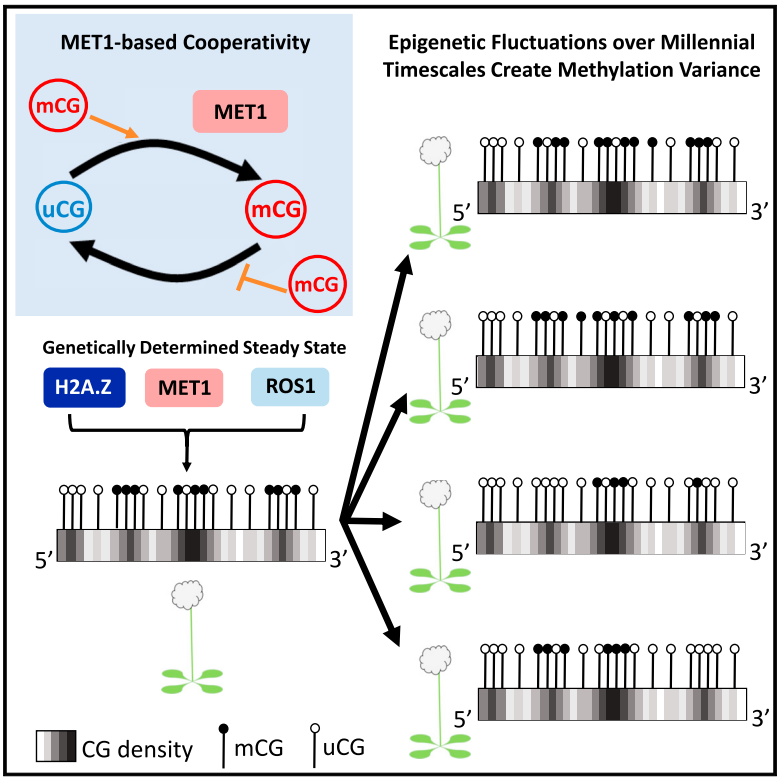
Cell Systems

Article

# Millennia-long epigenetic fluctuations generate intragenic DNA methylation variance in *Arabidopsis* populations

## Graphical abstract




## Highlights

- MET1 mediates a unified process of mCG establishment and maintenance within genes
- ROS1 and H2A.Z negatively regulate the epigenetic dynamics of genic mCG
- A mathematical model predicts genic mCG patterns and their population variance
- Genic mCG undergoes large epigenetic fluctuations that can last thousands of years

## Authors

Amy Briffa, Elizabeth Hollwey, Zaigham Shahzad, Jonathan D. Moore, David B. Lyons, Martin Howard, Daniel Zilberman

## Correspondence

martin.howard@jic.ac.uk (M.H.), daniel.zilberman@ist.ac.at (D.Z.)

## In brief

Briffa and Hollwey et al. find that *Arabidopsis thaliana* intragenic DNA methylation establishment, inheritance, and intergenerational variance constitute a unified process mediated by the methyltransferase MET1 and delimited by ROS1 and H2A.Z. Overall methylation patterns are genetically determined but undergo millennial-timescale epigenetic fluctuations that explain variation in natural populations.

Briffa et al., 2023, Cell Systems *14*, 953–967
November 15, 2023 
https://doi.org/10.1016/j.cels.2023.10.007

CellPress

Cell Systems

CellPress
OPEN ACCESS

Article

# Millennia-long epigenetic fluctuations generate intragenic DNA methylation variance in *Arabidopsis* populations

Amy Briffa,[1,5,6] Elizabeth Hollwey,[2,3,5] Zaigham Shahzad,[2,4] Jonathan D. Moore,[2] David B. Lyons,[2] Martin Howard,[1,*] and Daniel Zilberman[2,3,7,*]

[1]Department of Computational and Systems Biology, John Innes Centre, Norwich NR4 7UH, UK
[2]Department of Cell and Developmental Biology, John Innes Centre, Norwich NR4 7UH, UK
[3]Institute of Science and Technology, 3400 Klosterneuburg, Austria
[4]Department of Life Sciences, Syed Babar Ali School of Science and Engineering, Lahore University of Management Sciences, Lahore, Pakistan
[5]These authors contributed equally
[6]Present address: Babraham Institute, Babraham Research Campus, Cambridge CB22 3AT, UK
[7]Lead contact
*Correspondence: martin.howard@jic.ac.uk (M.H.), daniel.zilberman@ist.ac.at (D.Z.)
https://doi.org/10.1016/j.cels.2023.10.007

## SUMMARY

Methylation of CG dinucleotides (mCGs), which regulates eukaryotic genome functions, is epigenetically propagated by Dnmt1/MET1 methyltransferases. How mCG is established and transmitted across generations despite imperfect enzyme fidelity is unclear. Whether mCG variation in natural populations is governed by genetic or epigenetic inheritance also remains mysterious. Here, we show that MET1 *de novo* activity, which is enhanced by existing proximate methylation, seeds and stabilizes mCG in *Arabidopsis thaliana* genes. MET1 activity is restricted by active demethylation and suppressed by histone variant H2A.Z, producing localized mCG patterns. Based on these observations, we develop a stochastic mathematical model that precisely recapitulates mCG inheritance dynamics and predicts intragenic mCG patterns and their population-scale variation given only CG site spacing. Our results demonstrate that intragenic mCG establishment, inheritance, and variance constitute a unified epigenetic process, revealing that intragenic mCG undergoes large, millennia-long epigenetic fluctuations and can therefore mediate evolution on this timescale.

## INTRODUCTION

Although CG DNA methylation (mCG) patterns are vital for plant and animal development, and for human health,[1–3] how these patterns are set up and propagated is not fully understood. mCG is thought to be installed by dedicated *de novo* methyltransferases, and subsequently, epigenetically propagated by Dnmt1/MET1 (animals and plants) or Dnmt5 (fungi and other eukaryotes) "maintenance" methyltransferases.[4–9] These enzymes restore full methylation to hemi-methylated CG sites produced by DNA replication,[10–12] although mammalian Dnmt1 also has *de novo* activity,[13] while DNA demethylases can at any point actively remove mCG.[14,15] In some lineages, such as flowering plants, mCG patterns are epigenetically inherited across generations,[4,8,16–18] with mCG associated with variation in both gene expression and phenotype.[19–21]

However, as mCG epigenetic inheritance is imperfect, it is unclear over how many cell cycles mCG can encode additional information independent of the underlying genetic sequence and thus function as an epigenetic genotype. Conversely, it is unknown over what timescales mCG patterns might not be epigenetic but instead are phenotypes that are predictable from local and global genetic variation. Resolving these questions is fundamental to understanding mCG epigenetic dynamics and determining whether epigenetically inherited mCG patterns (epialleles) can be a basis for natural selection.[22]

In this context, computational modeling is an essential tool for mechanistic understanding of long-term epigenetic inheritance that allows access to experimentally inaccessible timescales. Computational models based on mammalian data indicate that strong cooperativity—nearby mCG promoting methylation gain and unmethylated CG (uCG) sites promoting methylation loss—can produce bistability: the capacity for stable epigenetic inheritance of either the methylated or the unmethylated (UM) state of a locus.[23–26] Dnmt1 activity indeed shows signs of cooperativity, with new methylation preferentially targeted to and maintained within regions of existing mCG.[13,26] However, how mCG inheritance unfolds over long periods of time has not been explored and whether a bistable—or any other—paradigm might mediate long-term epigenetic inheritance *in vivo* is unknown.

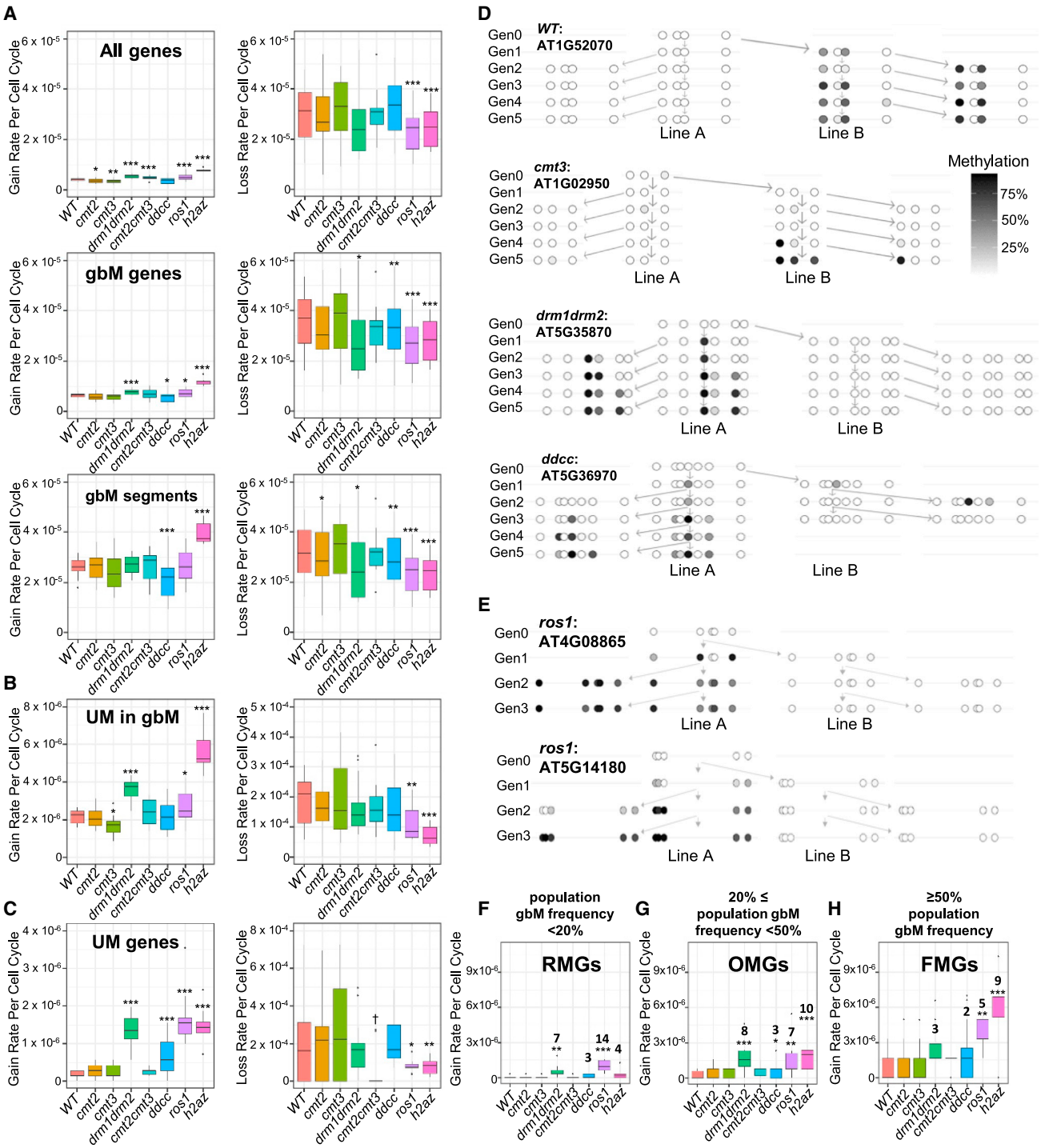



**Figure 1. GbM epigenetic dynamics are dominated by MET1 and delimited by ROS1 and H2AZ**

(A–C) Per-cell-cycle rates of mCG gain and loss at individual CG sites within indicated genomic regions. GbM genes were divided into methylated (gbM segments, A), and unmethylated (UM in gbM, B) regions. A significant difference from the WT rate is indicated by *p = 0.01–0.001, **p = 0.001–0.00001, and ***p < 0.00001 (Fisher's exact test), N = 9–36 (STAR Methods). The UM gene *cmt2cmt3* loss rate marked by (†) is based on very few observations and hence is not reliable (and is not significantly different from WT).

(D) GbM gains that expand to form new methylation clusters can be observed in previously entirely unmethylated genes in different genotypes, including *ddcc*. Each circle represents an individual CG pair, with darkness of fill indicating fractional mCG.

*(legend continued on next page)*

Flowering plants and most invertebrates have a type of intragenic mCG, called gene body methylation (gbM), that tends to occur in exonic nucleosomes of conserved, constitutively expressed genes,[27–30] and (in the model plant *Arabidopsis thaliana*) can prevent aberrant intragenic transcription and promote gene expression.[21,31] Plant gbM is characterized by the absence of overlapping non-CG methylation[32,33] and by a more equal balance between the rates of mCG loss and gain than that observed in transposons.[17,34] Compared with transposon methylation, the epigenetics of gbM are far more mysterious.[35] The CMT3 methyltransferase has been proposed as the main *de novo* enzyme, but the only direct evidence for this is from experiments in which CMT3 is ectopically overexpressed.[36,37] How apparently random mCG losses and gains at individual CG sites produce gbM patterns that remain coherent over long periods of time is unknown. Why gbM favors nucleosomes and exons is unclear, and why some genes are reproducibly methylated across time, others are reproducibly unmethylated, and yet others are variably methylated is mysterious.[19,38,39]

Here, through a combination of theoretical investigation, genetics, and population genetics, we elucidate the mechanistic basis of gbM epigenetic inheritance, the timescale over which it operates and its connection with genetic variation in *Arabidopsis*. We find that gbM establishment, maintenance, and even loss, constitute a unified, MET1-mediated process. MET1 activity is suppressed by the histone variant H2A.Z and, especially in UM genes, by the DNA demethylase ROS1. Any level of gbM can be stably inherited over tens of generations, but—contrary to existing models—gbM patterns are not governed by bistable epigenetic inheritance. Instead, our simulations show that, with a constant genetic background, gbM patterns eventually converge to a single DNA-sequence-dependent steady state. Nevertheless, gbM undergoes large stochastic epigenetic fluctuations that explain much of the observed population-scale gbM variance. These fluctuations can last for thousands of years in the absence of genetic change, thereby establishing gbM as an epigenetic genotype able to mediate evolution on this timescale.

## RESULTS

### Gene body methylation epigenetic dynamics are primarily mediated by MET1

To investigate the timescale and mechanism of gbM epigenetic inheritance, we grew *Arabidopsis* for up to six consecutive generations (Figure S1A) and obtained whole-genome bisulfite sequencing data. In addition to wild-type (WT) Col-0, we analyzed mutants of all non-MET1 methyltransferases in different combinations (*cmt2*, *cmt3*, *drm1drm2*, *cmt2cmt3*, *drm1drm2cmt2cmt3* [*ddcc*]). In *Arabidopsis*, the MET1 family is composed of four genes: *MET1*—the main, and potentially the only functional enzyme[40]—and three close homologs.[41–43] Thus, unless we specifically refer to the *MET1* gene, all conclusions below apply to the overall MET1 family.

We identified single CG site methylation gains and losses at each generation and calculated epimutation rates per cell cycle based on the published estimate of 34 cell cycles per generation[44] (Figures 1A–1C; Tables S1A and S1B). We used published data[45] to define gbM and UM segments within gbM genes, so that UM segments have small amounts of mCG in our datasets (Tables S1C and S1D), which allows calculation of loss rates. Our rates for all *Arabidopsis* genes agree with those previously published (Figure 1A; Table S1).[17] Notably, rates of mCG gain ($2.6 \times 10^{-5}$ per site per cell cycle) and loss ($3.2 \times 10^{-5}$ per site per cell cycle) are finely balanced in gbM segments, but the rate of gain in UM segments of gbM genes (UM in gbM, $2.2 \times 10^{-6}$ per site per cell cycle) is 80-fold lower than the loss rate ($1.8 \times 10^{-4}$ per site per cell cycle), which is consistent with the low mCG levels in UM segments. Although a recent study concluded that sparsely methylated regions (SPMRs) within gbM genes have an enhanced epimutation rate,[46] we find that epimutation rates are similar between SPMRs and methylated regions of gbM genes that lie outside SPMRs (Figures S1B–S1E; Table S1).

We find that rates of mCG gain and loss in gbM regions are similar to WT in all the methyltransferase mutants we tested, including the quadruple *ddcc* mutant where the only functional methyltransferase is MET1[40] (Figures 1A and 1B; Table S1). Reanalysis of published *cmt3* and *suvh4/5/6* triple mutant data[46] (SUVH4, 5, and 6 are histone methyltransferases that mediate DNA methylation by CMT2 and 3[47–49]) produced similar results (Figures S1F and S1G). Nevertheless, rates are somewhat changed in some mutants, most significantly a decreased gain rate in gbM segments of *ddcc* plants and elevated gain rates in all genes, gbM genes and in UM segments of gbM genes in *drm1drm2* mutants (Figures 1A and 1B; Table S1). These differences likely reflect small contributions of non-MET1 methyltransferases to gbM, which may contribute to gbM steady state over long periods of time. However, our data demonstrate that gbM gains and losses are primarily due to MET1 activity (Figures 1A and 1B; Table S1).

Methylation dynamics within gbM genes may differ from those of UM loci. To explore whether MET1 can establish methylation in such loci, we calculated mCG gain and loss rates in UM genes (defined using published data[45]; Figure 1C; Table S1). The WT loss rate in UM genes ($2.1 \times 10^{-4}$ per site per cell cycle) is similar to that in UM regions of gbM genes, but the gain rate ($1.7 \times 10^{-7}$ per site per cell cycle) is 13-fold lower than in UM regions of gbM genes, and 150-fold lower than in gbM segments, so that the UM gene loss rate is 1,200-fold higher than the gain rate. The epimutation rates in UM genes imply steady-state mCG of around 0.1% (STAR Methods), which is consistent with their overall lack of methylation.

(E) New methylated clusters occur in entirely unmethylated genes in *ros1* mutants, including in a gene (AT4G08865) that is almost never methylated in the population. Each circle represents an individual CG pair, with darkness of fill indicating fractional mCG.
(F–H) Per cell-cycle rates of mCG gain at individual CG sites in genes that are unmethylated (UM) in Col-0, grouped depending on their population gbM frequency into rarely methylated genes (RMGs) (F, <20% population gbM frequency), occasionally methylated genes (OMGs) (G, 20%–50% population gbM frequency), and frequently methylated genes (FMGs) (H, ≥50% population gbM frequency). A significant difference from the WT rate is indicated by *p = 0.01–0.001, **p = 0.001–0.00001, and ***p < 0.00001 (Fisher's exact test), N = 10–36. Numbers above bars indicate fold-change from WT for genotypes that are significantly changed in UM genes overall.

We observed no reduction of the UM gene gain rate in any of the analyzed mutants and identified new methylation events in all methyltransferase mutants, including the quadruple *ddcc* mutant (Figure 1C; Table S1A). Indeed, gain rates are strongly elevated in *drm1drm2* and *ddcc* plants (Figure 1C; Table S1A), suggesting that DRM methyltransferases indirectly suppress gains at UM genes. All genotypes except *cmt2*, including *ddcc*, contain at least one example of a single mCG gain within a UM gene expanding to produce a new mCG cluster (Figure 1D; Table S2A). These results demonstrate that methyltransferases other than MET1 are not required to initiate or expand gbM clusters. MET1 is therefore a *de novo* methyltransferase that can stochastically establish gbM in previously UM genes.

GbM patterns vary substantially between natural *Arabidopsis* accessions.[19,20,50,51] These differences could arise, at least in part, due to the stochastic *de novo* activity of MET1. To investigate this hypothesis, we determined whether UM genes that gained mCG in our experiments (regardless of whether this expanded into a new gbM cluster), or in analogous published experiments, are more likely to be methylated in the *Arabidopsis* population than all Col-0 UM genes. Col-0 UM genes that gained gbM in our WT data have gbM in 46% of accessions on average (Table S2B). The corresponding percentages for other genotypes are 50% for *cmt2*, 45% for *cmt3*, 42% for *cmt2cmt3*, 41% for published WT mutation accumulation line (MAL) data, 32% for *drm1drm2*, and 35% for *ddcc* (Table S2B). In comparison, all Col-0 UM genes are much less likely to contain gbM in the population (16%, $p < 2 \times 10^{-16}$, Table S2B) and overall, genes with population gbM frequencies between 10% and 90% are relatively uncommon.[21,52] This indicates that some (relatively rare) UM Col-0 genes are predisposed to gain gbM due to *de novo* MET1 activity and suggests that natural *Arabidopsis* gbM diversity reflects an accumulated pattern of stochastic differences that are initiated and epigenetically maintained by MET1.

## ROS1 preferentially prevents methylation in UM gene bodies

The above results raise the question of why some genes are more likely to experience *de novo* MET1 activity. Two of the methyltransferase mutants we analyzed, *drm1drm2* and *ddcc*, exhibit significantly increased rates of mCG gain in UM genes (Figure 1C; Table S1A). We hypothesized that this may be due to reduced expression of the DNA demethylase *ROS1* that occurs in DRM pathway mutants.[53,54] We therefore analyzed mCG in *ros1* mutants over consecutive generations. We did not observe a significant change in the gain rate in gbM segments (Figure 1A). The rate of mCG gain is elevated by about 25% in UM segments of gbM genes and increases 10-fold in UM genes, to about the WT level in UM segments of gbM genes (Figures 1B and 1C; Table S1A). The loss rate in *ros1* mutants is reduced by about 25% in gbM segments, by about 40% in UM segments of gbM genes, and by over 60% in UM genes (Figures 1A–1C; Table S1B). This indicates that ROS1 preferentially suppresses mCG in UM genes but also affects the mCG loss and gain rates in gbM genes.

New gbM clusters arose in UM genes of *ros1* mutants, but unlike in the other genotypes, these are not restricted to genes with gbM in other accessions (Table S2A). A new cluster arose in a gene (AT4G08865) that contains gbM in only 0.4% of accessions (Figure 1E), and clusters arose in AT4G34419 (gbM in 3.4% of accessions) and AT5G63715 (gbM in 2.6% of accessions; Table S2A). In contrast, no gene with a population gbM frequency below 13.4% (a group that includes most Col-0 UM genes[21,52]) gained a gbM cluster in other genotypes (Table S2A), suggesting that ROS1 maintains a large subset of genes in a perpetually UM state. Indeed, mean population gbM frequencies in Col-0 UM genes that gain any gbM in *ros1* (33%, n = 104) and low-ROS1 genotypes (32%, n = 113 in *drm1drm2* and 35%, n = 125 in *ddcc*) are significantly lower than such frequencies in other lines (41%–50%, $p < 7.46 \times 10^{-09}$; n = 857, Table S2B). These data show that in the absence of ROS1, gains occur at greater frequency in genes that are unlikely to have gbM in the population (i.e., most UM genes).

To further investigate the connection between ROS1 and likelihood of gbM gain, we subdivided Col-0 UM genes based on their population gbM frequency into rarely methylated (RMGs; gbM in <20% of accessions, n = 6,113), occasionally methylated (OMGs; gbM in >20% and <50% of accessions, n = 1,500) and frequently methylated genes (FMGs; gbM in >50% of accessions, n = 662). WT and *cmt* mutant lines (*cmt2*, *cmt3*, and *cmt2cmt3*) show a clear progression of mCG gain rates, which increase with population gbM frequency (Figures 1F–1H; Table S1A). In contrast, gain rates in *ros1* mutants are more uniform, so that the relative gain rate increase between WT and *ros1* is greatest in RMGs (14-fold) compared with FMGs (5-fold) (Figures 1F–1H; Table S1A). We therefore conclude that ROS1 regulates the relative probability of gbM gain.

## Histone variant H2A.Z broadly reduces the rate of gbM gains

The histone variant H2A.Z shows a strong, quantitative anticorrelation with gbM in *Arabidopsis* and is known to antagonize gbM and DNA methylation in general.[30,55–60] We therefore investigated whether H2A.Z, similar to ROS1, shapes gbM epigenetics by analyzing triple *hta8 hta9 hta11* (*h2az*) mutants over multiple generations. We observed a significant increase in the gbM gain rate in *h2az* compared with WT, with the strongest relative effect in UM genes (8.8-fold increase), followed by UM regions of gbM genes (2.6-fold increase) and gbM segments (1.5-fold increase; Figures 1A–1C; Table S1A). Loss rates overall decreased to about the same extent as in *ros1* mutants, with an even stronger decrease in UM regions of gbM genes, so that the *h2a.z* loss rate there and in UM genes is about the same (Figures 1A–1C; Table S1B). Thus, H2A.Z, like ROS1, preferentially lowers mCG gains and increases mCG losses in UM sequences. However, unlike ROS1, H2A.Z has a significant effect on mCG gain in gbM segments, and has a relatively stronger effect on epimutation rates in UM regions of gbM genes, whereas the effects of ROS1 and H2A.Z are about the same in UM genes (Figures 1A–1C; Tables S1A and S1B). Also, unlike in *ros1* mutants, Col-0 UM genes with methylation gains in *h2az* mutants have an average population gbM frequency (47%, n = 73) similar to that in WT data (Table S2B). Consistently, lack of H2A.Z causes a smaller (and non-significant) relative mCG gain rate increase in RMGs (4-fold) than in UM genes with higher population gbM frequencies (~9-fold; Figures 1F–1H; Table S1A). Therefore, H2A.Z does not preferentially influence RMGs as ROS1

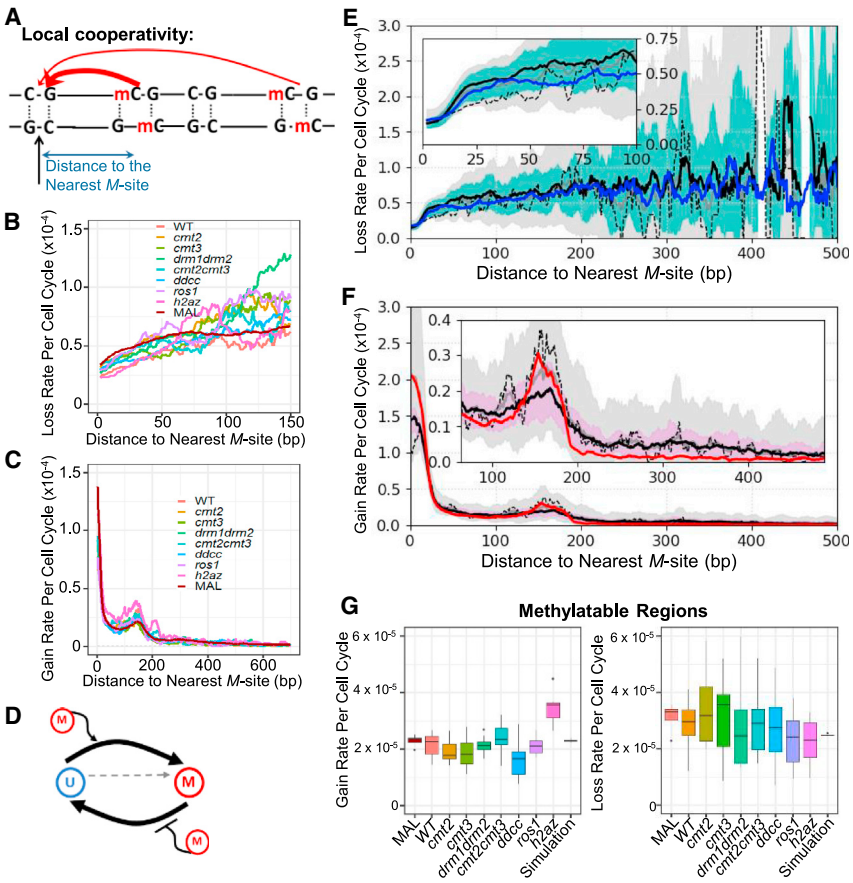



**Figure 2. Existing methylation shapes MET1-mediated mCG gain and loss**

(A) Schematic showing expected dependence of mCG gain rate on proximity to nearby methylated sites produced by a local cooperative interaction. Strength of the effect decreases with the distance between a mCG site and a target CG site.

(B and C) Data profiles of methylation loss (B) and gain (C) rates per cell cycle, plotted as a function of distance to the nearest mCG site, plotted over whole gbM genes. Gain rate plateaus at a level ($5 \times 10^{-7}$ per site per cell cycle; dashed line) comparable to the gain rate in entirely unmethylated genes (Table S1A). Rates shown as a 30 bp moving average. Different genotypes show similar patterns of cooperative gain and loss, as does our data and published data (MAL).

(D) Schematic of the simulated effective two state model.

(E and F) Simulated profiles of mCG loss (E, blue line) and gain (F, red line) rates per cell cycle, plotted as a function of distance to the nearest mCG, calculated over whole gbM genes, where each line is the mean of 4 simulated replicates (simulated over 30 generations from the Col-0 initial state). Solid black lines represent the loss (E) and gain (F) rate profiles averaged over published MAL data,[45] with the light-blue (E) and pink (F) bands corresponding to ±1 standard deviation (SD). Dotted black lines reproduce the Col-0 loss/gain rate profiles shown in (B) and (C). Solid gray lines are mean gain/loss rates, averaged over all WT datasets, with gray bands showing ±1 SD. Rates are shown as a 10 bp moving average. Model parameters given in Table S3C. Insets highlight loss rate at short length scales (E) and enhanced gain rate at a length scale of ~170 bp (F).

(G) Epimutation rates in simulation over modeled methylatable regions agree well with those from experimental data. Simulation rate analysis performed over 30 generations starting from Col-0 initial state using 4 simulation realizations and averaging, to closely resemble methodology used to calculate MAL rates. Model parameters given in Table S3A.

does but instead broadly suppresses gbM gain. Our results suggest that regions of high H2A.Z are incompatible with gbM due to decreased mCG gain and increased loss, and indicate that ROS1 and H2A.Z define the genic regions where MET1-mediated gbM epigenetic dynamics can unfold.

## Local cooperativity shapes MET1 *de novo* activity

We observed that gains of single mCG sites in UM genes were often followed either by reversion to a UM state or by rapid expansion of a gbM cluster (Figures 1D and 1E). This suggests MET1-mediated gbM has cooperative dynamics, where the rates of mCG change are influenced by nearby mCG sites (Figure 2A).[23–25,61] To examine this behavior, we plotted the likelihood of mCG loss and gain within gbM genes relative to the nearest mCG site (Figures 2B and 2C). In addition to our newly generated data, we analyzed data from previous studies that assessed mCG inheritance in WT *Arabidopsis* MALs over 30 generations (Table S1).[45,62] The published data produced the same general results as our WT data (Figures 2B and 2C; Table S1). Individual lines within the MAL datasets produced similar patterns, as did the two different MAL datasets overall (Figures S2A–S2C).

Rates of gain and loss are shaped by proximity to mCG sites (Figures 2B and 2C), with loss rates rising with increasing distance to other mCG sites (Figure 2B). The effect of nearby mCG on methylation gain is even clearer, with the likelihood of gain over 100-fold above background within 25 bp of a methylated site, before dropping sharply (Figure 2C). The methylation gain rate then rises again, peaking between 160 and 170 bp from the nearest mCG site, before plateauing to a background level ($5 \times 10^{-7}$ per site per cell cycle; indicated by the dotted line in Figure 2C) by ~1,000 bp away from the nearest mCG site. Analysis of loci polymorphic for gbM between datasets shows that the rate of mCG gain is indeed dependent on existing proximate mCG and is not an intrinsic property of a locus (Figure S2D; STAR Methods).[63] The profiles of transgenerational gain and loss rates are consistent with those recently reported for somatic *Arabidopsis* development,[64] and are very similar between WT and all methyltransferase mutants, including the quadruple *ddcc* mutant (Figures 2B and 2C), as well as *ros1* and *h2a.z* mutants (Figures 2B and 2C). These results indicate that nearby mCG strongly stimulates MET1 *de novo* activity as well as either promoting MET1 maintenance activity or inhibiting DNA demethylation (or both).

## A mathematical model can faithfully reproduce the observed steady-state gbM levels

To gain a quantitative understanding of gbM dynamics, we developed a computational stochastic model that contains four

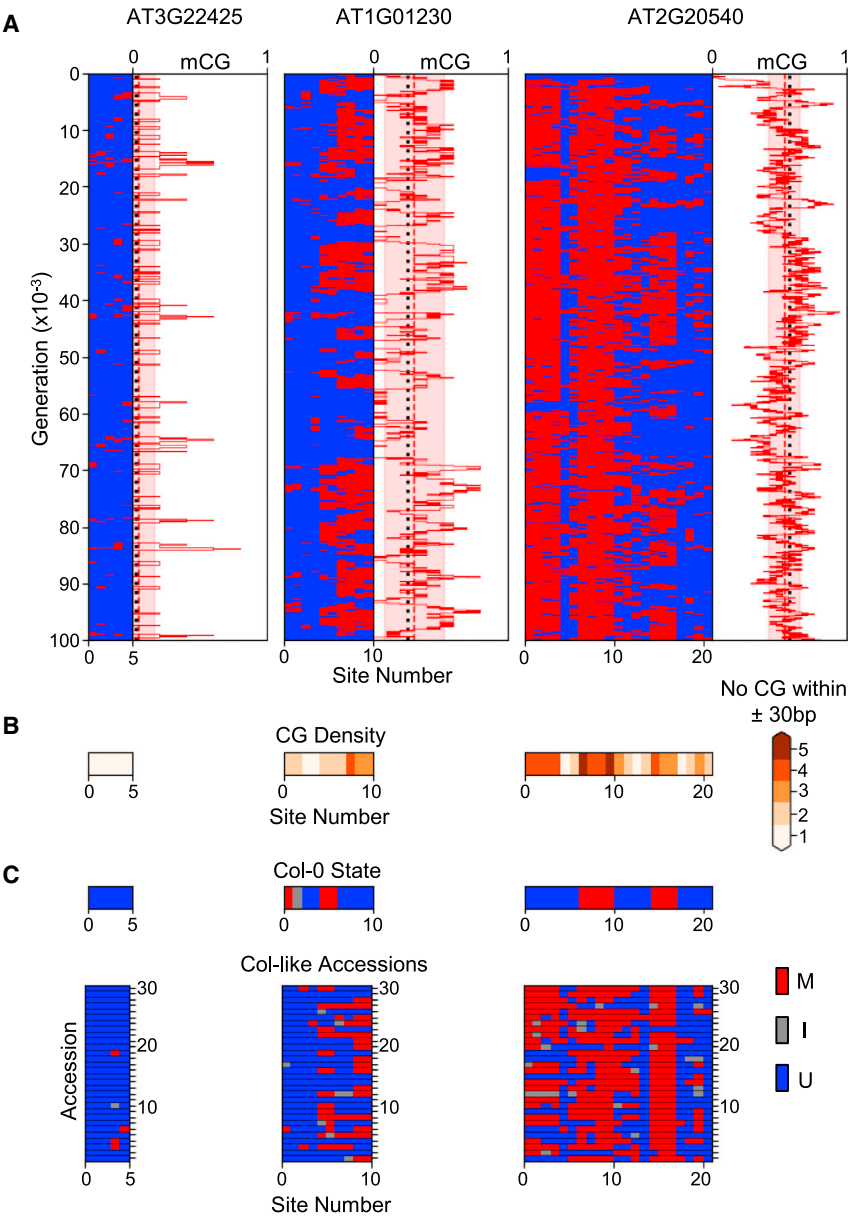



**Figure 3. Reproduction of steady-state gbM by a mathematical model**

(A) Simulated time courses for three example methylatable region loci of different sizes (5, 10, and 21 CG sites) reaching steady state from all-*U* initial states. Pink highlighted windows of mCG level indicate the typical magnitude of mCG fluctuations from steady state (calculated by SD of fluctuations), and the red dashed vertical line indicates the time-averaged steady-state mCG level (both calculated between 50,000 and 100,000 generations). Black dotted vertical line shows the average mCG level of 30 Col-like accessions. Model parameters in Table S3A.

(B) Local CG density in these same loci over a 60 bp window centered on each CG site.

(C) Methylation patterns in these loci for Col-0 and 30 Col-like accessions. *M*-site: red; *U*-site: blue; unknown methylation status, I: gray.

gain/loss processes: spontaneous *de novo*, cooperative *de novo*, background loss, and cooperative suppression of loss, with power-law cooperativity decay (STAR Methods; Figures 2D and S2E–S2H; Table S3). Simulations over 30 generations (1,020 cell cycles), to match the timescale of the published MAL data,[45,62] quantitatively recapitulate the experimental gain and loss profiles within gbM genes (Figures 2E and 2F), indicating that a simple mathematical model can encapsulate complex epimutation patterns.

Over short timescales, such as 30 generations, it is essential to simulate starting from the experimental Col-0 methylation state, because epimutation rates are modulated by proximate mCG (Figures 2B–2F). Due to the low rates of mCG loss and gain, a simulation over 30 generations only slightly alters the initial methylation pattern. However, as *de novo* methylation occurs in completely UM regions (Figure 1C), and new gbM clusters can appear and expand over just a few generations (Figure 1D), we tested the long-timescale behavior of the model to investigate whether it can reproduce the observed methylation patterns when run from a completely UM initial state (all-*U*) for 100,000 generations (3.4 million cell cycles; sufficient to reach steady state). Because our results indicate that ROS1 prevents a subset of genes from gaining gbM (Figures 1F–1H) and H2A.Z broadly influences gbM epimutation rates (Figures 1A–1C and 1F–1H), we simulated over genic regions that contain gbM in the *Arabidopsis* population and have low H2A.Z (Table S4; STAR Methods). The short-term gain and loss profiles continue to be reproduced well in these methylatable regions by the model (Figure S2I; Table S5, with the corresponding locus averaged epimutation rates shown in Figure 2G), similar to those over whole gbM genes (Figures 2E and 2F). Within such methylatable regions (three example loci of different lengths shown in Figure 3), overall methylation levels are reproduced accurately (Figure 4A): the distribution of individual locus mCG levels of 30 accessions with Col-like gbM corresponds well with the steady-state mCG levels produced by our model (simulations repeated 30 times to match the number of accessions; Figures 3 and 4A; see STAR Methods; Figures S2 and S3 for description of model construction and fitting).

To confirm that the steady-state methylation dynamics are independent of the initial methylation state, we also simulated starting from a fully methylated state (all-*M*; Figures 4A, 4B, and S3I). In both instances, the model is fully converged after around 10,000 generations (Figure 4B): the model therefore predicts that eventual mCG patterns are not bistable (discussed further in STAR Methods). Starting from an UM state, the first mCG is gained, on average, after 340 generations (Figure 4C),

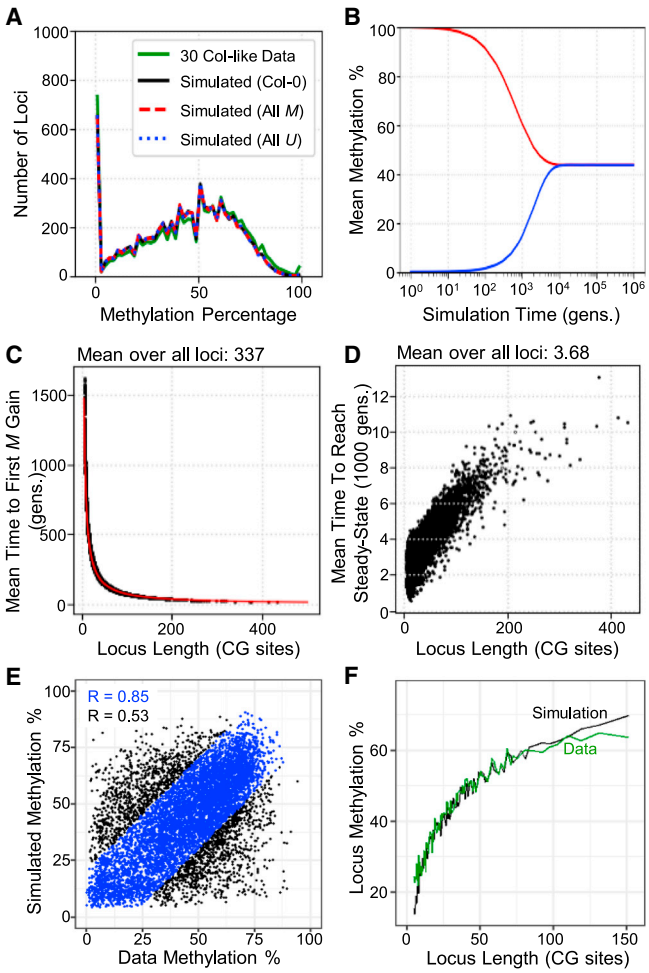



**Figure 4. Accurate prediction of a unique steady-state gbM level for each gene**

(A) Simulated steady-state methylatable regions mCG distribution, using three initial state choices: the experimental Col-0 methylation (black solid), completely unmethylated (blue dotted) and fully methylated (red dashed). All 3 simulations converge to the same steady state after 100,000 generations (normalized for 30 replicates, model parameters in Table S3A). Distribution of methylation levels over methylatable regions for high-coverage Col-like accessions shown in green (N = 30).

(B) Time to convergence of simulated methylatable regions model (in generations). Model is converged once simulations from the all-*M* initial state (red) and the all-*U* initial state (blue) reach the same steady state. At each time point, methylation level for each locus averaged across 740 replicates and then all averaged together. Model parameters in Table S3A.

(C) Mean time to first methylation gain from an all-*U* initial state shown for loci of different lengths, averaged over 740 locus replicates from methylatable regions model simulations (model parameters in Table S3A).

(D) Mean time to first reach steady state as a function of methylatable region locus length. Simulated over 100,000 generations from all-*U* initial state (averaged over 740 replicates), model parameters in Table S3A.

(E) Correlation of mCG levels at individual methylatable regions between simulations (averaged over 740 replicates) and data (averaged over 740 Col-like accessions; R = 0.53, n = 7,980). Well-fitting loci (<20% difference in methylation between data and simulation) are shown in blue (R = 0.85, n = 6,005), while poorly captured gene regions are shown in black. Simulations run as in (D).

(F) Longer methylatable regions are more highly methylated in both simulation results (black) and in the 740 Col-like accessions data (green). Loci were grouped into percentiles by CG site number, with 99% of percentiles shown and experimental methylation level averaged over N = 740. Simulated methylation level for each locus was averaged over 740 replicates after 100,000 generations starting in all-*U* initial state. Model parameters in Table S3A.

and a locus reaches steady state in around 3,700 generations on average (Figure 4D). The time needed depends strongly on locus size, so that loci with few CG sites reach steady state relatively quickly (~2,000 generations), whereas loci with >200 CG sites can take 10,000 generations or more (Figure 4D).

As good agreement is observed when calculating the epimutation rates over the methylatable regions using the data and the simulated methylation changes over 30 generations (Figure 2G; Tables S1 and S5), we estimated the steady-state gbM distributions implied by the rate changes in *cmt3*, *ddcc*, *ros1*, and *h2az* mutants (Figures S4A–S4D; Table S6; STAR Methods). Both *cmt3* and *ddcc* are predicted to cause only small decreases in steady-state gbM (Figures S4A and S4B; Table S6), further supporting the primary role of MET1 in shaping *Arabidopsis* gbM. By far the largest change is predicted in *h2az*, with steady-state mean gbM in methylatable regions increasing from 44% to 68% (Figure S4D; Table S6). This increase is predicted to unfold over almost 2,000 generations (Table S6), which explains why *h2az* mutants grown in the laboratory for a few generations show only small gbM increases[55] (Table S1D).

To test the predictive power of the model with data that was not used for fitting, we employed a much larger set of unique, high-coverage Col-like accessions (N = 740) and simulated from a completely UM initial state for 100,000 generations (repeated 740 times to match the number of accessions; Figure S4E). The values for each gene were averaged across accessions and compared with those equivalently averaged across the 740 simulated realizations. All replicates were simulated using identical CG site positions and the same methylatable regions annotation and parameter values (making them genetically identical), thus ensuring that methylation pattern differences between simulated replicates are purely epigenetic. We found that simulated mCG levels within 75% of gene regions agreed well with the data, having a Pearson's R = 0.85 (Figure 4E). When all gene regions were considered, the poorly captured gene regions reduce the correlation to R = 0.53. The model does best with regions that have classical gbM traits (longer, less CG-dense genes with low H2A.Z[35]) (Figure S4F). Mean mCG increases strongly with CG site number, which is captured well by the model (Figures 4F and S4G). This occurs because the overall cooperative feedback, and therefore the gain rate, increases with the number of CG sites a locus contains.

## The model accurately predicts steady-state gbM patterns at individual loci

So far, we have used the model to predict gbM levels and epimutation rates—the types of data used to fit the model. To test the model's generality, we examined its predictions for steady-state distributions of mCG sites within each gene (gbM patterns). The model quantitatively predicts the experimental 30-generation gain/loss rate profiles plotted as a function of distance to the nearest uCG site (Figure S5A, rather than the nearest mCG as in Figure S2I). The predicted distributions of

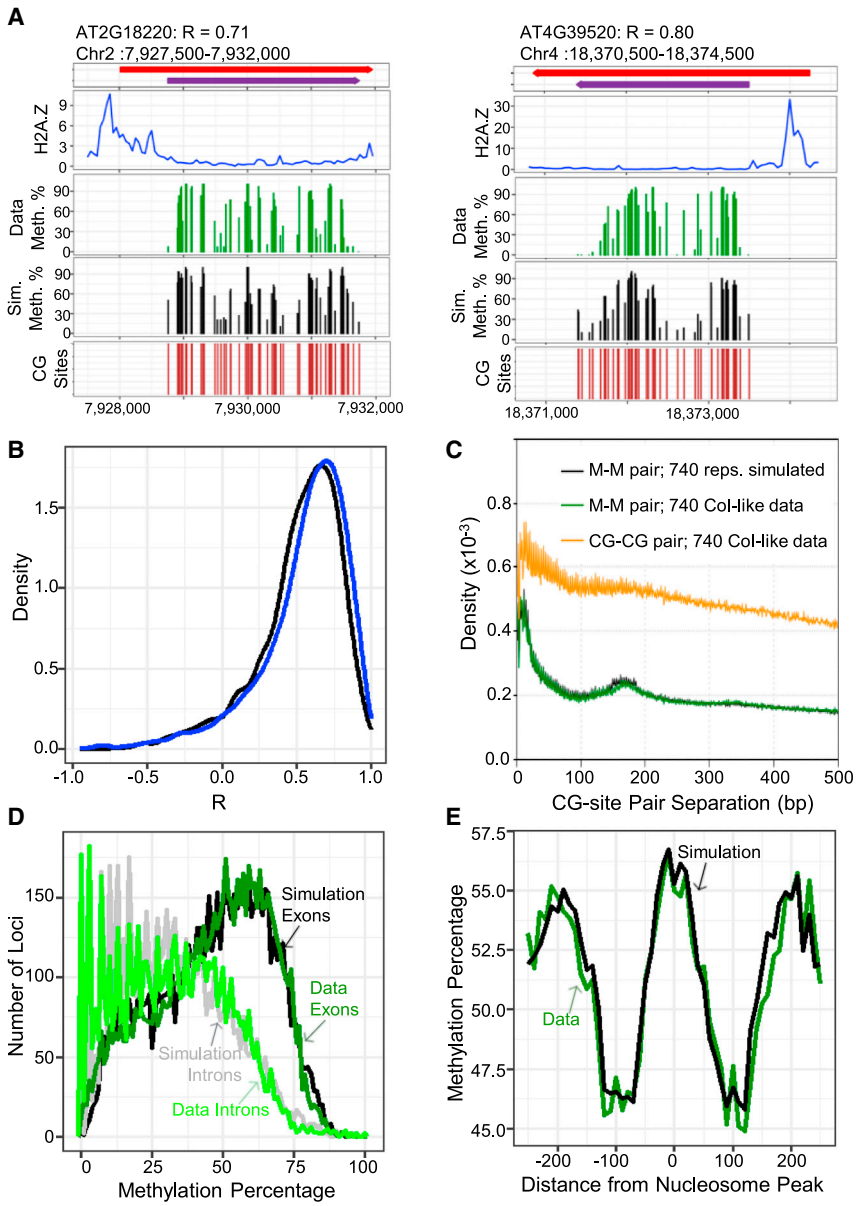



**Figure 5. GbM patterns are correctly reproduced by the mathematical model**

(A) Genome browser views of AT2G18220 (left) and AT4G39520 (right), with gene annotation in red and the modeled methylatable region in purple. Methylation of individual cytosines agrees well between the average of 30 Col-like accessions (data methylation, green) and the average of 30 model realizations (simulated methylation, black). Pearson's R of the correlation of these methylation patterns between data and simulated results is shown above each plot. Col-0 H2A.Z enrichment (blue) and positions of individual CGs within the modeled region (brown) are indicated. Simulated for 100,000 generations from all-*U* initial state, model parameters in Table S3A.

(B) Pearson's R values between the simulation and data were calculated for each methylatable region (as for those shown in A), and the distribution of these R values is shown. Loci where the overall methylation level is reproduced well (<20% difference in methylation between data and simulation) are shown in blue, as in Figure 4E (N = 6,005), while gene regions with a poorly captured overall methylation level are shown in black. Simulated as in (A).

(C) Distances between all M-M pairs within the same methylatable gene region were calculated (STAR Methods). Simulated (black) and observed (740 Col-like accessions, green) distributions of M-M separation distances are in good agreement, indicating accurate reproduction of methylation patterns. Distribution of separation distances of all CG-CG site pairs within the same methylatable gene regions shown in orange. Model simulation results for 740 locus replicates after 100,000 generations starting from all-*U* initial state. Model parameters in Table S3A.

(D) Modeled methylatable regions were divided into exons and introns, and average mCG was calculated for each region across Col-like accessions (data, dark green for exons and light green for introns, N = 30) and simulated realizations (simulation, black for exons and gray for introns, N = 30). Simulated for 100,000 generations from all-*U* initial state, model parameters in Table S3A.

(E) Methylation is enriched over nucleosomes in the simulation (black, simulated as in D, N = 30) and in the methylatable regions data (green, N = 30). Methylation patterns averaged over well-positioned nucleosomes as defined in Lyons and Zilberman.[65]

methylated/unmethylated neighboring CG site pairs are in good agreement with the data, with mCG-mCG separations enriched at short distances compared with uCG-mCG or uCG-uCG (Figure S5B). This is consistent with local cooperativity, which will favor clustering of methylated sites.

Overall, spatial mCG model predictions and observed patterns at individual targets are in good agreement (including at loci where the model is less successful at predicting the steady-state gbM level), revealing that methylation is enriched in areas of high local CG density (Figures 3, 5A–5C, and S3I). This is an expected feature of a cooperative process, as regions of high CG density can generate greater positive feedback between nearby mCG sites. Model predictions for mCG levels in exons and introns agree well with the data, with higher methylation observed for exons (Figure 5D), reflecting the greater CG density of exons (Figure S1B, right). The model also reproduces the reported gbM enrichment within nucleosomes (Figure 5E),[27,65] indicating that this enrichment stems from the known tendency of nucleosomes to center on CG-dense DNA,[66,67] where cooperative interactions drive higher mCG levels (Figures 2B–2F and S4G; confirmed by simulating steady-state mCG after randomizing CG site positions; Figures S5C–S5E; STAR Methods). Hence, the model accurately predicts steady-state gbM patterns as well as gbM levels within methylatable regions given only CG site spacing as input. The level of agreement genome-wide is remarkable given the model was never fit to predict gbM patterns.

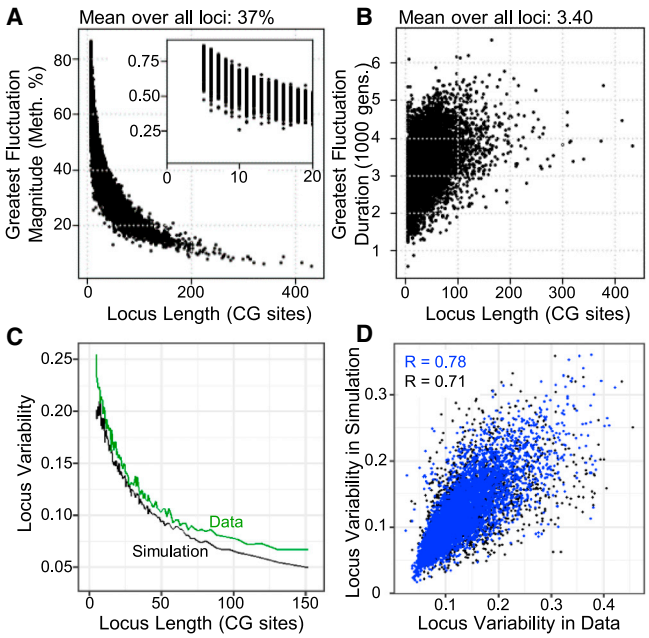



**Figure 6. Epigenetic fluctuations explain gbM pattern variation in *Arabidopsis* populations**

(A) Mean magnitude of greatest methylation fluctuation around mean steady-state mCG level (i.e., largest departure from steady-state mean methylation) plotted as a function of methylatable region length. Simulation over 100,000 generations from all-*U* initial state, over 740 replicates: first 50,000 generations of simulation used to equilibrate, with second 50,000 generations used to measure fluctuations (STAR Methods, model parameters in Table S3A).

(B) Mean duration (in 1,000s of generations) of the greatest methylation fluctuation away from steady state (i.e., duration of the fluctuation found to have the largest departure from steady-state mean methylation), plotted as a function of methylatable region length. Simulated as in (A).

(C) Shorter loci are more variable in both simulated results and data. Standard deviation of overall methylation levels of individual methylatable regions between 740 Col-like accessions (data, green) or simulated realizations (simulation, black) shown on y axis. Loci were grouped into percentiles of length, with 99% of percentiles shown. Simulations run for 100,000 generations starting from all-*U* initial state with 740 realizations for each locus. Model parameters in Table S3A.

(D) Correlation between the standard deviation of mCG levels of individual methylatable regions within 740 Col-like accessions (locus variability in data) and 740 simulated realizations (locus variability in simulation; R = 0.71, n = 7,980). Well-fitting loci (<20% difference in methylation between data and simulation) shown in blue, as in Figure 4E (R = 0.78, n = 6,005), while gene regions with a poorly captured overall methylation level shown in black. Simulations run over 100,000 generations from all-*U* initial state and averaged over 740 locus replicates (model parameters in Table S3A).

## GbM variation across natural accessions is accurately predicted by the model

Our model predicts a unique gbM steady state for each gene, but this state is subject to substantial fluctuations, which can include completely losing and regaining methylation (Figures 3 and S3I). Over the second half of 100,000 generation simulations (using the first half to ensure steady state has been reached), we found that the largest absolute fractional fluctuation experienced by each locus was 0.37 on average (a change of 1 representing a transition from a fully unmethylated to a fully methylated state, or vice versa), and lasted 3,400 generations on average (Figures 6A, 6B, S6A, and S6B). This measure indicates the largest departure from the steady-state mean methylation level of a given locus and illustrates that gbM epigenetic fluctuations can be very large and can last a long time. This stochastic variation strongly depends on the number of CG sites in the locus (Figure 6A), with bigger fluctuations for smaller loci, and the smallest loci remaining unmethylated for most of the simulation (Figures 3, 6A, 6B, and S6A–S6C). Large clusters of CG sites maintain an overall methylated state almost indefinitely, only occasionally losing a patch of methylation (Figures 3, 6A, S3I, and S6A–S6C). However, large fluctuations tend to last longer in loci with more CG sites (Figure 6B). Consistently, loci with few CG sites have more variable gbM across natural accessions, and this is captured well by the model (Figure 6C).

We note in particular how locus AT2G20540 (rightmost in Figure 3) has three clear mCG bands that correspond to regions of high local CG density, but in seven of the 30 accessions shown, and in Col-0, the leftmost band is absent (Figure 3C). In the simulation shown in Figure 3A, this band of methylation is spontaneously lost twice, each time remaining absent for several thousand generations before being re-nucleated. Such behavior is a hallmark of a cooperative process, as the feedback between nearby mCG sites reinforces and stabilizes a small initial random fluctuation.

Although the overall mCG levels of highly variable loci are predicted least accurately (Figure S4F), the variability itself is accurately predicted, with strong positive correlation between the variability of individual loci in the data and in the simulated results (R = 0.71; Figure 6D). Thus, the stochastic fluctuations of the model represent gbM variation across *Arabidopsis* accessions. The ability of our purely epigenetic model to accurately predict gbM variance across natural accessions indicates that gbM patterns in Col-like accessions (74% of assayed accessions, Figure S6D) are primarily different stochastic realizations of an identical epigenetic inheritance process.

## Modeling outlier gbM accessions

So far, we have modeled gbM with parameters that reflect the epigenetic dynamics of Col-0 and accessions with globally similar gbM. To explore the extent to which our model can reproduce gbM in accessions where it is substantially higher (Dör-10 and North Swedish accessions [NS, excluding Dör-10]) or lower (Can-0, UKID116, Cvi-0 and Relicts [RL, excluding Can-0 and Cvi-0]; Figures S6D and S7A), we focused on two model parameters: the overall strength of the cooperative interaction pathways relative to that for Col-0 ($r^*$, such that $r^* = 1$ for Col-0) and the level of spontaneous *de novo* activity ($r_0^+$; Figures S7B and S7C). Adjusting only the cooperativity strength (Figure S7B) produces reasonable fits in all cases (Figures 7A–7D), though sometimes considerably underestimates the number of fully UM loci, represented by the height of the spike at the origin (Figures 7C and 7D). In comparison, adjusting only the spontaneous *de novo* activity performs similarly for all but the most sparsely methylated accessions (Figures 7A–7D and S7C). The best fit to Can-0, UKID116 and Cvi-0, however, is found by adjusting both the cooperativity and the spontaneous *de novo* strength (Figure 7D). Overall, as the cooperative interaction is non-linear, smaller changes in its strength are required to alter the methylation level. Our results indicate that relatively small changes to

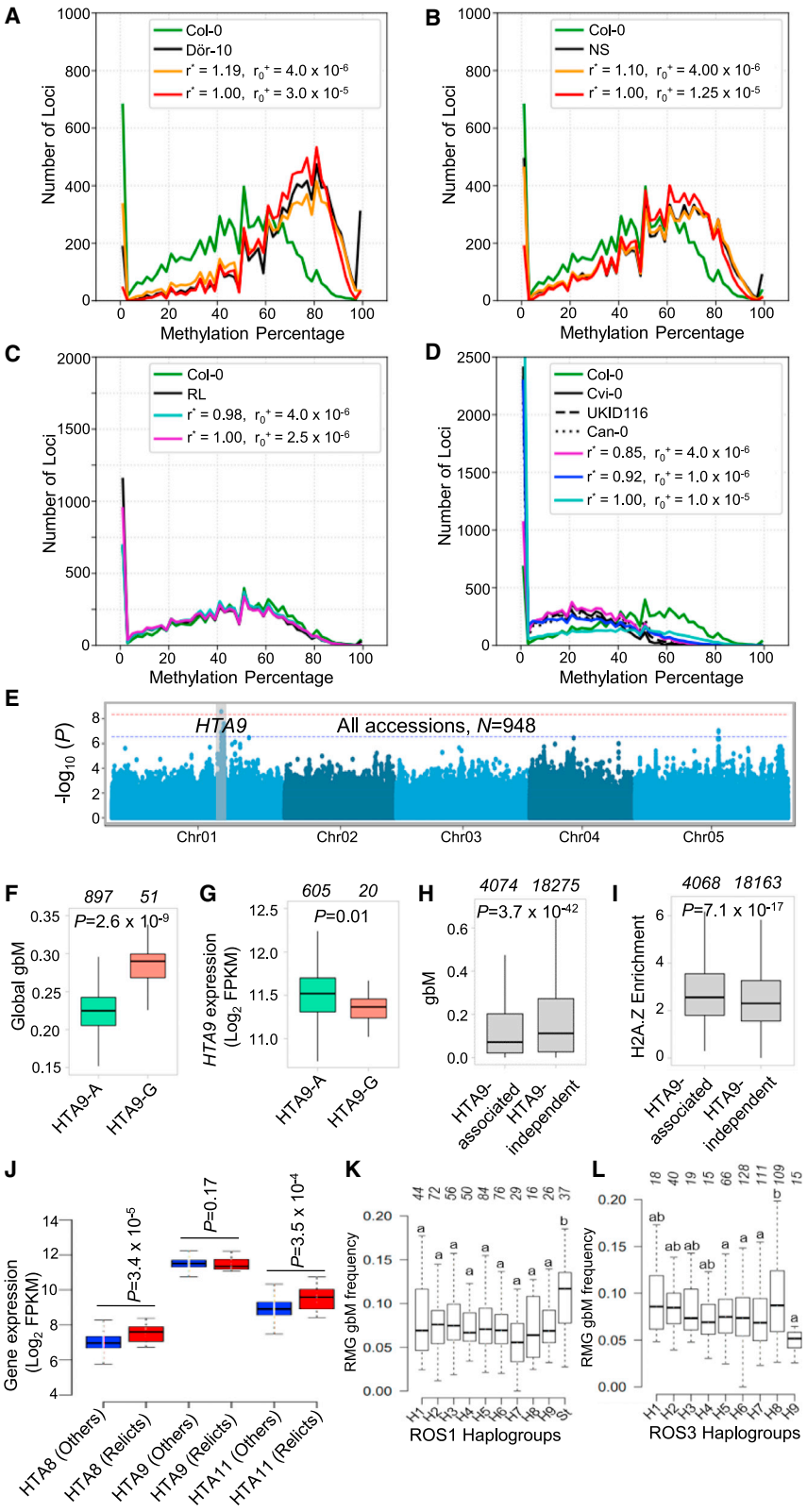



**Figure 7. Genetic variation is associated with global gbM variation in the population**

(A) Distribution of mCG levels for loci within Dör-10 (black) and simulated steady states for adjusted cooperative interaction strength ($r^*$, orange) or adjusted *de novo* activity level ($r_0^+$, red). Col-0 distribution also included (green, $r^* = 1.00$, $r_0^+ = 4.0 \times 10^{-6}$). Simulations run using methylatable regions annotation, for 100,000 generations starting from all-*U* initial state, normalized for 30 realizations. Unspecified model parameters as in Table S3A.

(B) Distribution of mCG levels for loci within Northern Swedish accessions (NS, black) and simulated steady states for adjusted cooperative interaction strength ($r^*$, orange) or adjusted *de novo* activity level ($r_0^+$, red). Col-0 distribution also included (green). Simulations as in (A).

(C) Distribution of mCG levels for loci within Relict accessions (RL, black) and simulated steady states for adjusted cooperative interaction strength ($r^*$, turquoise) or adjusted *de novo* activity level ($r_0^+$, purple). Col-0 distribution also included (green). Simulations as in (A).

(D) Distribution of mCG levels for loci within Cvi-0 (solid black), UKID116 (dashed), Can-0 (dotted), and simulated steady states for adjusted cooperative interaction strength ($r^*$, purple) or both adjusted cooperative interaction strength and *de novo* activity level ($r^*$ and $r_0^+$, blue). Adjusting only the *de novo* activity level ($r_0^+$, turquoise) and the Col-0 distribution (green) are also included. Simulations as in (A).

(E) *HTA9* was identified as a significant quantitative trait locus at both false discovery rate 0.05 (dashed blue line) and the Bonferroni threshold ($\alpha = 0.05$; dashed red line) with GWA analysis using global gbM levels as the phenotype. y axis indicates the $-\log_{10}$ of the p values of association between SNPs and global gbM variation.

(F) Accessions harboring minor *HTA9-G* allele exhibit significantly enhanced global gbM levels compared with those with *HTA9-A* allele ($p = 2.6 \times 10^{-9}$, mixed linear model GWA). Number of accessions in each group indicated on top.

(G) Expression of *HTA9* is lower in accessions with the *HTA9-G* allele ($p = 0.01$, Wilcoxon rank sum test). FPKM, fragments per kilobase of transcript per million mapped reads. Number of accessions in each group indicated on top.

(H) Genes whose gbM levels are associated with *HTA9* SNPs (*HTA9*-associated) have significantly lower gbM than *HTA9*-independent genes ($p = 3.7 \times 10^{-42}$, two-tailed t test). Number of accessions in each group indicated on top.

(I) Levels of H2A.Z enrichment in Col-0 are significantly higher in *HTA9*-associated genes ($p = 7.1 \times 10^{-17}$, two-tailed t test). Number of genes indicated on top.

(J) Expression of *HTA8*, *HTA9*, and *HTA11* in relicts (N = 20) and other accessions (N = 595). p values correspond to Wilcoxon rank sum test.

(K and L) Frequency of gbM within RMGs in ROS1 haplogroups (K) and ROS3 haplogroups (L) based on the amino acid sequence, and an additional ROS1-St group that contains accessions with a premature stop codon. Different letters indicate significant differences at $p < 0.05$ (one-way ANOVA), i.e., groups denoted with "a" are statistically different from "b" and ab ones are different from neither a nor b. The number of accessions is indicated for each haplogroup (top).

the gbM system can accommodate the entire range of gbM levels observed across the *Arabidopsis* population.

### H2AZ and MET1 polymorphism is associated with high gbM in natural populations

The necessity to modify model parameters to reproduce the methylation patterns of outlier accessions implies genetic differences in gbM factors that ultimately impinge on MET1 or active demethylation. To reveal such factors, we performed mixed model genome-wide association (GWA) analysis using the global gbM level of each accession as the phenotype. At the most stringent statistical threshold, this identified one locus (Figures 7E and S7D–S7F), which contains *HTA9*, one of three genes encoding the H2A.Z protein in *Arabidopsis*.[68] The minor *HTA9* allele (*HTA9-G*) is found exclusively in the north, mainly in northern Sweden (Figure S7G; Table S7A), a region where accessions tend to have high gbM.[19,50] Indeed, accessions harboring the *HTA9-G* allele exhibit greatly enhanced global gbM (Figure 7F) and show significantly lower expression of *HTA9* (Figure 7G). In addition to global gbM, we detected *HTA9* associations with gbM variation in 4,074 individual genes (Figures S7H–S7J; Table S7B). These genes have relatively low gbM across accessions (Figure 7H) and exhibit significantly enhanced deposition of H2A.Z in Col-0 (Figure 7I). Therefore, genes with high H2A.Z appear to be more sensitive to the effects of *HTA9* genetic variation. *HTA9* expression is not significantly different between low-gbM RL and other accessions (Figures 7J and S6D), but *HTA8* and *HTA11*, the other two *Arabidopsis* H2A.Z genes, have >40% higher expression in RL (Figure 7J), which may partly account for the low gbM in these accessions.

At a lower statistical threshold, we also identified a region containing *MET1* associated with global gbM levels of accessions (Figure S7D). A *MET1* haplotype (H9) is associated with high global gbM, and the two largest *MET1* haplotypes (H7 and H10) have significantly different global gbM (Figure S7K; Table S7A). Taken together, our results indicate that H2A.Z variation is an important driver of global gbM variation in natural *Arabidopsis* populations, and MET1 variation also likely contributes to natural gbM variation.

### ROS1 pathway polymorphism is associated with RMG gbM in natural populations

We did not detect associations between ROS1 genetic polymorphism and global gbM levels of accessions. However, our genetic results (Figures 1A–1C and 1F–1H) predict that natural accessions with reduced ROS1 activity should have an overabundance of gbM in genes that are rarely methylated in the population (RMGs), which would not necessarily translate into substantially elevated global gbM. To test this hypothesis, we examined the association between ROS1 amino acid polymorphism and RMG gbM frequency across accessions. We defined nine ROS1 haplotypes, and a tenth group containing alleles with premature stop codons (Figure S7L; Table S7A). Accessions carrying a premature stop codon (ROS1-St) display a high frequency of RMG gbM (Figure 7K). Furthermore, GWA analysis using the number of methylated RMGs as the phenotype identified several single nucleotide polymorphisms (SNPs) around ROS1 that are marginally associated with RMG gbM frequency (Figure S7M). This analysis also detected a significant association (SNP Chr5:23536319; Figure S7N) near ROS3, an RNA-binding protein that functions in the ROS1 pathway to mediate active DNA demethylation.[69] Based on amino acid sequence variation, we defined nine ROS3 haplogroups (H1–H9). GWA SNP Chr5:23536319 is linked with ROS3-H8, which is associated with high RMG gbM frequency (Figures 7L and S7O; Table S7A). Dör-10, the accession with the highest gbM (Figures S6D and S7A), harbors both the ROS1-St and ROS3-H8 alleles, as well the HTA9-G allele (Figures 7F, 7K, and 7L; Table S7A)—an allelic combination that may account for the extraordinarily high Dör-10 gbM levels (Figures S6D and S7A). Taken together, these results indicate that the ROS1 pathway protects UM genes from methylation in the population, and ROS1 and H2A.Z together define the scope and scale of *Arabidopsis* gbM in nature.

## DISCUSSION

Our results demonstrate that gbM epigenetic dynamics are dominated by MET1. In addition to its canonical semiconservative maintenance activity,[9] MET1 has *de novo* activity that is stimulated by proximate mCG (Figures 1, 2C, and S2D). Proximate mCG also boosts the efficiency of MET1-mediated maintenance (Figure 2B). A balance of *de novo* and maintenance not only maintains existing gbM but initiates and expands new gbM clusters (Figures 1C, 1D, 2B–2F, 3, 4C, and 4D). Rates of mCG loss and gain, including gain rates in completely UM sequences, are barely altered either in our *cmt3* and *cmt2cmt3* data, or in published *cmt3* and *suv4/5/6* data (Figures 1A–1C and S1G). This argues against the proposal that CMT3 plays a central role in gbM establishment.[70] Although a role for CMT3 in gbM maintenance is supported by the very low gbM observed in plant species that lack CMT3,[36] our results indicate that *Arabidopsis cmt3* mutants should have only very modestly altered steady-state gbM (Figure S4A; Table S6). This suggests either that CMT3 has a stronger influence on gbM in other species, or that the correlation between CMT3 loss and low gbM is not causal. It is also possible that the effect of CMT3 loss on gbM dynamics increases over time, perhaps due to global epigenetic alterations that unfold over many generations.

Our results also show that MET1 activity within genes is regulated by ROS1 and H2A.Z (Figures 1A–1C), but their effects are different. ROS1 activity is strongest in UM genes that are rarely methylated in the population, weakens as population gbM frequency rises, and is weakest in gbM genes (Figures 1A–1C and 1F–1H). This suggests that ROS1 (likely in collaboration with other factors) determines the population gbM frequency of each gene. H2A.Z also preferentially affects mCG epimutation in UM sequences (Figures 1A–1C), likely because these have high H2A.Z levels.[55,60] However, H2A.Z reduces gains and enhances losses in all sequences we analyzed (Figures 1A–1C and 1F–1H), and therefore appears to broadly suppress gbM. This suggests that gbM may only stably exist in sequences with low H2A.Z, but because DNA methylation reduces H2A.Z abundance,[60] H2A.Z and gbM may have a complex, dynamic relationship. This is not included in the model, where we treat H2A.Z as a static background that constrains gbM regions. However, this is one of several factors that could dynamically alter on long timescales, including mCG-induced mutation of CG sites.[71] The observation that H2A.Z affects mCG epimutation rates also relates to our earlier conclusion that H2A.Z has a global but small effect on

DNA methylation levels.[55] The *h2az* epimutation rate changes require many generations to reach a new gbM steady state (Table S6), so that laboratory measurements after a few generations show small effects in *h2az* mutants, whereas our population genetic analyses indicate that even a modest change in H2A.Z expression can cause a major long-term alteration of the gbM landscape (Figures 7E–7J). Indeed, the long timescales involved in these processes underline the importance of modeling, where they are easily accessible.

The gbM epigenetic dynamics we describe unify mCG establishment, maintenance and loss, and predict a unique steady state for each sequence (i.e., absence of bistability). Our model does not contain distinct establishment and maintenance phases: the apparent distinction between the two is driven by the sensitivity of *de novo* methylation and semiconservative maintenance to proximate mCG. Without nearby mCG to stimulate cooperative effects, gbM establishment in an UM region is very rare, much rarer than cooperative *de novo* mCG addition. Nonetheless, establishment and maintenance are a single, continuous process. Any level of methylation is stable over a few generations due to high maintenance fidelity and low *de novo* activity (even with cooperativity). However, this stability is ephemeral. Due to the slight imbalance between *de novo* addition and maintenance failure, the methylation level will drift toward a unique steady state (Figures 3, 4A, 4B, and S3I) over several thousand generations (Figure 4D). For shorter loci, the steady state is highly unstable, as they cycle through methylation establishment, maintenance, and loss (Figure 3). Thus, our model differs fundamentally from previous models that predict bistability.[23–25,72] Because cooperative feedbacks in our model stabilize the methylated, but not the UM state, most loci that can support gbM are predicted to be methylated at any moment. This is consistent with typical genic gbM population frequencies of either >90% or <10%,[21,52] with our data indicating that the <10% genes are kept UM by ROS1 and H2A.Z.

Our model flows directly from MET1 properties: the frequency and spatial distribution of *de novo* activity and maintenance failure. Because the reported *de novo* and maintenance failure rates for mammalian Dnmt1 are both about a 1,000-fold greater,[26,73] and its cooperativity patterns might also differ, the epigenetics we describe may be profoundly unlike those of mammalian mCG. However, as our model successfully predicts gbM features that are common across plants and invertebrates, including enrichment in exons and nucleosomes[27–29,65] (Figures 5D and 5E), it likely generally describes MET1/Dnmt1-mediated gbM epigenetics in these lineages.

Our results have important implications for the relative contributions of genetics vs. epigenetics to gbM pattern variation across different timescales. Our hypothesis that the same processes shape gbM over short and long timescales is validated by our success in predicting long-term steady-state gbM patterns using a model fundamentally based on short-term epigenetic dynamics (Figures 3 and 5). Over short timescales, such as 30 generations, our model recapitulates the observed epigenetic gbM dynamics (Figures 2E–2G and S2I). However, over thousands of generations, steady-state gbM levels and overall patterns are genetically determined by the interaction of global regulatory factors—H2A.Z, ROS1, ROS3, MET1, and likely others—with local CG site number and density. Nevertheless, around this gbM steady state, there are continuous stochastic fluctuations at each gene, which can be large and last for thousands of generations (Figures 3, 6A, and 6B). As demonstrated by comparing multiple realizations simulated with identical underlying genetics (i.e., identical CG site positions and parameter values specifying the methylation system), these intrinsic fluctuations are both generated and inherited epigenetically, and they closely match the gbM differences within the *Arabidopsis* population (Figure 6D). Thus, stochastic epigenetic inheritance generates most of the observed gbM variation between natural *Arabidopsis* accessions. This confirms the hypothesis that *Arabidopsis* gbM variation is primarily epigenetic[20] and indicates that gbM is an epigenetic genotype that can mediate phenotypic evolution in the *Arabidopsis* population.[21] Against the backdrop of our currently changing climate, understanding and protecting the diversity of natural populations is vital: our results show that diversity loss may not only be genetic[74] but may also have an important and hitherto unappreciated epigenetic component.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Arabidopsis thaliana: Biological materials and growth conditions
- METHOD DETAILS
  - Leaf genomic DNA isolation, library preparation, bisulfite conversion, and sequencing
  - Sequence Alignments and Segmentation
  - Methylation Calling
  - Calculation of Gain and Loss Rates
  - Population gbM frequency
  - Epimutation gain profiles within polymorphic loci
  - Choice of accessions for modelling
  - Methylatable Gene Regions
  - Correlation of spatial patterns of methylation within simulated and experimental loci
  - Analysis of methylation patterns over well-positioned nucleosomes
  - Analysis of sparsely methylated regions
  - Reanalysis of published methylation data
  - Genome-wide association mapping
  - Haplotype analyses
  - Modelling gene body methylation dynamics
  - Construction of Two-State Model
  - Gillespie simulation: cooperative gains
  - Gillespie simulation: cooperative maintenance
  - Active demethylation
  - Functional form of cooperative interactions
  - Calculation of diploid methylation gain and loss rates
  - Simulated gain and loss rates

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2023.10.007.

## ACKNOWLEDGMENTS

We would like to thank Xiaoqi Feng, Ander Movilla Miangolarra, and Suzanne de Bruijn for discussions. This work was supported by BBSRC Institute Strategic Programme GEN (BB/P013511/1) to M.H. and D.Z. and by a European Research Council grant MaintainMeth (725746) to D.Z.

## AUTHOR CONTRIBUTIONS

A.B. developed the mathematical model, performed analysis, and wrote the paper. E.H. performed DNA methylation experiments and analysis and wrote the paper. Z.S. performed population DNA methylation analyses and wrote the paper. J.D.M. performed DNA methylation analyses and wrote the paper. D.B.L. performed nucleosome positioning and DNA methylation analyses. M.H. conceived the research plan, supervised analysis, and wrote the paper. D.Z. conceived the research plan, supervised experiments and analysis, and wrote the paper.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 3, 2022
Revised: July 18, 2023
Accepted: October 13, 2023
Published: November 8, 2023

## REFERENCES


1. Greenberg, M.V.C., and Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. Nat. Rev. Mol. Cell Biol. *20*, 590–607. https://doi.org/10.1038/s41580-019-0159-6.
2. Kumar, S., and Mohapatra, T. (2021). Dynamics of DNA methylation and its functions in plant growth and development. Front. Plant Sci. *12*, 596236. https://doi.org/10.3389/fpls.2021.596236.
3. Zhang, H., Lang, Z., and Zhu, J.K. (2018). Dynamics and function of DNA methylation in plants. Nat. Rev. Mol. Cell Biol. *19*, 489–506. https://doi.org/10.1038/s41580-018-0016-z.
4. Catania, S., Dumesic, P.A., Pimentel, H., Nasif, A., Stoddard, C.I., Burke, J.E., Diedrich, J.K., Cook, S., Shea, T., Geinger, E., et al. (2020). Evolutionary persistence of DNA methylation for millions of years after ancient loss of a de novo methyltransferase. Cell *180*, 263–277.e20. https://doi.org/10.1016/j.cell.2019.12.012.
5. Chen, Z.X., and Riggs, A.D. (2011). DNA methylation and demethylation in mammals. J. Biol. Chem. *286*, 18347–18353. https://doi.org/10.1074/jbc.R110.205286.
6. Huff, J.T., and Zilberman, D. (2014). Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. Cell *156*, 1286–1297. https://doi.org/10.1016/j.cell.2014.01.029.
7. Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat. Rev. Genet. *11*, 204–220. https://doi.org/10.1038/nrg2719.
8. Petryk, N., Bultmann, S., Bartke, T., and Defossez, P.A. (2021). Staying true to yourself: mechanisms of DNA methylation maintenance in mammals. Nucleic Acids Res. *49*, 3020–3032. https://doi.org/10.1093/nar/gkaa1154.
9. Tirot, L., Jullien, P.E., and Ingouff, M. (2021). Evolution of CG methylation maintenance machinery in plants. Epigenomes *5*, 19. https://doi.org/10.3390/epigenomes5030019.
10. Dumesic, P.A., Stoddard, C.I., Catania, S., Narlikar, G.J., and Madhani, H.D. (2020). ATP hydrolysis by the SNF2 domain of Dnmt5 is coupled to both specific recognition and modification of hemimethylated DNA. Mol. Cell *79*, 127–139.e4. https://doi.org/10.1016/j.molcel.2020.04.029.
11. Hashimoto, H., Horton, J.R., Zhang, X., Bostick, M., Jacobsen, S.E., and Cheng, X. (2008). The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. Nature *455*, 826–829. https://doi.org/10.1038/nature07280.
12. Wang, J., Catania, S., Wang, C., de la Cruz, M.J., Rao, B., Madhani, H.D., and Patel, D.J. (2022). Structural insights into DNMT5-mediated ATP-dependent high-fidelity epigenome maintenance. Mol. Cell *82*, 1186–1198.e6. https://doi.org/10.1016/j.molcel.2022.01.028.
13. Haggerty, C., Kretzmer, H., Riemenschneider, C., Kumar, A.S., Mattei, A.L., Bailly, N., Gottfreund, J., Giesselmann, P., Weigert, R., Brändl, B., et al. (2021). Dnmt1 has de novo activity targeted to transposable elements. Nat. Struct. Mol. Biol. *28*, 594–603. https://doi.org/10.1038/s41594-021-00603-8.
14. Wei, A., and Wu, H. (2022). Mammalian DNA methylome dynamics: mechanisms, functions and new frontiers. Development *149*, dev182683. https://doi.org/10.1242/dev.182683.
15. Zhang, H., Gong, Z., and Zhu, J.K. (2022). Active DNA demethylation in plants: 20 years of discovery and beyond. J. Integr. Plant Biol. *64*, 2217–2239. https://doi.org/10.1111/jipb.13423.
16. Fitz-James, M.H., and Cavalli, G. (2022). Molecular mechanisms of transgenerational epigenetic inheritance. Nat. Rev. Genet. *23*, 325–341. https://doi.org/10.1038/s41576-021-00438-5.
17. van der Graaf, A., Wardenaar, R., Neumann, D.A., Taudt, A., Shaw, R.G., Jansen, R.C., Schmitz, R.J., Colomé-Tatché, M., and Johannes, F. (2015). Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. Proc. Natl. Acad. Sci. USA *112*, 6676–6681. https://doi.org/10.1073/pnas.1424254112.
18. Vidalis, A., Živković, D., Wardenaar, R., Roquis, D., Tellier, A., and Johannes, F. (2016). Methylome evolution in plants. Genome Biol. *17*, 264. https://doi.org/10.1186/s13059-016-1127-5.
19. Kawakatsu, T., Huang, S.C., Jupe, F., Sasaki, E., Schmitz, R.J., Urich, M.A., Castanon, R., Nery, J.R., Barragan, C., He, Y., et al. (2016). Epigenomic diversity in a global collection of Arabidopsis thaliana accessions. Cell *166*, 492–505. https://doi.org/10.1016/j.cell.2016.06.044.
20. Schmitz, R.J., Schultz, M.D., Urich, M.A., Nery, J.R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R.B., Chen, H., Schork, N.J., et al. (2013). Patterns of population epigenomic diversity. Nature *495*, 193–198. https://doi.org/10.1038/nature11968.
21. Shahzad, Z., Moore, J.D., Choi, J., and Zilberman, D. (2021). Epigenetic inheritance mediates phenotypic diversity in natural populations. Preprint at bioRxiv. https://doi.org/10.1101/2021.03.15.435374.
22. Baduel, P., and Colot, V. (2021). The epiallelic potential of transposable elements and its evolutionary significance in plants. Philos. Trans. R. Soc. Lond. B Biol. Sci. *376*, 20200123. https://doi.org/10.1098/rstb.2020.0123.
23. Haerter, J.O., Lövkvist, C., Dodd, I.B., and Sneppen, K. (2014). Collaboration between CpG sites is needed for stable somatic inheritance

of DNA methylation states. Nucleic Acids Res. *42*, 2235–2244. https://doi.org/10.1093/nar/gkt1235.

24. Lövkvist, C., Dodd, I.B., Sneppen, K., and Haerter, J.O. (2016). DNA methylation in human epigenomes depends on local topology of CpG sites. Nucleic Acids Res. *44*, 5123–5132. https://doi.org/10.1093/nar/gkw124.
25. Sontag, L.B., Lorincz, M.C., and Georg Luebeck, E. (2006). Dynamics, stability and inheritance of somatic DNA methylation imprints. J. Theor. Biol. *242*, 890–899. https://doi.org/10.1016/j.jtbi.2006.05.012.
26. Wang, Q., Yu, G., Ming, X., Xia, W., Xu, X., Zhang, Y., Zhang, W., Li, Y., Huang, C., Xie, H., et al. (2020). Imprecise DNMT1 activity coupled with neighbor-guided correction enables robust yet flexible epigenetic inheritance. Nat. Genet. *52*, 828–839. https://doi.org/10.1038/s41588-020-0661-y.
27. Chodavarapu, R.K., Feng, S., Bernatavichute, Y.V., Chen, P.Y., Stroud, H., Yu, Y., Hetzel, J.A., Kuo, F., Kim, J., Cokus, S.J., et al. (2010). Relationship between nucleosome positioning and DNA methylation. Nature *466*, 388–392. https://doi.org/10.1038/nature09147.
28. Feng, S., Cokus, S.J., Zhang, X., Chen, P.Y., Bostick, M., Goll, M.G., Hetzel, J., Jain, J., Strauss, S.H., Halpern, M.E., et al. (2010). Conservation and divergence of methylation patterning in plants and animals. Proc. Natl. Acad. Sci. USA *107*, 8689–8694. https://doi.org/10.1073/pnas.1002720107.
29. Lewis, S.H., Ross, L., Bain, S.A., Pahita, E., Smith, S.A., Cordaux, R., Miska, E.A., Lenhard, B., Jiggins, F.M., and Sarkies, P. (2020). Widespread conservation and lineage-specific diversification of genome-wide DNA methylation patterns across arthropods. PLoS Genet. *16*, e1008864. https://doi.org/10.1371/journal.pgen.1008864.
30. Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science *328*, 916–919. https://doi.org/10.1126/science.1186366.
31. Choi, J., Lyons, D.B., Kim, M., Moore, J.D., and Zilberman, D. (2020). DNA methylation and histone H1 jointly repress transposable elements and aberrant intragenic transcripts. Mol. Cell *77*, 310–323.e7. https://doi.org/10.1016/j.molcel.2019.10.011.
32. Takuno, S., and Gaut, B.S. (2012). Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly. Mol. Biol. Evol. *29*, 219–227. https://doi.org/10.1093/molbev/msr188.
33. Zemach, A., and Zilberman, D. (2010). Evolution of eukaryotic DNA methylation and the pursuit of safer sex. Curr. Biol. *20*, R780–R785. https://doi.org/10.1016/j.cub.2010.07.007.
34. Lyons, D.B., Briffa, A., He, S., Choi, J., Hollwey, E., Colicchio, J., Anderson, I., Feng, X., Howard, M., and Zilberman, D. (2023). Extensive de novo activity stabilizes epigenetic inheritance of CG methylation in Arabidopsis transposons. Cell Rep. *42*, 112132. https://doi.org/10.1016/j.celrep.2023.112132.
35. Muyle, A.M., Seymour, D.K., Lv, Y., Huettel, B., and Gaut, B.S. (2022). Gene body methylation in plants: mechanisms, functions, and important implications for understanding evolutionary processes. Genome Biol. Evol. *14*, evac038. https://doi.org/10.1093/gbe/evac038.
36. Bewick, A.J., Ji, L., Niederhuth, C.E., Willing, E.M., Hofmeister, B.T., Shi, X., Wang, L., Lu, Z., Rohr, N.A., Hartwig, B., et al. (2016). On the origin and evolutionary consequences of gene body DNA methylation. Proc. Natl. Acad. Sci. USA *113*, 9111–9116. https://doi.org/10.1073/pnas.1604666113.
37. Wendte, J.M., Zhang, Y., Ji, L., Shi, X., Hazarika, R.R., Shahryary, Y., Johannes, F., and Schmitz, R.J. (2019). Epimutations are associated with CHROMOMETHYLASE 3-induced de novo DNA methylation. eLife *8*, e47891. https://doi.org/10.7554/eLife.47891.
38. Niederhuth, C.E., Bewick, A.J., Ji, L., Alabady, M.S., Kim, K.D., Li, Q., Rohr, N.A., Rambani, A., Burke, J.M., Udall, J.A., et al. (2016). Widespread natural variation of DNA methylation within angiosperms. Genome Biol. *17*, 194. https://doi.org/10.1186/s13059-016-1059-0.
39. Zhang, Y., Wendte, J.M., Ji, L., and Schmitz, R.J. (2020). Natural variation in DNA methylation homeostasis and the emergence of epialleles. Proc. Natl. Acad. Sci. USA *117*, 4874–4884. https://doi.org/10.1073/pnas.1918172117.
40. He, L., Huang, H., Bradai, M., Zhao, C., You, Y., Ma, J., Zhao, L., Lozano-Durán, R., and Zhu, J.K. (2022). DNA methylation-free Arabidopsis reveals crucial roles of DNA methylation in regulating gene expression and development. Nat. Commun. *13*, 1335. https://doi.org/10.1038/s41467-022-28940-2.
41. Tirot, L., Bonnet, D.M.V., and Jullien, P.E. (2022). DNA methyltransferase 3 (MET3) is regulated by Polycomb group complex during Arabidopsis endosperm development. Plant Reprod. *35*, 141–151. https://doi.org/10.1007/s00497-021-00436-x.
42. Quadrana, L., Bortolini Silveira, A., Mayhew, G.F., LeBlanc, C., Martienssen, R.A., Jeddeloh, J.A., and Colot, V. (2016). The Arabidopsis thaliana mobilome and its impact at the species level. eLife *5*, e15716. https://doi.org/10.7554/eLife.15716.
43. Jullien, P.E., Susaki, D., Yelagandula, R., Higashiyama, T., and Berger, F. (2012). DNA methylation dynamics during sexual reproduction in Arabidopsis thaliana. Curr. Biol. *22*, 1825–1830. https://doi.org/10.1016/j.cub.2012.07.061.
44. Watson, J.M., Platzer, A., Kazda, A., Akimcheva, S., Valuchova, S., Nizhynska, V., Nordborg, M., and Riha, K. (2016). Germline replications and somatic mutation accumulation are independent of vegetative life span in Arabidopsis. Proc. Natl. Acad. Sci. USA *113*, 12226–12231. https://doi.org/10.1073/pnas.1609686113.
45. Schmitz, R.J., Schultz, M.D., Lewsey, M.G., O'Malley, R.C., Urich, M.A., Libiger, O., Schork, N.J., and Ecker, J.R. (2011). Transgenerational epigenetic instability is a source of novel methylation variants. Science *334*, 369–373. https://doi.org/10.1126/science.1212959.
46. Hazarika, R.R., Serra, M., Zhang, Z., Zhang, Y., Schmitz, R.J., and Johannes, F. (2022). Molecular properties of epimutation hotspots. Nat. Plants *8*, 146–156. https://doi.org/10.1038/s41477-021-01086-7.
47. Du, J., Zhong, X., Bernatavichute, Y.V., Stroud, H., Feng, S., Caro, E., Vashisht, A.A., Terragni, J., Chin, H.G., Tu, A., et al. (2012). Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. Cell *151*, 167–180. https://doi.org/10.1016/j.cell.2012.07.034.
48. Du, J., Johnson, L.M., Jacobsen, S.E., and Patel, D.J. (2015). DNA methylation pathways and their crosstalk with histone methylation. Nat. Rev. Mol. Cell Biol. *16*, 519–532. https://doi.org/10.1038/nrm4043.
49. Rajakumara, E., Law, J.A., Simanshu, D.K., Voigt, P., Johnson, L.M., Reinberg, D., Patel, D.J., and Jacobsen, S.E. (2011). A dual flip-out mechanism for 5mC recognition by the Arabidopsis SUVH5 SRA domain and its impact on DNA methylation and H3K9 dimethylation in vivo. Genes Dev. *25*, 137–152. https://doi.org/10.1101/gad.1980311.
50. Dubin, M.J., Zhang, P., Meng, D., Remigereau, M.S., Osborne, E.J., Paolo Casale, F., Drewe, P., Kahles, A., Jean, G., Vilhjálmsson, B., et al. (2015). DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. eLife *4*, e05255. https://doi.org/10.7554/eLife.05255.
51. Pignatta, D., Erdmann, R.M., Scheer, E., Picard, C.L., Bell, G.W., and Gehring, M. (2014). Natural epigenetic polymorphisms lead to intraspecific variation in Arabidopsis gene imprinting. eLife *3*, e03198. https://doi.org/10.7554/eLife.03198.
52. Muyle, A., Ross-Ibarra, J., Seymour, D.K., and Gaut, B.S. (2021). Gene body methylation is under selection in Arabidopsis thaliana. Genetics *218*. https://doi.org/10.1093/genetics/iyab061.
53. Huettel, B., Kanno, T., Daxinger, L., Aufsatz, W., Matzke, A.J.M., and Matzke, M. (2006). Endogenous targets of RNA-directed DNA methylation and Pol IV in Arabidopsis. EMBO J. *25*, 2828–2836. https://doi.org/10.1038/sj.emboj.7601150.
54. Penterman, J., Uzawa, R., and Fischer, R.L. (2007). Genetic interactions between DNA demethylation and methylation in Arabidopsis. Plant Physiol. *145*, 1549–1557. https://doi.org/10.1104/pp.107.107730.

55. Coleman-Derr, D., and Zilberman, D. (2012). Deposition of histone variant H2A.Z within gene bodies regulates responsive genes. PLoS Genet. *8*, e1002988. https://doi.org/10.1371/journal.pgen.1002988.

56. Conerly, M.L., Teves, S.S., Diolaiti, D., Ulrich, M., Eisenman, R.N., and Henikoff, S. (2010). Changes in H2A.Z occupancy and DNA methylation during B-cell lymphomagenesis. Genome Res. *20*, 1383–1390. https://doi.org/10.1101/gr.106542.110.

57. Edwards, J.R., O'Donnell, A.H., Rollins, R.A., Peckham, H.E., Lee, C., Milekic, M.H., Chanrion, B., Fu, Y., Su, T., Hibshoosh, H., et al. (2010). Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. Genome Res. *20*, 972–980. https://doi.org/10.1101/gr.101535.109.

58. Murphy, P.J., Wu, S.F., James, C.R., Wike, C.L., and Cairns, B.R. (2018). Placeholder nucleosomes underlie germline-to-embryo DNA methylation reprogramming. Cell *172*, 993–1006.e13. https://doi.org/10.1016/j.cell.2018.01.022.

59. To, T.K., Nishizawa, Y., Inagaki, S., Tarutani, Y., Tominaga, S., Toyoda, A., Fujiyama, A., Berger, F., and Kakutani, T. (2020). RNA interference-independent reprogramming of DNA methylation in Arabidopsis. Nat. Plants *6*, 1455–1467. https://doi.org/10.1038/s41477-020-00810−z.

60. Zilberman, D., Coleman-Derr, D., Ballinger, T., and Henikoff, S. (2008). Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. Nature *456*, 125–129. https://doi.org/10.1038/nature07324.

61. De Riso, G., Fiorillo, D.F.G., Fierro, A., Cuomo, M., Chiariotti, L., Miele, G., and Cocozza, S. (2020). Modeling DNA methylation profiles through a dynamic equilibrium between methylation and demethylation. Biomolecules *10*, 1271. https://doi.org/10.3390/biom10091271.

62. Becker, C., Hagmann, J., Müller, J., Koenig, D., Stegle, O., Borgwardt, K., and Weigel, D. (2011). Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature *480*, 245–249. https://doi.org/10.1038/nature10555.

63. Goeldel, C., and Johannes, F. (2023). Stochasticity in gene body methylation. Curr. Opin. Plant Biol. *75*, 102436. https://doi.org/10.1016/j.pbi.2023.102436.

64. Pisupati, R., Nizhynska, V., Mollá Morales, A., and Nordborg, M. (2023). On the causes of gene-body methylation variation in Arabidopsis thaliana. PLoS Genet. *19*, e1010728. https://doi.org/10.1371/journal.pgen.1010728.

65. Lyons, D.B., and Zilberman, D. (2017). DDM1 and Lsh remodelers allow methylation of DNA wrapped in nucleosomes. Elife *6*, e30674. https://doi.org/10.7554/eLife.30674.

66. Tillo, D., and Hughes, T.R. (2009). G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics *10*, 442. https://doi.org/10.1186/1471-2105-10−442.

67. Zhang, T., Zhang, W., and Jiang, J. (2015). Genome-wide nucleosome occupancy and positioning and their impact on gene expression and evolution in plants. Plant Physiol. *168*, 1406–1416. https://doi.org/10.1104/pp.15.00125.

68. Yi, H., Sardesai, N., Fujinuma, T., Chan, C.W., Veena, and Gelvin, S.B. (2006). Constitutive expression exposes functional redundancy between the Arabidopsis histone H2A gene HTA1 and other H2A gene family members. Plant Cell *18*, 1575–1589. https://doi.org/10.1105/tpc.105.039719.

69. Zheng, X., Pontes, O., Zhu, J., Miki, D., Zhang, F., Li, W.X., Iida, K., Kapoor, A., Pikaard, C.S., and Zhu, J.K. (2008). ROS3 is an RNA-binding protein required for DNA demethylation in Arabidopsis. Nature *455*, 1259–1262. https://doi.org/10.1038/nature07305.

70. Bewick, A.J., and Schmitz, R.J. (2017). Gene body DNA methylation in plants. Curr. Opin. Plant Biol. *36*, 103–110. https://doi.org/10.1016/j.pbi.2016.12.007.

71. Xia, J., Han, L., and Zhao, Z. (2012). Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome. BMC Genomics *13* (*Suppl 8*), S7. https://doi.org/10.1186/1471-2164-13-S8-S7.

72. Zagkos, L., Auley, M.M., Roberts, J., and Kavallaris, N.I. (2019). Mathematical models of DNA methylation dynamics: implications for health and ageing. J. Theor. Biol. *462*, 184–193. https://doi.org/10.1016/j.jtbi.2018.11.006.

73. Ginno, P.A., Gaidatzis, D., Feldmann, A., Hoerner, L., Imanci, D., Burger, L., Zilbermann, F., Peters, A.H.F.M., Edenhofer, F., Smallwood, S.A., et al. (2020). A genome-scale map of DNA methylation turnover identifies site-specific dependencies of DNMT and TET activity. Nat. Commun. *11*, 2680. https://doi.org/10.1038/s41467-020-16354-x.

74. Exposito-Alonso, M., Booker, T.R., Czech, L., Gillespie, L., Hateley, S., Kyriazis, C.C., Lang, P.L.M., Leventhal, L., Nogues-Bravo, D., Pagowski, V., et al. (2022). Genetic diversity loss in the Anthropocene. Science *377*, 1431–1435. https://doi.org/10.1126/science.abn5642.

75. Shahryary, Y., Symeonidi, A., Hazarika, R.R., Denkena, J., Mubeen, T., Hofmeister, B., van Gurp, T., Colomé-Tatché, M., Verhoeven, K.J.F., Tuskan, G., et al. (2020). AlphaBeta: computational inference of epimutation rates and spectra from high-throughput DNA methylation data in plants. Genome Biol. *21*, 260. https://doi.org/10.1186/s13059-020-02161-6.

76. Chan, S.W.-L., Henderson, I.R., Zhang, X., Shah, G., Chien, J.S.-C., and Jacobsen, S.E. (2006). RNAi, DRD1, and histone methylation actively target developmentally important non-CG DNA methylation in Arabidopsis. PLoS Genet. *2*, e83. https://doi.org/10.1371/journal.pgen.0020083.

77. Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics *10*, 232. https://doi.org/10.1186/1471-2105-10−232.

78. Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. *40*, D1202–D1210. https://doi.org/10.1093/nar/gkr1090.

79. 1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at; 1001 Genomes Consortium (2016). 1,135 Genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. Cell *166*, 481–491. https://doi.org/10.1016/j.cell.2016.05.063.

80. Cheng, C.Y., Krishnakumar, V., Chan, A.P., Thibaud-Nissen, F., Schobel, S., and Town, C.D. (2017). Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. Plant J. *89*, 789–804. https://doi.org/10.1111/tpj.13415.

81. Taudt, A., Roquis, D., Vidalis, A., Wardenaar, R., Johannes, F., and Colomé-Tatché, M. (2018). METHimpute: imputation-guided construction of complete methylomes from WGBS data. BMC Genomics *19*, 444. https://doi.org/10.1186/s12864-018-4641-x.

82. Seren, Ü., Vilhjálmsson, B.J., Horton, M.W., Meng, D., Forai, P., Huang, Y.S., Long, Q., Segura, V., and Nordborg, M. (2012). GWAPP: a web application for genome-wide association mapping in Arabidopsis. Plant Cell *24*, 4793–4805. https://doi.org/10.1105/tpc.112.108068.

83. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B Methodol. *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

84. Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. *28*, 2731–2739. https://doi.org/10.1093/molbev/msr121.

85. Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. *81*, 2340–2361. https://doi.org/10.1021/j100540a008.

86. Song, J., Teplova, M., Ishibe-Murakami, S., and Patel, D.J. (2012). Structure-based mechanistic insights into DNMT1-mediated maintenance DNA methylation. Science *335*, 709–712. https://doi.org/10.1126/science.1214453.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| **Critical commercial assays** | | |
| DNeasy plant mini kit | Qiagen | 69104 |
| Ultra II DNA Library prep kit | NEB | E7645L |
| EZ DNA Methylation Lightning Kit | Zymo | D5046 |
| Methylated NEBNext Multiplex Oligos | NEB | E7535L |
| **Deposited data** | | |
| Raw and analyzed BS-seq data | This paper | GEO: GSE204837 |
| H2A.Z ChIP-seq data | Coleman-Derr and Zilberman[55] | GEO: GSE39045 |
| 30 generation MAL BS-seq data | Schmitz et al.[45] | SRA: SRA035939 |
| 30 generation MAL BS-seq data | Becker et al.[62] | ENA: PRJEB2678 |
| Generational bs-seq data | Hazarika et al.[46] | GEO: GSE178684 |
| Generational bs-seq data | Shahryary et al.[75] | GEO: GSE64463 |
| 1001 Epigenomes bs-seq data | Kawakatsu et al.[19] | GEO: GSE43857 |
| RNA sequencing data | Kawakatsu et al.[19] | GEO: GSE80744 |
| Nucleosome positions | Lyons and Zilberman[65] | GEO: GSE96994 |
| **Experimental models: Organisms/strains** | | |
| cmt2-3 | NASC | NASC ID: N683827 |
| cmt2-4 | NASC | NASC ID: N689216 |
| cmt3-11 | NASC | NASC ID: N16392 |
| drm1-2drm2-2 | Chan et al.[76] | N/A |
| drm1-2drm2-2cmt3-11 | Chan et al.[76] | N/A |
| hta8-1hta9-1hta11-1 | Coleman-Derr and Zilberman[55] | N/A |
| ros1-3 | Penterman et al.[54] | N/A |
| **Software and algorithms** | | |
| Stochastic modelling code | This paper | https://doi.org/10.5281/zenodo.8332883 |
| BSMAP 2.90 | Xi and Li[77] | N/A |
| Bisulfite alignment pipeline | Lyons and Zilberman[65] | N/A |
| R | https://www.r-project.org/ | N/A |
| SeqMonk | https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/ | N/A |

### RESOURCE AVAILABILITY

#### Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contacts, experimental, bisulfite analysis and population genetics: Daniel Zilberman (daniel.zilberman@ist.ac.at).

#### Materials availability

This study did not generate new materials.

#### Data and code availability

All newly generated bisulfite-seq data are available in GEO under accession GSE204837. Additionally, this paper analyzes existing, publicly available data (see key resources table above).

All original stochastic modelling code is available on GitHub: https://github.com/BriffaAKR/gbM_modelling.git, and is publicly available at Zenodo: 10.5281/zenodo.8332883

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

## EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

### Arabidopsis thaliana: Biological materials and growth conditions

Arabidopsis thaliana plants were sown on soil, stratified for 2 days at 4°C and grown in a growth chamber under long day conditions (16 hours light, 8 hours dark), using the generational pattern described in Figure S1A. Two replicate leaves from the same plant at G0 were tested in each genotype, to confirm that differences in sequencing and bisulfite conversion did not produce errors. For WT (Col-0), *cmt2* (*cmt2-3*), *cmt3* (*cmt3-11*) and *drm1drm2* (*drm1-2drm2-2*), plants were grown for six total generations (G0-G5), in two separate branching lineages, giving a total of 20 individual triplet comparisons. For *cmt2cmt3* (*cmt2-4cmt3-11*), *h2az* (*hta8-1hta9-1hta11-1*) and *ros1* (*ros1-3*) genotypes, plants were grown for 4 total generations (G0-G3), again in two separate branching lineages (giving a total of 12 individual comparisons). For *ddcc* plants (*drm1-2drm2-2cmt2-4cmt3-11*), three separate G0 plants were grown for four generations (Lines A-C), each in two branching lineages, resulting in a total of 36 individual comparisons. *cmt2cmt3* and *ddcc* lines were generated by crossing *cmt2-4* (SALK_201637) and *cmt3-11* (CS16392) lines, and *cmt2-4* and *ddc*[76] lines respectively and selfing to obtain homozygous progeny.

## METHOD DETAILS

### Leaf genomic DNA isolation, library preparation, bisulfite conversion, and sequencing

Genomic DNA (gDNA) was extracted from 1-month-old *Arabidopsis thaliana* rosette leaves with the DNeasy plant mini kit (Qiagen, cat. no. 69104) per the manufacturer's instructions. Libraries were prepared from roughly 500 ng of purified gDNA that was sheared to approximately 400 bp on a Diagenode Bioruptor Pico water bath sonicator. Libraries were produced with the Ultra II DNA Library prep kit according to manufacturer's instructions (New England Biolabs, cat. no. E7645L). Bisulfite conversion of DNA was carried out according to manufacturer's protocol (Zymo, EZ DNA Methylation Lightning Kit, cat. no. D5046). DNA was converted twice to ensure complete bisulfite conversion of unmethylated cytosine. NEBNext Multiplex Oligos with methylated adaptors (cat. no. E7535L) were used for generating multiplexed libraries during PCR amplification of libraries. Sequencing was carried out as single-end 75 bp reads on Illumina NextSeq 550 at the John Innes Centre.

### Sequence Alignments and Segmentation

RNA sequencing data for 625 Arabidopsis accessions was retrieved from GEO: GSE80744.[19] Bisulfite sequence reads were accessed for the mutation accumulation lines (MAL),[45,62] 1001 methylomes,[19] and Hazarika[46] experiments from the Sequence Read Archive (SRA). In-house, Hazarika and published MAL sequence reads were aligned to the Arabidopsis TAIR10 genome reference sequence,[78] using an in-house alignment pipeline as previously described.[65] 1001 Methylomes sequence reads were aligned using BSMAP 2.90[77] and known SNPs and indels[79] were masked. Genes and transposons were annotated using the Araport11 annotation.[80] Methylomes were segmented into Unmethylated, gbM and TE-like methylated segments as previously described[31] (Table S4B). For the Col-0 data, a segmentation model was created by segmenting the combined reads from all of the Generation 3 samples in data from Schmitz et al.[45] Genes were defined as gbM genes if they both contained a gbM segment and any TE-like segment was both less than a quarter the length of the gbM segment and smaller than 3 CG sites (n=14,581, Table S4C). Genes were defined as unmethylated genes if they contained no gbM segment or TE-like segment and did not overlap with any gene that did (n=12,045).

### Methylation Calling

Methylation status of individual CG sites was called by comparing the counts of aligned reads indicating methylated and unmethylated status at the site. Fisher's Exact test ($p<0.005$) was used to determine whether there was sufficient read coverage at the site to distinguish the site from a fully unmethylated site with an error rate similar to the methylation rate observed in the chloroplast of the sample in question (as an estimate of bisulfite conversion inefficiency), or from a fully methylated site with a similar error rate. For sites where these tests indicated coverage was sufficient, a binomial test ($p<0.05$) was used to identify sites with significantly more methylated reads than expected from an unmethylated site. These sites were considered to be methylated. Sites which did not have significantly more methylated reads than an unmethylated site but did have sufficient reads to pass the Fisher's exact test were considered to be unmethylated.

### Calculation of Gain and Loss Rates

For the MAL, sites with significantly more methylated reads than would be expected for an unmethylated site, but with less than 45% reads methylated, were classified as partially methylated, generally treated as missing data and assumed to consist of somatic changes and heterozygously methylated sites. A consensus methylation call was made for each site based on concordance between methylation calls for the two sibling replicates representing each line. A 'parental consensus' methylation state was calculated for each CG site by taking the majority methylation state among those generation 3 lines where sibling replicates agreed. For MAL, methylation state changes over 30 generations were identified by comparing parental consensus states with the states in individual lines where sibling replicates agreed, and masking known DMRs[45] so that we could focus on spontaneous changes at individual sites. One MA line from the Becker et al. dataset (Line 79) was an outlier in comparison to other MAL results and was therefore excluded. Over the course of 30 generations, most changes will have segregated out into the homozygously methylated or

homozygously unmethylated state (see below). This, and the exclusion of sites that are heterozygously methylated at the beginning of the experiment, using the partial methylation cutoff and the 'parental consensus' means that MAL gains reflect those that change over the course of 30 generations from homozygously unmethylated to homozygously methylated and losses reflect those that change over thirty generation from homozygously methylated to homozygously unmethylated. The number of gains and losses was then divided by 30 to produce a per generation rate.

For our bisulfite sequencing data, sites with significantly more methylated reads than would be expected for an unmethylated site, but with less than 25% reads methylated, were classified as partially methylated, and generally treated as missing data, again to exclude somatic methylation gain. Changes over sets of three generations were identified, using a methylation level cutoff of 70% in the parent (for losses) and offspring (for gains) to exclude segregating heterozygotes. This more severe methylation cutoff (70%) was required due to the increased noise in data sampled at individual generations, in comparison to the MAL data sampled across 30 generations. To reduce noise, and further exclude segregating heterozygously methylated sites, only sites where parental and offspring statuses were maintained in the six closest relatives were considered. The result of this is that gains reflect those which change over one generation from homozygously unmethylated to homozygously methylated and losses reflect those which change over one generation from homozygously methylated to homozygously unmethylated. This method will not capture all changes, but will underestimate the rate by a factor of two. To illustrate, a gain represents a site that begins as homozygously unmethylated (U,U) in Generation 0. Occurrence of a gain will result in the site being heterozygously methylated (U,M) at Generation 1 (we assume that changes happen on only one chromosome due to the low rate of epimutation). Mendelian segregation predicts three potential outcomes in Generation 2: 25% of all (U,M) sites will return to being (U,U), and will not result in a gain in methylation. 25% of all (U,M) sites will become (M,M), fixing the gain in the population. Using our method, only these changes will be detected. The final 50% of (U,M) sites will remain (U,M) in Generation 2. These gains will not be detected, as their methylation level in Generation 2 is not expected to be above 70%, the cutoff used here. Half of these (U,M) sites will eventually, over multiple generations, become fixed as (M,M) and therefore result in methylation gain. However, our method does not detect these gains as they are not yet fixed in the population, and we therefore multiply by two the number of gains that we do detect to include these sites. The branching structure of our lineage means that multiple comparisons can be drawn for each genotype using different plants (N=12-36, Figure S1A). For example, comparisons in WT consist of the following (where 'Gen' stands for generation, 'Rep' for replicate, and 'sib' for sibling – the plant that does not contribute to the next generation):

1. Gen0Rep1-Gen1LineA-Gen2LineA
2. Gen0Rep1-Gen1LineA-Gen2LineAsib
3. Gen0Rep1-Gen1LineB-Gen2LineB
4. Gen0Rep1-Gen1LineB-Gen2LineBsib
5. Gen0Rep2-Gen1LineA-Gen2LineA
6. Gen0Rep2-Gen1LineA-Gen2LineAsib
7. Gen0Rep2-Gen1LineB-Gen2LineB
8. Gen0Rep2-Gen1LineB-Gen2LineBsib
9. Gen1LineA-Gen2LineA-Gen3LineA
10. Gen1LineA-Gen2LineA-Gen3LineAsib
11. Gen2LineA-Gen3LineA-Gen4LineA
12. Gen2LineA-Gen3LineA-Gen4LineAsib
13. Gen3LineA-Gen4LineA-Gen5LineA
14. Gen3LineA-Gen4LineA-Gen5LineAsib
15. Gen1LineB-Gen2LineB-Gen3LineB
16. Gen1LineB-Gen2LineB-Gen3LineBsib
17. Gen2LineB-Gen3LineB-Gen4LineB
18. Gen2LineB-Gen3LineB-Gen4LineBsib
19. Gen3LineB-Gen4LineB-Gen5LineB
20. Gen3LineB-Gen4LineB-Gen5LineBsib

Individual comparisons between plants with rates that were strong outliers (values +/- 1.5 S.D. from the average) were excluded.

For all methods, rates of change per site were estimated by dividing inferred gains/losses by the number of *U* calls/*M* calls in the previous generation. The resulting rates per generation were converted to per cell cycle by a fixed estimate of 34 germ cell cycles per generation.[44]

## Population gbM frequency

Population gbM frequency of each gene was obtained from Shahzad et al.[21] and were calculated as described there. Briefly, the population gbM frequency represents the number of accessions having gbM at a given gene as a percentage of the total available calls of gbM, teM or unmethylated across 948 *Arabidopsis* accessions. In order to investigate the frequency of methylation gains in different groups of UM genes, we subdivided Col-0 UM genes based on their population gbM frequency. We first excluded genes that contained TE-like methylation in greater than 1% of accessions and then subdivided the remaining UM genes into rarely methylated

(RMGs; gbM in <20% of accessions, n=6113), occasionally methylated (OMGs; gbM in >20% and <50% of accessions, n=1500) and frequently methylated genes (FMGs; gbM in >50% of accessions, n=662)

### Epimutation gain profiles within polymorphic loci

Substantial gbM polymorphism exists within various epimutation datasets. This allowed us to analyze gain rates with respect to distance to the nearest mCG and compare these to gain rates in relation to the same sites when they are unmethylated (and exist within genes that lack gbM). Genes were identified which were unmethylated (UM) in the published data (MAL[45]) and gene body methylated (gbM) within at least one dataset of our newly generated data (n=2766; Figure S2D, left). On the right, we show the inverse analysis: genes were identified which were gbM in the MAL data and unmethylated within at least one dataset of our newly generated data (n=1615; Figure S2D, right). Analogously to Figure 2C, the gain rate distribution for each sample is plotted with respect to the position of the nearest M site of the gbM dataset (e.g., the MAL dataset for the panel on the right).

### Choice of accessions for modelling

Col-like accessions were defined as accessions where the proportion of called sites called as M, in segments annotated as gbM in Col-0 (Schmitz parental consensus data set), is within 1 SD of the mean among all samples (N=891). CG sites where SNPs resulted in the loss of a CG site with respect to Col-0 were removed from the model and data. The model was run with identical initial conditions, with the variation between accessions arising from stochasticity. Accessions' global mean gbM was calculated by identifying the region of broadly gbM-methylatable genome space, defined as regions that are covered by gbM segments (defined as described above in the section sequence alignments and segmentation) in at least 5 *Arabidopsis* accessions, and counting CG sites in that space called as methylated in the accession by binomial test, as a proportion of sites called as methylated or unmethylated. Individual accessions' global mean gbM was calculated excluding any portion of the broadly gbM-methylatable space that is overlapped by a teM segment in the accession in question.

To fit the model, accessions with more than 60% of CG sites called as either methylated or unmethylated (adequate coverage) were divided into groups based on their global mean gbM. The 10 most hypo- and hyper-gbM accessions (~1%) were identified as potential outliers and removed from the broader group. The remainder of accessions were divided into deciles by global mean gbM. The 10 accessions with highest proportion of CG sites called as methylated or unmethylated (best coverage) in each of the three deciles closest to Col-0 (Deciles 3, 4 and 5) were chosen to determine model parameters (N=30).

For a wider comparison of the model, simulations were compared to all unique, high-coverage Col-like accessions. Duplicate samples of individual accessions were removed (N=798). Accessions with less than 50% of CG sites called as either methylated or unmethylated were considered to have inadequate coverage and were removed, resulting in a final accession set for modelling of N=740.

To analyze accessions with non-Col-like levels of methylation, two groups were chosen: Northern Swedish accessions (NS), which have elevated gbM, and Relict (RL) accessions, which have unusually low levels of gbM. Dör-10, a Northern Swedish accession which has the highest known level of gbM was analyzed separately, as were two Relict accessions (Can-0 and Cvi-0), which have extremely low levels of gbM, and a third accession (UKID116), which has similarly low levels of gbM.

### Methylatable Gene Regions

Genes for modelling were selected from the Araport 11 annotation, annotated as protein coding – (n=27,473). Genes that contained TE-like methylation in greater than 1% of accessions were discarded to retain non-TE genes only (n=19,082). The segments chosen for modelling comprise the region in each gene between the first and last CG sites overlapped by gbM segments spanning at least 3 CG sites, in at least 5% of the accessions from the 891 which constitute the Col-0 like data set, resulting in the regions that were broadly methylatable (n=13,138). These segments were further trimmed to remove the ends of segments where H2A.Z ChIP-Seq signal, smoothed over 5 adjacent 50bp bins, was above 1.2. H2A.Z ChIP data was obtained from Coleman-Derr and Zilberman.[55] Segments were also removed if the mean H2A.Z ChIP-Seq signal along the remaining segment exceeded 1.2, resulting in the H2A.Z-low methylatable annotation (n=8,843). Our modelling method considers methylation levels in each locus in isolation of the surrounding DNA and therefore overlapping genes (where gbM in one gene may affect gbM in the other, overlapping gene) cannot be appropriately modelled. To prevent this, if segments overlapped by more than 20%, the smaller segment was discarded (n=416), as were any gene regions containing fewer than 5 CG sites (n= 447). This left us with 7980 loci in our final modelling dataset (Table S4A).

### Correlation of spatial patterns of methylation within simulated and experimental loci

We calculated the methylation percentage of each CG within a locus in the simulated data (30 iterations, modelled as in Figure 4A from the all-U initial state) and the experimental data (30 accessions). For each locus, we generated Pearson's linear correlation coefficients (R) describing how well the patterns of simulated and observed per-site mCG levels correlated across a single locus. These R values are included in the genome browser examples in Figure 5A. The distribution of these coefficients for all loci (n=7837) in our final analysis is shown in Figure 5B, with modelled genes for which the overall methylation level is poorly captured shown in black and well-modeled genes in blue, defined as in Figure 4E.

### Analysis of methylation patterns over well-positioned nucleosomes

The locations of well-positioned nucleosomes as defined in Lyons and Zilberman[65] were obtained. Methylation patterns in data, averaged across 30 Col-like accessions, were plotted over these nucleosomes and showed enrichment over the nucleosome, as has been previously described to occur as a core feature of gbM in plants and animals.[27,65] Simulated methylation patterns, generated only from the number and position of CG sites, using the all-U initial state and simulated for 100,000 generations were also plotted over these regions (black) and saw a similar enrichment over the nucleosome as the data, despite being generated with no information on nucleosome positioning.

### Analysis of sparsely methylated regions

Co-ordinates of sparsely methylated regions within gbM genes (SPMRs) and the gene IDs for the gbM genes concerned (Hazarika gbM genes) were downloaded from Hazarika et al.[46] Regions within the Hazarika gbM genes that were not SPMRs were defined as non-SPMRs. Non-SPMRs were bimodal with respect to mCG and so were partitioned into methylated (M_non_SPMRs) and unmethylated (U_non_SPMRs) (Figures S1B–S1D). 91% of M_non_SPMR CG sites and 92% of SPMR CG sites fall within our gbM segments. Epimutation rates in both published MAL data and our bisulfite sequencing data were calculated in SPMRs, M_non_SPMRs and U_non_SPMRs (Figure S1E; Table S1).

### Reanalysis of published methylation data

Bisulfite sequencing data was downloaded from Hazarika et al.[46] and methylation levels were calculated by averaging the percentage of C reads/total reads at each site over each genomic region. Samples are of highly variable coverage (4-83X). In the original analysis, in order to make single site methylation calls in all samples, methylation states of individual cytosines were imputed based on the methylation status of nearby cytosines.[81] This method may not be appropriate to make single site mCG calls in sparsely methylated genes, especially to subsequently identify rare epimutations within an otherwise unchanged methylation pattern of neighboring sites. We therefore excluded samples of coverage <10X (WT Line 1 Generation 11, *suvh4/5/6* Line 4 Generations 5 and 13, *suvh4/5/6* Line 8 Generation 8 and 9) and assigned methylation status without imputation in all remaining samples. Sites with significantly more methylated reads than would be expected for an unmethylated site (see *Methylation Calling*), but with less than 25% reads methylated, were classified as partially methylated, and generally treated as missing data. A 'parental consensus' methylation state was calculated for each CG site by identifying shared *M* or *U* calls in the three samples at the earliest generations for each genotype. Methylation state changes were identified by comparing parental consensus states with the states in individual samples and converted to per generation rates depending on the number of generations since the parental consensus. Rates were calculated in different genomic regions.

### Genome-wide association mapping

Three types of genome-wide association (GWA) analyses were performed to identify single nucleotide polymorphisms (SNPs) associated with the natural variation of gbM in *Arabidopsis* accessions. Genome-wide average gbM levels of accessions were used to identify the genetic factors influencing global gbM. Global gbM levels are influenced by sequencing coverage and thus analyses were carried out using multiple cutoffs for sequencing coverage.

Additionally, gbM levels of individual genes were used for GWA analysis to identify SNPs linked with local gbM variation. Furthermore, gbM frequency in rarely methylated genes (RMGs) was used for GWA analysis. RMGs are defined as genes with gbM in <20% of accessions and which have mean H2A.Z ChIP-seq signal <4 in Col-0 gene bodies. GWA mapping was performed using 1001 genomes SNP data[79] with an accelerated mixed model (AMM)[82] implemented in PyGWAS: Python library for running GWAS (version 1.7.4). SNPs with Minor Allele Frequency (MAF) >5% in the population were considered. 0.05 False Discovery Rate (FDR) correction[83] was implemented to account for multiple tests and identify SNPs associated with gbM variation.

### Haplotype analyses

ROS1 and ROS3 nucleotide sequences of *Arabidopsis* accessions were retrieved from http://signal.salk.edu/atg1001/3.0/gebrowser.php.[79] Sequences were translated with the EMBOSS Transeq pipeline, and aligned using MEGA5.[84] Amino acid polymorphisms were identified and accessions identical (100%) for predicted full length protein sequences were classified as a haplogroup. Haplogroups comprising less than 15 accessions were discarded from association analyses to have a reasonable number within each haplogroup. The ROS1 stop codon group includes all the accessions with a premature stop codon irrespective of the stop codon position and predicted protein amino acid sequence identity. Associations between ROS1 and ROS3 haplogroups and gbM variation were examined using the linear model ANOVA.

### Modelling gene body methylation dynamics

Methylation dynamics are modelled over both short timescales (30 plant generations) and long timescales (100,000 plant generations, corresponding to over three million cell cycles). Despite the remarkably high fidelity of the maintenance pathway, over these long timescales maintenance failure events will accumulate. We study the role of cooperative *de novo* and cooperative maintenance pathways (indicated by experimentally observed gain/loss rate profiles, Figures 2B, 2C, and S2A–S2D) in generating stable

methylation dynamics consistent with the experimentally observed methylation patterns. The term 'cooperative' is used to denote, for example, that the probability of a specific unmethylated CG site (uCG) gaining methylation is enhanced by the presence of other nearby methylated sites (mCG).

A single copy of each diploid chromosome is modeled through multiple cell cycles, assuming that the methylation dynamics are unaffected by whether that chromosome is currently in a diploid or haploid cell environment. Below we discuss in detail how this choice relates to the experimental setups. Furthermore, we model only the CG sites and assume that the methylation dynamics of each gene are independent from all others so that each gene can be modelled individually.

We assume that methylation gain and loss is an intrinsically stochastic process, in which case unmethylated sites (and even regions) can, at times, exist in locations that in principle can be methylated. We therefore created an annotation of such 'methylatable-regions', identified using the frequency of methylation across *Arabidopsis* accessions and the Col-0 H2A.Z thresholds, as described in detail in the section: methylatable gene regions. In brief, we selected accessions with overall gbM levels similar to Col-0 (N=891)[19] and used these to define genes that could contain gbM. Accessions were considered similar to Col-0 if they have a global gbM level within one standard deviation of the mean, because independent Col-0 samples vary substantially within this interval (Figure S6D). Gene ends were then removed using the location of methylation in these accessions, and the level of H2A.Z in Col-0. This procedure produced a single, continuous, methylatable region per gene in 7980 genes (Table S4A). Unless otherwise stated, all simulations are performed using this methylatable-regions annotation. Methylatable-regions consisting of only four CG-sites or fewer are not simulated. Subtleties occurring when genes overlap are also discussed above in the section methylatable gene regions. For comparison we also simulate gain/loss-rate profiles and steady-state methylation levels when applying the cooperative feedback model to a whole gbM genes annotation, as discussed in the section: simulations using alternative annotations. The sections model fitting and model predictions provide details of the model fitting and subsequent predictions.

The methylation dynamics are simulated stochastically using the direct Gillespie algorithm,[85] as described below, with the cooperativity implemented using an independent interaction between, for example, each target uCG and every mCG. All timescales and rates are defined relative to the cell cycle duration (here taken to be the time between successive DNA replication events).

### Construction of Two-State Model

The model does not keep track of the individual methylation status of the top and bottom strand cytosines of each CG site. Instead, a single CG site is defined to be in one of three states: fully-methylated (*M*), hemi-methylated (*H*) or un-methylated (*U*). Initially we assume that there is no active demethylation occurring in gene bodies (e.g., through the DNA glycosylase pathways, such as the DNA demethylase ROS1). We later revisit this assumption, however, in light of potential ROS1 activity. The possible methylation gain and loss transitions between the three states are illustrated and defined in Figure S2E.

During replication, all *M* sites are converted to *H* sites. On average, we assume that half of the pre-existing *H* sites will be methylated on the stand of DNA corresponding to the lineage that we follow; these therefore remain as *H*, whereas the remainder of the pre-existing *H* sites transform to *U* sites.

It is known that the MET1 maintenance pathway is extremely efficient,[86] a result recapitulated by our analysis. In the analysis below, we assume that MET1-mediated re-methylation is not only highly efficient but occurs rapidly after replication on a timescale much faster than the cell cycle duration. This assumption simplifies our analysis. If, however, maintenance is slower (but still highly efficient within a cell cycle), our conclusions are nevertheless unchanged. We first define a maintenance failure rate, $f$, which is the probability per cell cycle that an *H* site generated during replication is not converted to an *M*. This gives a remethylation maintenance rate

$$r_{\mathrm{H}}^{+} = 1 - f = 1 - \epsilon\gamma,$$

with $0 < \epsilon \ll 1$, and where $0 < \gamma < 1$ is a further cooperative 'suppression-factor', which reduces the chance of a maintenance failure occurring if there are existing *M* sites nearby (see below). In the limit of $\gamma \to 1$, $\epsilon$ is the background maintenance failure rate that a single isolated *M* site immediately before replication, in an otherwise completely unmethylated gene, is not fully remethylated in the period between successive DNA replication events.

The *de novo* methylation of a *U* site, $r_{\mathrm{U}}^{+} \ll 1$, is composed of both a spontaneous and a cooperative pathway as discussed below. As expected, in the experimentally relevant region of parameter-space (i.e., very efficient maintenance by MET1, $\epsilon \ll 1$), *H* sites only exist transiently, up to one cell cycle, before being resolved to *U* or *M*. As a result, the intermediate state, *H*, can be integrated out to create a two-state model consisting of only *U* and *M* states with effective direct transitions between these two states. This simplification was confirmed to have negligible effect on the simulation results by first implementing the three-state model described above using a Gillespie algorithm that was interrupted at the end of every cell cycle to explicitly simulate every replication event. In the experimentally applicable region of parameter-space, the three-state and two-state models produced almost identical methylation dynamics. Crucially, however, as replication is no longer explicitly simulated for the two-state model, a single Gillespie time increment can span multiple cell cycles, speeding up the simulation by several orders of magnitude. This speed up was helpful in simulating methylation dynamics over very long time periods of $10^5$ plant generations.

In the two-state model, the effective $M \to U$ loss-rate, $r^{-}$, to first order is given by:

$$r^{-} = \mathrm{prob}(M \to U)$$

$$= p^{(\mathrm{Rpn})}(M \rightarrow H) \times p^{(\mathrm{Maint})}(H \rightarrow H) \times p^{(\mathrm{Rpn})}(H \rightarrow U)$$

$$\approx (1)(\epsilon\gamma)\left(\frac{1}{2}\right) = \frac{\epsilon\gamma}{2}$$

where $p^{(\mathrm{Rpn})}$ represents the probability of the specified event occurring at replication, and $p^{(\mathrm{Maint})}$ the probability of the specified event occurring during maintenance. Similarly, the effective gain-rate, to first order, is given by:

$$r^{+} = \mathrm{prob}(U \rightarrow M) = p^{(\mathrm{Maint})}(U \rightarrow H) \times p^{(\mathrm{Rpn})}(H \rightarrow H) \times p^{(\mathrm{Maint})}(H \rightarrow M) \approx \left(2r_0^{+} + 2r_{\mathrm{Coop}}^{+}\right)\left(\frac{1}{2}\right)(1) = r_0^{+} + r_{\mathrm{Coop}}^{+}$$

where $r_U^{+} = 2r_0^{+} + 2r_{\mathrm{Coop}}^{+}$ is decomposed into two components. Firstly, a constant spontaneous *de novo* gain-rate, $r_0^{+}$, which is assumed to be a uniform background across all CG sites. Secondly, $r_{\mathrm{Coop}}^{+}$ represents the cooperative gain rate. We discuss the precise implementation of this contribution in the next section. The explicit factor of 2 in the parameterization of $r_0^{+}$ and $r_{\mathrm{Coop}}^{+}$ is included to account for the fact that in each *U* site there are two possible unmethylated cytosine targets for the pathways to act on. Only the first order contributions to the effective rates are included, as self-consistently, we find that the fitted values for $r_{\mathrm{Coop}}^{+}$, $r_0^{+}$ and $\epsilon$ per cell cycle are all several orders of magnitude smaller than one.

In our previous work modelling decay of methylation levels at transposable elements (TEs) in various *A. thaliana* mutants[34] we assumed a constant methylation gain and loss rate per cell cycle without explicit cooperativity, despite the rapidly falling methylation level. In light of our current work, it seems likely that both cooperative *de novo* and cooperative maintenance will also shape MET1 activity in TEs. The previously calculated rates for TEs in Lyons et al.[34] therefore constitute effective 'average' methylation gain/loss rates (including cooperativity) during the decay dynamics.

### Gillespie simulation: cooperative gains

We simulate using the 'direct' Gillespie algorithm.[85] An intermediate cooperative gain propensity, $r_{Coop\ (i)}^{+}$, is calculated for every pair of *U* and *M* sites (*UM*-pair) in the current gene-region. The functional form of $r_{\mathrm{Coop}\ (i)}^{+}(x)$, depends on *x*, the base-pair separation between the two CG-sites in the *UM*-pair and is discussed in a later section. The cooperative gain propensity, $r_{\mathrm{Coop}}^{+}$, for a given *U* site is then found by summing all the individual contributions from each of the *UM*-pairs for that particular *U* site: $r_{\mathrm{Coop}}^{+} = \sum_{(i)} r_{\mathrm{Coop}\ (i)}^{+}(x)$. The overall gain propensity for that individual *U* site is then given by $r^{+} = r_0^{+} + r_{\mathrm{Coop}}^{+}$ as defined in the previous section, while *M* sites are assigned $r^{+} = 0$. The total gain propensity for the gene-region, $r_{\mathrm{total}}^{+}$, is the sum over all the individual site-propensities. Similarly, a loss propensity, $r^{-}$, is calculated for every individual *M* site, as described in detail in the next section, along with $r^{-} = 0$ for *U* sites. The individual-site loss propensities are also summed to give $r_{\mathrm{total}}^{-}$ and finally the total propensity for the entire gene-region is defined as $r_{\mathrm{total}} = r_{\mathrm{total}}^{+} + r_{\mathrm{total}}^{-}$.

At a time, *t*, the next event (i.e., gain or loss) will occur at time: $t + \Delta t$, where $\Delta t = \frac{\ln\left(\frac{1}{\mathrm{rand}_1}\right)}{r_{\mathrm{total}}}$ and $\mathrm{rand}_1$ is a uniformly distributed random number between 0 and 1. The site to be updated at this time is found using the propensity threshold: $r_{\mathrm{total}} * \mathrm{rand}_2$, where $\mathrm{rand}_2$ is a second random number from the same distribution. The smallest individual site-propensity for which the cumulative sum of site-propensities, up to and including that site-propensity, exceeds the threshold determines the identity of the next site to be updated (i.e., gain/loss of methylation) at time $t + \Delta t$.

We assume that there are no cooperative interactions between CG sites in different genes. For a *UM*-pair separated by $x = 300$ bp, the cooperative gain amplitude is over a factor of 1000 smaller than its maximum value, therefore, this assumption is unlikely to have a significant impact on the simulated dynamics.

### Gillespie simulation: cooperative maintenance

We model a cooperative maintenance pathway using a similar approach to that for the cooperative gains. We assume that the presence of other *M* sites in the locus increases the chance of a site being maintained after replication. Again, we approach this in a pairwise fashion, this time considering all *MM*-pairs in a locus. As the bare (i.e., non-cooperative) maintenance failure rate, $\epsilon$, is already so small, we multiply the contributions from the individual *MM*-pairs to the enhanced maintenance rate. This is to ensure that the total maintenance probability can never exceed 1 at any *M* site. Each *MM*-pair contributes a 'suppression-factor' to the maintenance failure rate: $(1 - \beta_{(i)}(x))$, (Figure S2H), where *x* is the base pair separation between the two sites of the *MM*-pair, and $0 < \beta_{(i)}(x) < 1$, for all *x*. The loss-rate at an individual *M* target-site is then calculated as $prob(M \rightarrow U) = r^{-} = \frac{\epsilon\gamma}{2}$, where:

$$\gamma = \prod_{(i)}\left(1 - \beta_{(i)}(x)\right)$$

The product is taken over all *MM*-pairs that contain the specific *M* target-site being maintained. The functional form of $\beta_{(i)}(x)$ is discussed below and is the same for every *MM*-pair in the locus. Simulations omitting cooperative maintenance produce a uniform

loss rate (equal to $\epsilon/2$), thus failing to reproduce the suppression of methylation losses observed at short-length scales (Figure S2K).

### Active demethylation

We constructed the model under the assumption of no active demethylation. However, it is possible that DNA demethylases, such as ROS1, could also target the modelled gene-regions. One scenario is that a ROS1 pathway could renormalize the $U \rightarrow M$ transition probability so that the values of the parameters $r_0^+$ and $r_{Coop}^+$ also include the action of ROS1 rapidly removing some hemi-methylation before the next replication event. In addition, ROS1 may be actively demethylating $M$ sites. This is largely captured by the current model by the existing maintenance failure pathway (i.e., some proportion of the parameters $\epsilon$ and $\beta_{(i)}$ could in principle arise from ROS1 action). An alternative parameterization for a ROS1 active demethylation pathway, however, might be to include a spontaneous (non-cooperative) $M \rightarrow U$ interaction (analogous to $r_0^+$ for the spontaneous gains). This is not included in the model as the current resolution of the loss-rate data is insufficient to support fitting to an additional parameter.

### Functional form of cooperative interactions

As discussed in detail below, the experimental methylation gains around a given mCG site fall away as an approximate power law (Figure S2F, see section simulated gain and loss rates for description of gain/loss rate profiles plotted as a function of distance to the nearest $M$-site). Consequently, the cooperative gain interaction strength for each uCG site is calculated using an (offset) power-law decay as a function of distance to existing mCG sites. By using a simple, monotonically decaying interaction with a single power-law, we could reproduce the rapid fall in gain rate at a length-scale of ~30 bp. This form, (equivalent to setting the parameter $\alpha_1 = 0$ in the equation defined in the following paragraph), however, fails to generate the second peak at ~170 bp (Figure S2J), demonstrating that this enhancement of the gain rate does not emerge from the intrinsic spatial distribution of CG sites. As ~170 bp corresponds to the nucleosome repeat length,[67] we hypothesize that the 3D chromatin conformation reduces the effective distance between the target uCG and promoting mCG sites, thus generating an enhanced cooperative interaction when their separation matches the nucleosome repeat length. This motivated including a secondary, weaker, component to the cooperative gain interaction, with a maximum amplitude at ~170 bp (Figure S2G).

Mathematically, the interactions are formulated as follows. The magnitude of the cooperative gain-rate for an individual $UM$-pair, $r^+_{\text{Coop (i)}}(x)$, is a function of $x$, the base-pair separation between the $UM$-pair: $x = |\boldsymbol{x_U} - \boldsymbol{x_M}|$. We split $r^+_{\text{Coop (i)}}(x)$ into a primary component, $r^{(1)}_{\text{Coop (i)}}(x)$, and a secondary component, $r^{(2)}_{\text{Coop (i)}}(x)$, for which the interaction has the same functional-form as the primary component but with the origin translated by $x_{nucl}$, a distance on the length scale of the nucleosome repeat length:

$$r^+_{\text{Coop (i)}}(x) = r^{(1)}_{\text{Coop (i)}}(x) + r^{(2)}_{\text{Coop (i)}}(x)$$

$$r^{(1)}_{\text{Coop (i)}}(x) = \begin{cases} \alpha_0, & 0 < x < x^+_{plat} \\ \alpha_0 \left| \dfrac{x^+_{plat} - x^+_{div}}{x - x^+_{div}} \right|^{\lambda^+}, & x \geq x^+_{plat} \end{cases}$$

$$r^{(2)}_{\text{Coop (i)}}(x) = \begin{cases} \alpha_0\alpha_1, & 0 < |x - x_{nucl}| < x^+_{plat} \\ \alpha_0\alpha_1 \left| \dfrac{x^+_{plat} - x^+_{div}}{|x - x_{nucl}| - x^+_{div}} \right|^{\lambda^+}, & |x - x_{nucl}| \geq x^+_{plat} \end{cases}$$

For $0 < x < x^+_{plat}$, for the primary component, and for $0 < |x - x_{nucl}| < x^+_{plat}$ for the secondary component, both have a constant magnitude described by $\alpha_0$ and $\alpha_0\alpha_1$, respectively. This parameterisation was chosen so that the strength of the secondary component could be scaled relative that of the primary component when introducing the interaction-strength dependence on CG-density (described below). All other parameters are assumed to be equivalent for the primary and secondary components of the cooperative gain interaction. The parameter $x_{plat}$ defines the end of this constant plateau, beyond which the interaction has a power-law decay. The location of the power-law divergence is specified by $x_{div}$, where $x_{div} < x_{plat}$. Finally, the power-law decay constant is specified by $\lambda^+$. As described above, the secondary interaction could occur due to the 3D organization of the DNA providing an alternative, shorter, interaction route between the two CG-sites in the $UM$-pair. The overall functional form of $r^+_{\text{Coop(i)}}(x)$ is shown in Figure S2G.

The model also includes mCG loss, which may occur either through active demethylation (e.g., via ROS1) or passively due to maintenance failure after DNA replication. We find that a cooperative maintenance mechanism (by which surrounding mCG sites reduce the probability of a maintenance failure at the target mCG site) must be included to qualitatively reproduce the loss rate profile. Excluding this interaction (equivalent to setting $\beta_0 = 0$ in the equation below), as expected, produces a simulated loss rate profile that is constant as function of distance to the nearest $M$-site (Figure S2K). This is inconsistent with the observed loss-rate profile. As with

the cooperative gains, we find the cooperative maintenance interaction to be well described by a power-law (Figure S2F). The cooperative maintenance interaction is, therefore, described analogously to the cooperative gain interaction (defined above):

$$\beta_{(i)}(x) = \begin{cases} \beta_0, & 0 < x < x^-_{plat} \\ \beta_0 \left| \dfrac{x^-_{plat} - x^-_{div}}{x - x^-_{div}} \right|^{\lambda^-}, & x \geq x^-_{plat}, \end{cases}$$

with the functional form $(1 - \beta_{(i)}(x))$ shown in Figure S2H. Choice of this functional form ensures convergence at long distances to the background (non-cooperative) methylation loss rate. We allow the amplitude ($\beta_0$), plateau length ($x^-_{plat}$), divergence location ($x^-_{div}$) and power-law decay constant ($\lambda^-$) parameters to all have values independent of those used for the cooperative gain interaction.

We chose a power-law to describe the decay of both the cooperative gain and cooperative maintenance interactions with increasing separation of the *UM*- or *MM*-pair. This is consistent with the linear form of the loss-rate over one order of magnitude, when plotted as a function of distance to the nearest *M*-site, using a log-log scale as shown in Figure S2F (here loss rate is calculated over the whole gene length). A linear best-fit (to the log-log transformed data $20 < x < 200$ bp) is also shown. Note that the gradient of this linear-fit to the overall loss-rate does not provide the value of the power-law $\lambda^-$, because $\lambda^-$describes the contribution to the total interaction that arises from only a single *MM*-pair. Hence, the entire shape of this interaction strength profile for a single *MM*-pair (Figure S2H) differs from the full losses profile (Figures 2E and S2I, right), due to the latter including interactions with multiple mCGs. The structure of the equivalently plotted gain-rate (again calculated over the whole gene length) is more complex as there is the secondary-interaction peak at $x \approx 170$ bp ($\log_{10}(x) \approx 2.2$). In this case, we therefore, make a linear fit to the log-log transformed gain-rate over only a narrow window of $25 < x < 47$ bp, where the primary cooperative gain pathway is dominant. The linear fit is plotted over the entire length-scale of the gain-rate data and at large x-values, beyond the range of the secondary interaction peak, it is also quite consistent with the gain rate data. We note however, that as the resolution for the gain and loss rates becomes low at larger length-scales (of the order $x > 500$bp), we cannot conclusively rule out the possibility of an exponentially decaying interaction strength, though fits to an exponential were not as good. Once again, the shape of this gains interaction strength profile for a single *UM*-pair (Figure S2G) differs from the full gains profile (Figures 2F and S2I, left), due to the latter including interactions with multiple mCGs.

For the model to correctly reproduce the observed distribution of steady-state methylation levels for gene-regions of varying length ($L_{locus}$), and average CG-density ($\rho_{CG}$), it was necessary to vary the strength of the cooperative gain interaction and of the cooperative maintenance interaction, as a function of CG-density. We define $L_{locus}$ to be the base pair distance between the first and last CG sites in the gene-region and chose the unconventional definition of $\rho_{CG} = \frac{N_{CG}-1}{L_{locus}}$, where $N_{CG}$ is the number of CG sites in the gene-region, so that $1/\rho_{CG}$ is equivalent to the average CG site spacing.

To introduce as few extra parameters as possible, we used the linear scaling:

$$\alpha_0(\rho_{CG}) = \begin{cases} \alpha_m \rho_{CG} + \alpha_c, & 0 < \rho_{CG} < \rho_o \\ \alpha_m \rho_0 + \alpha_c, & \rho_{CG} \geq \rho_0 \end{cases}$$

where $\alpha_m < 0$ so that the cooperative gain interaction strength decreases with increasing CG-density, and the maximum CG-density threshold, $\rho_0$, prevents a vanishing interaction strength. Similarly, we vary the strength of the cooperative maintenance interaction, $\beta_0$, as a function of CG-density using:

$$\beta_0(\rho_{CG}) = \begin{cases} \beta_m \rho_{CG} + \beta_c, & \beta_0 > 0 \\ 0, & \text{otherwise} \end{cases}$$

where $\beta_m < 0$. The cooperative maintenance strength is capped to have a minimum value of zero. Finally, when we model the outlier *Arabidopsis* accessions, we vary the strength of the cooperative gain and cooperative maintenance using a single scale-factor, $r^*$, so that $\alpha_c \rightarrow r^* \alpha_c$ and $\beta_c \rightarrow r^* \beta_c$.

In summary, the final model contains four gain/loss processes: spontaneous *de novo* (one parameter); cooperative *de novo* (eight parameters); background maintenance failure (one parameter); cooperative suppression of maintenance failure (five parameters), to give 15 parameters in total.

### Calculation of diploid methylation gain and loss rates

We simulate the methylation dynamics of a single chromosome. All bisulfite sequencing used here, however, measured the methylome of diploid leaf tissue. Each of the two copies forming a chromosome pair will have had a different trajectory, alternating between a haploid and diploid environment to reach the final leaf tissue that is sequenced. Our model contains no information on whether the simulated chromosome is currently in a haploid or diploid cell, or any other details about its environment. We therefore treat the two chromosome copies as indistinguishable.

The simulated gain and loss rates are compared to two different data sets: firstly, the existing Mutation Accumulation Line experiments (MAL),[45,62] and secondly the *WT* dataset presented in this work. These two approaches use contrastingly shaped

lineage-trees and consequently involve complementary assumptions. We convert the rates from both experimental data sets, and the simulated data, into the diploid gain/loss rate per cell cycle, i.e., for methylation gains, this is the rate that a homozygous *U* site, is converted to a homozygous *M* site.

As we now show, the diploid gain/loss rate is equivalent to the haploid gain/loss rate found by simulating a single chromosome. If we define *r* to be the haploid gain rate per plant generation for a single chromosome, then for a diploid cell, the total number of newly heterozygous-methylated sites generated by a methylation gain on one of the chromosomes in a single generation is 2*r*. In these experiments the plants are self-fertilized each generation, so for the heterozygous sites, there is a 25% chance that Mendelian segregation will result in a homozygous gain/loss, a 25% chance the gain/loss disappears so that the original homozygous state is retained, and a 50% chance that it remains heterozygous at each subsequent reproductive cycle. Eventually, therefore, there is 50% chance that a heterozygous gain is 'fixed' into a homozygous gain and a 50% chance it reverts to a homozygous *U* site. This factor of ½ cancels the initial factor of two and hence the diploid gain gate is equal to the haploid gain rate on a single chromosome.

We note that after the male and female cell-lineages diverge to form the reproductive tissues, the rate of producing new heterozygous gains doubles as there are now four chromosomes on which a gain could occur. For the purposes of this discussion, we chose to define the start of the generational cycle as the divergence point of the male and female lineages. Meiosis and fertilization then occur mid-way through this cycle. Neglecting the (negligible) possibility of a gain occurring at the same CG site in both the male and female lineage during a single generation, upon fertilization, there is a 50% probability that (in our followed lineage) the CG site reverts to homozygous unmethylated and a 50% probability that it persists as a heterozygous gain in the now single germline lineage. The total number of new heterozygous gains formed so far is therefore double the number of haploid gains. As there is only a single germline lineage for the remainder of the generational cycle, heterozygous gains continue to be formed at double the haploid gain rate. Consequently, the new heterozygous gains are produced at the same overall rate from before and after the male and female lineages diverges, and the first opportunity for self-fertilized Mendelian segregation then occurs part way through the following generational cycle (with losses behaving equivalently).

The mutation accumulation line experiments have a narrow and very deep tree, following 4 lines in the Schmitz et al. dataset[45] and 8 in Becker et al.[62] (we excluded 1 outlier line, as discussed above), all for 30 consecutive plant generations. Consider a heterozygous methylation gain occurring in the first generation. After five reproductive cycles, for example, there is a probability of only $(0.5)^5 \sim 3\%$ that the CG site will still be in a heterozygous state. We assume, however, that all gains and losses occurring during the observed 30 generations have fully segregated out at the point of bisulfite sequencing. With this approximation, our analysis of these 30 generation datasets directly measures the diploid gain/loss rates over 30 generations. Finally, this is converted to the per cell cycle gain/loss rate by dividing the gain/loss rate for the entire experiment by the number of intermediate generations (in this case, 30 generations) and by the number of cell divisions through the germline lineage over one generation. For the latter, we assume that there is a constant number of cell divisions of 34 per generation.[44]

In reality, a proportion of the sites will remain in the heterozygous state at the end of the experiment. To analyze the Schmitz et al. and Becker et al. datasets, both the Generation 0, parental germline state, and the Generation 30, final germline state, are reconstructed from sibling offspring lines, as described previously. Defining the generational cycle to begin at the divergence of the male and female lineages is therefore consistent with aligning it to the experimental measurement of germline methylation state. Heterozygous sites in the initial or final experimental state are most likely to be classified as indeterminate methylation (*I* sites). As gains/losses are only identified at CG sites where both the initial and final methylation state can be called as either *U* or *M*, heterozygous methylated sites in the final observed state are excluded from the measured gain/loss rates. Excluding heterozygous sites that were formed prior to Generation 0 from the analysis is consistent with our assumption that all gains and losses are initiated during the 30-generation experiment. Excluding all heterozygous sites remaining in the final state at Generation 30, however, produces a slight underestimate of the true gain and loss rates.

We now estimate an upper limit to the fraction of gains lost due to the assumption that all will have fully segregated by the end of the experiment. Again, we use '*r*' to represent the haploid gain rate per generation. Over the 30-generation experiment, 2*r* heterozygous gains will occur per generation. Theoretically, if we could wait until all the heterozygous gains generated during the 30 generations then fully segregated out, this would lead to 30*r* diploid gains in total. Over the 30-generation experiment, heterozygous sites formed during the $x^{th}$ generation will have $30 - x$ self-fertilized Mendelian segregation opportunities. On average, therefore, 100% of the heterozygous gains occurring in the $30^{th}$ generation will have been missed, and 50% from the $29^{th}$ generation and 25% from the $28^{th}$ etc. The total number of heterozygous gains yet to segregate is given by the sum: $2r\left(1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots.\right) = 4r$. Half of these are expected, on average, to be fixed into full diploid changes, so that 2*r* gains have been uncounted. This fraction is $\frac{2}{30} \sim 7\%$ of the total number of diploid gains over the entire 30 generations. An identical argument applies to the fraction of losses not accounted for. Finally, we note, however, that heterozygous sites do not explain all the CG sites identified experimentally as indeterminately methylated. The majority are likely due to somatic methylation gains. As they are concentrated close to existing methylation, we anticipate that they correspond to sites that are unmethylated in the germline but having a high gain probability. Excluding these sites, therefore results in a small underestimate of the gain rate.

In contrast, the bisulfite sequencing data presented in this work uses a much broader and shallower lineage tree. As described above in the section calculation of gain and loss rates, gain and loss rates are calculated by comparing sets of three samples, transitioning between (for gains): homozygously unmethylated, to heterozygously methylated, to homozygously methylated over three consecutive generations. Identified methylation gains/losses are homozygous (present on both chromosomes). This rate is therefore

then doubled to give the eventual diploid gain/loss rate per generation, reflecting the additional 50% of the heterozygous changes that will eventually become fixed in the changed state. Again, we convert to a rate per cell cycle.

These two contrasting approaches (first the 30-generation MAL analysis, and secondly the three-consecutive generation analysis presented in the work) rely on different sets of assumptions. Despite this, they produce broadly consistent gain/loss rate profiles (Figures 2E, 2F, and S2I) and overall epimutation rates (Figure 2G), increasing our confidence in these calculated rates.

### Simulated gain and loss rates

We focused on fitting the model to our analysis of four lines of the 30-generation data set of Schmitz et al.,[45] as this data set provided the greatest spatial resolution. We simulate a single copy of each chromosome for 30 generations, assuming 34 cell cycles per generation.[44] For a direct comparison to the experimental data, the simulation is also repeated for four independent replicates. The random-number seed was increased by one for every new gene-region and for each replicate. We note that as expected, simulating over a greater number of replicates reduces the fluctuations seen in the gain and loss rates, especially at long length-scales.

The simulation uses the experimental Col-0 state to assign CG sites an initial status of either *M* or *U*. To achieve the best resolution, we use the 'Col-0 consensus state', generated by collating high coverage bisulfite sequencing from twenty different Col-0 plants grown under consistent conditions. In this consensus state, 1.77% of sites have indeterminate (*I*) status, most likely indicating that the CG site either has a large proportion of somatic methylation changes, or that it is heterozygous in the germline. For consistency with the experimental analysis, we entirely exclude these *I* sites from the simulation for the fit to the experimental gains/losses data. Gains and losses are identified by comparing the final simulated state to initial state for each gene, e.g., to qualify as a gain, a site must have status *U* in the initial state and status *M* in the final state (as is the case for the experimental gain/loss analysis). Overall simulated epimutation rates within methylatable regions can then be calculated as for the experimental data (Figure 2G; Table S5).

To quantify how the gain and loss rates vary as a function of surrounding *M*-sites, we plot spatially resolved gain/loss rate distributions as a function of distance to the nearest *M*-site. The signature of cooperative gains feedback is then an enhanced gain-rate when the distance to the nearest *M*-site is small. Conversely, cooperative maintenance corresponds to a suppressed loss rate at a small distance to the nearest *M*-site. The total distribution for the number of gains as a function of base pair distance to the nearest *M*-site (in the same methylatable-gene-region) is compiled from all simulated genes and all replicates, along with the equivalent distribution for the number of losses. Gain (or loss) rates are calculated by dividing the distributions for number of gains (or losses) by the normalization: the distribution of the number of *U*-sites (or *M*-sites) sites in the initial state, again as a function of distance to the nearest *M*-site. Finally, we convert to an average gain and loss rate per cell cycle, as described above for the experimental case. Also, see above for further details of how we relate the modelled gain and loss rates per cell cycle to those measured experimentally. Both the modelled and experimental gain and loss rate distributions are plotted after smoothing over a 10 base pair window. In Figures 2E, 2F, S2I–S2M, and S5A 'All experimental data sets' includes 8 lines from the Becker et al. dataset,[62] 4 lines from the Schmitz et al. dataset[45] and the *WT* dataset presented in this work. We simulate the gain and loss rate distributions both for the whole gbM genes annotation (Figures 2E, 2F, and S2M) and for methylatable regions (Figures S2I and S2L).

### Simulated steady-state methylation patterns

The steady-state methylation patterns predicted by the model are investigated by simulating for 100,000 generations for the specified number of replicates. In many cases, our simulations use a fully unmethylated initial state, however, we also confirmed that the model has a unique steady-state by also simulating from an experimental Col-0 or fully methylated initial state (Figure 4A). We confirmed that a simulation time of 100,000 generations is ample for the model (parameterized according to Table S3A) to reach steady-state (Figure 4B). As the steady-state is independent of the initial state choice (Figures 4A and 4B), when using the experimental Col-0 initial state we could safely arbitrarily assign an initial state of *U* to the CG sites with unknown methylation status in the Col-0 consensus state.

We compare the modelled steady-state methylation states, for a certain number of replicates, to those observed for an equivalent number of *A. thaliana* accessions, using the 1001methylomes resource.[19] Each independently simulated replicate is taken to represent a single accession, to investigate the extent to which our stochastic model of the purely epigenetic aspects of methylation dynamics can account for the observed natural variation in gene body methylation level across Col-like accessions (defined above). All Col-like accessions are, therefore, assumed to have genetically equivalent methylation machinery (equivalent to modelling replicates with an identical set of parameters). This approach neglects all effects of selection pressure that may affect the observed methylation patterns and assumes that no significant changes occur to the methylation machinery, the H2A.Z distribution (and thus the boundaries of the methylatable-regions), and finally the CG-site positions, over the simulated timescales. Consequently, we neglect any SNPs occurring at CG-sites in the Col-like accessions or outlier accessions (which incidentally are rare, <2% in Cvi-0), instead assuming a CG-site structure identical to Col-0 for every modelled replicate. This approach allows us to isolate the purely epigenetic contributions to the methylation dynamics from those influenced by genetic mutations over much longer time-scales (such as changes to the underlying CG-site positions). We also neglect any possible outcrossing between different accessions over this timescale: such outcrossing may occasionally occur, but provided it is between the same accession or with another Col-like accession, then statistically equivalent methylomes will be combined, which will not change the steady-state methylation distributions.

For each methylation-state (either a simulated replicate, or measured *A. thaliana* accession), we calculate the mean methylation level, $\langle M \rangle$, for each methylatable-region, where $\langle M \rangle = \frac{N_M}{N_M+N_U}$, with $N_M$ and $N_U$ being the number of CG-sites identified with status

$M$ and $U$ respectively in that methylatable region. For the simulated states, this calculation is performed at the end of the simulated period (100,000 generations). The total number of CG-sites in the simulations of the methylatable region is $N_{CG} = N_M + N_U$, by definition. This is not the case for the observed methylation states, where some CG-sites are assigned a status $X$, either due to having an indeterminate methylation status (as described previously), or a SNP occurring at that site, so that $N_{CG} = N_M + N_X + N_U$. In the rare event that $N_M + N_U = 0$ for a methylatable region in an accession, we assign $\langle M \rangle = 0$. Histograms of the distribution of $\langle M \rangle$ (Figures 4A, 7A–7D, S3A–S3C, S3H, S4A–S4E, S4G, and S7A–S7C) are normalized to the number of replicates (either simulated or observed accessions).

The clear spike at $\langle M \rangle = 0$ (Figure 4A) corresponds to fully unmethylated regions. Due to the stochastic nature of the dynamics, at any time there will be a number of methylatable-regions that by chance are completely unmethylated (most likely to be methylatable-regions containing few CG-sites). The fraction of completely unmethylated methylatable-regions is directly controlled by the magnitude of the spontaneous *de novo* contribution: the higher the spontaneous *de novo* rate, the less likely it is that a locus will be completely unmethylated. The height of this peak can therefore be fit largely independently of the other parameters, thus providing a strong constraint on the strength of the spontaneous *de novo* rate. We note that the extra structure in the $\langle M \rangle$ distribution arises due to the fractional definition of methylation level: a value of $\langle M \rangle = 1/2$ can be obtained for any gene-region with an even value of $N_{CG}$, whereas a value of approximately $\langle M \rangle = 49/100$, for example, can only be obtained from a much more limited subset of gene-regions.

Finally, to study the variability of methylation levels between replicates, we also calculated the standard deviation of $\langle M \rangle$, $\sigma_{\langle M \rangle}$, over all replicates for every gene-region.

## Model fitting

We manually fitted the model (final parameterization given in Table S3A), simultaneously to two very different data sets (both calculated over methylatable regions): firstly, the short-timescale (30-generation) spatially resolved Col-0 gain/loss rate profiles (described above, Figures S2I and S2L), and secondly, the long-timescale steady-state distributions of $\langle M \rangle$ (described above, Figure 4A) compiled for 30 Col-like accessions. We chose to fit the model manually as these two data sets have very different, and difficult to quantify, uncertainty levels. Furthermore, we discovered that one standard measure of fit-quality: the R-value calculated over individual methylatable regions (Figure 4E), is dominated by the ~20% outlying loci for which the mean methylation level is captured poorly by the model. Indeed, very small improvements to this R-value are possible, at the expense of introducing a considerable systematic bias (away from the observed values) in the modelled methylation level of the majority of well-fitting loci. Although the generic model assumptions might be poor approximations for the ~20% of outlier loci in our manual fit, we elected not to implement an automatic fit, as that would require the arbitrary exclusion of these outlier loci, rather than selecting appropriate loci to model based purely on generic biological characteristics (such as mean H2A.Z level across the locus).

To perform the fit to the steady-state methylation levels, distributions of $\langle M \rangle$ (not shown) were separated into subsets of gene-regions, grouped according to both the length of the gene-region, $L_{locus}$, and the mean CG-density, $\rho_{CG}$, as defined previously. Subgroups were defined as: $L_{min} \leq L_{locus} < L_{max}$ for $L_{min} = \{0, 1\times10^3, 2\times10^3, 4\times10^3, 6\times10^3, 1\times10^4\}$ bp and $L_{max} = \{1\times10^3, 2\times10^3, 4\times10^3, 6\times10^3, 1\times10^4, 3\times10^4\}$ bp and $\frac{1}{\rho_{max}} \leq \frac{1}{\rho_{CG}} < \frac{1}{\rho_{min}}$, for $\frac{1}{\rho_{max}} = \{3, 6, 9, \ldots, 96, 99\}$ bp and $\frac{1}{\rho_{min}} = \{6, 9, 12, \ldots, 99, 102\}$ bp, with $\rho^{-1}$ increasing in units of 3 bp. We compared the simulated steady-state methylation distributions after 100,000 generations to 30 Col-like replicates, selected for their high sequencing coverage (described above).

Splitting the gene-regions into sub-groups of CG-density revealed the necessity to include a linear variation of the cooperative interaction strength as a function of mean CG-site density over the gene-region. A prior attempt to fit the model parameters using only the spatial gain/loss rate profiles omitted the above CG-site density-based correction to the cooperativity strength (i.e., $\alpha_m = \beta_m = 0$). For this parameterization (Table S3B, splitting the mCG level distributions by CG-site density (Figure S3H) revealed that the model systematically over-methylated CG-rich methylatable regions and under-methylated CG-poor methylatable regions. The consequence of the CG-site density dependent correction is to reduce the cooperative feedback per CG-site for loci of higher CG-site density. It is likely that the requirement for this correction is (at least in part) due to assuming that each mCG site independently enhances the cooperativity strength. In reality, however, the contribution of each individual site may be reduced when there are many mCG sites in close vicinity. Finally, we note that the need to vary the strength of the cooperative interaction amplitude as a function of mean CG site density cannot be justified from considering only the spatial gain/loss rate profiles.

The majority of parameters could be fit quite precisely (and are therefore well constrained) by the short-timescale spatially-resolved gain/loss rate profiles. Only the gradients of the CG-density corrections ($\alpha_m$, $\beta_m$, $\rho_0$ and corresponding adjustments to $\alpha_c$, $\beta_c$) relied solely on the long-timescale steady-state simulations. It was challenging to accurately capture the increase in mean methylation level with increasing locus length ($\langle M \rangle[L_{locus}]$). We found that slight adjustments to the power-law exponents ($\sim \pm 5\%$) were possible while still maintaining a good fit to the gain/loss rate distributions, while having a noticeable effect on $\langle M \rangle[L_{locus}]$. The other influential parameters were: $\epsilon$, $\alpha_c$ and $\alpha_1$, though the latter was well constrained by the gain/loss rate distributions (as were the plateau lengths and the positions of the power-law divergences), while the sensitivity to $\epsilon$ was similar to that of the power-laws. After fixing all other parameters, an initial estimate of $\alpha_m$, $\alpha_c$, $\beta_m$ and $\beta_c$ was found by independently manually fitting the individual sets of histograms generated with a fixed value of $\rho_{CG}$ to a value of $\alpha_0$ and $\beta_0$ before then using a linear fit to the resulting values of $\alpha_0(\rho_{CG})$ and $\beta_0(\rho_{CG})$ to extract corresponding values of $\alpha_m$, $\alpha_c$, $\beta_m$ and $\beta_c$.

We obtained an initial, order of magnitude, fit for the value of the spontaneous background *de novo* gain rate, $r_0^+$, from the long length-scale tails (at ~1000 bp distance to the nearest *M*-site) of the spatial gain-rate profile (Figure S2L). This was then refined to more precisely capture the height of the spike at $\langle M \rangle = 0$ in the distribution of steady-state methylation levels (Figure 4A).

We made an additional test of the parameter-sensitivity of the model, by repeating all simulations with each individual parameter increased and decreased by 10%. None of these parameter alterations qualitatively affected the model results. For the following parameters, however, noticeable quantitative shifts (reducing the fit quality) occur to the results of both the short and long-timescale simulations: $x_{plat}^+$, $x_{plat}^-$, $x_{div}^+$, $x_{div}^-$, $\lambda^-$, $\beta_c$. Similar quantitative shifts are seen for $x_{nucl}$ (in gain/loss rate simulations) and $\alpha_c, r_0^+$ (in steady-state simulations). All these parameters are therefore well constrained by a combination of both the gain/loss rate and steady-state methylation level fits. The remaining parameters: $\lambda^+$, $\alpha_1$, $\epsilon$, $\alpha_m$, $\beta_m$, $\rho_0$, are less well constrained at the level of a 10% shift of the parameter-value.

Finally, we investigated the sensitivity of our model to potential genetic diversity in the methylation machinery, by adjusting the overall strength of the cooperative interaction pathways (parameterization details discussed above) and/or spontaneous *de novo* strength. Simulated methylation patterns are compared to *A. thaliana* methylation-outlier accessions (see *Choice of accessions for modelling*), assuming the same methylatable-regions as identified for Col-like accessions (Figures 7A–7D and S7A–S7C). We note, however, that methylation-outlier accessions may also have accession-specific methylatable-regions.

## Model predictions

The fully parameterized model was simulated to steady-state (100,000 generations, all-*U* initial state) for 740 replicates and compared to all non-redundant Col-like accessions with sequencing coverage > 50% (see Figures S4E and S4G). The performance of the model at the individual gene level was tested by comparing the predicted and observed values of both $\langle M \rangle$ and $\sigma_{\langle M \rangle}$, both averaged over the 740 simulated/observed accession replicates respectively, for each individual methylatable region (Figures 4E and 6D). We note that at no point was the variation in methylation level (even for the initial 30 Col-like accessions), $\sigma_{\langle M \rangle}$, used in the fits.

To examine the spatial distribution of methylated/unmethylated sites, we grouped neighboring CG sites into pairs: either uCG-uCG (denoted *UU*), uCG-mCG (*UM*) or mCG-mCG (*MM*). Additionally, an *XX*-neighbour-pair is defined to be a pair of neighbouring CG sites, at least one of which cannot be identified as either *M* or *U* from the sequencing data. Note that there are no *XX*-neighbour pairs, by definition, for the simulated states. For each of these four (or three) groups, we calculated the observed (and simulated) distribution of pair-separations (measured in bp). All three simulated distributions are normalized to the total number of pairs (of any type), totaled over all 740 realizations. Equivalently, the four experimental distributions are all normalized to the total number of pairs (of any type), totaled over all 740 accessions. The three simulated distributions are shown directly (Figure S5B). For the experimental distributions, we add the *XX*-neighbour pairs distribution to each of the other three to produce the green bands. The bottom of each green band therefore corresponds to the case that none of the *XX*-neighbour pairs (if their methylation status were known) contribute to the depicted distribution. The top of each green band corresponds to the case that all the *XX*-neighbour pairs contribute to the depicted distribution. The simulated distributions are therefore expected to lie somewhere within the green bands. Although each *XX*-neighbour pair can only actually belong to one of the three distributions, the fraction of *XX*-neighbour pairs that corresponds to each distribution will vary considerably as a function of pair separation. This is because the sites with unknown methylation status are not evenly distributed, but instead are concentrated close to existing methylation. The predicted and observed distributions are in good agreement (Figure S5B). mCG-mCG separations are greatly enriched at short distances compared to uCG-mCG or uCG-uCG. This is a further demonstration of local cooperativity, which will favor clustering of methylated sites.

To investigate the spatial patterns across whole loci, we also calculated the long-range mCG-mCG (*MM*) correlation function for all CG-pairings within an individual locus (Figure 5C). Pairings of CG-sites between different methylatable-regions are not considered. We combine counts over all loci to compile a histogram of the CG-site separation (in bp) of every *MM*-pair. We note that any pairings containing a CG-site of unknown methylation status in an individual Col-like accession are excluded from the data analysis. The simulated histogram is then normalized by dividing the number of *MM*-pairs observed at a specific CG-site separation, with the total number of CG-site pairs of any type (i.e., *UU*, *UM*, or *MM*). The observed data histogram is normalized equivalently, this time dividing by the total number of CG-site pairs but excluding all those that contain at least one CG-site of unknown methylation status. This normalization choice means that the difference in amplitude between the two distributions at a given separation reflects the accuracy of the model's prediction of the total number of *MM*-pairs generated at that separation. The strong enhancement in the correlation function at short length-scales reflects the strong cooperativity of methylation gains, with methylation more frequently observed in regions of locally high CG-density (as expected for distance-dependent cooperative feedback interactions, Figures S2G and S2H). Additionally, a second peak in the correlation function is seen centered on 167 bp, corresponding to a periodicity of highly methylated regions consistent with the nucleosome repeat distance (and consistent with the methylation enhancement observed under nucleosomes, Figure 5E).

Finally, as an additional test of the model, we also compared the simulated and the experimental gain/loss rate distributions as a function of distance to the nearest *U*-site (Figure S5A) (calculated analogously to the distance to the nearest *M*-site distributions described previously, Figure S2I). We did not fit to these distributions.

## Simulations using alternative annotations

In addition to simulating the methylatable-regions, we also simulated whole gbM genes fitting the model to the Col-0 30-generation gain/loss rate profiles using a whole gbM gene annotation. Here we omitted the CG-density correction (i.e., $\alpha_m = \beta_m = 0$) with the

parameterization given in Table S3C. Using this annotation, we could also reproduce well the spatial gain/loss rate profiles (Figures 2E and 2F: loss, gain; Figure S2M: gain-tail), though the fit to the gain rate profile was noticeably poorer over length-scales of 200 to 400 bp. However, long-timescale (100,000 generations) simulations to steady-state revealed that with this fit to the whole gbM genes annotation gain/loss rate profiles, the model drastically over-methylates gbM genes (Figure S3A).

Conversely, using this same parameterization (Table S3C) to simulate only the regions of Col-0 observed to be highly methylated (gbM segments annotation) produced a severely under-methylated steady-state (Figure S3B). This result is consistent with stochastic gene-body methylation dynamics: at any point in time a significant fraction of CG-sites that in principle could be methylated at other times happen to currently be in the unmethylated state. Applying the model to only the regions currently methylated, then excludes many CG-sites that are actually playing an active role in gbM methylation dynamics. Due to the cooperative feedback processes, the steady-state methylation level predicted by the model is highly sensitive to the number and positions of CG-sites modelled. It is, therefore, necessary to identify which regions of the genes are 'permissive' to the cooperative feedback interactions, hence the need for a methylatable-regions annotation.

For the whole gbM genes annotation, steady-state (100,000 generations) simulation, we calculated the spatially resolved methylation level, averaged over all gbM genes (Figure S3D). Additionally, a single example gene is shown in Figure S3E, comparing the average methylation across 30 simulated realizations and the observed methylation, averaged across 30 Col-like accessions of each individual CG-site. Towards the center of the example gene, there exist several unmethylated CG-sites that are captured very well by the model. Averaging over all gbM genes (to produce Figure S3D), however, smooths out all these spatial methylation patterns (as the unmethylated sites appear in different locations in every gene), to produce a constant methylation level across the center gene-body. As a result, this gene-averaged methylation profile provides very limited information about the spatial methylation patterns. We do, however, see clear changes in methylation level at the 5' and 3' gene ends, indicating that this must be a generic feature of all modelled genes.

The experimental Col-0 gene-averaged methylation level (Figure S3D) declines towards the 3' end of the genes. This decline is reproduced well by the model, although the absolute level is too high. The model, however, fails to reproduce the observed, and stronger, 5' end decline, instead predicting a peak of mCG aligned with a corresponding enrichment of the CG dinucleotide density (Figures S3D and S3E). Due to the nonlinear cooperative interaction, the over-methylation of 5′ gene ends drives higher average methylation of the entire gene (Figures S3D and S3E). In addition, we note that, as expected, simulating unmethylated genes to steady-state (100,000 generations, using the model parameterization fitted to whole gbM genes, Table S3C) predicts a high methylation contrary to that observed (Figure S3C). These findings indicate that only certain regions of some genes are subject to methylation, and that the epigenetic dynamics that we measure, and model, only apply to these methylatable regions.

The regions that become over-methylated in our simulation (unmethylated genes and 5' gene ends) are enriched for histone variant H2A.Z (Figures S3D–S3F).[60] DNA methylation and H2A.Z are anticorrelated in plants and animals, and they can affect each other's distribution (Figures S3D–S3F).[30,55–60] Consistently, genes over-methylated by the model have relatively high H2A.Z, whereas the genes under-methylated by the model have relatively low H2A.Z compared to the accurately modelled genes (Figure S3G). This raises the possibility that gbM epigenetic dynamics are influenced by H2A.Z. Here, we take an empirical approach of excluding the H2A.Z rich gene-ends from the simulations, implemented via the methylatable-regions annotation described previously.

## Properties of modelled steady-state methylation

Given that models of methylation dynamics applied to mammalian systems[23] show unambiguous bistability of the steady-state methylation, it is instructive to consider the stability of the fluctuating methylation levels around steady-state for the model that we present here. The most immediate difference between these two types of models is in the nature of the feedback interactions. In the mammalian bistable model[23] strong non-linear feedbacks exist in both directions. Existing methylation further enhances methylation gains to stabilize a high mCG level, while unmethylated sites enhance methylation losses to stabilize a low mCG state. In contrast, both of the feedbacks in our present model (cooperative de novo, and cooperative maintenance, Figure 2D) reinforce a methylated state. There is no opposing feedback to reinforce an unmethylated state.

For the present model, the situation is complicated, however, as the scale of the feedbacks are strongly dependent on the number of CG sites within a locus, and their local density. Therefore, for long and CG-rich loci, there is strong feedback reinforcing mCG, thus producing a high mCG level (Figure 4F) with strongly suppressed fluctuations (Figure 6C). This can also be seen in the rightmost example gene of Figure 3, and gives rise to the narrow width of the steady-state mCG level distributions of long loci (Figure S4G, bottom). However, we note that if long loci contain sufficiently wide CG-poor regions, then these areas of much weaker cooperative feedback can produce a partial separation between CG-rich areas. Consequently, large patches of methylation can be lost and sustained in an unmethylated state, independently of neighboring regions which maintain a high mCG level (for example as seen in Figure 3, right).

The time taken to approach steady state for the longest loci is $\sim$ 10 thousand generations (Figure 4D). The simulations to produce the distributions in Figure S4G were run for 10 times this duration. Therefore, if there were any inherent bistability for long loci, there should have been ample time for at least some of the 740 replicates to explore the corresponding lowly methylated state. This would have produced a bimodal distribution of steady-state mCG levels for loci of similar CG-site number and density. Instead, the distributions show a clear single peak with a mean level above 50% (Figure S4G, bottom right plots), despite all simulations being initiated from a fully-unmethylated state. As the panels in Figure S4G group data for multiple genes, we also confirmed the equivalent distributions for a small number of example genes (not shown) are comparable to those of Figure S4G.

As locus length decreases, the mCG-reinforcing feedback decreases, and fluctuation magnitude increases. The distributions of steady-state mCG levels therefore become broader, with a decreased mean value (Figure S4G, middle rows). Very lowly methylated loci (or regions within larger loci) can have long lifetimes, as the rate of spontaneous *de novo* methylation is so low. Hence, patches of methylation can disappear for long times (see Figure 3, right). This spontaneous *de novo* gain rate is much lower than that for the cooperative gains, and therefore causes a separation of timescales, which makes the unmethylated state metastable rather than bistable, due to the lack of any feedback to stabilize it. In comparison, the state where a locus has only one or two CG sites methylated is very unstable. The precise dynamics here will be unique to each locus, given that the gain and loss rates are extremely dependent on the exact CG-site configuration of a locus. For very short loci, this generates a very wide distribution of steady-state mCG levels, with a strong skew towards lowly methylated states (as seen in the top plots of Figure S4G).

## Additional biological insights from modelling

To investigate the relative importance of the primary and secondary components of the cooperative gain interaction, we simulated the number of gains predicted by the model over 30 generations, in methylatable regions, in the absence of the secondary gain pathway (parameterization in Table S3A, but with the secondary cooperative gain component, $r^{(2)}_{\mathrm{Coop}}$, set to zero, Figure S2J). Similarly, the strength of the cooperative maintenance contribution was assessed by simulating the number of losses in methylatable regions predicted by the model in the absence of cooperative maintenance (parameterization in Table S3A, but with $\beta_0 = 0$, Figure S2K). This provides a uniform loss rate at the level of the background loss rate ($\epsilon$), therefore failing to capture the suppression of loss-rate observed at short-length scales. The impact of the various cooperative gain/loss interactions are summarized in Table S5 (again assessed over 30 generations). The cooperative suppression of losses accounts for a ~4-fold loss reduction over 30 generations, whereas cooperative promotion of gains accounts for ~6-fold increase in gains (Table S5). This illustrates that although the gains cooperativity is much stronger over short distances (Figures 2B and 2C), cooperative gain and loss dynamics make comparable contributions to gbM epigenetic inheritance and are therefore of roughly equal importance.

Fitting to the long length-scale tail ($x = 1000$ bp) of the spatially resolved gain-rate distribution provides a good estimate of the background spontaneous *de novo* gain rate. This is of interest because both the methylatable-regions annotation (Figure S2L), and the whole gbM genes annotation can be analyzed with this approach (Figure S2M). For the whole gbM genes annotation, the vast majority of unmethylated CG-sites at such large separations to the nearest *M*-site occur at the 5' and 3' ends (the portions excluded from the methylatable-regions annotation). The analysis, therefore, will give a reasonable estimate of the spontaneous *de novo* rate specifically in these unmethylated gene-ends. In contrast, the analysis for the methylatable-regions provides the spontaneous *de novo* rate within the methylatable-region. Upon fitting the parameters, we find that at this long length-scale, $x \sim 1000$ bp, the contribution from the cooperative gain interactions is negligible. Notably, comparing the spontaneous *de novo* rate calculated for these two regions reveals a ~10 fold difference: $4\times10^{-6}$ per site per cell cycle for methylatable regions (Table S3A), and $5\times10^{-7}$ per site per cell cycle for the gene-ends (found using the whole gbM genes annotation, Table S3C). Interestingly, the methylatable-regions spontaneous *de novo* rate is similar to the gain rate observed in the UM segments within gbM genes in *WT* ($3\times10^{-6}$ per site per cell cycle for the Schmitz dataset, Table S1). However, the spontaneous *de novo* rate found using the whole gbM genes annotation is instead comparable to the gain rate observed in WT UM genes ($6.4\times10^{-7}$ per site per cell cycle for the Schmitz dataset, Table S1). This raises the intriguing possibility that similar methylation dynamics might apply to both the unmethylated ends of gbM genes and to the entirety of UM genes.

## Analysis of UM gene steady state mCG

We have no evidence to support cooperative methylation dynamics in unmethylated genes. Therefore, here we consider the simplest scenario of a uniform, constant, spontaneous gain and loss rate. We denote $\alpha$ to be the haploid gain rate (uCG → mCG), and $\beta$ to be the haploid loss rate (mCG → uCG), where this total loss rate will be a composite of active demethylation and passive losses through maintenance failure. A heterozygous gain (or loss) can occur on either chromosome so that the diploid gain (or loss) rate will be double the haploid rate. We assume that in the natural population outcrossing events are negligibly rare, with plants reproducing by selfing (inbreeding). In which case, only 50% of heterozygous gains, will be converted to homozygous gains through Mendelian segregation (and equivalently for losses). The homozygous gain and loss rates are also therefore represented by $\alpha$ and $\beta$ respectively. The WT gain and loss rates per cell cycle for unmethylated genes are given in Figure 1C: $\alpha = 1.7\times10^{-7}$ and $\beta = 2.1\times10^{-4}$ per cell cycle. The average methylation level across a gene is then described by the ordinary differential equation:

$$\frac{dM}{dt} = \alpha U - \beta M,$$

with the steady-state methylation level given by $\alpha U - \beta M = 0$, where $M$ is the fraction of methylated CG sites, and $U$ is the fraction of unmethylated CG sites in the gene. Assuming that the fraction of hemi-methylated sites is negligible on account of the highly efficient maintenance,[34] such that $U = 1 - M$, the steady-state methylation level, $M^*$, becomes:

$$M^* = \frac{\alpha}{\alpha+\beta}.$$

This provides a steady-state methylation of $M^* = 8 \times 10^{-4}$, corresponding to a methylation level of ~ 0.1%.

### Assessing the scale of simulated methylation fluctuations

The fluctuations in methylation level were investigated by simulating 740 replicates in methylatable regions, starting from the all-*U* initial state each time. The time of the first methylation gain was recorded for each locus, and averaged over the 740 replicates. This is compared to the theoretically expected value of: $1/(N_{CG}\, r_0^+)$ cell cycles, or $1/(N_{CG}\, r_0^+\, n_{cc})$generations (red curve in Figure 4C). A steady-state methylation level was found for each replicate of each locus, by first simulating (from the all-*U* initial state) for an equilibration-time of 50,000 generations (ample time to reach steady-state, Figure 4B). The simulation was then continued for another 50,000 generations (denoted by $T$ below), over which the overall methylation level of the locus, $\langle M \rangle$, was time-averaged to find the steady-state methylation level, $\overline{\langle M \rangle}$, where:

$$\overline{\langle M \rangle} = \frac{1}{T}\int_0^T \langle M \rangle\, dt.$$

The time for an individual locus to first reach steady-state is then defined as the time the methylation level of that locus first reached the value of $\overline{\langle M \rangle}$. This calculation is repeated independently for each replicate of each locus, with the mean and standard deviation over replicates then shown in Figures 4D and S3J respectively. An estimate of the typical magnitude of fluctuations away from the mean steady-state methylation level is found from the standard deviation of the methylation level over time, $\Sigma$, where:

$$\Sigma^2 = \frac{1}{T}\int_0^T \langle M \rangle^2 dt - (\overline{\langle M \rangle})^2$$

Again, this quantity is calculated individually for each replicate of each locus, with the mean over replicates shown in Figure S6C. Finally, we study the 'greatest' fluctuation occurring in the second 50,000 generations of the above 100,000 generation simulations. Here we use 'greatest' to refer to the fluctuation showing the largest deviation in methylation level from the time-averaged mean. Both the magnitude and duration of this fluctuation are recorded for each replicate of each locus, with the mean and standard deviation over replicates shown in Figures 6A and S6A respectively for the magnitude, and in Figures 6B and S6B respectively for the duration. Hence, the magnitude of the greatest fluctuation refers to the largest departure from the time-averaged mGG level over the second 50,000 generation simulation period (where mCG level refers to the mean mCG fraction over an individual locus). Similarly, the duration of the greatest fluctuation refers to the total time (in generations) over which the mCG level departed from the mean, for the specific fluctuation found to be the greatest magnitude. We note that the magnitude of the greatest fluctuation will depend on the length of the time-interval, $T$, that is studied: the longer the time-interval, the higher the likelihood that more extreme fluctuations will be observed.

### Fits to mutant mean gain and loss rates over methylatable regions

The large fluctuations and insufficient range, particularly in the observed mutant loss rate distributions (as a function of distance to the nearest *M*-site), preclude a full fitting to the gain/loss rate distributions (Figures 2B and 2C). Instead, we therefore used the observed mean gain/loss rate over methylatable regions (Table S1). Although the simulated mean gain/loss rates for the model fit to Col-0 are close to those found for our *WT* data set (Figure 2G; Tables S1 and S5), there are still ~10% discrepancies. We therefore calculated a target mean gain and loss rate for each mutant (Table S6) by applying the same relative change from the simulated mean *WT* rates as is found from comparing the mean rate for each mutant to the corresponding newly measured mean *WT* rates (as these were all measured using a consistent experimental design). The simulated gain/loss rate distributions were found by fitting the overall cooperativity strength ($r^*$), and the background loss rate ($\epsilon$), to the target mean gain/loss rates.

As relatively recently generated mutants are used, we assume that their methylation state is still very close to that of the Col-0 consensus state. Consequently, as before, we use the Col-0 consensus state for the initial state and simulate for 30 generations, using 100 replicates to find the simulated mean gain/loss rates over methylated regions. Parameters are as for the full model (Table S3A) but with the overall cooperativity strength ($r^*$) and background loss rate ($\epsilon$) adjusted according to Table S6. For *h2az*, caps were introduced to the cooperative interaction to ensure that $0 < r^+_{\mathrm{Coop}} < 1$ and: $0 < \gamma < 1$. Using these parameter values, the steady-state methylation level distribution, and the time to reach steady-state were then found for each mutant, similarly to the description above for Col-like accessions, though only using 30 modelled replicates for each simulation and from the Col-0 consensus initial state.

Finally, we note that for the *ros1* mutant, we could alternatively produce a very similar fit to that produced by altering the overall cooperativity strength by adjusting the spontaneous *de novo* rate ($r_0^+$) instead of the overall cooperativity strength. These results are not shown, as they did not appreciably alter the steady-state state methylation level. We emphasize that all fits to mutant rates are tentative due to the restricted amount of data available for the fit (only a single average gain and loss rate per mutant, as opposed to the spatially resolved Col-0 gain/loss rate distributions and the steady-state methylation levels available for the *WT* population). Long-timescale (steady-state) simulations only reflect the direct consequences of the altered short-timescale (30 generation) methylation dynamics. Time to steady-state calculations are estimates, as it is unknown which side of the steady-state each locus is in the initial state and therefore loci are assumed to transition in the direction of the overall mean of the distribution. Any potential indirect effects, such as altered methyltransferase targeting are not accounted for.

### Simulations using randomised CG-site positions

The stochastic model indicates that both the observed enhancement of mCG level in exons vs introns (Figure 5D) and the mCG enhancement under well-positioned nucleosomes (Figure 5E) are both driven by the local CG density.

To confirm this, we conducted additional simulations using randomized CG-positions to act as a control. Methylatable regions were simulated as described previously. The start and end genomic positions of each methylatable region were enforced exactly as previously. Additionally, the experimentally observed number of CG sites (for Col-0) was used for each locus. Within every locus, the experimentally prescribed number of CG sites were randomly allocated a position within that locus, with the constraint that no two CG sites could overlap (each CG site occupies exactly 2 bp). Stochastic simulations were then performed as previously (model parameters in Table S3A): 30 replicates were simulated from a fully unmethylated initial state for 100,000 generations and the final methylation state recorded.

Randomizing CG site positions within each individual locus produces a relatively modest adjustment to the CG-CG site pair cross correlation function (see Figure 5C for experimental CG site positions compared to Figure S5C for randomized CG sites). The corresponding change to the simulated M-M pair cross correlation function is minimal (Figure S5C, blue for experimental CG site positions and black for randomized CG sites). This is to be expected given that the gain and loss interaction profiles are unchanged, while the changes to the CG site distribution are subtle.

The effect of randomizing CG sites on the mCG level of exons vs introns however is stark. The mCG level distributions for exons and introns (averaged over the 30 simulated replicates) were calculated as for Figure 5D but with a slight modification. In cases where no CG sites lie within either the annotated exons or annotated introns for a particular locus, the locus was then excluded (n=1368, unlike for Figure 5D), to ensure an equal number of exon and intron regions were compared. Randomizing the CG site positions removes the enhanced CG density found for exons (Figure S1B, right). The resulting mCG level distributions for exons and introns then become equivalent (Figure S5D), confirming that the enhancement of mCG level in exons compared to introns was indeed driven by the greater CG site density of exons.

Similarly, the mCG level averaged over well positioned nucleosomes (after using randomized CG site locations) was calculated equivalently to in Figure 5E. The fluctuations in the simulated mCG level in Figure S5E no longer reproduce the mCG oscillations observed for Col-0. Instead, the simulated mCG level now fluctuates erratically (Figure S5E). Randomly generated CG site distributions create regions of higher or lower CG site density. However, these regions will no longer be aligned with the nucleosome positions, which normally center on CG-dense DNA, and will no longer be correlated with exons/introns. We, therefore, interpret the peaks and troughs in simulated mCG level (Figure S5E) as reflecting these regions of randomly generated high and low CG density.

## QUANTIFICATION AND STATISTICAL ANALYSIS

All stochastic modelling was performed using python via the distribution: anaconda3-5.2.0, which included the following libraries: dask; matplotlib; numpy; pandas; scipy. Analysis of aligned bisulfite data was performed using R-3.6.0, including the following libraries: tidyr, GenomicRanges, data.table scales, stringr, ggplot2, and SeqMonk.

All details of statistical tests, including replicate numbers and p-values are included in the appropriate Figures, Figure legends and Tables, with further information provided in the methods details.