2011.
5. Albert Jeltsch. Beyond watson and crick: Dna methylation and molecular enzymology of dna methyltransferases. *Chembiochem*, 3(4):274, 2002.
6. Robert J Klose and Adrian P Bird. Genomic dna methylation: the mark and its mediators. *Trends in biochemical sciences*, 31(2):89–97, 2006.
7. Michela Curradi, Annalisa Izzo, Gianfranco Badaracco, and Nicoletta Landsberger. Molecular mechanisms of gene silencing mediated by dna methylation. *Molecular and cellular biology*, 22(9):3157–3173, 2002.
8. Vishal Nanavaty, Elizabeth W Abrash, Changjin Hong, Sunho Park, Emily E Fink, Zhuangyue Li, Thomas J Sweet, Jeffrey M Bhasin, Srinidhi Singuri, Byron H Lee, et al. Dna methylation regulates alternative polyadenylation via ctcf and the cohesin complex. *Molecular cell*, 78(4):752–764, 2020.
9. Peter A Jones and Shirley M Taylor. Cellular differentiation, cytidine analogs and dna methylation. *Cell*, 20(1):85–93, 1980.
10. Arthur D Riggs. X inactivation, differentiation, and dna methylation. *Cytogenetic and Genome Research*, 14(1):9–25, 1975.
11. Peter A Jones and Gangning Liang. Rethinking how dna methylation patterns are maintained. *Nature Reviews Genetics*, 10(11):805–811, 2009.
12. En Li, Timothy H Bestor, and Rudolf Jaenisch. Targeted mutation of the dna methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–926, 1992.
13. Robin Holliday and John E Pugh. Dna modification mechanisms and gene activity during development. *Science*, 187(4173):226–232, 1975.
14. Julia Arand, David Spieler, Tommy Karius, Miguel R Branco, Daniela Meilinger, Alexander Meissner, Thomas Jenuwein, Guoliang Xu, Heinrich Leonhardt, Verena Wolf, et al. In vivo control of cpg and non-cpg dna methylation by dna methyltransferases. *PLoS genetics*, 8 (6):e1002750, 2012.
15. Andrea Hermann, Rachna Goyal, and Albert Jeltsch. The dnmt1 dna-(cytosine-c5)-methyltransferase methylates dna processively with high preference for hemimethylated target sites. *Journal of Biological Chemistry*, 279(46):48350–48359, 2004.
16. Giedrius Vilkaitis, Isao Suetake, Saulius Klimašauskas, and Shoji Tajima. Processive methylation of hemimethylated cpg sites by mouse dnmt1 dna methyltransferase. *Journal of Biological Chemistry*, 280(1):64–72, 2005.
17. Željko M Svedružić and Norbert O Reich. Mechanism of allosteric regulation of dnmt1's processivity. *Biochemistry*, 44(45):14977–14988, 2005.
18. Magnolia Bostick, Jong Kyong Kim, Pierre-Olivier Estève, Amander Clark, Sriharsa Pradhan, and Steven E Jacobsen. Uhrf1 plays a role in maintaining dna methylation in mammalian cells. *Science*, 317(5845):1760–1764, 2007.
19. Jafar Sharif, Masahiro Muto, Shin-ichiro Takebayashi, Isao Suetake, Akihiro Iwamatsu, Takaho A Endo, Jun Shinga, Yoko Mizutani-Koseki, Tetsuro Toyoda, Kunihiro Okamura, et al. The sra protein np95 mediates epigenetic inheritance by recruiting dnmt1 to methylated dna. *Nature*, 450(7171):908–912, 2007.
20. Laura B Sontag, Matthew C Lorincz, and E Georg Luebeck. Dynamics, stability and inheritance of somatic dna methylation imprints. *Journal of theoretical biology*, 242(4):890–899, 2006.
21. Rachna Goyal, Richard Reinhardt, and Albert Jeltsch. Accuracy of dna methylation pattern preservation by the dnmt1 methyltransferase. *Nucleic acids research*, 34(4):1182–1188, 2006.
22. Jan O Haerter, Cecilia Lövkvist, Ian B Dodd, and Kim Sneppen. Collaboration between cpg sites is needed for stable somatic inheritance of dna methylation states. *Nucleic acids research*, 42(4):2235–2244, 2014.
23. Shijie C Zheng, Martin Widschwendter, and Andrew E Teschendorff. Epigenetic drift, epigenetic clocks and cancer risk. *Epigenomics*, 8(5):705–719, 2016.
24. You Song, Honglei Ren, and Jinzhi Lei. Collaborations between cpg sites in dna methylation. *International Journal of Modern Physics B*, 31(20):1750243, 2017.
25. Loukas Zagkos, Mark Mc Auley, Jason Roberts, and Nikos I Kavallaris. Mathematical models of dna methylation dynamics: Implications for health and ageing. *Journal of theoretical biology*, 462:184–193, 2019.
26. Jocelyn Charlton, Timothy L Downing, Zachary D Smith, Hongcang Gu, Kendell Clement, Ramona Pop, Veronika Akopian, Sven Klages, David P Santos, Alexander M Tsankov, et al. Global delay in nascent strand dna methylation. *Nature structural & molecular biology*, 25 (4):327–332, 2018.
27. Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, et al. The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, 28(10):1045–1048, 2010.
28. Jocelyn Charlton, Timothy Downing, Zachary Smith, Hongcang Gu, Kendell Clement, Ramona Pop, Veronika Akopian, Sven Klages, David Santos, Alexander Tsankov, Bernd Timmermann, Michael Ziller, Evangelos Kiskinis, Andreas Gnirke, and Alexander Meissner. Global delay in nascent strand dna methylation. *Nature Structural & Molecular Biology*, 25 (4):327–332, 2018.
29. Luis Busto-Moner, Julien Morival, Honglei Ren, Arjang Fahim, Zachary Reitz, Timothy L Downing, and Elizabeth L Read. Stochastic modeling reveals kinetic heterogeneity in post-replication dna methylation. *PLOS Computational Biology*, 16(4):e1007195, 2020.
30. Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.
31. Andri Bezzola, Benjamin B Bales, Richard C Alkire, and Linda R Petzold. An exact and efficient first passage time algorithm for reaction–diffusion processes on a 2d-lattice. *Journal of Computational Physics*, 256:183–197, 2014.
32. William G Jacoby. Loess:: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Studies*, 19(4):577–613, 2000.
33. Shicheng Guo, Dinh Diep, Nongluk Plongthongkum, Ho-Lim Fung, Kang Zhang, and Kun Zhang. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma dna. *Nature genetics*, 49(4):635–642, 2017.
34. Leonid Mirny, Michael Slutsky, Zeba Wunderlich, Anahita Tafvizi, Jason Leith, and Andrej Kosmrlj. How a protein searches for its site on dna: the mechanism of facilitated diffusion. *Journal of Physics A: Mathematical and Theoretical*, 42(43):434013, 2009.
35. R Scott Hansen, Sean Thomas, Richard Sandstrom, Theresa K Canfield, Robert E Thurman, Molly Weaver, Michael O Dorschner, Stanley M Gartler, and John A Stamatoyannopoulos. Sequencing newly replicated dna reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences*, 107(1):139–144, 2010.
36. Tyrone Ryba, Dana Battaglia, Benjamin D Pope, Ichiro Hiratani, and David M Gilbert. Genome-scale analysis of replication timing: from bench to bioinformatics. *Nature protocols*, 6(6):870–895, 2011.
37. Claire Marchal, Takayo Sasaki, Daniel Vera, Korey Wilson, Jiao Sima, Juan Carlos Rivera-Mulia, Claudia Trevilla-García, Coralin Nogues, Ebtesam Nafie, and David M Gilbert.

Genome-wide analysis of replication timing by next-generation sequencing with e/l repliseq. *Nature protocols*, 13(5):819–839, 2018.
38. Qian Du, Saul A Bert, Nicola J Armstrong, C Elizabeth Caldon, Jenny Z Song, Shalima S Nair, Cathryn M Gould, Phuc-Loi Luu, Timothy Peters, Amanda Khoury, et al. Replication timing and epigenome remodelling are associated with the nature of chromosomal rearrangements in cancer. *Nature communications*, 10(1):1–15, 2019.
39. Chenhuan Xu and Victor G Corces. Genome-wide mapping of protein-dna interactions on nascent chromatin. *Methods in molecular biology (Clifton, NJ)*, 1766:231, 2018.
40. Constance Alabert, Teresa K Barth, Nazaret Reverón-Gómez, Simone Sidoli, Andreas Schmidt, Ole N Jensen, Axel Imhof, and Anja Groth. Two distinct modes for propagation of histone ptms across the cell cycle. *Genes & development*, 29(6):585–590, 2015.
41. Nazaret Reverón-Gómez, Cristina González-Aguilera, Kathleen R Stewart-Morgan, Nataliya Petryk, Valentin Flury, Simona Graziano, Jens Vilstrup Johansen, Janus Schou Jakobsen, Constance Alabert, and Anja Groth. Accurate recycling of parental histones reproduces the histone modification landscape during dna replication. *Molecular cell*, 72(2):239–249, 2018.
42. Pauline Vasseur, Saphia Tonazzini, Rahima Ziane, Alain Camasses, Oliver J Rando, and Marta Radman-Livaja. Dynamics of nucleosome positioning maturation following genomic replication. *Cell reports*, 16(10):2651–2665, 2016.
43. Mónica P Gutiérrez, Heather K MacAlpine, and David M MacAlpine. Nascent chromatin occupancy profiling reveals locus-and factor-specific chromatin maturation dynamics behind the dna replication fork. *Genome research*, 29(7):1123–1133, 2019.
44. Chenhuan Xu and Victor G Corces. Nascent dna methylome mapping reveals inheritance of hemimethylation at ctcf/cohesin sites. *Science*, 359(6380):1166–1170, 2018.
45. Timothy H Bestor and Vernon M Ingram. Two dna methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with dna. *Proceedings of the National Academy of Sciences*, 80(18):5559–5563, 1983.
46. Sabrina Adam, Hiwot Anteneh, Maximilian Hornisch, Vincent Wagner, Jiuwei Lu, Nicole E Radde, Pavel Bashtrykov, Jikui Song, and Albert Jeltsch. Dna sequence-dependent activity and base flipping mechanisms of dnmt1 regulate genome-wide dna methylation. *Nature communications*, 11(1):1–15, 2020.
47. Kiyoto Kamagata, Eriko Mano, Kana Ouchi, Saori Kanbayashi, and Reid C Johnson. High free-energy barrier of 1d diffusion along dna by architectural dna-binding proteins. *Journal of molecular biology*, 430(5):655–667, 2018.
48. Petter Hammar, Prune Leroy, Anel Mahmutovic, Erik G Marklund, Otto G Berg, and Johan Elf. The lac repressor displays facilitated diffusion in living cells. *Science*, 336(6088):1595–1598, 2012.
49. David M Suter. Transcription factors and dna play hide and seek. *Trends in cell biology*, 30 (6):491–500, 2020.
50. Xiaoli Liu, Qinqin Gao, Pishun Li, Qian Zhao, Jiqin Zhang, Jiwen Li, Haruhiko Koseki, and Jiemin Wong. Uhrf1 targets dnmt1 for dna methylation through cooperative binding of hemimethylated dna and methylated h3k9. *Nature communications*, 4(1):1–13, 2013.
51. Yuichi Mishima, Laura Brueckner, Saori Takahashi, Toru Kawakami, Junji Otani, Akira Shinohara, Kohei Takeshita, Ronald Garingalao Garvilles, Mikio Watanabe, Norio Sakai, et al. Enhanced processivity of dnmt1 by monoubiquitinated histone h3. *Genes to cells*, 25(1): 22–32, 2020.
52. Robert M Vaughan, Bradley M Dickson, Matthew F Whelihan, Andrea L Johnstone, Evan M Cornett, Marcus A Cheek, Christine A Ausherman, Martis W Cowles, Zu-Wen Sun, and Scott B Rothbart. Chromatin structure and its chemical modifications regulate the ubiquitin ligase substrate selectivity of uhrf1. *Proceedings of the National Academy of Sciences*, 115 (35):8775–8780, 2018.
53. Sarah P Otto and Virginia Walbot. Dna methylation in eukaryotes: kinetics of demethylation and de novo methylation during the life cycle. *Genetics*, 124(2):429–437, 1990.
54. Jing Liao, Rahul Karnik, Hongcang Gu, Michael J Ziller, Kendell Clement, Alexander M Tsankov, Veronika Akopian, Casey A Gifford, Julie Donaghey, Christina Galonska, et al. Targeted disruption of dnmt1, dnmt3a and dnmt3b in human embryonic stem cells. *Nature genetics*, 47(5):469–478, 2015.
55. Albert Jeltsch and Renata Z Jurkowska. New concepts in dna methylation. *Trends in biochemical sciences*, 39(7):310–318, 2014.
56. Ivan Hemeon, Jemy A Gutierrez, Meng-Chiao Ho, and Vern L Schramm. Characterizing dna methyltransferases with an ultrasensitive luciferase-linked continuous assay. *Analytical chemistry*, 83(12):4996–5004, 2011.
57. Anahita Tafvizi, Fang Huang, Alan R Fersht, Leonid A Mirny, and Antoine M van Oijen. A single-molecule characterization of p53 search on dna. *Proceedings of the National Academy of Sciences*, 108(2):563–568, 2011.

Extended Methods

Maximum Likelihood Estimation of Remethylation Rates. The Repli-BS dataset(26) contains read-depths that vary widely both across CpGs and across measured timepoints. The data can be expressed as $\{N^0_{ij}, N^1_{ij}\}$, or the number of unmethylated (0) and methylated (1) reads observed at site i at timepoint j. We assume that the probability of finding a methylated read at site i and at time t_j is

$$p(1|k_i, f_i, t_j) = f_i(1 - \exp(-k_i t_j)) \tag{8}$$

where k_i is the rate of accumulation of methylation post-replication and f_i is the steady-state fraction of cells in the population methylated at site i. Parameters are estimated from the data by Maximum Likelihood Estimation (MLE), as in (29). In the MLE pipeline, we use a per-site minimum read-depth threshold to increase the baseline accuracy of inferred parameters. In the current work, the threshold was set as follows: each CpG required a minimum of 5 reads (methylated or unmethylated) at both timepoint $t = 0$ and combined timepoints $t > 0$ (for a total minimum of 10 reads over all timepoints). Furthermore, each CpG required a minimum of 5 methylated reads in total (over all measured timepoints). (We performed detailed testing and uncertainty quantification for various methods of setting thresholds previously (29), finding that a read-depth-based threshold gave good balance between genomic coverage and uncertainty in estimates). For each site, we compute the Likelihood Function, $L(\theta\|x_i)$, where θ contains the parameters $\theta = \{k, f\}$ and x_i is the observed data at site i, given the Poisson process model. We observed three cases: (1) the parameters are identifiable. This occurs when the likelihood function contains a global maximum in $\theta = \{k, f\}$; the location of the maximum gives the inferred parameter values. The majority of sites in the dataset fall in this category; (2) The parameters are practically unidentifiable. This occurs when a global maximum exists, but the confidence interval is very wide. We discard sites for which this occurs, requiring that the 75% confidence interval on k (obtained by the profile likelihood method) fall within the range of values $[0.01 - 10]\mathrm{h}^{-1}$; (3) The rate k is unidentifiable, but its value can be bounded. This occurs frequently, as many sites show full methylation already at the first experimental timepoint, indicating a post-replication methylation rate that is faster than the time-resolution of the experiment, given the one-hour pulse length. We observe this in a likelihood function $L(\theta\|x_i)$ that plateaus at high rate values. In these cases, we use a procedure to estimate a lower bound for k_i: we determine the 25% lower bound of the confidence interval on k, relative to the plateaued value of $L(\theta\|x_i)$. We choose this value for the lower bound on k, reasoning that it estimates the location where the likelihood function is no longer increasing significantly.

Stochastic Simulation. The models are written in the form of stochastic biochemical kinetic reaction models (or in the case of the processive mechanism, a stochastic reaction-diffusion model). The models concentrate solely on the process of maintenance methylation, i.e., the remethylation process that occurs in less than 24 hours, while ignoring other processes such as demethylation. Both models treat DNA as a 1D strand with CpG sites that can be unmethylated (u), hemi-methylated (h), or methylated (m). Sites are assumed to be unmethylated or hemimethylated immediately after replication, with hemimethylated sites susceptible to remethylation by the enzyme. The distributive and processive models differ in how the enzyme moves to new hemimethylated sites after catalyzing methylation at one site. The Distributive mechanism treats individual CpG sites as fully independent. The DNMT1 will be disassociated from DNA after it methylated a hemimethylated site. By contrast, in the Processive model, after methylation DNMT1 can diffuse towards its neighbor sites either upstream or downstream. We assume the enzyme travels with 1D diffusion coefficient D and unbinds with rate koff. E denotes a free enzyme (not bound to DNA). Eh is a complex formed when E is bound to h (similar for m). The model reactions are as follows.

Distributive Model.

$$\begin{aligned} &\mathrm{E} + \mathrm{h} \rightleftharpoons \mathrm{Eh} && k_{1f}, k_{1r} \\ &\mathrm{Eh} \longrightarrow \mathrm{E} + \mathrm{m} && k_{cat} \end{aligned}$$

After the catalytic step, unbinding is assumed to happen rapidly, therefore the catalytic step and unbinding are assumed to occur together in one step, with rate k_{cat}.

Processive Model.

$$\begin{aligned} &\mathrm{E} + \mathrm{h} \rightleftharpoons \mathrm{Eh} && k_{1f}, k_{1r} \\ &\mathrm{Eh} \longrightarrow \mathrm{Em} && k_{cat} \\ &\mathrm{Em}_i + \mathrm{h}_j \longrightarrow \mathrm{m}_i + \mathrm{Eh}_j && D \\ &\mathrm{Em} \longrightarrow \mathrm{E} + \mathrm{m} && k_{off} \end{aligned}$$

In the processive model, the diffusive step allows the enzyme to reach a hemimethylated CpG from a nearby CpG in Em state. The likelihood of this occuring before unbinding with rate k_{off} depends on the properties of diffusion (see below).

Model Parameters. Model parameters were chosen to be approximately in line with experimentally measured values for DNMT1 where possible (56), while also showing reasonable agreement with our experiment-derived methylation rates and rate correlation (in the case of the Processive model). Parameters used in simulations are shown in Table S1.

Simulation algorithms. For the distributive model, the standard Gillespie Stochastic Simulation Algorithm was used (30). For the processive model, we combine stochastic reactions with a 1D lattice model of diffusion. The lattice spaciging is 1 basepair. We apply a First Passage Time Kinetic Monte Carlo algorithm inspired by (31). This allows us to simulate diffusive steps efficiently, since we do not need to track each diffusive hop with basepair resolution. Rather, the algorithm requires sampling from the First Passage Time Distribution for an enzyme to reach a nearby CpG by diffusion. This allows us to simulate large stretches (tens of thousands) of CpGs.
In the simulations, the enzyme is allowed to diffuse bidirectionally along DNA. Consider an enzyme with a starting position located at a CpG in the Em state. Let there be two nearest neighbor CpGs in the h state (which are thus candidates to be subsequently targeted by the enzyme, according to the Processive model reactions above). We refer to the diffusion *Domain* as the 1D region between left and right neighbors. Let d_L be the distance (in bp) between the left neighbor and the starting position, and similarly d_R for the right neighbor. Then the next event that can occur for this Em CpG is one of: (1) the enzyme diffuses distance d_L to the left neighbor with diffusion coefficient D. The current CpG undergoes $\mathrm{Em} \longrightarrow \mathrm{m}$ and the left neighbor undergoes $\mathrm{h} \longrightarrow \mathrm{Eh}$; (2) Similar for the right neighbor; or (3) The enzyme unbinds from DNA before reaching either left or right neighbor at rate k_{off}, so $\mathrm{Em} \longrightarrow \mathrm{E}+\mathrm{m}$. For the simulation, we must compute the waiting time distribution (or First Passage Time Distribution $f(t)$) for the enzyme to exit the domain by one of these events. We must also compute the exit probabilities, i.e. the probability of events (1),(2) or (3) occuring, which we refer to as P_L, P_R, P_{off}, and $P_L + P_R + P_{off} = 1$. The distribution $F(t)$ and the exit probabilities can be solved analytically (see below). These events, which involve diffusion, have non-exponential waiting times, and we refer to them as FPT (first passage time) events. The other reactions (k_{1f}, k_{1r}, k_{cat}) are assumed to proceed via standard Markovian (memoryless) kinetics, with exponential waiting times, and are referred to as KMC (kinetic Monte Carlo) events. Following (31), a sketch of the algorithm for the processive simulation is then:

Initialize N CpGs at post-replication time $t_{p.r.} = 0$ in u or h state by sampling from input methylation landscape

while $t_{p.r.} < t_{out}$ (t_{out} is the desired readout time)

Determine which CpG undergoes next event and which reaction occurs at that site by:

- For all CpGs in h or Eh, determine next KMC event and time to event, τ_{KMC}, by the Gillespie algorithm (30).
- *Build new diffusion domains*. For all CpGs newly in Em state, determine domain length L=$d_R + d_L$ and $x_0 = d_L$. Sample τ^i_{FPT} from $F(t)$ and store in event queue. Compute corresponding exit probabilities for each ith site, P^i_L, P^i_R, P^i_{off}.
- *Update diffusion domains.* For all CpGs remaining in Em state since last event, determine whether left or right neighbor has changed. If so, rebuild diffusion domain and resample τ^i_{FPT} from $F(t)$ and replace in event queue.
- Select the minimum FPT time, and the corresponding site j for the next FPT event, $\tau_{FPT} = \tau^j_{FPT} = \min_{\mathrm{i}} \tau^{\mathrm{i}}_{\mathrm{FPT}}$.

if $\tau_{KMC} < \tau_{FPT}$, set $\tau_{next} = \tau_{KMC}$. Update system state according to KMC event.

else $\tau_{next} = \tau_{FPT}$. Sample from the exit probabilities for the jth site to determine outcome and update system state according to FPT event.

end if

Update the system time: $t_{p.r.} = t_{p.r.} + \tau_{next}$.

Update FPT queue by $\tau^i_{FPT} = \tau^i_{FPT} - \tau_{next}$. (Remove the site which underwent diffusion previously from event queue).

end while

Record system state.

Analytic 1D Diffusion Model.

Table 1. Table of Parameters

Parameter	Symbol	Units	Value	Source or Comment
Catalytic rate, $h \to m$	k_{cat}	hr^{-1}	40	(56)
Michaelis Constant	K_m	μM	0.8	(56)
Stochastic Michaelis Constant	$K_m^{\text{stoch.}}$	#	11500	estimated from K_m, assuming reaction volume $\approx 10^{15}$L
Enzyme unbinding from h-CpG	k_{1r}	hr^{-1}	5	estimated from rate correlation max (Processive)
Enzyme binding to h-CpG	k_{1f}	$(\# \cdot hr)^{-1}$	0.039	from $K_m^{\text{stoch.}}$, k_{1r}
Enzymes per h initially	$E_0/h(t=0)$	unitless	1/45	estimated by assuming $t_{50}^{M.M.} \approx 0.5$ h based on Repli-BS
Initial hemimethylated substrate #	$S_0 = h(t=0)$	#	≈ 20000	arbitrary, balance of simulation time and region coverage
Enzyme 1-D diffusion coefficient	D	$bp^2 s^{-1}$	10^6	estimate from other DNA-binding proteins (57)
Enzyme drop-off from DNA	k_{off}	hr^{-1}		
Ratio D/k_{off}	D/k_{off}	bp^2	1300	estimated from rate correlation decay (Processive)

Semi-infinite Domain. Following the model described in the main text, consider two CpG sites, linear distance d apart. Let $P(x,t)$ be the probability density of a particle (enzyme) to be at location x in 1D at time t. For simple diffusion (no unbinding) the governing equation

$$\frac{\partial P(x,t)}{\partial t} = D\frac{\partial^2 P(x,t)}{\partial x^2}$$

with initial condition $P(x=0,t=0)=1$ and absorbing boundary condition $P(x=d,t>0)=0$ gives the density

$$\tilde{P}(x,t) - \tilde{P}(x-2d,t)$$

where $\tilde{P}(x,t) = \frac{1}{\sqrt{4\pi Dt}}\exp\left(-x^2/4Dt\right)$ is the solution to the standard 1D diffusion model on an infinite domain (no absorbing boundary condition).
When unbinding is included in the model, the governing equation becomes:

$$\frac{\partial P(x,t)}{\partial t} = D\frac{\partial^2 P(x,t)}{\partial x^2} - k_{off}P(x,t)$$

The solution to this PDE with the same absorbing condition at distance d as above gives

$$P(x,t) = \frac{e^{-k_{off}t}}{\sqrt{4\pi Dt}}\left(e^{-x^2/4Dt} - e^{-(x-2d)^2/4Dt}\right).$$

We then have the survival probability for the protein to remain bound on DNA at time t:

$$S(t) = \int_{-\infty}^{d} P(x,t)dx = e^{-kt}\text{erf}(\frac{d}{\sqrt{4Dt}})$$

leading to an expression for the total probability flux to "exit" the DNA (either by unbinding or by absorption at $x=d$):

$$F(t) = -\frac{\partial S}{\partial t} = e^{-kt}\left(k\text{erf}(\frac{d}{\sqrt{4Dt}}) + \frac{d}{t}\frac{1}{\sqrt{4\pi Dt}}e^{-d^2/4Dt}\right)$$

The first term is the flux to exit by unbinding, the second is the flux to the absorbing state at distance d, i.e. $F(t) = F_{unbind}(t) + F_{absorb}(t)$. Total probability to reach the absorbing state at distance d by diffusion, integrating the absorbing flux,

$$R(d) = \int_0^{\infty} F_{absorb}(t)dt = e^{-d\sqrt{k/D}}$$

Thus, the probability for the enzyme to reach a site at distance d away from the initial site before unbinding is given by $e^{-d\sqrt{k/D}}$.

Length of Processive Runs. To estimate the length of processive runs based on the above semi-infinite domain model, we make use of the infinite series:

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \tag{9}$$

$$\sum_{n=0}^{\infty} nx^n = \frac{x}{(1-x)^2}. \tag{10}$$

Assume a domain of uniformly-spaced CpGs, distance d apart, and the enzyme starts at the edge of the domain (i.e., so that it is likely to move unidirectionally). Let $\theta = e^{-d\sqrt{k/D}}$ be the probability that the enzyme will successfully make each step, i.e., to reach each successive CpG. Then the probability of the enzyme making n processive steps before falling off is

$$P(n) = \frac{\theta^n}{\sum_{n=0}^{\infty}\theta^n} = (1-\theta)\theta^n. \tag{11}$$

Then the expected number of steps is

$$< n >= \sum_{n=0}^{\infty} nP(n) = (1-\theta)\sum_{n=0}^{\infty} n\theta^n = \frac{\theta}{1-\theta} \tag{12}$$

and the expected run-length is $d < n >$. For adjacent CpGs ($d = 2$) and taking $\sqrt{k/D} = 0.028$, this gives an expected processive run legtnh of 34.7 bp.

Finite Domain. In the simulations, we need to consider not just pairs of CpGs distance d apart, but consider that an enzyme bound to DNA may have two neighbors within reach in h state, left and right. Let the total domain length be L=$d_R + d_L$, and the enzyme starting position is $x_0 = d_L$. We have a diffusion model, with unbinding, with initial condition $P(x = x_0, t = 0) = 1$ and absorbing boundary conditions at both ends, $P(x = 0, t > 0) = 0$; $P(x = L, t > 0) = 0$. The Fourier series solution is

$$\begin{aligned} P(x,t) &= \sum_{n=1}^{\infty} P_n(x,t) \\ P_n(x,t) &= \frac{2}{L}\sin\frac{n\pi x_0}{L}\sin\frac{n\pi x}{L}\exp\left(-\left(\frac{Dn^2\pi^2}{L^2}+k\right)t\right) \end{aligned}$$

The survival probability of the enzyme on DNA at location x (i.e. before either unbinding or reaching left or right neighbor) is

$$\begin{aligned} S(t) &= \sum_{n=1}^{\infty} S_n(t) \\ S_n(t) &= \int_0^L P_n(x,t)dx \\ &= \frac{2}{n\pi}\sin\frac{n\pi x_0}{L}(1-(-1)^n)\exp\left(-\left(\frac{Dn^2\pi^2}{L^2}+k\right)t\right) \end{aligned}$$

The first passage time distribution $f(t)$ can be obtained by normalizing the time-dependent probability flux $F(t) = -\frac{\partial S(t)}{\partial t}$. For purposes of the simulation, we sample from $f(t)$ making use of the C.D.F. for $f(t)$. This C.D.F. is given by:

$$C(t) = 1 - S(t).$$

To determine the exit probabilities as a function of time, (i.e., the relative likelihood of exiting left, right, or to solution), we make use of the time-dependent probability fluxes to each end-state: $F_L(t), F_R(t), F_{solution}(t)$. These are obtained as:

$$\begin{aligned} F_L(t) &= DP(x = 1, t) \\ F_R(t) &= DP(x = L-1, t) \\ F_{solution}(t) &= kS(t). \end{aligned}$$