Rate, spectrum, and evolutionary dynamics of spontaneous epimutations

Adriaan van der Graaf[a,1], René Wardenaar[a,1], Drexel A. Neumann[b], Aaron Taudt[c], Ruth G. Shaw[d], Ritsert C. Jansen[a], Robert J. Schmitz[b,2], Maria Colomé-Tatché[c,2], and Frank Johannes[a,2]

[a]Groningen Bioinformatics Centre, University of Groningen, 9747 AG Groningen, The Netherlands; [b]Department of Genetics, University of Georgia, Athens, GA 30602; [c]European Institute for the Biology of Aging, University of Groningen, University Medical Centre Groningen, 9713 AV Groningen, The Netherlands; and [d]Department of Ecology, Evolution and Behavior, University of Minnesota, Minneapolis, MN 55455

Edited by James A. Birchler, University of Missouri–Columbia, Columbia, MO, and approved April 14, 2015 (received for review December 19, 2014)

Stochastic changes in cytosine methylation are a source of heritable epigenetic and phenotypic diversity in plants. Using the model plant *Arabidopsis thaliana*, we derive robust estimates of the rate at which methylation is spontaneously gained (forward epimutation) or lost (backward epimutation) at individual cytosines and construct a comprehensive picture of the epimutation landscape in this species. We demonstrate that the dynamic interplay between forward and backward epimutations is modulated by genomic context and show that subtle contextual differences have profoundly shaped patterns of methylation diversity in *A. thaliana* natural populations over evolutionary timescales. Theoretical arguments indicate that the epimutation rates reported here are high enough to rapidly uncouple genetic from epigenetic variation, but low enough for new epialleles to sustain long-term selection responses. Our results provide new insights into methylome evolution and its population-level consequences.

epigenetics | epimutation | DNA methylation | evolution | *Arabidopsis*

Plant genomes make extensive use of cytosine methylation to control the expression of transposable elements (TEs) and genes (1). Despite its tight regulation, methylation losses or gains at individual cytosines or clusters of cytosines can emerge spontaneously, in an event termed "epimutation" (2, 3). Many examples of segregating epimutations have been documented in experimental and wild populations of plants and in some cases contribute to heritable variation in phenotypes independently of DNA sequence variation (4, 5). These observations have led to much speculation about the role of DNA methylation in plant evolution (6–8), and its potential in breeding programs (9). In the model plant *Arabidopsis thaliana*, spontaneous methylation changes at CG dinucleotides accumulate in a rapid but nonlinear fashion over generations (2, 3, 10), thus pointing to high forward–backward epimutation rates (11). Precise estimates of these rates are necessary to be able to quantify the long-term dynamics of epigenetic variation under laboratory or natural conditions, and to understand the molecular mechanisms that drive methylome evolution (12–14). Here we combine theoretical modeling with high-resolution methylome analysis of multiple independent *A. thaliana* mutation accumulation (MA) lines (15), including measurements of methylation changes in continuous generations, to obtain robust estimates of forward and backward epimutation rates.

Results

We joined whole-genome MethylC-seq (16) data from two earlier MA studies (2, 3) with extensive multigenerational MethylC-seq measurements from three additional MA lines (Fig. 1*A* and *SI Appendix*, Tables S1–S6). The first of these new MA lines (MA1_3) was propagated for 30 generations and includes measurements for 13 (nearly) consecutive generations (Fig. 1*A*). The other two MA lines (MA2_3) were propagated for 17 generations and were measured every four generations on average (Fig. 1*A*). These new data therefore allowed us to track epimutation dynamics over a large number of generations and at high temporal resolution. We constructed base pair-resolution methylation maps for all sequenced individuals (*SI Appendix*). To obtain a measure of genome-wide methylation divergence between any two individuals in a given MA pedigree, we calculated the proportion of differentially methylated cytosines in sequence contexts CG, CHG, and CHH (where H can be any base but G). For these calculations we used a set of consensus cytosines for which all individuals in the pedigrees had coverage of more than three reads (*SI Appendix*). This read coverage cutoff was found to be sufficient for robust downstream analyses (*SI Appendix*, Figs. S1 and S2). Consistent with previous reports (2, 3, 10), genome-wide methylation divergence at CG dinucleotides increased with divergence time in all pedigrees (Fig. 1*B*), but not in sequence contexts CHG and CHH (*SI Appendix*, Fig. S3). This distinction reflects intrinsic differences in the maintenance pathways that target these three contexts (1) and possibly also increased measurement error and cellular heterogeneity for non-CG methylation (*SI Appendix*, Fig. S4).

Neutral Epimutation Model. To quantify CG methylation divergence in the MA lines as a function of divergence time (measured in generations) and forward–backward epimutation rates, we developed a theoretical model similar to those used in the analysis of regular systems of inbreeding (*Materials and Methods* and *SI Appendix*). Briefly, the model assumes that an unmethylated

Significance

Changes in the methylation status of cytosine nucleotides are a source of heritable epigenetic and phenotypic diversity in plants. Here we derive robust estimates of the rate at which cytosine methylation is spontaneously gained (forward epimutation) or lost (backward epimutation) in the genome of the model plant *Arabidopsis thaliana*. We show that the forward–backward dynamics of selectively neutral epimutations have a major impact on methylome evolution and shape genome-wide patterns of methylation diversity among natural populations in this species. The epimutation rates presented here can serve as reference values in future empirical and theoretical population epigenetic studies in plants.

Author contributions: R.J.S., M.C.-T., and F.J. designed research; A.v.d.G., R.W., D.A.N., A.T., R.J.S., M.C.-T., and F.J. performed research; D.A.N., R.G.S., R.C.J., and R.J.S. contributed new reagents/analytic tools; A.v.d.G., R.W., A.T., M.C.-T., and F.J. analyzed data; and M.C.-T. and F.J. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The sequence reported in this paper has been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE64463).

[1]A.v.d.G. and R.W. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: schmitz@uga.edu, m.colome.tatche@umcg.nl, or frank@johanneslab.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1424254112/-/DCSupplemental.

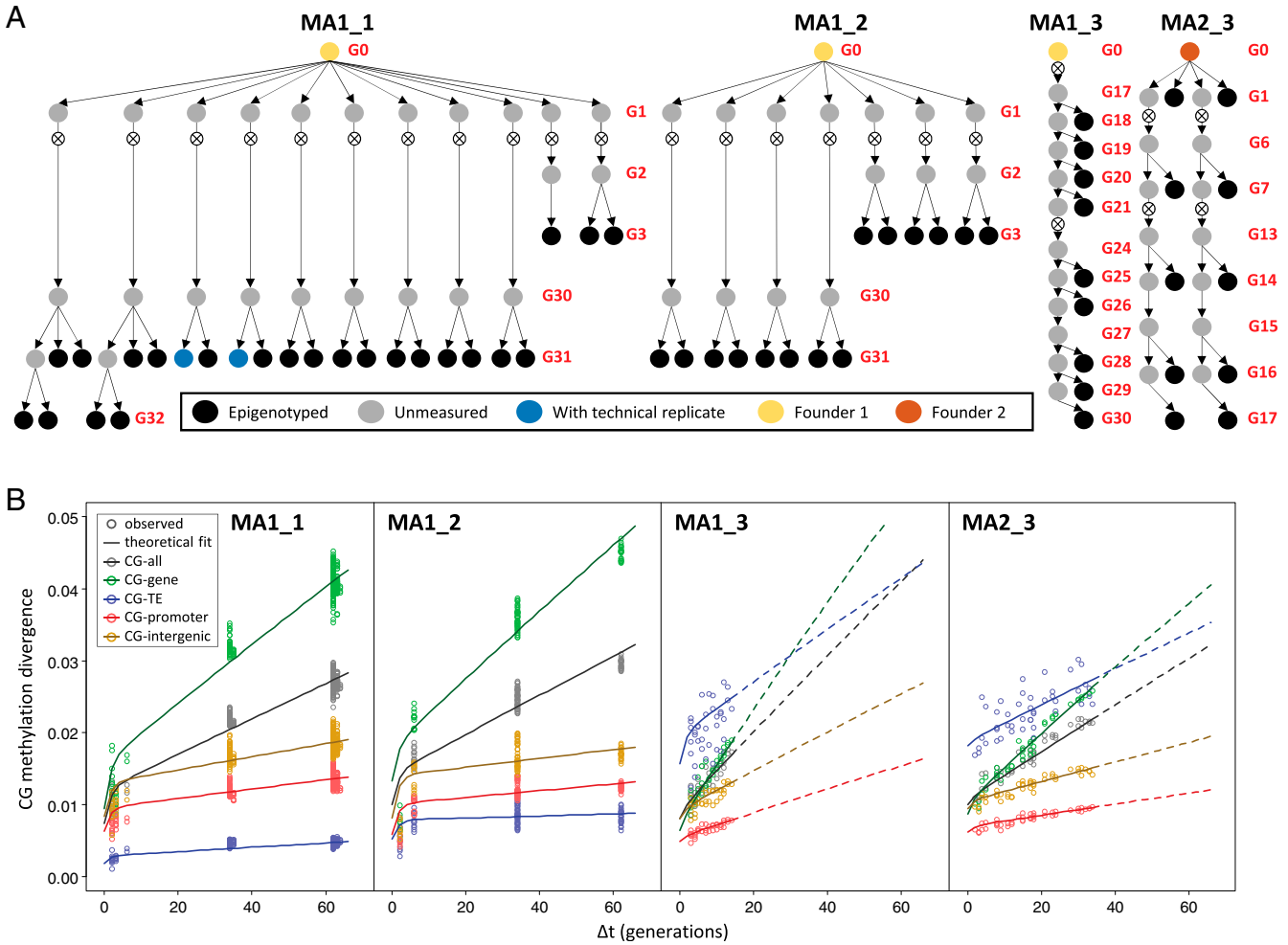


Fig. 1. (*A*) Overview of pedigrees of mutation accumulation lines (MA lines). Red numbers indicate the number of generations from common founder. The MA1_1 and MA1_2 lines were originally created by Shaw et al. (15) and their methylomes were presented in Becker et al. (2) and Schmitz et al. (3), respectively. The MA1_3 and MA2_3 were generated in this study. (*B*) Measured CG methylation divergence (circles) with corresponding theoretical fits (lines) as a function of divergence time between two individuals in a given pedigree (Δt = total divergence time in generations between any two individuals). Dashed lines indicate extrapolation from the fitted model. Divergence values for CG-TE in datasets MA1_3 and MA2_3 were susceptible to low sequencing depth in these experiments and showed increased measurement noise (*SI Appendix*). Model-based analysis of all MA pedigrees revealed that the highly nonlinear divergence until generation eight is due to the fixation of segregating epi-heterozygote founder loci (*SI Appendix*), rather than the result of recurrent cycles of forward and backward epimutations as previously suggested (11).

cytosine (c^u) can become methylated (c^m) with probability α and likewise a methylated cytosine can become unmethylated with probability β. We arbitrarily define α as the forward and β as the backward epimutation rate per generation per haploid methylome. Transitions of diploid epigenotypes (c^mc^m, c^uc^u, or c^mc^u) from one generation to the next are modeled through a transition matrix where the elements of the matrix are determined by the Mendelian segregation of epialleles (c^m or c^u) and the rates α and β. Consistent with the MA experimental design, selection on epigenotypes during inbreeding is assumed to be absent, so that the cumulative divergence among lines is driven solely by neutral epigenetic drift. Estimates for the unknown epimutation rates are obtained by fitting our model to the CG methylation divergence data of each MA pedigree separately (*Materials and Methods*).

Estimates of Global CG Epimutation Rates. As shown in Fig. 1*B*, our model provides an excellent fit to the data, which suggests that the observed divergence patterns among the MA lines are largely the result of the transgenerational accumulation of selectively neutral epimutations. Model-based estimates for the forward and backward CG epimutation rates (CG-all) were $2.56 \cdot 10^{-4}$ and $6.30 \cdot 10^{-4}$, respectively (Table 1 and *SI Appendix*, Table S7). These estimates are similar to the value provided by Schmitz et al. (3) ($4.46 \cdot 10^{-4}$) but illustrate that methylation loss at CG dinucleotides is globally three times as likely as methylation gain. The ratio of loss to gain (β/α), also known as the mutational bias parameter, is an important quantity: It determines the CG methylation content of the *A. thaliana* genome over evolutionary timescales. Assuming that the *A. thaliana* methylome is at equilibrium, the estimated CG forward–backward epimutation rates imply that—in the absence of selection or gene conversion—about 30% of all CG sites should be methylated and about 70% are unmethylated, which is consistent with actual measurements (16, 17).

Estimates of Annotation-Specific CG Epimutation Rates. We examined the extent to which CG epimutation rates depend on genomic context. To do this, we separated all CGs according to annotation (gene bodies, promoters, TEs, and intergenic regions; see *SI Appendix*). Although annotation-specific CG epimutation rates were approximately within the same order of magnitude (Table 1 and *SI Appendix*, Table S7), subtle differences in these rates had a substantial impact on differential divergence of CG methylation

Table 1. Estimates of forward and backward epimutation rates

Context	α	Range (α)		β	Range (β)		β/α	Range (β/α)	
CG-all	$2.56 \cdot 10^{-4}$	$2.08 \cdot 10^{-4}$	$3.69 \cdot 10^{-4}$	$6.30 \cdot 10^{-4}$	$3.23 \cdot 10^{-4}$	$1.13 \cdot 10^{-3}$	2.36	1.55	3.24
CG-gene	$3.48 \cdot 10^{-4}$	$2.77 \cdot 10^{-4}$	$4.87 \cdot 10^{-4}$	$1.47 \cdot 10^{-3}$	$9.46 \cdot 10^{-4}$	$2.45 \cdot 10^{-3}$	4.24	2.84	5.10
CG-TE*	$3.24 \cdot 10^{-4}$	$1.68 \cdot 10^{-4}$	$4.80 \cdot 10^{-4}$	$1.20 \cdot 10^{-5}$	$7.76 \cdot 10^{-6}$	$1.62 \cdot 10^{-5}$	0.040	0.034	0.046
CG-promoter	$5.17 \cdot 10^{-5}$	$2.92 \cdot 10^{-5}$	$9.33 \cdot 10^{-5}$	$5.88 \cdot 10^{-4}$	$1.33 \cdot 10^{-4}$	$1.40 \cdot 10^{-3}$	11.4	4.16	15.08
CG-intergenic	$1.15 \cdot 10^{-4}$	$6.13 \cdot 10^{-5}$	$1.70 \cdot 10^{-4}$	$3.25 \cdot 10^{-4}$	$6.36 \cdot 10^{-5}$	$7.69 \cdot 10^{-4}$	2.83	0.47	4.80

We assume that an unmethylated cytosine (c^u) can become methylated (c^m) with probability α, and likewise a methylated cytosine can become unmethylated with probability β. We arbitrarily define α as the forward and β as the backward epimutation rate per generation per haploid methylome. Shown are model-based estimates for α and β as an average of the MA1_1, MA1_2, MA1_3, and MA2_3 datasets, as well as the range of these estimates across datasets (range). The asterisk indicates that the average estimate was based only on the MA1_1 and the MA1_2 data (*SI Appendix*). These estimates can be considered robust, because the different MA pedigrees varied considerably in terms of plant material, growth conditions, and sequencing approach (*SI Appendix*, Table S1).

across annotation categories (Fig. 1*B*). The highest combined forward and backward rates were found for CGs in gene bodies (CG-gene), which were $3.48 \cdot 10^{-4}$ and $1.47 \cdot 10^{-3}$, respectively (Table 1 and *SI Appendix*, Table S7). By contrast, the lowest rates were found for CGs in TEs (CG-TEs, forward: $3.24 \cdot 10^{-4}$ and backward: $1.20 \cdot 10^{-5}$). As a result of these low epimutation rates, methylation divergence for CG-TEs was much less pronounced (Fig. 1*B*), resembling the divergence patterns seen for CHG and CHH contexts (*SI Appendix*, Fig. S3). This observation suggests that CG-TEs come under the influence of silencing pathways that primarily target neighboring CHHs and CHGs (18–20). Indeed, CG-TE was the only annotation category in which the ratio of backward to forward epimutation rates was less than unity (Table 1 and *SI Appendix*, Table S7), which implies that gain of methylation is strongly favored over methylation loss.

Genome Architecture and Chromatin Environment Predict CG Methylation Divergence Patterns Along Chromosomes. Because CG epimutation rates are annotation-specific, we predicted that methylation divergence closely tracks annotation density along chromosomes. To test this, we moved in a 1-Mb sliding window along the genome (step size 100 kb) and calculated the divergence between MA lines as expected from our model after 31 generations of independent selfing (Fig. 2*B* and *SI Appendix*). Our calculations predicted that CG-methylation divergence is low in TE-rich pericentromeric regions and high in gene-rich chromosome arms (Fig. 2*B* and *SI Appendix*, Figs. S5 and S6). Remarkably, these predictions strongly agreed with the observed divergence patterns at the genome-wide scale ($R^2 = 0.74$, $P < 0.0001$).

An alternative, or complementary, explanation is that the annotation-specific divergence patterns are simply a reflection of the genome-wide distribution of heterochromatic domains, which would explain the clear partitioning between pericentromeres and euchromatin. To test this directly, we reanalyzed recent ChIP-seq data on histone variant H2A.W (21), a proxy for heterochromatin, and estimated epimutation rates for CGs in regions that were either enriched or depleted for H2A.W (*SI Appendix*). We used these rates in combination with the genome-wide density distribution of H2A.W to derive predictions of CG-methylation divergence patterns. Our analysis revealed that, at the genome-wide scale, heterochromatin-based predictions were approximately equivalent to annotation-based predictions ($R^2 = 0.72$, $P < 0.0001$, *SI Appendix*, Fig. S5), suggesting that chromatin environment is a sufficient and parsimonious explanation for the observed divergence patterns along chromosomes. These results further indicate that the maintenance of methylation at CG dinucleotides is slightly more error-prone in regions of open chromatin compared with more compact regions, probably as a by-product of active transcription.

The Spectrum of Neutral Epimutations Shapes CG Methylation Diversity in Natural Populations. An intriguing question is to what extent the epimutation landscape in the MA lines provides insights into the mechanisms that shape CG methylation diversity in *A. thaliana* natural populations, which are the outcome of long and complex evolutionary processes. To assess this we reanalyzed MethylC-seq data from a large number of accessions collected from across the Northern Hemisphere (22) (*SI Appendix*, Table S8). We focused on a subset of 133 accessions that met our quality criteria and calculated CG-methylation diversity in a 1-Mb sliding window using the same protocol as with the MA lines (*SI Appendix*). Although the natural accessions were clearly more diverse (Fig. 2*B*), genome-wide diversity patterns were highly similar to those seen in the MA lines (weighted $R^2 = 0.624$, $P < 0.0001$, Fig. 2*C* and *SI Appendix*, Fig. S7), particularly in pericentromeric regions ($R^2 = 0.899$, $P < 0.0001$) and to a slightly lesser extent in chromosome arms ($R^2 = 0.525$, $P < 0.0001$). These observations are consistent with a recent report by Hagmann et al. (23). Moreover, CG-methylation divergence among the MA lines was also moderately correlated with sequence diversity in the accessions, explaining over 25% of the genome-wide SNP distribution (weighted $R^2 = 0.254$, $P < 0.0001$, Fig. 2*D* and *SI Appendix*, Fig. S8).

It is unlikely that global patterns of CG-methylation diversity among natural accessions are the result of selection acting over broad genomic regions, because the same patterns are quickly established in isogenic MA lines in the course of only 31 generations under constant environmental conditions. Rather, our results suggest that these patterns reflect major structural properties of the *A. thaliana* genome, which modulate the ratio of forward–backward epimutation rates, and thus determine the accumulation dynamics of neutral epimutations over time. It is therefore not surprising that the reorganization of genomes during macroevolution is necessarily accompanied by a repatterning of methylation divergence among lineages or species (24), insofar that such structural changes alter genome-wide annotation densities and their accompanying chromatin environment. However, structural changes of this type are less prevalent in the course of microevolution; hence, neutral epimutations are probably the single most important factor in shaping methylome diversity in populations over short to intermediate evolutionary timescales.

Discussion

CG Epimutation Rates Are High Enough to Rapidly Uncouple Genetic and Epigenetic Variation over Evolutionary Timescales. Our analysis shows that CG epimutations are about five orders of magnitude more frequent than genetic mutations in *A. thaliana* [$\sim 10^{-4}$ compared with $\sim 10^{-9}$ (25)] and are subject to forward–backward dynamics that are rarely observed for genetic loci. Because of these properties, it is intuitively obvious that these epimutation dynamics will lead to an uncoupling of epigenetic from genetic variation over relatively short evolutionary timescales (26). Simple deterministic models show that in a strictly selfing system without selection it would require only about 800 generations to reduce correlations between genotype and epigenotype from unity to below 0.5,

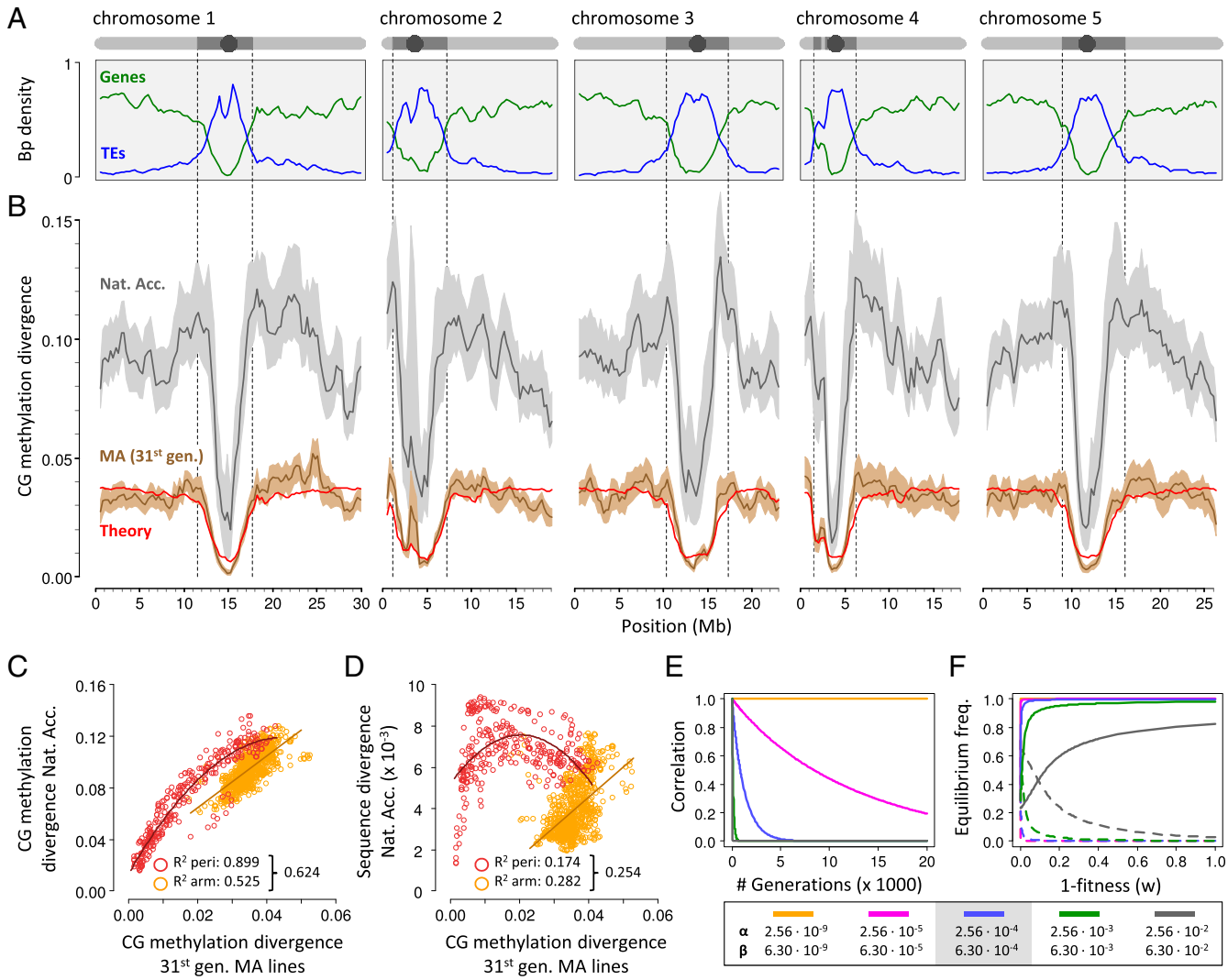


Fig. 2. (*A*) Genome-wide gene (green) and TE (blue) density as well as a schematic representation of chromosomes (circle, centromere; dark gray, pericentromeric region; light gray, arm). (*B*) Genome-wide CG methylation divergence patterns among the 31st generation MA lines (MA1_1 and MA1_2) and the natural accessions (brown and gray, respectively). The red line indicates the theoretical prediction of divergence based on the estimated epimutation rates per annotation weighted by local annotation densities. Genome-wide divergence patterns in the MA lines are strongly correlated with the diversity patterns in the natural accessions (*C*), as well as with sequence diversity in the accessions (*D*). (*E*) The relationship between genetic and epigenetic variation as a function of time and different values of the epimutation rates α and β in a strictly selfing system without selection; *x* axis, time in generations; *y* axis, expected correlation between genotype and epigenotype of two perfectly linked loci (recombination fraction = 0). The estimated CG epimutation rate (blue line) is high enough to efficiently uncouple genetic from epigenetic variation over relatively short timescales (*SI Appendix*). (*F*) Equilibrium frequency (*y* axis) of epigenotypes $c^m c^m$ (solid) and $c^u c^u$ (dashed) in a strictly selfing system as a function of fitness (*x* axis) and different forward–backward epimutation rates (colored lines). The fitness of epigenotypes $c^u c^u$, $c^u c^m$ and $c^m c^m$ was defined by *w*, $0.5(1+w)$ and 1, respectively (*SI Appendix*). The estimated CG epimutation rate (blue line) is low enough to yield epimutation-selection equilibria close to those found for DNA sequence mutation rates, even under weak selection regimes (i.e., small fitness differentials between $c^u c^u$ and $c^m c^m$). This means that CG-type epialleles should be stable enough to effectively respond to long-term selection, provided they affect fitness.

and only about 2,700 generations to reduce it to below 0.1 (Fig. 2*E* and *SI Appendix*), and this breakdown is expected to be even faster in outcrossing systems. This rapid uncoupling may explain why variation in DNA methylation in *A. thaliana* populations is only partly associated with *cis*- and *trans*-acting DNA sequence variants (22), and thus sheds new light on the molecular mechanisms that drive the coevolution of genomes and epigenomes. One situation in which genotype-epigenotype associations are expected to be more prevalent is when natural accessions have only newly diverged from a common ancestor, as it may be the case in recently founded local populations. This prediction can be tested using genome-wide association study-based *cis*- and *trans*-mapping analysis across different groups of accessions that vary along a gradient of genetic relatedness and (or) geographic locations. Recent whole-genome and methylome datasets of *A. thaliana* local populations collected in North America (23) and Sweden (27) may be applicable for that purpose.

CG Epimutation Rates Are Low Enough for New Epialleles to Sustain Long-Term Selection Responses. Although our results provide strong evidence that global patterns of CG-methylation diversity among *A. thaliana* natural accessions are mainly influenced by the accumulation of selectively neutral epimutations, targeted selection of epialleles at specific loci may still be an important process. Particularly in chromosome arms, the MA divergence patterns were only moderately correlated with those of the natural accessions, suggesting the involvement of other factors such as direct selection on CG methylation states and (or) selection via DNA sequence

variants that indirectly regulate CG methylation in *cis* or *trans*. Indeed, the observation that methylation profiles of orthologous genes is often highly conserved across species (28) indicates that some epigenetic states are subject to strong evolutionary constraints. For epigenetic selection to be effective, epimutations need to be sufficiently stable (29), and a lack of stability has been cited as one reason why epigenetic inheritance has no potent role in evolution or in the heritability of complex traits (30). Contrary to these conclusions, simple deterministic selection models show that newly arising epimutations are stable enough to respond effectively to long-term selection, even under weak selection regimes, yielding epimutation-selection equilibria that are close to those expected for DNA sequence mutation rates (Fig. 2*F* and *SI Appendix*).

Reference Values for Future Population Epigenetic Studies. In light of our estimates of forward–backward epimutation rates, future work should examine the effect of selection in more complex population genetic models that account for finite population sizes, migration, and drift such as those proposed by Charlesworth and Jain (13). Recently, Wang and Fan (14) devised a neutrality test based on single methylation polymorphism data using a modified version of Tajima's *D*. We caution that care needs to be taken when supplying epimutation rates to this or similar tests. Incorrect assumptions about the ratio of forward and backward rates can lead to widely misleading conclusions regarding the role of selection on CG methylation. If one assumes that forward and backward rates are equivalent, TE-associated CGs would most likely be detected as being under strong selection, and pericentromeric regions would seem to have undergone selective sweeps. However, if one considers that spontaneous methylation gain is about 30 times more likely than methylation loss (see Table 1), equilibrium levels of CG-methylation diversity in TEs would seem to be entirely consistent with neutrality. Hence, the context- or annotation-specific epimutation rates provided here should serve as useful reference values when inferring signatures of epigenetic selection in *A. thaliana* and possibly in other plant species.

Materials and Methods

Below we provide a brief description of the theoretical model and our estimation approach. For a more detailed explanation we refer the reader to *SI Appendix*.

Derivation of Neutral Epimutation Model. Let c^u and c^m denote an unmethylated and a methylated cytosine, respectively, and $\alpha = Pr(c^u \to c^m)$ and $\beta = Pr(c^m \to c^u)$ be the probabilities that a cytosine gains or loses methylation during or before gamete formation, which can include gains or losses of DNA methylation in somatic tissues from which the gametic cells were derived. We arbitrarily call α the forward and β the backward epimutation rate per generation per haploid methylome. We modeled the epigenotype frequencies at the *j*th cytosine using a Markov chain with three states: c^uc^u, c^uc^m, and c^mc^m. Taking into account Mendelian segregation of epialleles c^m and c^u together with rates α and β, we derived the epigenotype transition matrix **T** after one selfing generation:

	c^uc^u	c^mc^u	c^mc^m
c^uc^u	$(1-\alpha)^2$	$2(1-\alpha)\alpha$	α^2
c^uc^m	$\frac{1}{4}(\beta+1-\alpha)^2$	$\frac{1}{2}(\beta+1-\alpha)(\alpha+1-\beta)$	$\frac{1}{4}(\alpha+1-\beta)^2$
c^mc^m	β^2	$2(1-\beta)\beta$	$(1-\beta)^2$

This formulation does not account for higher-order epimutation events, because such events are expected to be rare for small epimutation rates. Following Markov chain theory, the epigenotype frequencies at cytosine *j* in the MA population after *t* generations of single seed descent, π_{tj}, can be expressed as $\pi_{tj} = \pi_{0j}\mathbf{P}\mathbf{V}^t\mathbf{P}^{-1}$, where *P* is the eigenvector of matrix **T** and **V** is a diagonal matrix of the eigenvalues of matrix **T**. Using Mathematica 10.0 (Wolfram Research, Inc.) we derived analytical solutions for the elements of π_{tj}, which are functions of *t*, α, β as well as the initial frequency vector π_{0j}. These analytical solutions have no easy form and are therefore omitted here for brevity. At equilibrium, the $\pi_{\infty j}$ represent the expected epigenotype frequencies at cytosine *j* among the MA lines after a (hypothetical) infinite number of selfing generations $(t=\infty)$, and were obtained by calculating $\lim_{t\to\infty}\pi_{tj}$:

$$\pi_{\infty j}(c^uc^u) = \frac{\beta\left((1-\beta)^2-(1-\alpha)^2-1\right)}{(\alpha+\beta)\left((\alpha+\beta-1)^2-2\right)}$$

$$\pi_{\infty j}(c^uc^m) = \frac{4\alpha\beta(\alpha+\beta-2)}{(\alpha+\beta)\left((\alpha+\beta-1)^2-2\right)}$$

$$\pi_{\infty j}(c^mc^m) = \frac{\alpha\left((1-\alpha)^2-(1-\beta)^2-1\right)}{(\alpha+\beta)\left((\alpha+\beta-1)^2-2\right)}.$$

For any $0<\alpha,\beta<1$, these equilibrium solutions are independent of the initial epigenotype proportions π_{0j} in the common founder, and depend only on the rates α and β. The rate at which the epigenotype proportions converge to these equilibrium values depends on the relative magnitude of the forward and backward rates.

Modeling Methylation Divergence. To derive analytical formulas for methylation divergence, we score the methylation divergence between two independently selfed lines at every cytosine with the following distance matrix:

	c^uc^u	c^mc^u	c^mc^m
c^uc^u	0	$\frac{1}{2}$	1
c^uc^m	$\frac{1}{2}$	0	$\frac{1}{2}$
c^mc^m	1	$\frac{1}{2}$	0

Let t_1 and t_2 denote the number of generations between two individuals at generations G_m and G_n and their most recent common founder at generation G_f, respectively (i.e., $t_1 = G_m - G_f$, $t_2 = G_n - G_f$, Fig. 1*A*). Let $\pi_{t_ij}|c^mc^m$ be the vector of epigenotype frequencies at the *j*th cytosine after t_i selfing generations from G_f, conditional on the fact that the most recent common founder epigenotype was c^mc^m: $\pi_{t_ij}|c^mc^m = (0,0,1)\cdot T^{t_i}$. Using this equation and the methylation divergence scoring table above, the divergence between these two lines at this locus can be calculated as

$$d_{t_1t_2j}|c^mc^m = \frac{1}{2}\sum_{k=1}^{4}\left(\pi_{t_1j}(P1_k)|c^mc^m\cdot\pi_{t_2j}(P2_k)|c^mc^m\right) + 1\sum_{k=1}^{2}\left(\pi_{t_1j}(Q1_k)|c^mc^m\cdot\pi_{t_2j}(Q2_k)|c^mc^m\right),$$

with $Q1=\{c^uc^u, c^mc^m\}$, $Q2=\{c^mc^m, c^uc^u\}$, $P1=\{c^uc^u, c^uc^m, c^uc^m, c^mc^m\}$, and $P2=\{c^uc^m, c^uc^u, c^mc^m, c^uc^m\}$. The simple multiplication of these frequencies follows from the fact that the selfing lines are conditionally independent. The divergence over all loci for which the most recent common founder at G_f was c^mc^m is

$$d_{G_f,t_1t_2}|c^mc^m = \sum_j d_{t_1t_2j}|c^mc^m = N^{mm}_{G_f}\cdot d_{t_1t_2j}|c^mc^m,$$

where $N^{mm}_{G_f}$ are the number of methylated cytosines at G_f. The global (or total) DNA methylation divergence along the genome can be calculated as

$$D_{G_f,t_1t_2} = d_{G_f,t_1t_2}|c^mc^m + d_{G_f,t_1t_2}|c^uc^m + d_{G_f,t_1t_2}|c^uc^u,$$

where $d_{G_f,t_1t_2}|c^uc^m$ and $d_{G_f,t_1t_2j}|c^uc^u$ are derived using similar arguments as for $d_{G_f,t_1t_2}|c^mc^m$. We prefer to express the global methylation divergence as a proportion of all of the cytosines, in which case

$$D^*_{G_f,t_1t_2} = \frac{D_{G_f,t_1t_2}}{N}.$$

Using the above derived equilibrium epigenotype frequencies, it can be shown that the equilibrium divergence is

$$D^*_{\infty}=\frac{2\alpha\beta\left(\left[(1-\beta)^2-(1-\alpha)^2\right]^2-2\left[\alpha+\beta-1\right]^2+3\right)}{(\alpha+\beta)^2\left((\alpha+\beta-1)^2-2\right)^2}.$$

Model Fitting and Parameter Estimation. For each pedigree we had a number M of line comparisons and we denoted the observed methylation divergence between each of them as O_{G_f,t_1t_2i}, with $i=\{1,2,\ldots,M\}$, and G_f, t_1, and t_2 the times of and from their most recent common founder, respectively. We assumed that these observations were generated from the proposed epimutation model but contained some unknown measurement error. Hence, we had

$$O_{G_f,t_1t_2i}=c+D^*_{G_f,t_1t_2}+\epsilon_i,$$

where c is the intercept, $D^*_{G_f,t_1t_2}$ is the theoretical global divergence measure introduced above, and ϵ is a random measurement error term. For the MA1_1 population the value of c was approximated using the methylation divergence between technical replicates. For the other three populations no technical replicates were available and c was estimated along with the other parameters (*SI Appendix*, Fig. S9). To obtain parameter estimates we minimized $r^2=\sum_i\left(O_{G_f,t_1t_2i}-D^*_{G_f,t_1t_2}\right)^2$, which is a problem in multivariate nonlinear regression. This involves finding solutions to $\nabla r^2=0$, which can be obtained numerically. Extensive simulations showed that our estimation method performs well, even with relatively large measurement error (*SI Appendix*, Fig. S10).

ACKNOWLEDGMENTS. We thank B. Charlesworth and J. Hadfield for their comments during a seminar at the University of Edinburgh. This work was supported by grants from the Netherlands Organization for Scientific Research (to R.C.J., R.W., A.v.d.G., F.J., and M.C.-T.), a University of Groningen Rosalind Franklin Fellowship (to M.C.-T.), National Institutes of Health Grant R00GM100000, and National Science Foundation Grant IOS-1339194 (to R.J.S.).

1. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 11(3):204–220.
2. Becker C, et al. (2011) Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature* 480(7376):245–249.
3. Schmitz RJ, et al. (2011) Transgenerational epigenetic instability is a source of novel methylation variants. *Science* 334(6054):369–373.
4. Richards EJ (2006) Inherited epigenetic variation—revisiting soft inheritance. *Nat Rev Genet* 7(5):395–401.
5. Cortijo S, et al. (2014) Mapping the epigenetic basis of complex traits. *Science* 343(6175):1145–1148.
6. Kalisz S, Purugganan MD (2004) Epialleles via DNA methylation: Consequences for plant evolution. *Trends Ecol Evol* 19(6):309–314.
7. Weigel D, Colot V (2012) Epialleles in plant evolution. *Genome Biol* 13(10):249.
8. Diez CM, Roessler K, Gaut BS (2014) Epigenetics and plant genome evolution. *Curr Opin Plant Biol* 18:1–8.
9. Springer NM (2013) Epigenetics and crop improvement. *Trends Genet* 29(4):241–247.
10. Jiang C, Mithani A, Belfield EJ, Mott R, Hurst LD, Harberd NP (2014) Environmentally responsive genome-wide accumulation of de novo *Arabidopsis thaliana* mutations and epimutations. *Genome Res* 24(11):1821–1829.
11. Becker C, Weigel D (2012) Epigenetic variation: Origin and transgenerational inheritance. *Curr Opin Plant Biol* 15(5):562–567.
12. Hunter B, Hollister JD, Bomblies K (2012) Epigenetic inheritance: What news for evolution? *Curr Biol* 22(2):R54–R56.
13. Charlesworth B, Jain K (2014) Purifying selection, drift, and reversible mutation with arbitrarily high mutation rates. *Genetics* 198(4):1587–1602.
14. Wang J, Fan C (2015) A neutrality test for detecting selection on DNA methylation using single methylation polymorphism frequency spectrum. *Genome Biol Evol* 7(1):154–171.
15. Shaw RG, Byers DL, Darmo E (2000) Spontaneous mutational effects on reproductive traits of *Arabidopsis thaliana*. *Genetics* 155(1):369–378.
16. Lister R, et al. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* 133(3):523–536.
17. Cokus SJ, et al. (2008) Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature* 452(7184):215–219.
18. Nuthikattu S, et al. (2013) The initiation of epigenetic silencing of active transposable elements is triggered by RDR6 and 21-22 nucleotide small interfering RNAs. *Plant Physiol* 162(1):116–131.
19. Zemach A, et al. (2013) The *Arabidopsis* nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell* 153(1):193–205.
20. Creasey KM, et al. (2014) miRNAs trigger widespread epigenetically activated siRNAs from transposons in *Arabidopsis*. *Nature* 508(7496):411–415.
21. Yelagandula R, et al. (2014) The histone variant H2A.W defines heterochromatin and promotes chromatin condensation in *Arabidopsis*. *Cell* 158(1):98–109.
22. Schmitz RJ, et al. (2013) Patterns of population epigenomic diversity. *Nature* 495(7440):193–198.
23. Hagmann J, et al. (2015) Century-scale methylome stability in a recently diverged *Arabidopsis thaliana* lineage. *PLoS Genet* 11(1):e1004920.
24. Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D (2014) Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. *PLoS Genet* 10(11):e1004785.
25. Ossowski S, et al. (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327(5961):92–94.
26. Johannes F, Colot V, Jansen RC (2008) Epigenome dynamics: A quantitative genetics perspective. *Nat Rev Genet* 9(11):883–890.
27. Dubin MJ, et al. (2015) DNA methylation variation in *Arabidopsis* has a genetic basis and shows evidence of local adaptation. arXiv:1410.5723.
28. Takuno S, Gaut BS (2013) Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proc Natl Acad Sci USA* 110(5):1797–1802.
29. Furrow RE (2014) Epigenetic inheritance, epimutation, and the response to selection. *PLoS ONE* 9(7):e101559.
30. Slatkin M (2009) Epigenetic inheritance and the missing heritability problem. *Genetics* 182(3):845–850.

Supplementary Information Appendix

Rate, spectrum and evolutionary dynamics of spontaneous epimutations

Adriaan van der Graaf[1†], René Wardenaar[1†], Drexel A. Neumann[2], Aaron Taudt[3], Ruth G. Shaw[4], Ritsert C. Jansen[1], Robert J. Schmitz[2*], Maria Colomé-Tatché[3*], and Frank Johannes[1*]

[1] University of Groningen, Groningen Bioinformatics Centre, The Netherlands
[2] University of Georgia, Department of Genetics, USA
[3] University Medical Centre Groningen (UMCG) and the University of Groningen, European Institute for the Biology of Aging (ERIBA), The Netherlands
[4] University of Minnesota, Department of Ecology, Evolution and Behavior, USA

† Equal contributions
* To whom correspondence should be addressed
E-mail: schmitz@uga.edu, m.colome.tatche@umcg.nl, f.johannes@rug.nl

Contents

Supplementary methods **3**
1.1 Mutation accumulation (MA) lines 3
1.1.1 Datasets and labeling 3
1.1.2 MA1_1 and MA1_2 data 3
1.1.3 MA1_3 and MA2_3 data 3
1.2 Methylation maps and methylation divergence in the MA lines . 4
1.2.1 Data preprocessing, alignment and copy removal 4
1.2.2 Methylation calling 4
1.2.3 Consensus positions and methylation divergence 5
1.2.4 Removal of lines for robust estimates 6
1.2.5 Annotation 6
1.3 Theoretical model 6
1.3.1 Epigenotype frequencies among MA lines at finite t and at equilibrium 7
1.3.2 DNA methylation divergence among MA lines at finite t and at equilibrium 9
1.3.3 Estimation approach 12
1.3.4 Effect of heterozygosity in the founder 14
1.3.5 Evaluation of coverage cutoffs on epimutation rate estimates 15
1.4 Simulation study 15
1.5 Expected CG methylation divergence genome-wide 16
1.5.1 Correlation between divergence and read coverage 17
1.5.2 Expected CG methylation divergence genome-wide accounting for heterochromatic domains 18
1.6 CG methylation divergence and sequence diversity among natural accessions 20
1.6.1 CG methylation divergence 20
1.6.2 Sequence divergence 20
1.6.3 CG methylation divergence in the MA lines and in the natural accessions 21
1.6.4 Relationship between CG methylation divergence in the MA lines and sequence divergence in the natural accessions 21
1.7 Uncoupling of epigenetic from genetic variation 22
1.8 Effects of selection on epialleles 25

Supplementary figures **29**

Supplementary tables **39**

Supplementary methods

1.1 Mutation accumulation (MA) lines

1.1.1 Datasets and labeling

A summary of the 4 datasets used in this study can be found at **Table S2**. Their pedigrees can be found in **Figure 1a**.

1.1.2 MA1_1 and MA1_2 data

Populations MA1_1 and MA1_2 were first described in Becker et al. [1] and Schmitz et al. [2], respectively. MA1_1 was retrieved from the European Nucleotide Archive under accession number ERP000902. MA1_2 was retrieved from National Center for Biotechnology Information Sequence Read Archive, accession SRA035939. The experimental material, conditions and sequencing approach are summarized in Table S1.

1.1.3 MA1_3 and MA2_3 data

Plants were grown on soil under 16 hours of light and 8 hours of dark cycles. Leaf tissue was taken at the same time of day from individual plants and flash-frozen in liquid nitrogen for later DNA extraction. Leaf tissue was ground with mortar and pestle under liquid nitrogen and genomic DNA was extracted from ground tissue using Qiagen DNeasy kit (Qiagen) according to manufacturer's instructions. One microgram gDNA was sonicated in $130\mu l$ nuclease free water to 200bp fragments using a Covaris S-2 sonicator (Covaris Inc.). Large fragment removal was accomplished by mixing magnetic purification beads (MPB - see Purification methods) with sonicated gDNA at a volume ratio of 0.6X and incubation for 10 minutes at room temperature. After 10 minutes samples were moved to a magnetic plate to separate beads from solution. Supernatant was removed from beads and added to a volume of MPB at a 0.8 ratio compared to original gDNA solution volume to create a total MPB to gDNA volume ratio of 1.4; tubes containing beads and large fragment DNA were discarded. Samples were then cleaned per described purification methods (See below).

Fragmented gDNA from sonication were repaired using End-It DNA End-Repair

Kit (Epicentre) according to manufacturer's instructions. A-tails were added to blunt ended fragments during a 30 minute 37°C incubation using Klenow 3'-5' exonuclease and dA-Tailing Buffer (New England Biolabs). Methylated NEXTflex DNA adapters (Bioo Scientific) were ligated via 16 hour, 16°C incubation with T4 DNA ligase and ligation buffer (New England Biolabs). Post-ligation, samples underwent two clean up procedures and then were subject to bisulfite conversion using a MethylCode kit per manufacturer's instructions. Libraries were amplified with eight cycles of PCR using Kapa HiFi Uracil+ Hotstart (Kapa Biosystems).

MPB were mixed with samples at a volume ratio of 1.4:1 prior to adapter ligation and 1.0:1.0 post adapter ligation. To wash, MPB solution was mixed with samples by pipetting and left at room temperature for 10 minutes. After, samples were moved to magnet to isolate magnetic beads from solution. Supernatant was removed and beads were washed with two cycles of 80% ethanol. Samples were removed from magnets and DNA was eluted off beads into 10mM Tris-HCl pH 8.0. and left at room temperature for 10 minutes. Sequencing was performed on an Illumina NextSeq 500. Libraries were sequenced to 75 and 150 bp using the single-end format. The experimental material, conditions and sequencing approach are summarized in Table S1.

1.2 Methylation maps and methylation divergence in the MA lines

1.2.1 Data preprocessing, alignment and copy removal

MA1_1 consists of paired end sequenced data, while MA1_2, MA1_3 and MA2_3 consist of single end data (**Table S2**). Adapters were removed and reads were quality trimmed based on a PHRED score of 5, using the standard features of the cutadapt tool version 1.2 [3]. Reads of less than 20 basepairs after preprocessing were removed from further analysis.

Bisulfite read alignment was performed by BS-Seeker 2 [4], using bowtie 1.0.0 as short read aligner [5]. Reads were aligned to the TAIR10 genome [6], allowing for a maximum of 4 mismatches per read. To counter PCR bias in library creation, only reads mapping to a unique first position in the genome were retained, one read was kept mapping to the same strand and same first 5' position. If reads with differing lengths mapped to the same first position and strand, the longest read was retained.

1.2.2 Methylation calling

For each sample we determined the bisulfite conversion rate using data from the unmethylated chloroplast chromosome. We calculated the bisulphite conversion

rate as:

$$BCR = 1 - \frac{\sum C_{\text{ref}}.C_{\text{read}}}{\sum C_{\text{ref}}.C_{\text{read}} + \sum C_{\text{ref}}.T_{\text{read}}}, \quad (1)$$

where $C_{\text{ref}}.C_{\text{read}}$ are the cytosines in the read sequence that were mapped to a cytosine in the reference (non-converted) and $C_{\text{ref}}.T_{\text{read}}$ are the thymine in the read sequence that were mapped to a cytosine in the reference (converted).

The conversion rates for each of the sequencing experiments are shown in **Tables S3, S4, S5 and S6**, and ranged from 96.828 to 99.965 (mean = 99.629). For the jth cytosine (c_j) we calculated the probability to be unmethylated based on the observation of k_j methylated reads out of a total of n_j reads. The probability that c_j is unmethylated is given by the binomial probability mass function:

$$Pr(k_j) = \frac{n_j!}{k_j!(n_j - k_j)!} R^{k_j}(1 - R)^{n_j - k_j}, \quad (2)$$

where $R = 1 - BCR$. Adjusting for multiple testing, c_j was finally called as methylated (c^m) or unmethylated (c^u) based on a genome-wide P-value cutoff corresponding to a FDR [7] of 0.05. Since epi-heterozygotes ($c^m c^u$) are difficult to call from these data, we assume that all detected c^m correspond to epigenotype $c^m c^m$ and all c^u correspond to epigenotype $c^u c^u$. Later analysis will attempt to infer the genome-wide proportion of epi-heterzygote loci $c^m c^u$ (see Section 1.3.3).

1.2.3 Consensus positions and methylation divergence

When epigenotyping multiple samples, experimental and technical variation resulted in differences in read coverage. When comparing individuals in a pedigree we only considered cytosines which were covered by more than three reads in all measured individuals. A coverage cutoff of more than three reads was used in previous methylome studies by Becker et al. [1] and Schmitz et al. [2]. Using more stringent coverage cutoffs yielded similar results (see section 1.3.5). These positions will subsequently be referred to as *consensus positions* or *consensus cytosines*. In order to calculate the methylation divergence between every pair of lines, we compared the methylation status of every consensus cytosine between the pair. At every cytosine j we attributed a divergence $d_j = 0$ if both cytosines had the same methylation status, and a divergence $d_j = 1$ if their methylation status was different. The pairwise methylation divergence was calculated as the sum over the divergence at every consensus cytosine, divided by the total number of consensus cytosines

$$MD = \frac{1}{N} \sum_{j=1}^{N} d_j. \quad (3)$$

In the case where only a part of the genome was considered we computed the methylation divergence in the selected area as

$$MD = \frac{1}{N'} \sum_{j=1}^{N'} d_j, \tag{4}$$

where the positions $j \in N'$ correspond to the consensus cytosines in the selected region.

1.2.4 Removal of lines for robust estimates

Line 69 from the original Becker et al. [1] (MA1_1) and Schmitz et al. [2] (MA1_2) populations was removed in this analysis, as the authors observed an unusual high methylation divergence for this line compared to the other lines. Sequencing revealed that line 69 contained a mutation in MATERNAL EFFECT EMBRYO ARREST57 (MEE57), possibly involved in cytosine methylation maintenance [1]. One plant of the 3^{rd} generation in MA1_1 also had unusually elevated CG divergence. This plant was considered an outlier and was removed from further analysis.

1.2.5 Annotation

We considered four annotations: gene, transposable element (TE), promoter and intergenic. Gene annotations were determined using positions identified as genes by TAIR10. TE annotations were identified using Quenesville annotations from TAIR [8]. Promoter annotations were identified as 1.5 kb upstream of the transcription start site of genes. Positions with no annotations on both strands were called intergenic. Positions with multiple annotations (on the same strand or opposite strand) were not considered in any annotation category. We assigned all consensus cytosines to their corresponding annotation.

1.3 Theoretical model

We modeled the methylation divergence between any two individuals in a pedigree using a Markov chain [9–12]. Let c^u and c^m denote an unmethylated and a methylated cytosine, respectively, and $\alpha = \Pr(c^u \to c^m)$ and $\beta = \Pr(c^m \to c^u)$ be the probabilities that a cytosine gains or loses methylation during or prior to gamete formation, which can include gains or losses of DNA methylation in somatic tissues from which the gametic cells were derived. We arbitrarily call α the forward and β the backward epimutation rate per generation per haploid methylome. In a diploid methylome, the gametes that can be produced from the three possible epigenotypes at a single cytosine, together with their probabilities, are:

Epigenotype	Gametes	
	c^u	c^m
$c^u c^u$	$1-\alpha$	α
$c^m c^u$	$\frac{1}{2}(\beta+(1-\alpha))$	$\frac{1}{2}(\alpha+(1-\beta))$
$c^m c^m$	β	$1-\beta$

We modeled the epigenotype frequencies at the jth cytosine using a Markov chain with three states: $c^u c^u$, $c^u c^m$ and $c^m c^m$. Taking into account Mendelian segregation of alleles together with the forward and backward epimutation rates per cytosine, we derived the epigenotype transition matrix $\mathbf{T}$ after one selfing generation, which is equal to:

	$c^u c^u$	$c^m c^u$	$c^m c^m$
$c^u c^u$	$(1-\alpha)^2$	$2(1-\alpha)\alpha$	α^2
$c^m c^u$	$\frac{1}{4}(\beta+1-\alpha)^2$	$\frac{1}{2}(\beta+1-\alpha)(\alpha+1-\beta)$	$\frac{1}{4}(\alpha+1-\beta)^2$
$c^m c^m$	β^2	$2(1-\beta)\beta$	$(1-\beta)^2$

(5)

This formulation does not account for higher order epimutation events, because such events are expected to be rare for small epimutation rates. The epigenotype frequencies at cytosine j in the MA population, π_{tj}, after t generations of single seed descent can be expressed as:

$$\pi_{tj} = \pi_{0j} P \mathbf{V}^t P^{-1}, \tag{6}$$

where π_{0j} are the (unobserved) epigenotype frequencies at the founder plant, P is the eigenvector of matrix $\mathbf{T}$ and $\mathbf{V}$ is a diagonal matrix of the distinct eigenvalues of matrix $\mathbf{T}$.

1.3.1 Epigenotype frequencies among MA lines at finite t and at equilibrium

Using Mathematica 10.0 (Wolfram Research, Inc.), we derived analytical solutions for the elements of π_{tj}, which are functions of t, α, β as well as the initial frequency vector π_{0j}. Solutions are given by Eq. 8. At equilibrium, the $\pi_{\infty j}$ represent the expected epigenotype frequencies at cytosine j among the MA lines after a (hypothetical) infinite number of selfing generations ($t=\infty$). They were obtained by calculating $\lim_{t\to\infty} \pi_{tj}$:

$$\begin{aligned}
\pi_{\infty j}(c^u c^u) &= \frac{\beta\left((1-\beta)^2-(1-\alpha)^2-1\right)}{(\alpha+\beta)\left((\alpha+\beta-1)^2-2\right)} \\
\pi_{\infty j}(c^u c^m) &= \frac{4\alpha\beta(\alpha+\beta-2)}{(\alpha+\beta)\left((\alpha+\beta-1)^2-2\right)} \\
\pi_{\infty j}(c^m c^m) &= \frac{\alpha\left((1-\alpha)^2-(1-\beta)^2-1\right)}{(\alpha+\beta)\left((\alpha+\beta-1)^2-2\right)}.
\end{aligned} \tag{7}$$

$$\pi_{tj}(c^u c^u) = \frac{2^{-t-1}}{(\alpha+\beta)(\alpha+\beta+1)\left(\alpha^2+2(\beta-1)\alpha+(\beta-2)\beta-1\right)}$$
$$\left(2\left(2\beta(\alpha+\beta)\left(\beta^2+(\alpha-2)\beta+\alpha-1\right)\left((\alpha+\beta-1)^2\right)^t+2^t\beta\left((\alpha-3\beta-1)\left(\alpha^2+2(\beta-1)\alpha+(\beta-2)\beta-1\right)(-\alpha-\beta+1)^t-\alpha^3+\beta^3+\alpha^2+\alpha\beta^2-\beta^2+\alpha-\alpha^2\beta-3\beta-1\right)\right)\pi_{0j}(c^m c^m)\right.$$
$$+\left((\alpha+\beta)\left(\beta^3-(\alpha+3)\beta^2-(\alpha-6)\alpha\beta+\beta+\alpha((\alpha-3)\alpha+1)+1\right)\left((\alpha+\beta-1)^2\right)^t+2^t\left((-\alpha-\beta+1)^t(\alpha-\beta)(-\alpha+3\beta+1)\left(\alpha^2+2(\beta-1)\alpha+(\beta-2)\beta-1\right)-2\beta(\alpha+\beta+1)\left((\alpha-1)^2-\beta^2+2\beta\right)\right)\right)\pi_{0j}(c^u c^m)$$
$$\left.+2\left(2\alpha(\alpha+\beta)(\beta+\alpha(\alpha+\beta-2)-1)\left((\alpha+\beta-1)^2\right)^t+2^t\left(-\alpha(\alpha-3\beta-1)\left(\alpha^2+2(\beta-1)\alpha+(\beta-2)\beta-1\right)(-\alpha-\beta+1)^t-\beta(\alpha+\beta+1)\left((\alpha-1)^2-\beta^2+2\beta\right)\right)\right)\pi_{0j}(c^u c^u)\right)$$
$$\pi_{tj}(c^u c^m) = \frac{2^{-t}}{(\alpha+\beta)(\alpha+\beta+1)\left(\alpha^2+2(\beta-1)\alpha+(\beta-2)\beta-1\right)}$$
$$\left(4\beta\left(2^t\left(\alpha(\alpha+\beta-2)(\alpha+\beta+1)-(-\alpha-\beta+1)^t(\alpha-\beta)\left(\alpha^2+2(\beta-1)\alpha+(\beta-2)\beta-1\right)\right)-\left((\alpha+\beta-1)^2\right)^t(\alpha+\beta)\left(\beta^2+(\alpha-2)\beta+\alpha-1\right)\right)\pi_{0j}(c^m c^m)\right.$$
$$+\left(2^{t+1}\left((\alpha-\beta)^2\left(\alpha^2+2(\beta-1)\alpha+(\beta-2)\beta-1\right)(-\alpha-\beta+1)^t+2\alpha\beta(\alpha+\beta-2)(\alpha+\beta+1)\right)-\left((\alpha+\beta-1)^2\right)^t(\alpha+\beta)\left(\beta^3-(\alpha+3)\beta^2-(\alpha-6)\alpha\beta+\beta+\alpha((\alpha-3)\alpha+1)+1\right)\right)\pi_{0j}(c^u c^m)$$
$$\left.+4\alpha\left(2^t\left((\alpha-\beta)\left(\alpha^2+2(\beta-1)\alpha+(\beta-2)\beta-1\right)(-\alpha-\beta+1)^t+\beta(\alpha+\beta-2)(\alpha+\beta+1)\right)-\left((\alpha+\beta-1)^2\right)^t(\alpha+\beta)(\beta+\alpha(\alpha+\beta-2)-1)\right)\pi_{0j}(c^u c^u)\right),$$
$$\pi_{tj}(c^m c^m) = \frac{2^{-t-1}}{(\alpha+\beta)(\alpha+\beta+1)\left(\alpha^2+2(\beta-1)\alpha+(\beta-2)\beta-1\right)}$$
$$\left(2\left(2\beta(\alpha+\beta)\left(\beta^2+(\alpha-2)\beta+\alpha-1\right)\left((\alpha+\beta-1)^2\right)^t+2^t\left((3\alpha-\beta+1)\beta\left(\alpha^2+2(\beta-1)\alpha+(\beta-2)\beta-1\right)(-\alpha-\beta+1)^t+\alpha\left((\alpha-2)\alpha-(\beta-1)^2\right)(\alpha+\beta+1)\right)\right)\pi_{0j}(c^m c^m)\right.$$
$$+\left((\alpha+\beta)\left(\beta^3-(\alpha+3)\beta^2-(\alpha-6)\alpha\beta+\beta+\alpha((\alpha-3)\alpha+1)+1\right)\left((\alpha+\beta-1)^2\right)^t+2^t\left(2\alpha\left((\alpha-2)\alpha-(\beta-1)^2\right)(\alpha+\beta+1)-(-\alpha-\beta+1)^t(\alpha-\beta)(3\alpha-\beta+1)\left(\alpha^2+2(\beta-1)\alpha+(\beta-2)\beta-1\right)\right)\right)\pi_{0j}(c^u c^m)$$
$$\left.+2\left(2\alpha(\alpha+\beta)(\beta+\alpha(\alpha+\beta-2)-1)\left((\alpha+\beta-1)^2\right)^t+2^t\alpha\left(-(3\alpha-\beta+1)\left(\alpha^2+2(\beta-1)\alpha+(\beta-2)\beta-1\right)(-\alpha-\beta+1)^t+\alpha^3-\beta^3-\alpha^2-\alpha\beta^2+\beta^2-3\alpha+\alpha^2\beta+\beta-1\right)\right)\pi_{0j}(c^u c^u)\right). \quad (8)$$

∞

One can observe that, for any $0 < \alpha, \beta < 1$, these equilibrium solutions are independent of the initial epigenotype proportions π_{0j} at the common founder individual, therefore only the rates α and β determine the methylation frequencies at a given cytosine at equilibrium. In the absence of backward epimutations ($\beta = 0$) but in the presence of forward epimutations ($\alpha \neq 0$), $\pi_{\infty j}(c^u c^u) = \pi_{\infty j}(c^u c^m) = 0$ and $\pi_{\infty j}(c^m c^m) = 1$, therefore all the MA lines would show a methylated cytosine at the location. Conversely, for $\beta \neq 0$ but $\alpha = 0$, $\pi_{\infty j}(c^u c^u) = 1$ and $\pi_{\infty j}(c^u c^m) = \pi_{\infty j}(c^m c^m) = 0$, so the reverse is true. Hence, the importance of a model that takes both forward and backward epimutations into account. Finally, if $\alpha = 0$ and $\beta = 0$, there are no epimutation events and the dynamics are according to known Mendelian inbreeding theory [11].

The rate at which the epigenotype proportions converge to these equilibrium values follows Eq. 8, and it therefore depends on the strength of the forward and backward rates.

1.3.2 DNA methylation divergence among MA lines at finite t and at equilibrium

At every cytosine, the methylation divergence between two independently selfed lines can be scored with the following distance matrix:

	$c^u c^u$	$c^m c^u$	$c^m c^m$
$c^u c^u$	0	$\frac{1}{2}$	1
$c^m c^u$	$\frac{1}{2}$	0	$\frac{1}{2}$
$c^m c^m$	1	$\frac{1}{2}$	0

Let t_1 and t_2 denote the number of generations between two individuals at generations G_m and G_n and their most recent common founder at generation G_f, respectively (i.e. $t_1 = G_m - G_f$, $t_2 = G_n - G_f$, **Figure 1a**, **Figure SI-1**). Let $\pi_{t_i j}|c^m c^m$ be the vector of epigenotype frequencies at the jth cytosine after t_i selfing generations from G_f, conditional on the fact that the most recent common founder epigenotype was $c^m c^m$:

$$\pi_{t_i j}|c^m c^m = (0, 0, 1) \cdot T^{t_i}. \tag{9}$$

Using Eq. 9 and the methylation divergence scoring table above, the divergence between the two lines at this locus can be calculated as:

$$\begin{aligned} d_{t_1t_2j}|c^mc^m &= 1/2 \cdot \pi_{t_1j}(c^uc^u)|c^mc^m \cdot \pi_{t_2j}(c^mc^u)|c^mc^m \\ &+ 1/2 \cdot \pi_{t_1j}(c^mc^u)|c^mc^m \cdot \pi_{t_2j}(c^uc^u)|c^mc^m \\ &+ 1/2 \cdot \pi_{t_1j}(c^mc^u)|c^mc^m \cdot \pi_{t_2j}(c^mc^m)|c^mc^m \\ &+ 1/2 \cdot \pi_{t_1j}(c^mc^m)|c^mc^m \cdot \pi_{t_2j}(c^mc^u)|c^mc^m \\ &+ 1 \cdot \pi_{t_1j}(c^uc^u)|c^mc^m) \cdot \pi_{t_2j}(c^mc^m)|c^mc^m \\ &+ 1 \cdot \pi_{t_1j}(c^mc^m)|c^mc^m) \cdot \pi_{t_2j}(c^uc^u)|c^mc^m. \end{aligned}$$

The simple multiplication of these frequencies follows from the fact that the selfing lines are conditionally independent. The divergence over all loci for which the most recent common founder at G_f was c^mc^m is:

$$d_{G_f,t_1t_2}|c^mc^m = \sum_j d_{t_1t_2j}|c^mc^m = N_{G_f}^{mm} d_{t_1t_2j}|c^mc^m,$$

where $N_{G_f}^{mm}$ are the number of consensus methylated cytosines at G_f.

Using similar arguments, the global (or total) DNA methylation divergence along the genome can be calculated as:

$$D_{G_f,t_1t_2} = d_{G_f,t_1t_2}|c^mc^m + d_{G_f,t_1t_2}|c^mc^u + d_{G_f,t_1t_2}|c^uc^u,$$

where

$$\begin{aligned} d_{G_f,t_1t_2}|c^mc^u &= N_{G_f,c^mc^u}\Big(\frac{1}{2} \cdot \pi_{t_1j}(c^uc^u)|c^mc^u \cdot \pi_{t_2j}(c^mc^u)|c^mc^u \\ &+ \frac{1}{2} \cdot \pi_{t_1j}(c^mc^u)|c^mc^u \cdot \pi_{t_2j}(c^uc^u)|c^mc^u \\ &+ \frac{1}{2} \cdot \pi_{t_1j}(c^mc^u)|c^mc^u \cdot \pi_{t_2j}(c^mc^m)|c^mc^u \\ &+ \frac{1}{2} \cdot \pi_{t_1j}(c^mc^u)|c^mc^m \cdot \pi_{t_2j}(c^mc^u)|c^mc^u \\ &+ 1 \cdot \pi_{t_1j}(c^uc^u)|c^mc^u \cdot \pi_{t_2j}(c^mc^m)|c^mc^u \\ &+ 1 \cdot \pi_{t_1j}(c^mc^m)|c^mc^u \cdot \pi_{t_2j}(c^uc^u)|c^mc^u\Big), \end{aligned}$$

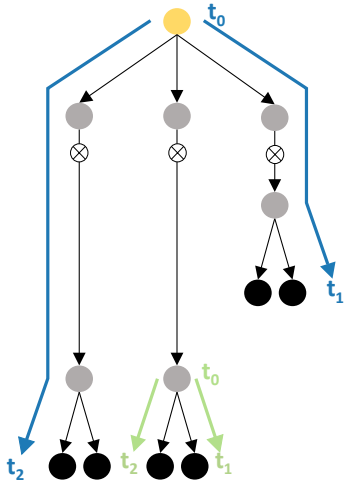


Figure SI-1: For two different pairs (blue and green) of individuals in a pedigree, we depict the time of their most recent common founder $t_0 = G_f$, the time of the first individual (t_1) and the time of the second individual (t_2) in the comparison. Both t_1 and t_2 are calculated from t_0. In the case of t_1 and t_2 (green) this common ancestor is at time t_0 (green), and in the case of t_1 and t_2 (blue) this common ancestor is at time t_0 (blue). This means that t_1 and t_2 do not necessarily correspond to the generation time within the pedigree relative to the founder $G0$ of the pedigree.

and

$$
\begin{aligned}
d_{G_f,t_1t_2j}|c^uc^u \;=\; & N_{G_f,c^uc^uc^u}\Big(\frac{1}{2}\cdot\pi_{t_1j}(c^uc^u)|c^uc^u\cdot\pi_{t_2j}(c^mc^u)|c^uc^u \\
+\; & \frac{1}{2}\cdot\pi_{t_1j}(c^mc^u)|c^uc^u\cdot\pi_{t_2j}(c^uc^u)|c^uc^u \\
+\; & \frac{1}{2}\cdot\pi_{t_1j}(c^mc^u)|c^uc^u\cdot\pi_{t_2j}(c^mc^m)|c^uc^u \\
+\; & \frac{1}{2}\cdot\pi_{t_1j}(c^mc^m)|c^uc^u\cdot\pi_{t_2j}(c^mc^u)|c^uc^u \\
+\; & 1\cdot\pi_{t_1j}(c^uc^u)|c^uc^u\cdot\pi_{t_2j}(c^mc^m)|c^uc^u \\
+\; & 1\cdot\pi_{t_1j}(c^mc^m)|c^uc^u\cdot\pi_{t_2j}(c^uc^u)|c^uc^u\Big).
\end{aligned}
$$

One can calculate the number of $c^u c^u$, $c^u c^m$ and $c^m c^m$ consensus cytosines at the (unobserved) most recent common founder by using

$$(N_{G_f}^{^{uu}}, N_{G_f}^{^{um}}, N_{G_f}^{^{mm}}) = N \cdot \pi_0 \cdot T^{t_0}, \tag{10}$$

where π_0 are the overall genome proportions of $c^u c^u$, $c^u c^m$ and $c^m c^m$ consensus cytosines at the (unobserved) founder of the pedigree, and N are the total number of consensus cytosines in the genome. We prefer to express the global methylation divergence as a proportion of all the cytosines, in which case:

$$D^*_{G_f,t_1t_2} = \frac{D_{G_f,t_1t_2}}{N}. \tag{11}$$

Using the above derived equilibrium epigenotype frequencies, we can also calculate the expected DNA methylation divergence as $t \to \infty$. Since the epigenotype frequencies at equilibrium are independent of the initial frequencies at locus j, we can calculate the equilibrium divergence as:

$$\begin{aligned} D_\infty &= N^{uu} d_{\infty j} + N^{mu} d_{\infty j} + N^{mm} d_{\infty j} \\ &= d_{\infty j}(N^{uu} + N^{mu} + N^{mm}) \\ &= d_{\infty j} N, \end{aligned}$$

where N^{uu}, N^{um} and N^{mm} are the number of consensus unmethylated, epiheterozygous and methylated cytosines in the genome of the founder of the pedigree. Hence

$$D^*_\infty = d_{\infty j} \frac{N}{N} = d_{\infty j},$$

where

$$\begin{aligned} d_{\infty j} &= \pi_{\infty j}(c^m c^u) \cdot (\pi_{\infty j}(c^u c^u) + \pi_{\infty j}(c^m c^m)) + 2 \cdot \pi_{\infty j}(c^u c^u) \cdot \pi_{\infty j}(c^m c^m) \\ &= \frac{2\alpha\beta\left(\left[(1-\beta)^2 - (1-\alpha)^2\right]^2 - 2\left[\alpha+\beta-1\right]^2 + 3\right)}{(\alpha+\beta)^2\left((\alpha+\beta-1)^2 - 2\right)^2}. \end{aligned}$$

1.3.3 Estimation approach

We considered data from 4 MA pedigrees (**Figure 1a**) to obtain 4 sets of estimates of the model parameters. For every MA pedigree, the DNA methylomes of each line were determined at cytosine resolution (section 1.2.2). As outlined in section 1.2.3, we used the scoring matrix above to measure, empirically, the genome-wide divergence between any unique pair of individuals in a pedigree as the mean divergence over all considered consensus cytosines (Eq. 3). We then classified the cytosines according to their context (CG, CHH, CHG) and calculated the divergence per context (Eq. 4) between any pair (**Figure S3**).

We further subdivided the CG context into different genome annotations (genes (CG-gene), transposable elements (CG-TE), promoters of genes (CG-promoter) and intergenic regions (CG-intergenic)), and calculated the methylation divergence per pair for a given annotation (Eq. 4) (**Figure 1b**).

Since only CG methylation divergence accumulated over time (**Figure S3**), only model fits to context CG were of interest. We sought to obtain epimutation rate estimates for all CGs (CG-all) as well as for CGs in different annotation contexts (CG-gene, CG-TE, CG-promoter and CG-intergenic). In order to ensure that epimutation rate estimates for CG-all are representative of the whole genome and more comparable across MA datasets (**Figure S3**), we randomly sub-sampled the consensus CGs (whithin each MA dataset) in such a way that they reflect the annotation proportion of the *A. thaliana* reference genome (TAIR 10). Because the total number of consensus cytosines is very large, the sampling error (i.e. standard error) arising from this sub-sampling procedure is negligible; and hence, sampling was only performed once. We denote the non-representative set of consensus cytosines by CG-all†, and the representative set of consensus cytosines by CG-all. **Table S2** reports the number of consensus cytosines corresponding to CG-all† and CG-all.

For each pedigree we had a number M of line comparisons and we denoted the observed methylation divergence between each of them as O_{G_f,t_1t_2i}, with $i = \{1, 2, \ldots, M\}$, and G_f, t_1 and t_2 the times of and from the most recent common founder, respectively (**Figure SI-1**). We assumed that these observations were generated from the proposed epimutation model but contained some unknown measurement error. Hence, we had

$$O_{G_f,t_1t_2i} = c + D^*_{G_f,t_1t_2} + \epsilon_i,$$

where c is the intercept, $D^*_{G_f,t_1t_2}$ is the theoretical global divergence measure introduced above, which is a function of α and β as well as N^{uu}, N^{um} and N^{mm}, and ϵ is a random measurement error term. For the MA1_1 population the value of c was approximated using the methylation divergence between technical replicates. For the other three populations no technical replicates were available and c was estimated along with the other parameters (**Figure S9**).
To obtain values for the (unobserved) number of methylated, unmethylated and epi-heterozygous consensus cytosines in the genome of the common founder of each pedigree (N^{mm}, N^{uu} and N^{mu}), we assumed the methylome of *A. thaliana* to be at equilibrium. We then measured the number of consensus cytosines that had been called as methylated (n^{mm}_k) and unmethylated (n^{uu}_k) at every plant k in the pedigree (section 1.2.2). Calling epi-heterozygotes is difficult as they typically manifest as an intermediate methylation signal, and counts for epi-heterozygous loci are included in n^{mm}. Hence, we assumed that an unknown

fraction γ of the total n^{mm} were actually epi-heterozygous. We defined

$$\begin{aligned} N^{uu} &= \frac{1}{n}\sum_{k=1}^{n} n_k^{uu} \\ N^{um} &= \frac{\gamma}{n}\sum_{k=1}^{n} n_k^{mm} \\ N^{mm} &= \frac{1-\gamma}{n}\sum_{k=1}^{n} n_k^{mm}, \end{aligned} \tag{12}$$

where n is the total number of sequenced plants in a pedigree and γ is the proportion of epi-heterozygous cytosines that have been called as methylated. The parameter γ was estimated along with α, β and c. To obtain parameter estimates we sought to minimize

$$r^2(\alpha,\beta,\gamma,c) = \sum_i (O_{G_f,t_1t_2i} - D^*_{G_f,t_1t_2})^2. \tag{13}$$

Minimizing r^2 is a problem in non-linear regression. This involves finding solutions to $\nabla r^2(\alpha,\beta,\gamma,c) = \mathbf{0}$, which can be obtained numerically. Since we assumed the *A. thaliana* genome to be at equilibrium, we searched for solutions such that the equilibrium proportions of methylated and unmethylated cytosines were in the following ranges:

$$\begin{aligned} N^{mm}(\infty) &\in [\min_k((1-\gamma)\, n_k^{mm}), \max_k((1-\gamma)\, n_k^{mm})] \\ N^{uu}(\infty) &\in [\min_k(n_k^{uu}), \max_k(n_k^{uu})]. \end{aligned} \tag{14}$$

The values of α and β can potentially be very low. Hence, in order to minimize r^2, we performed an extensive grid search over the values of α, β, γ and (when needed) c (2212 points both for the grids of α and β (range $10^{-11} - 10^{-2}$), 452 points for the grid of γ (range $10^{-7} - 0.999$) and 218 points for the grid of c (range $10^{-10} - 0.2$)). The minimization procedure was carried out in C++. The values for the estimates in every MA population, for CG context in every different annotation, are reported in **Table S7**.

1.3.4 Effect of heterozygosity in the founder

We assume that the methylome of the common founder plant consists of a proportion $\pi_{c^u c^m} = \gamma N^{mm}/N$ of epiheterozygote cytosines. Following Mendelian segregation these epiheterozygotes will become fixed in approximately 8 generations of selfing: if we fix the epimutation rates to zero, then at every new selfing generation 1/2 of epiheterozygotes will remain epiheterozygote, while 1/4 will become homozygous $c^m c^m$ and 1/4 will become homozygous $c^u c^u$. These dynamics can be seen in the methylation divergence line (**Figure 1b**) as a non-linear increase at the initial $\sim$ 8 generations of selfing. As expected, the

non-linearity is more pronounced for the MA populations for which our estimated proportion of epiheterozygote cytosines in the common founder is larger (MA1_1 and MA1_2).

Due to the residual epimutation events the absorption proportions are very close to, but not exactly equal to, 1/2, 1/4, 1/4 (Eq. 5), but the non-linear signature in the methylation divergence line can still be seen. With non-zero epimutation rates the states $c^u c^u$ and $c^m c^m$ are non absorbing, therefore the system will always contain a proportion of epiheterozygote cytosines even at fixation (Eq. 7).

1.3.5 Evaluation of coverage cutoffs on epimutation rate estimates

We evaluated whether more stringent coverage cutoffs would affect our epimutation rate estimates. We reanalyzed all the data using cutoffs >4, >5, >6, $\cdots$, >8. As **Figures S1** and **S2** show, more stringent cutoffs had negligable impact, yielding very similar CG-methylation divergence patterns over generation time, and resulted in nearly identical epimutation rates.

1.4 Simulation study

We evaluated the performance of our estimation procedure using extensive simulations. We generated 10 pedigrees of the form of MA1_1 as well as 10 pedigrees of the form MA1_3. The simulations took the following steps:

- **Simulation of founder methylome:** We simulated the methylome of the MA founder plant in accordance with the size, sequence composition and methylation content of the *A. thaliana* genome.
- **Induction of random epimutations according to known forward-backward rates:** We moved along the genome from cytosine to cytosine, and changed every methylated cytosine (c^m) to its unmethylated form (c^u) with fixed probability β following a Bernoulli random process. Similarly, we changed every unmethylated cytosine (c^u) to its methylated form (c^m) with fixed probability α.
- **Generation of recombinant gametes and next generation zygotes:** For chromosome i we drew a random number of breakpoints B_i from a Poisson distribution of the form

$$Pr(B_i = k) = \frac{\frac{3}{2}^k e^{-\frac{3}{2}}}{k!}, \tag{15}$$

  where $k = 1, 2, \cdots, n$ and the constant 3/2 is the number of recombination events typically observed on an *A. thaliana* chromosome. We placed the

B_i random breakpoints uniformally between the start and the stop base-pair positions of chromosome i. Hence, for simplicity, this approach assumes no crossover interference. Homologous chromosomal segments were swapped inbetween breakpoints to obtain recombinant gametic products. Random (haploid) gametes were jointed to produce the zygote of the next generation. In accordance with the MA experimental design, we generated two siblings at each generation: one sibling was used for (*in silico*)BS-seq measurements and the other sibling was propogated by selfing to obtain the next generation.

- **Final pedigrees:** The two previous steps were iterated along the seperate branches of the pedigree until the final MA generation was obtained.

- **Distance matrices and measurement error:** Using the (*in silico*) BS-seq measurements at the sampled generation times, we created distance matrices using the same protocol as in the analysis of the real data (see Section 1.3.2). Keeping with notation introduced in (Eq. 11), the ith entry of this matrix is given $D^*_{G_f,t_1t_2i}$. We generated random error around these distance measures using the following model

$$O_{G_f,t_1t_2i} = c + D^*_{G_f,t_1t_2} + \epsilon_i,$$

where c was fixed to the value obtained from the data of Becker et al. [1] and

$$\epsilon_i \sim N(0, \sigma^2).$$

The normality assumption for ϵ was justified based on careful inspection of the estimated error distribution in the data of Becker et al. [1] as well as the data of pedigree MA1_3. To assess the robustness of our estimation approach, we simulated a range of error variances $\sigma_i^2, 1.5\times\sigma_i^2, \cdots, 2.5\times\sigma_i^2$, where σ_i^2 is the error variance estimated in the real data of Becker et al. [1] or in the real data of pedigree MA1_3.

- **Estimation procedure:** The simulated values O_{G_f,t_1t_2i} served as input for estimating forward-backward epimutation rates as described in Section 1.3.3.

- **Results: Figure S10b** shows that the simulated data produces realistic divergence patterns. **Figure S10a** summarizes our parameter estimates from the simulated data. We find that the unknown parameters α and β are estimated very well, even when measurement error exceeds the observed measurement error in the real data.

1.5 Expected CG methylation divergence genome-wide

Since the CG epimutation rates are annotation-specific we expected that the genome-wide CG methylation divergence patterns would closely track the an-

notation density along the genome. In order to test this we calculated the expected CG methylation divergence for 1 Mb windows along the genome (step size: 100 kb), with the goal to compare this result with the observed CG methylation divergence per window. For this analysis we considered only pairs of lines that had been independently propagated for 31 generations (i.e. $t_0 = G0 = 0,\ t_1 = 31,\ t_2 = 31$) from the MA populations MA1_1 and MA1_2, since these are the pairs that had diverged from each other for the longest time.

For every 1 Mb window (step size: 100 kb) we calculated the proportion of consensus CG cytosines in every annotation (p_{g}, p_{TE}, p_{in} and p_{pr}, where g stands for genes, TE for transposable elements, in for intergenic and pr for promoter). We only considered CGs with one annotation category (no overlap with multiple annotations). The expected window-methylation-divergence is given by

$$\begin{aligned} D_w &= \Big(p_{\mathrm{g}} \cdot D^{*\mathrm{g}}_{G_0,t_1=31,t_2=31} + p_{\mathrm{TE}} \cdot D^{*\mathrm{TE}}_{G_0,t_1=31,t_2=31} \\ &+ \quad p_{\mathrm{pr}} \cdot D^{*\mathrm{pr}}_{G_0,t_1=31,t_2=31} + p_{\mathrm{in}} \cdot D^{*\mathrm{in}}_{G_0,t_1=31,t_2=31} \Big) \end{aligned} \tag{16}$$

where $D^{*h}_{G_0,t_1=31,t_2=31}$ is the annotation specific methylation divergence for annotation h, with the most common founder being the founder of the pedigree G_0, and with $t_1 = t_2 = 31$ (Eq. 11), calculated using the epimutation rates from **Table S7** for the populations MA1_1 and MA1_2. The expected methylation divergence over all pairs of individuals is depicted in **Figure 2b** in red. As expected, the methylation divergence was low in TE-rich pericentromeric regions and high in gene-rich arms.

We compared the expected methylation divergence values to the measured methylation divergence per window, calculated as the sum of the methylation divergence for every consensus cytosine in one window, divided by the number of consensus cytosines in the window (section 1.2.3, Eq. 4). In **Figure 2b** we present the average observed divergence over all pairs (brown line) together with its range (brown area). The observed CG methylation divergence among the 31st generation MA lines tracks the expected divergence (red line) genome-wide very well.

1.5.1 Correlation between divergence and read coverage

We calculated the mean CG coverage for the consensus CGs of the 31st generation MA lines for each genomic window (size: 1 Mb; step size: 100 kb) in order to test whether CG read coverage was a good predictor of the observed CG methylation divergence among the 31st generation MA lines. We fitted a linear model for both pericentromeric and chromosomal arm windows:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{17}$$

where y_i is the average of the observed CG methylation divergence in window i and x_i is the mean CG coverage of the consensus CGs of the 31st generation MA lines.
The results (**Figure S6**) show that less than 2% of the genome-wide CG methylation divergence could be explained by the genome-wide read coverage of consensus CGs (genome-wide) (weighted $R^2 = ((1.89 \cdot 10^{-2}, 5.70 \cdot 10^{-3}, 0.106, 3.76 \cdot 10^{-3}, 4.62 \cdot 10^{-2}, 7.82 \cdot 10^{-2})$ for (all chromosomes, chromosomes 1 - 5), respectively) indicating that the contribution of CG read coverage to the observed CG methylation divergence of the 31st generation MA lines is minimal.

1.5.2 Expected CG methylation divergence genome-wide accounting for heterochromatic domains

We tested whether accounting for chromatin structure provided a correlation between expected and observed divergence patterns genome-wide. To that end, we re-analyzed ChIP-seq data for the histone variant H2A.W, that has been shown to mark regions of heterochromatin in *A. thaliana* [13]. We obtained epimutation rate estimates per annotation for presence / absence of this epigenetic mark and correlated the new expected CG methylation divergence profiles with the observed ones.

Detection of heterochromatic domains

Histone variant H2A.W was used as marker for heterochromatin as described in [13]. The data with accession number GSM1232780 was downloaded and reads aligned to the TAIR10 reference genome using bowtie-1.1.1 [5], allowing up to 2 mismatches and keeping only uniquely aligned reads (parameters: bowtie -n 2 -m 1). Duplicate reads were removed with samtools [14].

We binned the genome into 200 bp non-overlapping windows and counted the number of reads that mapped in each bin [15]. We assumed that bins with high number of reads corresponded to enriched regions, while bins with a low number of reads corresponded to non-enriched regions. To classify the bins in that way we used a Hidden Markov Model. We modeled the emission density for the unmodified (U) state as a zero-inflated negative binomial distribution

$$f(x, \theta_U = (r, p, \beta)) = \beta I_{x=0} + (1 - \beta)\frac{\Gamma(r + x)}{\Gamma(r)x!}p^r(1 - p)^x, \tag{18}$$

and the emission density for the modified (M) state as a negative binomial distribution

$$f(x, \theta_M = (r, p)) = \frac{\Gamma(r + x)}{\Gamma(r)x!}p^r(1 - p)^x, \tag{19}$$

where x is the number of reads in one bin, r and p are the probability and the dispersion parameter of the negative binomial distribution, respectively, Γ is the gamma function, $I_{x=0}$ is an indicator function (delta function) and β is the

inflation parameter for zero counts.

To improve computational efficiency we split the zero inflated negative binomial distribution into the zero-inflated component (delta-distribution) and the negative binomial component, and we used them as different emission probabilities corresponding to two different hidden states (zero-reads state, low-reads state). We used the Baum-Welch algorithm [16] to obtain parameter estimates for these 3 distributions (zero-reads state, low-reads state, modified state), and we used the forward-backward [17] algorithm to calculate, at every bin, the probability of belonging to every one of the states. We called a bin modified if the calculated probability of it being modified was larger than 0.5, and unmodified otherwise. Consecutive bins that were labeled as enriched were merged to produce a genomic map of enriched regions.

Epimutation rates per annotation in H2A.W enriched and non-enriched domains

We subdivided the consensus CGs into two categories: H2A.W-enriched and H2A.W-non-enriched. We calculated the CG methylation divergence between pairs in populations MA1_1 and MA1_2 as described in section (section 1.2.3) separately for every new category.

We estimated the expected epimutation rates for every category. Following the same procedure as in section 1.3.3, we obtained estimates for α, β, γ and (in population MA1_2) c by minimizing

$$r^2(\alpha,\beta,\gamma,c) = \sum_j (O^{\mathrm{a}}_{t_0 t_1 t_2 j} - GD^{*}_{t_0 t_1 t_2})^2, \tag{20}$$

where $a = \{$CG_H2A.W-enriched, CG_H2A.W-non-enriched$\}$. The minimization procedure was done as described in section 1.3.3.

Genome-wide methylation divergence

We calculated the expected CG methylation divergence per 1 Mb window (step size 100 kb) for individuals at the 31st generation from MA1_1 and MA1_2. For every window we calculated the proportion of consensus cytosines with H2A.W enrichment, and calculated the overall window divergence as

$$D_w \quad = \quad \left(p_{\mathrm{H2A.W}} \cdot D^{*\mathrm{H2A.W}}_{G_0,t_1=31,t_2=31} + p_{\mathrm{noH2A.W}} \cdot D^{*\mathrm{noH2A.W}}_{G_0,t_1=31,t_2=31} \right), \tag{21}$$

where $D^{*a}_{G_0,t_1=31,t_2=31}$ is the specific genome-wide methylation divergence for enrichment status a, with the most common founder being the founder of the pedigree G_0, and $t_1 = t_2 = 31$ (Eq. 11), and with epimutation rates estimated in the section above. The averaged expected methylation divergence over all pairs of individuals is depicted in **Figure S5b** in green. The correlation between expected and observed methylation divergence is $R^2 = 0.72$ ($p < 0.0001$).

1.6 CG methylation divergence and sequence diversity among natural accessions

1.6.1 CG methylation divergence

We reanalyzed 140 natural accessions from Schmitz et al. [18] for which MethylC-seq data was available (**Table S8**). The MethylC-seq data was downloaded from NCBI GEO (GSE43857). Both leaf and mixed stage inflorescence tissue (i.e. bud) were used for MethylC-seq. The authors showed that accessions were grouped by their genotype and not their tissue type when clustering was based on the methylation levels of CG-DMRs or C-DMRs. The opposite was observed when clustering was performed using RNA-seq data. The authors concluded that DNA methylation is less dynamic than gene expression patterns and only plays a minor role in stages of development or cell types. We therefore decided to use the data from both tissues.

For the analysis of the CG methylation divergence we only selected CG cytosines which were present in the reference genome (not CGs which are a result of SNPs) and that had a coverage of larger than three reads (same as for the MA lines section 1.2.3). If both leaf and bud data were available for a given accession we chose the dataset with the highest number of covered CGs (reference CGs with coverage of larger than three). Among the 140 accessions we selected 133 accessions with the highest number of covered CGs (we excluded seven accessions with a low number of covered CGs; 5% of total; **Table S8**).

For every 1 Mb window in the genome (step size 100 kb) we calculated the CG methylation divergence between pairs of accessions in a similar way as we did for pairs of MA individuals (section 1.2.3). However, we did not require the consensus cytosines to have sufficient coverage in all the accessions, as this would lead to a very small number of selected positions. Instead, we only required that the selected cytosines had enough coverage for every individual in the pair. Therefore, for each pair of accessions and each window in the genome we calculated the proportion of CGs which are differentially methylated, based on the consensus positions between the pair. Subsequently, the average divergence per window was calculated as the mean proportion of differentially methylated CGs among all pairs of accessions (8778 pairs in total).

The number of consensus CG cytosines (reference CGs with coverage larger than 3) between each pair of accessions ranged from 1,072,792 (19.3 % of all ref. CGs) to 3,372,515 (60.6 % of all ref. CGs) (Mean: 1,840,570; 33.1 % of all ref. CGs).

1.6.2 Sequence divergence

In order to determine the sequence divergence between these natural accessions we considered the genomic sequences of 129 of these accessions (**Table S8**; 8256

pairwise combinations). The sequence divergence was determined at every 1 Mb genomic window (step size: 100 kb) as the mean proportion of nucleotides that differ between any pair of accessions:

$$\text{seq}_{div} = \frac{1}{8256} \sum_{\text{pairs}} \frac{\#\text{allele differences per pair}}{\text{window length}}. \tag{22}$$

1.6.3 CG methylation divergence in the MA lines and in the natural accessions

In order to explore the relationship between the observed divergence in the natural accessions and the observed divergence in the MA lines (for MA1_1 and MA1_2, comparisons at $t_0 = 0, t_1 = 31, t_2 = 31$) we used regression analysis. The windows in the pericentromeric regions show a clear non-linear (quadratic) relation, whereas the windows located in the chromosomal arms show a linear relation (**Figure S7**). We therefore fitted a linear regression model for the windows that are located within chromosomal arms:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{23}$$

and a quadratic model for the windows located within pericentromeric regions:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \tag{24}$$

where y_i is the average of the observed methylation divergence in window i in the natural accessions, and x_i is the average of the observed methylation divergence in window i for the MA lines.

Figure S7 shows the results for all chromosomes (top left) as well as for each chromosome separately. Although the natural accessions show a higher CG methylation divergence (**Figure 2b**), over 60% of the genome-wide CG methylation divergence in the natural accessions could be explained by the genome-wide CG methylation divergence among the 31st generation MA lines (weighted $R^2 = (0.624, 0.578, 0.745, 0.617, 0.610, 0.630)$ for (all chromosomes, chromosomes 1 - 5) respectively). Results show a stronger prediction for pericentromeric windows ($R^2 = (0.899, 0.988, 0.615, 0.918, 0.953, 0.983)$ for (all chromosomes, chromosomes 1 - 5) respectively) compared to chromosomal arm windows ($R^2 = (0.525, 0.471, 0.807, 0.484, 0.504, 0.502)$ for (all chromosomes, chromosomes 1 - 5) respectively).

1.6.4 Relationship between CG methylation divergence in the MA lines and sequence divergence in the natural accessions

In order to determine to what extent the sequence divergence among the natural accessions could be explained by the CG methylation divergence among the 31st

generation MA lines, we used regression analysis. We fitted a linear model for the windows that are located within chromosomal arms:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{25}$$

and a quadratic model for the windows located within pericentromeric regions:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \tag{26}$$

where y_i is the average of the sequence divergence (Eq. 22) in window i in the natural accessions, and x_i is the average of the observed methylation divergence in window i for the MA lines.

Results show that the CG methylation divergence among the MA lines correlates moderately with the sequencing divergence among the natural accessions (**Figure S8**). About 25% of the sequence divergence among the natural accessions could be explained by the CG methylation divergence among the 31st generation MA lines across the genome (weighted $R^2 = (0.254, 0.540, 0.566, 0.102, 0.246, 0.283)$ for (all chromosomes, chromosomes 1 - 5), respectively).

1.7 Uncoupling of epigenetic from genetic variation

We consider a simple determistic single-locus model with sequence alleles A and B. Either allele can be in two epiallelic states: it can either be unmethylated u or methylated m. Hence, there are 10 possible genotype-epigenotype associations $\{A^uA^u, A^uA^m, A^mA^m, A^uB^u, A^uB^m, A^mB^u, A^mB^m, B^uB^u, B^uB^m, B^mB^m\}$. We define the following epimutation events: $\alpha = \Pr(A^u \to A^m) = \Pr(B^u \to B^m)$ and $\beta = \Pr(A^m \to A^u) = \Pr(B^m \to B^u)$, implying that epimutations are not allele-specific. For simplicity we suppose that sequence mutations of the form $A \rightleftharpoons B$ are absent. In a strictly selfing system without selection or drift, changes in genotype-epigenotype associations are driven entirely by forward-backward epimuation events. We model this processes using a non-absorbing Markov chain with 10 states and transition matrix T given by Eq. 29.
We start from the most extreme scenario where allele A is associated with epiallele u and allele B with epiallele m at $t = 0$. Hence, in this scenario, DNA sequence variation is completely confounded with epigenetic variation, so that the initial genotype-epigenotype frequencies are

$$\begin{aligned}
&\pi_0(A^uA^u) = 1/2,\\
&\pi_0(B^mB^m) = 1/2,\\
&\pi_0(A^uA^m) = \pi_0(A^mA^m) = 0,\\
&\pi_0(A^uB^u) = \pi_0(A^uB^m) = \pi_0(A^mB^u) = \pi_0(A^mB^m) = 0\\
&\pi_0(B^uB^u) = \pi_0(B^uB^m) = 0.
\end{aligned} \tag{27}$$

Frequency changes over generation time are given by

$$\pi_t = \pi_0 \cdot T^t. \tag{28}$$

In order to calculate the correlation between genetic (G) and epigenetic (EG) variation as a function of generation time, we score genotypes and epigenotypes using the following values:

genotype	**value**
AA	-1
AB	0
BB	$+1$

epigenotype	**value**
uu	-1
um	0
mm	$+1$

We calculated the correlation as a function of generation time, α and β using the formula

$$\rho_{G,EG}(t,\alpha,\beta) = \frac{\mathrm{cov}(G,EG)}{\mathrm{var}(G)\mathrm{var}(EG)} = \frac{\mathrm{var}(G\cdot EG) - \mathrm{var}(G)\mathrm{var}(EG)}{(\mathrm{var}(G)^2 - \mathrm{var}(G^2))(\mathrm{var}(EG)^2 - \mathrm{var}(EG^2))},$$

where

$$\begin{aligned}
\mathrm{var}(G\cdot EG) &= \pi_t(A^uA^u) - \pi_t(A^mA^m) - \pi_t(B^uB^u) + \pi_t(B^mB^m),\\
\mathrm{var}(G) &= -\pi_t(A^uA^u) - \pi_t(A^uA^m) - \pi_t(A^mA^m) + \pi_t(B^uB^u) + \pi_t(B^uB^m) + \pi_t(B^mB^m),\\
\mathrm{var}(G^2) &= \pi_t(A^uA^u) + \pi_t(A^uA^m) + \pi_t(A^mA^m) + \pi_t(B^uB^u) + \pi_t(B^uB^m) + \pi_t(B^mB^m),\\
\mathrm{var}(EG) &= -\pi_t(A^uA^u) - \pi_t(A^uB^u) - \pi_t(B^uB^u) + \pi_t(A^mA^m) + \pi_t(A^mB^m) + \pi_t(B^mB^m),\\
\mathrm{var}(EG^2) &= \pi_t(A^uA^u) + \pi_t(A^uB^u) + \pi_t(B^uB^u) + \pi_t(A^mA^m) + \pi_t(A^mB^m) + \pi_t(B^mB^m).
\end{aligned}$$

	$A^u A^u$	$A^u A^m$	$A^m A^m$	$A^u B^u$	$A^u B^m$	$A^m B^u$	$A^m B^m$	$B^u B^u$	$B^u B^m$	$B^m B^m$
$A^u A^u$	$(1-\alpha)^2$	$2(1-\alpha)\alpha$	α^2	0	0	0	0	0	0	0
$A^u A^m$	$\frac{1}{4}(-\alpha+\beta+1)^2$	$\frac{1}{2}(\alpha-\beta+1)(-\alpha+\beta+1)$	$\frac{1}{4}(\alpha-\beta+1)^2$	0	0	0	0	0	0	0
$A^m A^m$	β^2	$2(1-\beta)\beta$	$(1-\beta)^2$	0	0	0	0	0	0	0
$A^u B^u$	$\frac{1}{4}(1-\alpha)^2$	$\frac{1}{2}(1-\alpha)\alpha$	$\frac{\alpha^2}{4}$	$\frac{1}{2}(1-\alpha)^2$	$\frac{1}{2}(1-\alpha)\alpha$	$\frac{1}{2}(1-\alpha)\alpha$	$\frac{\alpha^2}{2}$	$\frac{1}{4}(1-\alpha)^2$	$\frac{1}{2}(1-\alpha)\alpha$	$\frac{\alpha^2}{4}$
$A^u B^m$	$\frac{1}{4}(1-\alpha)^2$	$\frac{1}{2}(1-\alpha)\alpha$	$\frac{\alpha^2}{4}$	$\frac{1}{2}(1-\alpha)\beta$	$\frac{1}{2}(1-\alpha)(1-\beta)$	$\frac{\alpha\beta}{2}$	$\frac{1}{2}\alpha(1-\beta)$	$\frac{\beta^2}{4}$	$\frac{1}{2}(1-\beta)\beta$	$\frac{1}{4}(1-\beta)^2$
$A^m B^u$	$\frac{\beta^2}{4}$	$\frac{1}{2}(1-\beta)\beta$	$\frac{1}{4}(1-\beta)^2$	$\frac{1}{2}(1-\alpha)\beta$	$\frac{\alpha\beta}{2}$	$\frac{1}{2}(1-\alpha)(1-\beta)$	$\frac{1}{2}\alpha(1-\beta)$	$\frac{1}{4}(1-\alpha)^2$	$\frac{1}{2}(1-\alpha)\alpha$	$\frac{\alpha^2}{4}$
$A^m B^m$	$\frac{\beta^2}{4}$	$\frac{1}{2}(1-\beta)\beta$	$\frac{1}{4}(1-\beta)^2$	$\frac{\beta^2}{2}$	$\frac{1}{2}(1-\beta)\beta$	$\frac{1}{2}(1-\beta)\beta$	$\frac{1}{2}(1-\beta)^2$	$\frac{\beta^2}{4}$	$\frac{1}{2}(1-\beta)\beta$	$\frac{1}{4}(1-\beta)^2$
$B^u B^u$	0	0	0	0	0	0	0	$(1-\alpha)^2$	$2(1-\alpha)\alpha$	α^2
$B^u B^m$	0	0	0	0	0	0	0	$\frac{1}{4}(-\alpha+\beta+1)^2$	$\frac{1}{2}(\alpha-\beta+1)(-\alpha+\beta+1)$	$\frac{1}{4}(\alpha-\beta+1)^2$
$B^m B^m$	0	0	0	0	0	0	0	β^2	$2(1-\beta)\beta$	$(1-\beta)^2$

(29)

Analytical solutions for $\rho_{G,EG}$ exist but have no easy form and are therefore omitted here. **Figure 2e** plots the correlation between genetic and epigenetic variation for different values of the epimutation rates α and β. Because of the initial proportions (Eq. 27), the correlation is always maximized at $t = 0$. For the epimutation rates estimated in the MA populations (corresponding to the blue line in **Figure 2e**), our model predicts that the correlation between genetic and epigenetic variation would drop below 0.5 in only 828 generations of selfing, and below 0.1 in 2737 generations of selfing. This breakdown is expected to be even faster in outbreeding systems.

1.8 Effects of selection on epialleles

Consider a single locus with epigenotypes $c^m c^m$, $c^m c^u$ and $c^u c^u$, the let the corresponding population frequencies at time t be $\pi_t(c^u c^u)$, $\pi_t(c^m c^u)$ and $\pi_t(c^m c^m)$. We define the fitness of each epigenotype using

Epigenotype	$c^u c^u$	$c^u c^m$	$c^m c^m$
Fitness	w	$\frac{1+w}{2}$	1

where $0 \leq w \leq 1$. Further, we suppose that epialleles can be subject to forward epimutations ($\alpha = Pr(c^u \rightarrow c^m)$) as well as backward epimutations ($\alpha = Pr(c^m \rightarrow c^u)$). Hence, the epigenetic composition of the population at time $t + 1$ is determined both by the relative fitness of the epigenotypes as well as by forward-backward epimutation rates. In a strictly selfing system, the epigenotype frequencies follow the recursion

$$\begin{aligned}
\pi_{t+1}(c^u c^u) &= \frac{\pi_t(c^u c^u)(1-\alpha)^2 w}{\overline{w}} + \frac{\pi_t(c^u c^m)\frac{1}{8}(-\alpha+\beta+1)^2(1+w)}{\overline{w}} \\
&+ \frac{\pi_t(c^m c^m)\beta^2}{\overline{w}} \\
\pi_{t+1}(c^u c^m) &= \frac{\pi_t(c^u c^u)2(1-\alpha)\alpha w}{\overline{w}} + \frac{\pi_t(c^u c^m)\frac{1}{4}(\alpha-\beta+1)(-\alpha+\beta+1)(1+w)}{\overline{w}} \\
&+ \frac{\pi_t(c^m c^m)2(1-\beta)\beta}{\overline{w}} \\
\pi_{t+1}(c^m c^m) &= \frac{\pi_t(c^u c^u)\alpha^2 w}{\overline{w}} + \frac{\pi_t(c^u c^m)\frac{1}{8}(\alpha-\beta+1)^2(1+w)}{\overline{w}} \\
&+ \frac{\pi_t(c^m c^m)(1-\beta)^2}{\overline{w}}
\end{aligned}$$

where $\overline{w}$ is the mean fitness in the population at time t. The equilibrium frequencies of each epigenotype can be derived by calculating $\pi_\infty = \lim_{t\rightarrow\infty} \pi_t$, where $\pi_t = (\pi_t(c^u c^u), \pi_t(c^u c^m), \pi_t(c^m c^m))$. These solutions exist in symbolic form but are too complex to be shown here. The equilibrium epigenotype frequencies are the result of epimutation-selection balance, so that for $w = 1$ (i.e.

no selection) we obtain the equlibrium frequencies given by Eq. 7.

Fig. 2f of the main text plots the values of $\pi_\infty(c^m c^m)$ and $\pi_\infty(c^u c^u)$ as a function of w, and for different values of the epimutation rates (different colors). The most important observation is that - for the epimutation rates estimated in our study ($\sim 10^{-4}$) - the equilibrium frequencies of epigenotype $c^u c^u$ and $c^m c^m$ are close to those expected for epimutation rates that are in the order of magnitude of DNA sequence mutations ($\sim 10^{-9}$).This suggests that CG-type epialleles are expected to respond effectively to selection provided they affect fitness. Moreover, it is intuitively obvious that the unfavored epigenotype $c^u c^u$ will never get purged from the population, even for very small values for α and β or for very high selection pressures (i.e. small w), as backward epimutations ($c^m \rightarrow c^u$) will continually produce unfavorable epiallele c^u. Using more complex population genetic models, Charlesworth and Jain [19] showed that for relatively high forward-backward rates (in the order of 10^{-2}) unfavorable epialleles will persist in populations at intermediate frequencies even in the presence of very strong selection. With the forward-backward rates reported in our study unfavorable epialleles are expected to be rather rare.

Bibliography

[1] Claude Becker et al. "Spontaneous epigenetic variation in the Arabidopsis thaliana methylome". en. In: *Nature* 480.7376 (Dec. 2011), pp. 245–249.

[2] Robert J. Schmitz et al. "Transgenerational Epigenetic Instability Is a Source of Novel Methylation Variants". In: *Science (New York, N.Y.)* 334.6054 (Oct. 2011), pp. 369–373.

[3] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". en. In: *EMBnet.journal* 17.1 (May 2011), pp. 10–12.

[4] Weilong Guo et al. "BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data". en. In: *BMC Genomics* 14.1 (Nov. 2013), p. 774.

[5] Ben Langmead et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". eng. In: *Genome Biology* 10.3 (2009), R25.

[6] Seung Yon Rhee et al. "The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community". In: *Nucleic Acids Research* 31.1 (Jan. 2003), pp. 224–228.

[7] Yoav Benjamini and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency". In: *The Annals of Statistics* 29.4 (Aug. 2001). Mathematical Reviews number (MathSciNet) MR1869245, Zentralblatt MATH identifier01829051, pp. 1165–1188.

[8] Buisine N Quesneville H. and Colot V. "Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets". In: *Genomics* 91 (2008), pp. 467–475.

[9] M. S. Barlett and J. B. S. Haldane. "The theory of inbreeding with forced heterozygosis". In: *J. Genet.* 31 (1934), pp. 327–340.

[10] R. A. Fisher. *The theory of inbreeding*. Oliver and Boyd, Edinburgh, 1949.

[11] K. W. Broman. "Genotype Probabilities at Intermediate Generations in the Construction of Recombinant Inbred Lines". In: *Genetics* 190 (2012), 403–412.

[12] F. Johannes and M. Colomé-Tatché. "Quantitative epigenetics through epigenomic perturbation of isogenic lines". In: *Genetics* 188 (2011), 215–227.

[13] Ramesh Yelagandula et al. "The Histone Variant H2A.W Defines Heterochromatin and Promotes Chromatin Condensation in Arabidopsis". In: *Cell* 158 (2014), pp. 98–109.

[14] Li H. et al. "The Sequence alignment/map (SAM) format and SAMtools". In: *Bioinformatics* 25 (2009), pp. 2078–2079.

[15] Lawrence M et al. "Software for Computing and Annotating Genomic Ranges". In: *PLoS Computational Biology* 9 (2013), e1003118.

[16] L. Baum et al. "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains". In: *Ann Math Stat* 41 (1970), 164–171.

[17] L. Rabiner. "A tutorial on hidden markov models and selected applications in speech recognition". In: *Proc IEEE* 77 (1989), 257–286.

[18] Robert J. Schmitz et al. "Patterns of population epigenomic diversity". en. In: *Nature* 495.7440 (Mar. 2013), pp. 193–198.

[19] B. Charlesworth and K. Jain. "Purifying Selection, Drift, and Reversible Mutation with Arbitrarily High Mutation Rates". In: *Genetics* 198 (2014), pp. 1587–1602.

Figure S1

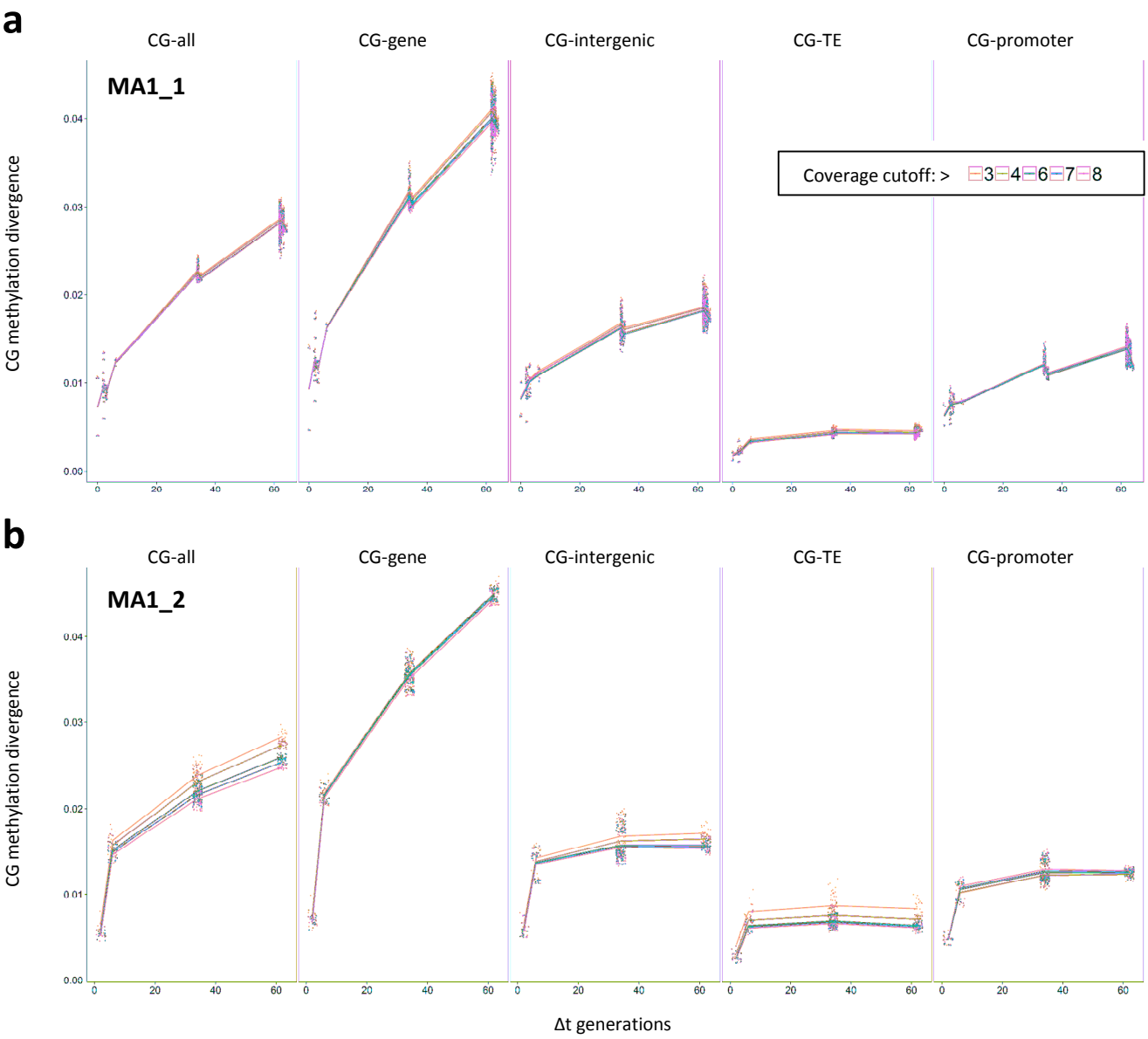
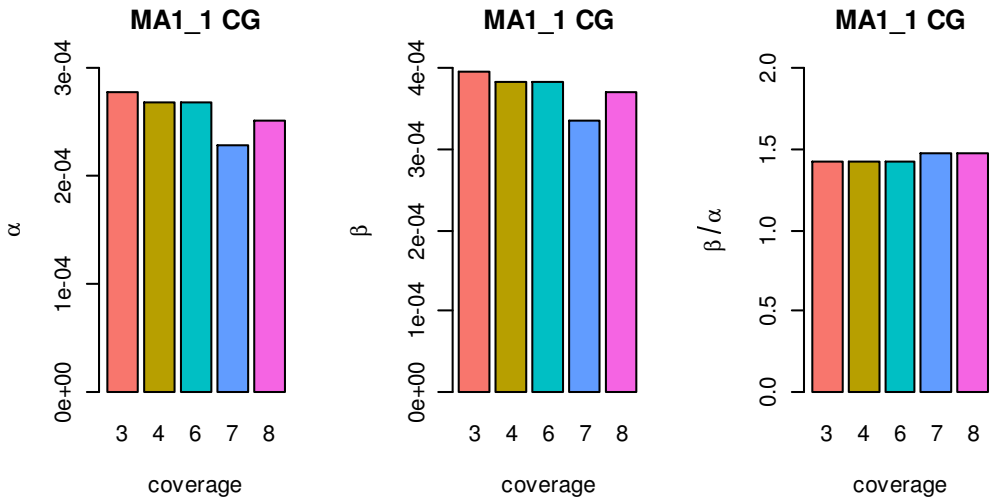


Figure S1: **Effect of coverage cutoff on CG-methylation divergence:** (**a**) Dataset MA1_1 for different annotation categories (CG-all, CG-gene, CG-intergenic, CG-TE and CG-promoter). Colors corresponding to coverages cutoffs > 3, > 4, > 6, > 7 and > 8 are indicated in the legend. (**b**) Same but for dataset MA1_2.

Figure S2

a



b

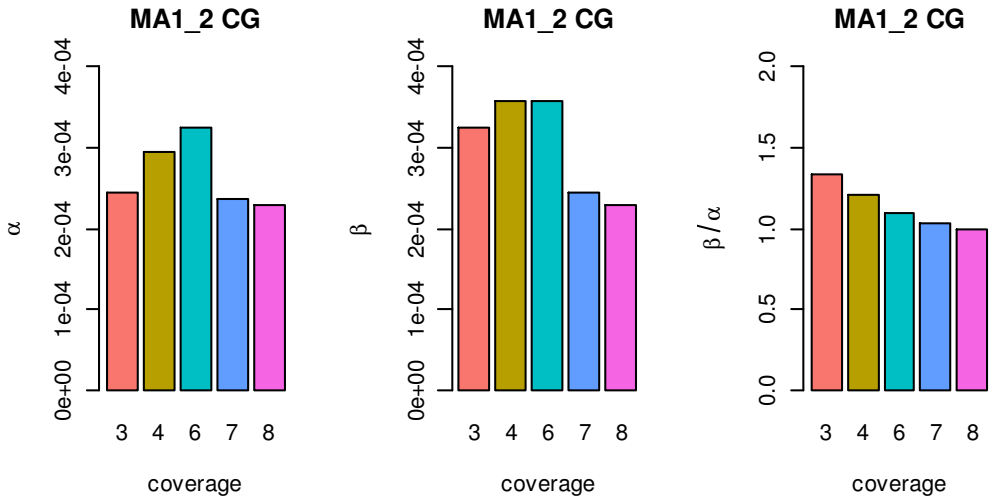
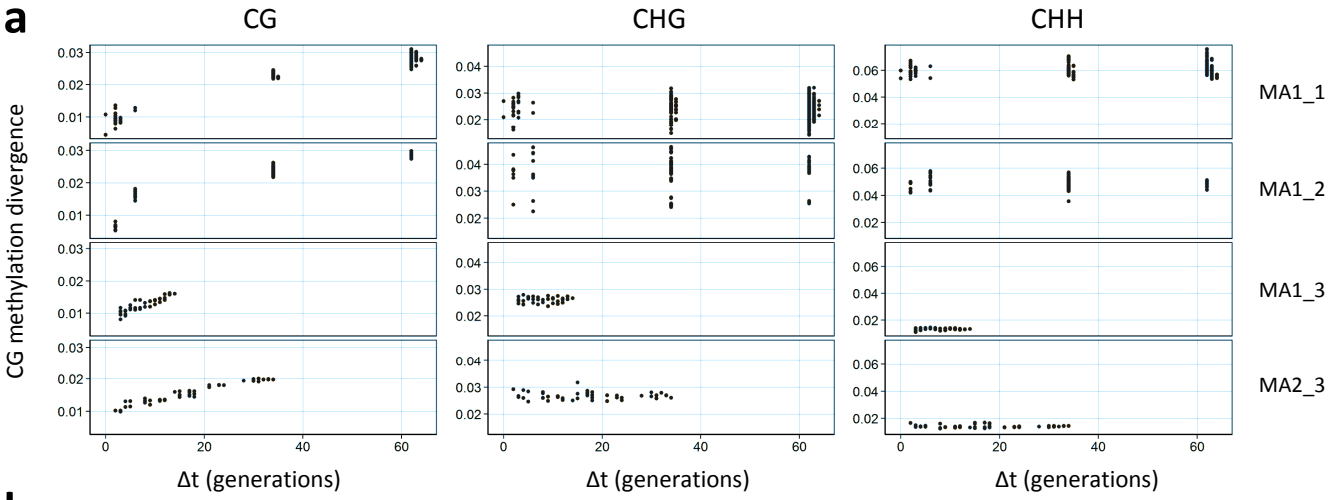


Figure S2: **Estimates of epimutation rates for CG-all with different coverage cutoffs:**
(**a**) Estimates for α, β, and β/α for dataset MA1_1 with coverage cutoffs > 3, > 4, > 6, > 7, > 8.
(**b**) Same but for dataset MA1_2.

Figure S3



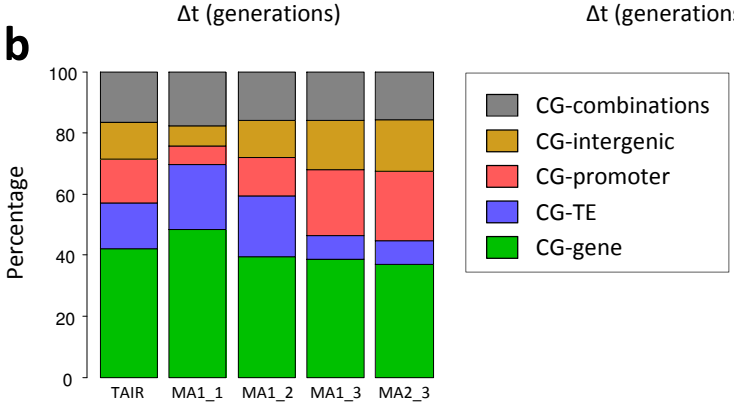


Figure S3: CG methylation divergence and annotation proportions (consensus positions) in all pedigrees: (**a**) Divergence in CG, CHG and CHH sequence context. (**b**) Annotation proportions of the non-normalized consensus CGs of all four pedigrees. The most left bar shows the annotation proportions (or composition) of the CGs of the reference genome (TAIR 10). The CG-combinations category contains CGs with multiple annotations (combination of genes, TEs and / or promoters)

Figure S4

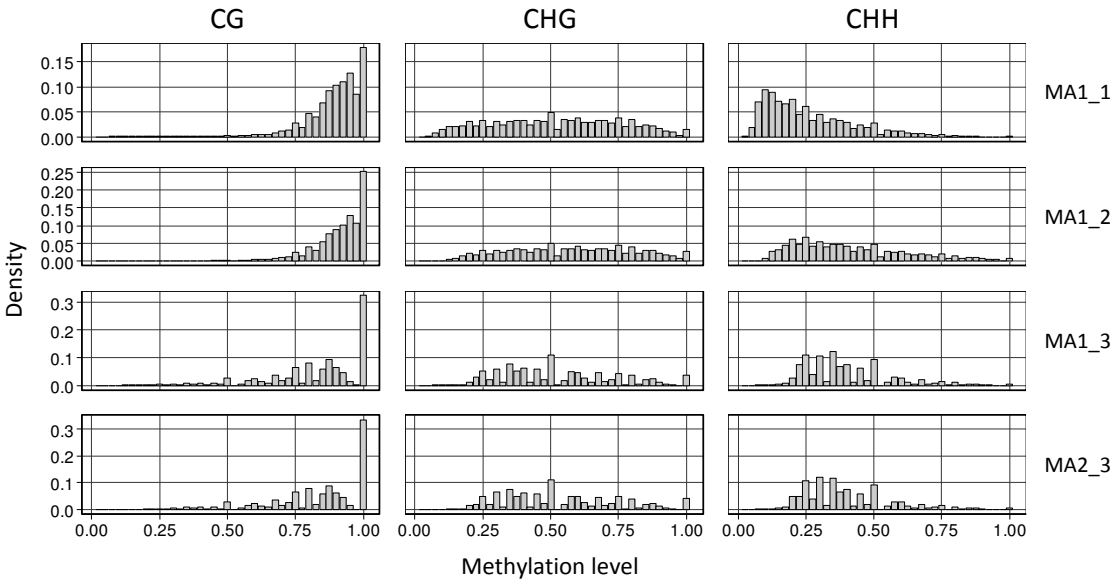


Figure S4: Methylation level for cytosines that have been called methylated, across contexts, for all pedigrees: Number of methylated reads over total number of reads (methylation level) per methylated cytosine for the different sequence contexts CG, CHG and CHH. For CHG and CHH the low methylation levels at methylated cytosines are an indication for cellular heterogeneity.

Figure S5

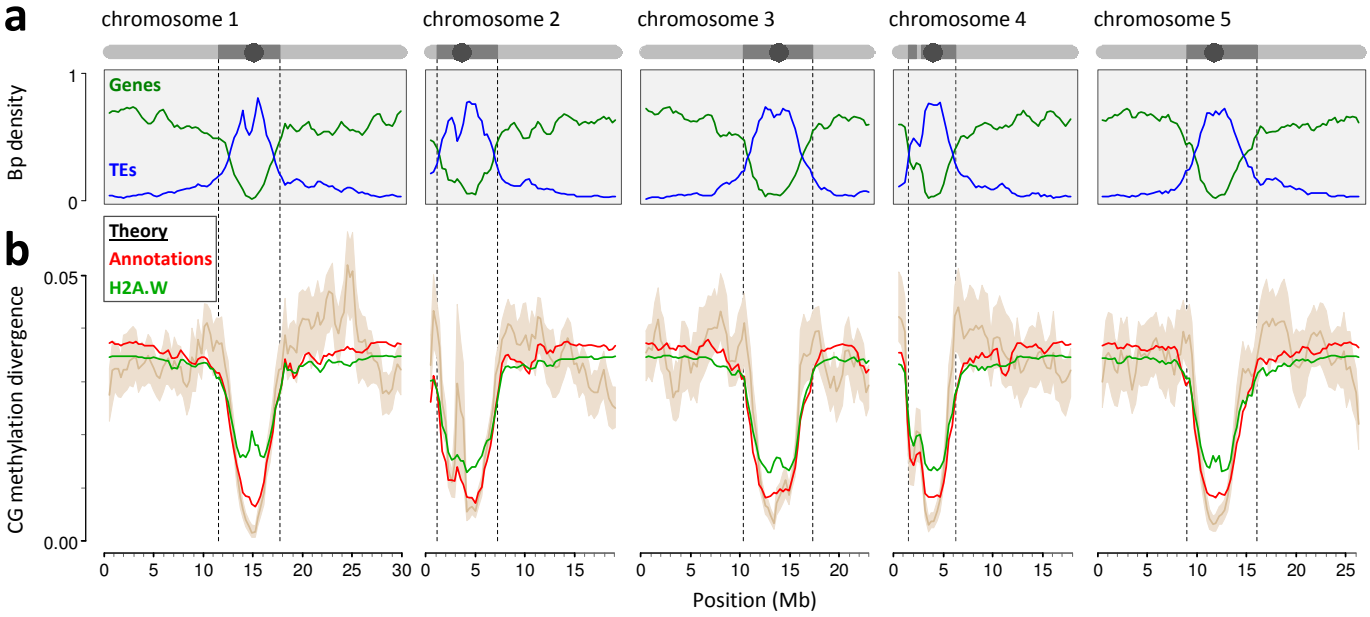


Figure S5: **Predicted CG methylation divergence profiles based on annotations and histone H2A.W:** (**a**) Genome-wide gene (green) and TE (blue) density as well as a schematic representation of chromosomes (dot: centromere; dark grey: pericentromeric region; light gray: arm) (**b**) Observed methylation divergence patterns among the 31st generation MA lines (MA1_1 and MA1_2) (brown). The red line indicates the theoretical prediction based on the estimated epimutation rates per annotation weighted by local annotation densities. The green line indicates the theoretical prediction of divergence based on the estimated epimutation rates per presence/absence of the histone variant H2A.W.

Figure S6

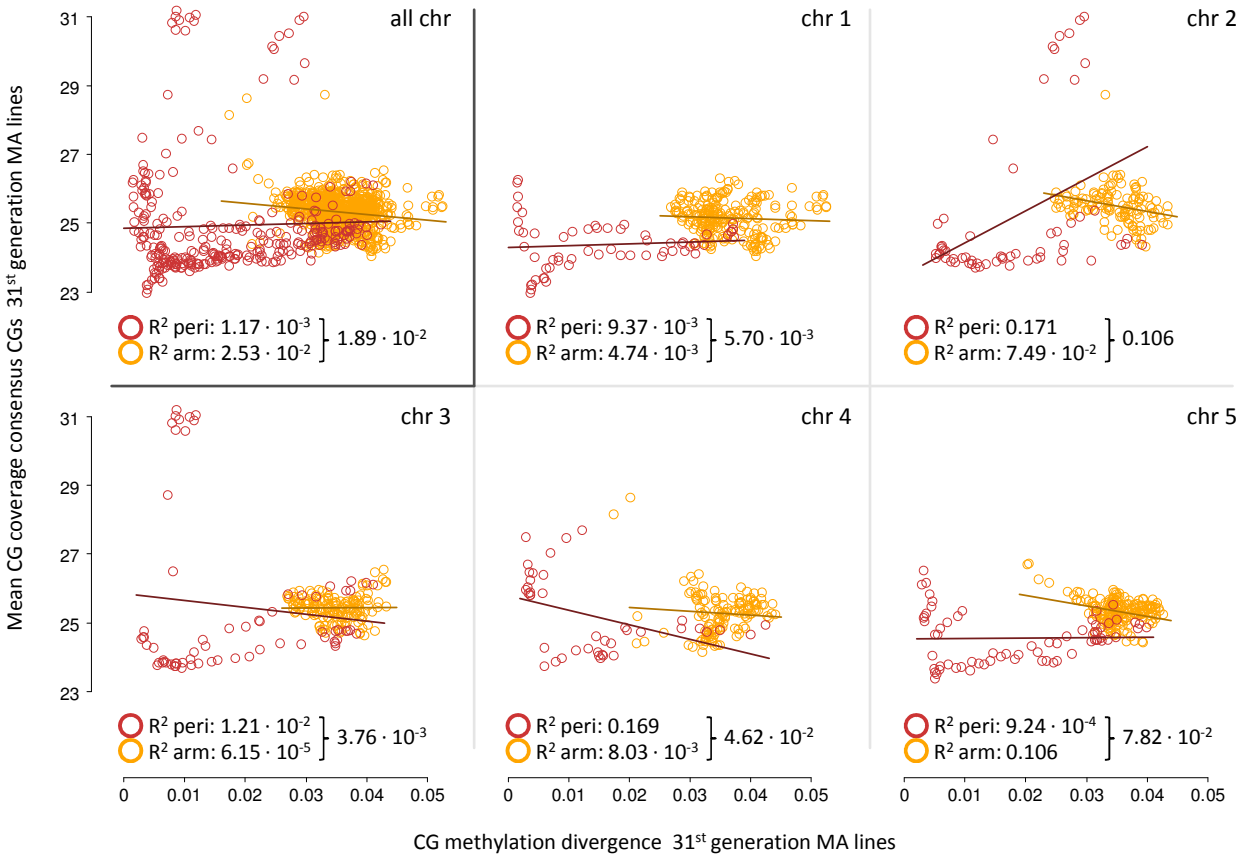


Figure S6: **Correlation between CG methylation divergence among 31st generation MA lines and the mean CG coverage of the consensus CGs of the 31st generation MA lines:** Each dot represents the value of a genomic window (size: 1 Mb; step size: 100 kb). Both the correlations between arm windows (orange) and pericentromeric windows (brown) were determined using a linear model. The correlation is shown for all chromosomes (all chr) as well as for each chromosome separately (chr 1 - 5). Only a low percentage of the genome-wide variation of the mean CG coverage of the 31st generation MA lines could be predicted by the CG methylation divergence among the 31st generation MA lines (< 2%). The chromosomal arm windows show a higher correlation compared to the pericentromeric windows (all chromosomes).

Figure S7

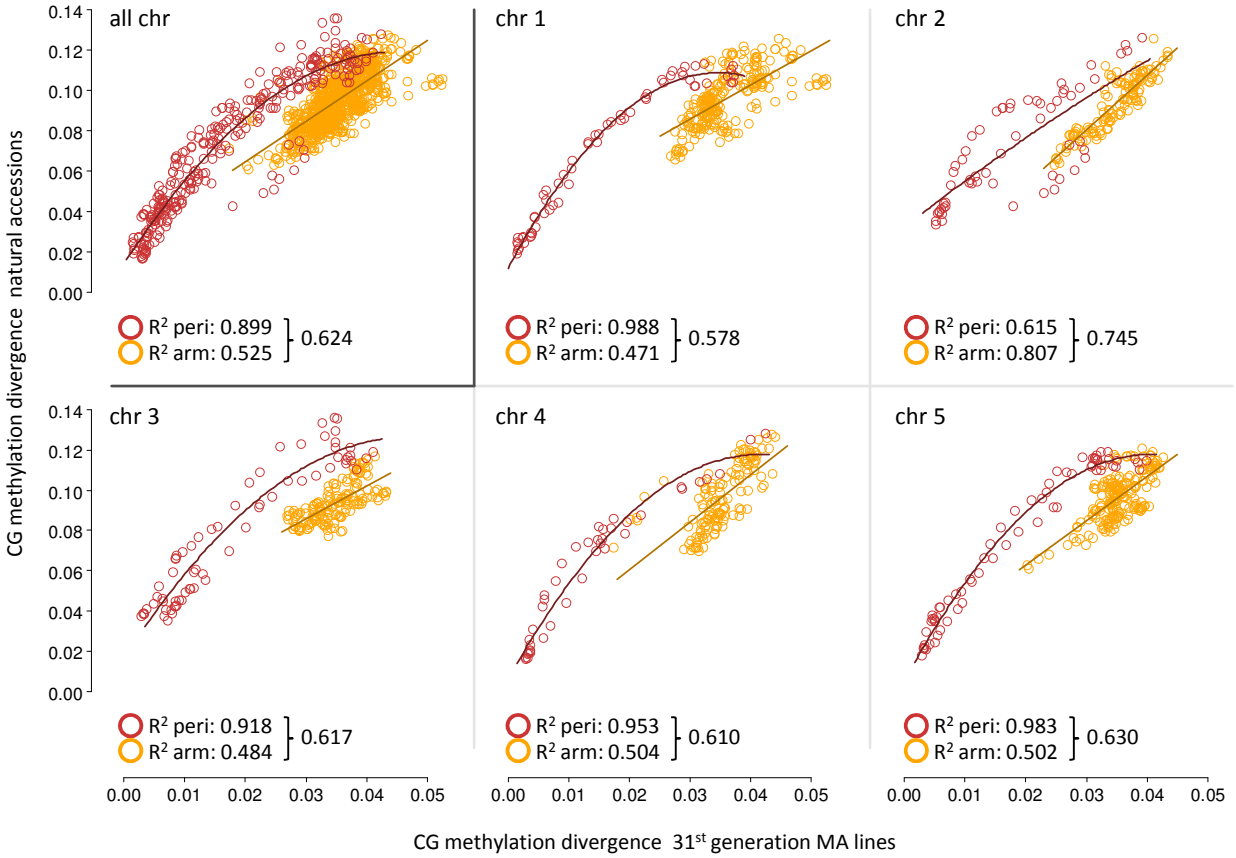


Figure S7: **Correlation between CG methylation divergence among 31st generation MA lines and CG methylation divergence among the natural accessions:** Each dot represents the value of one genomic window (size: 1 Mb; step size: 100 kb). The correlation between arm windows (orange) is determined using a linear model and the correlation of the pericentromeric windows (brown) is determined using a quadratic model. The correlation is shown for all chromosomes (all chr) as well as for each chromosome separately (chr 1 - 5). A high percentage of the genome-wide variation of the CG methylation divergence among the natural accessions could be predicted by the CG methylation divergence among the 31st generation MA lines (~ 60%). The pericentromeric windows show a higher correlation compared to the chromosomal arm windows.

Figure S8

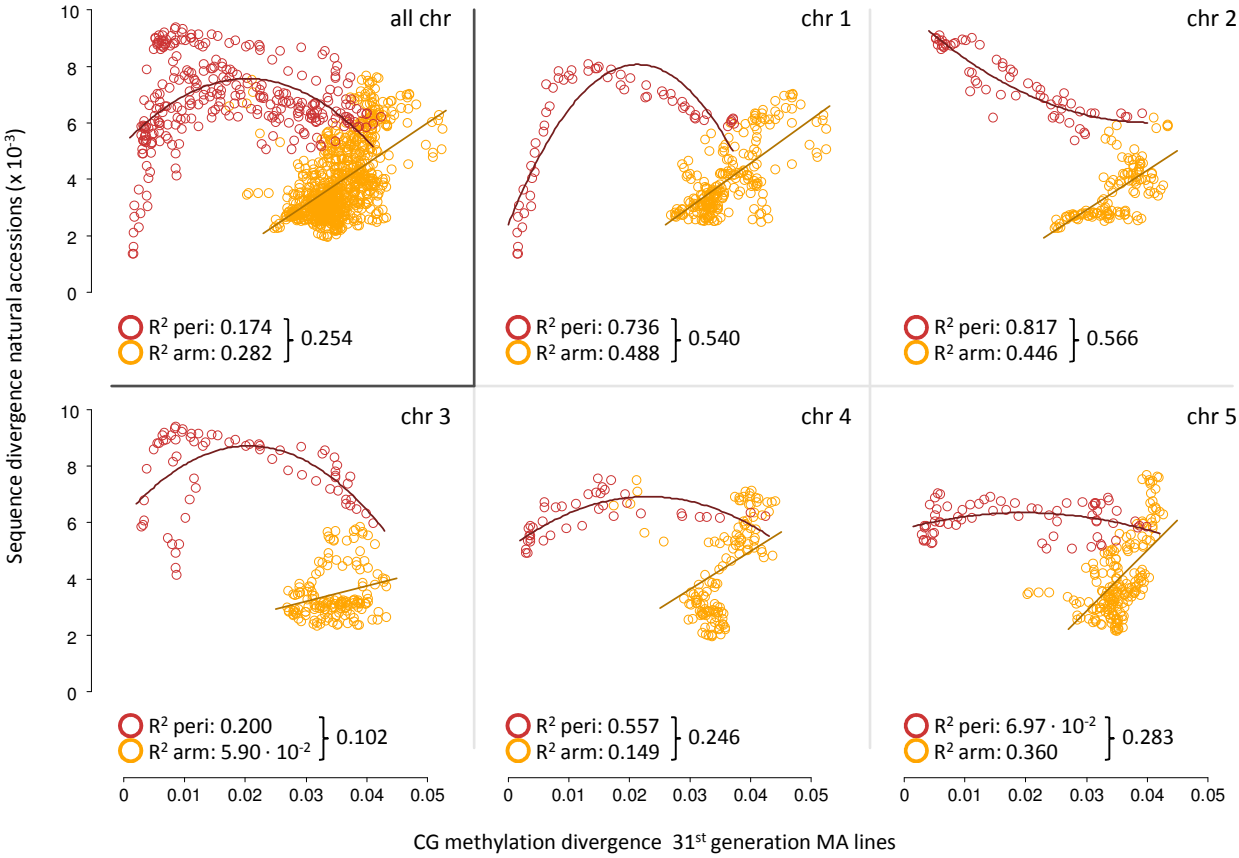


Figure S8: **Correlation between CG methylation divergence among 31st generation MA lines and the sequence divergence among the natural accessions:** Each dot represents the value of a genomic window (size: 1 Mb; step size: 100 kb). The correlation between arm windows (orange) is determined using a linear model and the correlation of the pericentromeric windows (brown) is determined using a quadratic model. The correlation is shown for all chromosomes (all chr) as well as for each chromosome separately (chr 1 - 5). A moderate percentage of the genome-wide variation of the sequence divergence among the natural accessions could be predicted by the CG methylation divergence among the 31st generation MA lines (~ 25%). The chromosomal arm windows show a higher correlation compared to the pericentromeric windows (all chromosomes).

Figure S9

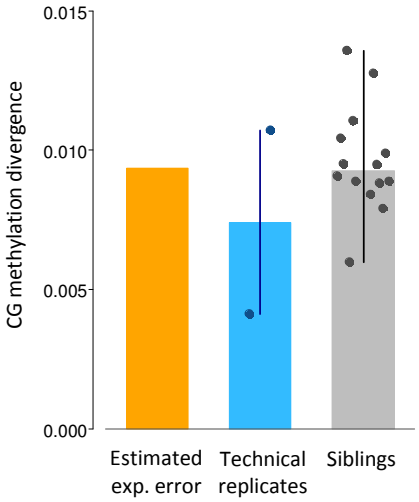


Figure S9: Estimated experimental error, CG methylation divergence between technical replicates and CG methylation divergence between siblings of pedigree MA1_1: The estimated experimental error is similar to the divergence between technical replicates and the divergence between siblings. The CG methylation divergence between technical replicates is however based on only two data points. Only two lines of this pedigree have a technical replicate (see Figure 1a).

Figure S10

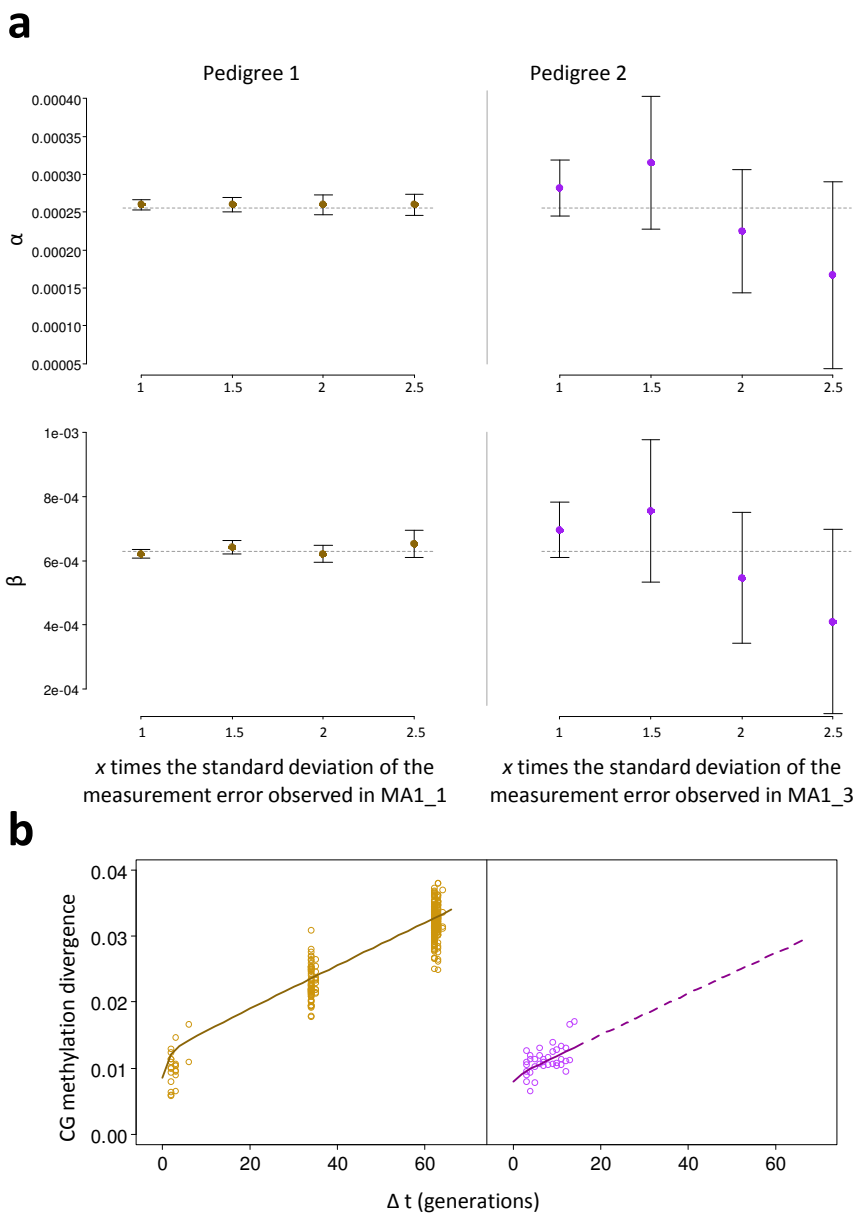


Figure S10: Simulation results: **(a)** Shown are the estimates for α and β based on simulated data. On the left, the mean ± 1 sd of 10 replicate pedigrees of the form of MA1_1 with four different levels of measurement error (x-axis). On the right, the mean ± 1 sd of 10 replicate pedigrees of the form MA1_3 with four different levels of measurement error (x-axis). The true (simulated) values for α and β are shown as horizontal dashed lines. **(b)** Simulated CG methylation divergence for one simulation of pedigree type MA1_1 (left) and for one simulation of pedigree type MA1_3 (right). Data was simulated with an error drawn from a normal distribution with twice the standard deviation of the measurement error observed in MA1_1 and MA1_3.

MA study	Growth conditions	DNA		
		Tissue selection	Extraction	Sonication
Becker et al. 2011	16h days, 8h nights Seed stratified 6 days with 150mM GA	Individual plants Leaf tissue 21-day rosettes	Dneasy (Qiagen)	Covaris 0.5-1µg DNA Sonicated to 300bp via 120s in frequency sweeping, intensity5, duty cycle 10%, 200 cycle burst, time 80 seconds Purified post sonication with Qiaquick PCR Purification columns
Schmitz et al. 2011	16h days, 8h nights	Individual plants Leaf tissue 10-leaf stage	Dneasy (Qiagen)	Covaris 2µg Sonicated to 100bp via cycle #6, duty cycle 20%, cycles/burst 200, time 60 seconds Purified post sonication with Qiagen Dneasy minielute columns
de Graaf et al. 2015	16h days, 8h nights	Individual plants Leaf tissue Random time points	Dneasy (Qiagen)	Covaris 1µg Sonicated to 200bp, intensity 5, duty cycle 10%, 200 cycle burst, time 180 seconds Purified post sonication with AMPure Bead purification

MA study	Library construction			Sequencing
	Library reagents	Bisulphite	Ampification	
Becker et al. 2011	NEBNext DNA Sample Prep Reagent Set 1 Illumina Early Acces Methylation Adapter Oligo Mix	Conversion done with EpiTect Plus DNA Bisulfite (x2) Conversion rate calculated with unmethylated lambda spike-in and unmethylated chloroplast	18 cycles Pfu Cx HotStart Polymerase	Illumina GAIIx 101 bp PE
Schmitz et al. 2011	NebNext DNA Sample Prep Reagent Set 1 Illumina methylated adapters	Conversino done with MethylCode Kit Conversion rate calculated with unmethylated chloroplast	4 cycles Pfu Cx Turbo	HiSeq 2000 101 bp SE
de Graaf et al. 2015	Individual kits/enzymes per step* BIOO NEXTseq adapters	Conversion done MethylCode Kit Conversion rate calculated with unmethylated chloroplast	8 cycles Kapa HiFi Uracil + HotStart	Illumina Nextseq 75 bp SE & 150 bp SE

*End-it Repair Kit; Klenow 3-5' Exonuclease + dA-tailing buffer; T4 Ligase

Table S1: Experimental material, conditions and sequencing approach

Table S2

		MA1_1		MA1_2		MA1_3		MA2_3	
Sequencing and mapping									
Library construction		paired		single		single		single	
Read length (bp)		101		101		150		75 / 150*	
# Samples		27		14		9		10	
Mean coverage		20.889		25.978		6.978		6.808	
Sd coverage		4.733		8.206		0.893		1.119	
CGs	**Ref.**	**Consensus**	**R-Con.**	**Consensus**	**R-Con.**	**Consensus**	**R-Con.**	**Consensus**	**R-Con.**
# CG-all	5,567,714	2,109,186	878,337	2,097,449	1,835,941	2,438,920	1,233,759	2,149,623	1,113,008
# CG-gene	2,339,713	1,022,985	369,103	828,736	771,516	945,097	518,462	795,074	467,719
# CG-TE	840,096	448,216	132,530	418,258	277,020	186,159	186,159	167,939	167,939
# CG-promoter	798,297	125,936	125,936	263,237	263,237	524,531	176,896	484,397	159,583
# CG-intergenic	687,178	141,317	106,986	254,288	223,627	394,802	150,279	366,270	135,570
# CG-gene & TE	34,748	12,790	5,481	14,073	11,458	13,250	7,699	12,294	6,946
# CG-gene & promoter	728,445	325,030	114,916	251,856	240,203	295,342	161,417	247,778	145,619
# CG-TE & promoter	141,961	30,714	22,395	64,881	46,811	77,415	31,457	73,796	28,378
# CG-gene & TE & promoter	6,276	2,198	990	2,120	2,069	2,324	1,390	2,075	1,254

*Line MA2_3_G1_48_r1 was sequenced to 150 bases, the remaining lines were sequenced to 75 bases.

Table S2: Sequencing and mapping summary of MA lines analyzed in this study: Reported are the sequencing specifications, the mapping statistics and the number of consensus CGs used for the analysis (i.e. CGs covered by at least four reads in all individuals of the pedigree). **Consensus** stands for all consensus CGs (CG-all†), while **R-Con.** Stands for consensus CGs that are representative of the genome (CG-all). With gray is indicated to which annotation category the reference consensus CGs were scaled (lowest consensus to reference ratio). The four annotation categories at the bottom are combinations of the non-overlapping annotation categories (e.g. a TE can be located inside a promoter of a gene). When non of the three annotation categories (gene, TE or promoter) were overlapping a CG the CG was classified as intergenic.

Sample	Coverage		Conversion	Cov >= 1				Cov >= 4			
	Mean	Median	(%)	ALLC	CG	CHG	CHH	ALLC	CG	CHG	CHH
MA1_1_G3_26_r1	16.656	10	99.961	29,462,182	4,339,872	4,755,029	20,367,281	18,776,620	3,071,352	3,340,812	12,364,456
MA1_1_G3_87_r1	22.852	17	99.961	33,928,758	4,813,806	5,270,315	23,844,637	23,996,656	3,724,476	4,080,215	16,191,965
MA1_1_G3_87_r2	16.003	9	99.965	28,293,985	4,204,684	4,606,380	19,482,921	17,811,172	2,933,994	3,188,296	11,688,882
MA1_1_G31_29_r1	16.693	12	99.923	33,194,288	4,784,016	5,216,799	23,193,473	22,256,801	3,608,285	3,898,640	14,749,876
MA1_1_G31_29_r2	28.161	24	99.961	37,989,370	5,211,646	5,695,614	27,082,110	29,550,725	4,443,053	4,839,839	20,267,833
MA1_1_G31_29_r3	16.406	13	99.879	34,005,767	4,865,232	5,302,249	23,838,286	22,970,675	3,693,540	3,989,682	15,287,453
MA1_1_G31_39_r1	25.615	21	99.948	37,148,122	5,140,990	5,620,089	26,387,043	28,154,612	4,289,212	4,677,286	19,188,114
MA1_1_G31_39_r2	14.491	12	99.890	34,491,202	4,915,141	5,350,027	24,226,034	22,908,083	3,687,748	3,972,900	15,247,435
MA1_1_G31_49_r1	18.281	15	99.923	35,122,574	4,981,263	5,426,753	24,714,558	24,990,389	3,965,535	4,282,960	16,741,894
MA1_1_G31_49_r2	23.583	21	99.951	37,598,293	5,181,968	5,657,877	26,758,448	28,651,232	4,356,713	4,730,773	19,563,746
MA1_1_G31_59_r1	23.081	21	99.951	37,899,945	5,208,730	5,687,571	27,003,644	29,328,940	4,434,256	4,815,007	20,079,677
MA1_1_G31_59_r2	14.928	10	99.940	31,545,662	4,628,104	5,030,840	21,886,718	20,238,599	3,375,808	3,612,522	13,250,269
MA1_1_G31_79_r1	23.641	20	99.941	37,051,446	5,130,015	5,602,188	26,319,243	27,564,703	4,221,495	4,588,546	18,754,662
MA1_1_G31_79_r2	28.607	20	99.910	35,411,557	4,986,708	5,443,229	24,981,620	25,574,015	4,006,825	4,347,455	17,219,735
MA1_1_G31_89_r1	19.473	16	99.908	35,389,225	4,967,822	5,442,342	24,979,061	25,177,771	3,904,350	4,279,798	16,993,623
MA1_1_G31_89_r2	16.578	10	99.927	31,256,639	4,572,776	4,979,191	21,704,672	19,703,468	3,243,322	3,495,905	12,964,241
MA1_1_G31_99_r1	22.645	17	99.915	34,817,092	4,928,405	5,384,558	24,504,129	24,866,620	3,902,946	4,244,211	16,719,463
MA1_1_G31_99_r2	15.008	8	99.909	29,891,567	4,416,785	4,807,347	20,667,435	18,158,712	3,021,161	3,249,787	11,887,764
MA1_1_G31_99_r3	27.057	21	99.929	35,861,305	5,027,081	5,492,922	25,341,302	26,570,769	4,109,313	4,470,907	17,990,549
MA1_1_G31_109_r1	22.363	18	99.901	35,580,177	4,989,970	5,458,556	25,131,651	25,801,832	3,984,900	4,357,000	17,459,932
MA1_1_G31_109_r2	15.720	10	99.944	31,936,749	4,646,469	5,058,308	22,231,972	20,143,341	3,313,519	3,566,774	13,263,048
MA1_1_G31_119_r1	20.647	16	99.933	33,832,556	4,814,857	5,267,434	23,750,265	23,417,667	3,681,373	4,025,336	15,710,958
MA1_1_G31_119_r2	21.894	15	99.917	34,118,197	4,868,007	5,308,364	23,941,826	23,417,253	3,738,160	4,046,784	15,632,309
MA1_1_G32_39_r1	20.489	15	99.855	33,154,808	4,749,555	5,194,662	23,210,591	22,642,026	3,589,337	3,914,233	15,138,456
MA1_1_G32_39_r2	19.758	11	99.935	28,690,843	4,245,217	4,648,985	19,796,641	18,848,743	3,071,881	3,340,289	12,436,573
MA1_1_G32_49_r1	18.067	10	99.950	28,177,081	4,176,802	4,581,694	19,418,585	18,196,474	2,976,257	3,238,787	11,981,430
MA1_1_G32_49_r2	13.852	7	99.955	26,785,663	4,022,174	4,401,228	18,362,261	16,294,097	2,719,217	2,943,544	10,631,336
Mean	**20.094**	**14.778**	**99.929**	**33,430,928**	**4,771,041**	**5,210,761**	**23,449,126**	**23,185,629**	**3,669,186**	**3,982,900**	**15,533,544**

Table S3: **Sequencing table of MA1_1**

Sample	Coverage		Conversion	Cov >= 1				Cov >= 4			
	Mean	Median	(%)	ALLC	CG	CHG	CHH	ALLC	CG	CHG	CHH
MA1_2_G3_1_r1	31.679	32	97.569	37,354,800	5,152,818	5,611,114	26,590,868	30,523,584	4,591,265	4,943,517	20,988,802
MA1_2_G3_1_r2	19.216	18	99.292	38,501,502	5,094,737	5,590,734	27,816,031	28,946,679	3,986,515	4,393,552	20,566,612
MA1_2_G3_12_r1	36.382	34	99.234	39,790,135	5,295,911	5,792,491	28,701,733	34,105,140	4,860,332	5,262,935	23,981,873
MA1_2_G3_12_r2	37.985	37	99.327	40,829,648	5,353,232	5,868,403	29,608,013	37,095,032	5,072,843	5,534,575	26,487,614
MA1_2_G3_19_r1	30.859	29	99.105	37,411,687	5,139,008	5,594,488	26,678,191	29,792,218	4,472,183	4,809,504	20,510,531
MA1_2_G3_19_r2	19.325	17	99.200	37,307,399	4,968,969	5,450,250	26,888,180	26,283,165	3,636,053	4,000,933	18,646,179
MA1_2_G31_29_r1	33.332	34	99.224	38,878,190	5,244,597	5,728,572	27,905,021	32,476,710	4,743,531	5,127,518	22,605,661
MA1_2_G31_29_r2	19.567	18	99.223	38,497,144	5,119,546	5,608,760	27,768,838	29,536,953	4,158,621	4,556,591	20,821,741
MA1_2_G31_49_r1	31.214	31	96.828	38,862,688	5,233,510	5,716,033	27,913,145	31,927,444	4,667,813	5,044,041	22,215,590
MA1_2_G31_49_r2	16.728	15	98.803	33,178,162	4,729,337	5,116,675	23,332,150	23,363,794	3,620,822	3,888,341	15,854,631
MA1_2_G31_59_r1	31.728	31	99.240	39,119,941	5,255,227	5,740,776	28,123,938	32,386,887	4,725,448	5,102,150	22,559,289
MA1_2_G31_59_r2	17.136	14	99.045	35,569,962	4,852,855	5,293,011	25,424,096	23,666,321	3,443,337	3,747,072	16,475,912
MA1_2_G31_119_r1	32.589	32	99.188	38,807,243	5,239,245	5,721,775	27,846,223	32,290,786	4,721,528	5,104,800	22,464,458
MA1_2_G31_119_r2	10.799	7	98.774	26,073,472	3,971,940	4,239,029	17,862,503	14,804,786	2,449,963	2,575,257	9,779,566
Mean	**26.324**	**24.929**	**98.861**	**37,155,855**	**5,046,495**	**5,505,151**	**26,604,209**	**29,085,679**	**4,225,018**	**4,577,913**	**20,282,747**

Table S4: **Sequencing table of MA1_2**

Sample	Coverage		Conversion	Cov >= 1				Cov >= 4			
	Mean	Median	(%)	ALLC	CG	CHG	CHH	ALLC	CG	CHG	CHH
MA1_3_G18_12_r1	5.890	6	99.727	40,759,062	5,279,295	5,806,968	29,672,799	33,297,923	4,216,761	4,715,428	24,365,734
MA1_3_G19_12_r1	6.280	6	99.729	40,716,782	5,265,717	5,798,343	29,652,722	33,753,337	4,218,754	4,760,432	24,774,151
MA1_3_G20_12_r1	6.593	6	99.734	40,577,670	5,259,325	5,786,095	29,532,250	32,860,283	4,170,087	4,652,896	24,037,300
MA1_3_G21_12_r1	8.592	9	99.776	41,078,114	5,329,038	5,853,612	29,895,464	38,406,182	4,944,751	5,482,508	27,978,923
MA1_3_G25_12_r1	6.493	6	99.814	40,840,032	5,286,609	5,816,809	29,736,614	34,874,690	4,382,073	4,929,478	25,563,139
MA1_3_G26_12_r1	6.788	7	99.788	40,892,933	5,302,145	5,827,860	29,762,928	35,386,038	4,520,785	5,036,694	25,828,559
MA1_3_G28_12_r1	8.221	8	99.791	41,059,633	5,326,196	5,851,081	29,882,356	38,032,639	4,890,271	5,424,849	27,717,519
MA1_3_G29_12_r1	6.706	7	99.806	40,846,724	5,289,061	5,818,916	29,738,747	35,186,081	4,437,553	4,983,544	25,764,984
MA1_3_G30_12_r1	7.233	7	99.799	40,976,906	5,313,784	5,839,128	29,823,994	36,675,833	4,699,079	5,227,507	26,749,247
Mean	**6.977**	**6.889**	**99.774**	**40,860,873**	**5,294,574**	**5,822,090**	**29,744,208**	**35,385,890**	**4,497,790**	**5,023,704**	**25,864,395**

Table S5: **Sequencing table of MA1_3**

Sample	Coverage		Conversion	Cov >= 1				Cov >= 4			
	Mean	Median	(%)	ALLC	CG	CHG	CHH	ALLC	CG	CHG	CHH
MA2_3_G1_48_r1	7.054	6	99.774	40,918,808	5,314,845	5,847,209	29,756,754	33,354,647	4,259,677	4,742,075	24,352,895
MA2_3_G7_48_r1	8.114	8	99.741	41,071,574	5,328,432	5,852,771	29,890,371	38,124,124	4,906,474	5,439,235	27,778,415
MA2_3_G14_48_r1	5.464	5	99.807	40,500,037	5,233,029	5,765,970	29,501,038	30,969,921	3,833,302	4,333,334	22,803,285
MA2_3_G16_48_r1	5.438	5	99.820	40,515,279	5,248,484	5,775,535	29,491,260	30,936,244	3,919,377	4,376,817	22,640,050
MA2_3_G17_48_r1	7.850	8	99.802	41,013,776	5,319,582	5,844,797	29,849,397	37,392,356	4,802,175	5,332,538	27,257,643
MA2_3_G1_82_r1	5.778	6	99.733	40,414,234	5,223,678	5,757,359	29,433,197	31,350,557	3,893,899	4,401,810	23,054,848
MA2_3_G7_82_r1	6.696	6	99.718	40,728,655	5,277,447	5,805,537	29,645,671	34,156,878	4,330,058	4,836,989	24,989,831
MA2_3_G14_82_r1	8.543	9	99.743	41,100,802	5,333,016	5,857,242	29,910,544	38,610,555	4,981,330	5,515,480	28,113,745
MA2_3_G16_82_r1	6.043	6	99.774	40,635,214	5,251,536	5,785,524	29,598,154	32,936,864	4,092,170	4,632,970	24,211,724
MA2_3_G17_82_r1	7.096	7	99.747	40,887,940	5,302,518	5,829,028	29,756,394	35,586,184	4,550,710	5,070,174	25,965,300
Mean	**6.808**	**6.600**	**99.766**	**40,778,632**	**5,283,257**	**5,812,097**	**29,683,278**	**34,341,833**	**4,356,917**	**4,868,142**	**25,116,774**

Table S6: **Sequencing table of MA2_3**

Table S7

Context	Dataset	α	β	β/α	$p_{um}(0)$	c
CG-all†	MA1_1	$2.35 \cdot 10^{-4}$	$3.43 \cdot 10^{-4}$	1.46	$8.86 \cdot 10^{-3}$	$7.41 \cdot 10^{-3}$
	MA1_2	$2.23 \cdot 10^{-4}$	$3.01 \cdot 10^{-4}$	1.35	$8.12 \cdot 10^{-3}$	$1.00 \cdot 10^{-2}$
	MA1_3	$3.23 \cdot 10^{-4}$	$1.33 \cdot 10^{-3}$	4.10	$2.70 \cdot 10^{-3}$	$7.55 \cdot 10^{-3}$
	MA2_3	$1.76 \cdot 10^{-4}$	$7.58 \cdot 10^{-4}$	4.30	$2.41 \cdot 10^{-3}$	$9.64 \cdot 10^{-3}$
	Mean	$\mathbf{2.39 \cdot 10^{-4}}$	$\mathbf{6.82 \cdot 10^{-4}}$	**2.85**	$\mathbf{5.52 \cdot 10^{-3}}$	$\mathbf{8.65 \cdot 10^{-3}}$
CG-all	MA1_1	$2.08 \cdot 10^{-4}$	$3.23 \cdot 10^{-4}$	1.55	$9.16 \cdot 10^{-3}$	$7.42 \cdot 10^{-3}$
	MA1_2	$2.28 \cdot 10^{-4}$	$3.62 \cdot 10^{-4}$	1.59	$8.46 \cdot 10^{-3}$	$1.00 \cdot 10^{-2}$
	MA1_3	$3.69 \cdot 10^{-4}$	$1.13 \cdot 10^{-3}$	3.05	$3.32 \cdot 10^{-3}$	$8.09 \cdot 10^{-3}$
	MA2_3	$2.18 \cdot 10^{-4}$	$7.07 \cdot 10^{-4}$	3.24	$2.50 \cdot 10^{-3}$	$1.00 \cdot 10^{-2}$
	Mean	$\mathbf{2.56 \cdot 10^{-4}}$	$\mathbf{6.30 \cdot 10^{-4}}$	**2.36**	$\mathbf{5.86 \cdot 10^{-3}}$	$\mathbf{8.88 \cdot 10^{-3}}$
CG-gene	MA1_1	$2.77 \cdot 10^{-4}$	$1.00 \cdot 10^{-3}$	3.62	$1.24 \cdot 10^{-2}$	$9.51 \cdot 10^{-3}$
	MA1_2	$3.33 \cdot 10^{-4}$	$9.46 \cdot 10^{-4}$	2.84	$9.23 \cdot 10^{-3}$	$1.33 \cdot 10^{-2}$
	MA1_3	$4.87 \cdot 10^{-4}$	$2.45 \cdot 10^{-3}$	5.03	$2.30 \cdot 10^{-3}$	$6.36 \cdot 10^{-3}$
	MA2_3	$2.93 \cdot 10^{-4}$	$1.49 \cdot 10^{-3}$	5.10	$3.11 \cdot 10^{-3}$	$8.73 \cdot 10^{-3}$
	Mean	$\mathbf{3.48 \cdot 10^{-4}}$	$\mathbf{1.47 \cdot 10^{-3}}$	**4.24**	$\mathbf{6.75 \cdot 10^{-3}}$	$\mathbf{9.47 \cdot 10^{-3}}$
CG-TE	MA1_1	$4.80 \cdot 10^{-4}$	$1.62 \cdot 10^{-5}$	$3.38 \cdot 10^{-2}$	$1.94 \cdot 10^{-3}$	$1.86 \cdot 10^{-3}$
	MA1_2	$1.68 \cdot 10^{-4}$	$7.76 \cdot 10^{-6}$	$4.63 \cdot 10^{-2}$	$5.33 \cdot 10^{-3}$	$5.18 \cdot 10^{-3}$
	MA1_3	$8.47 \cdot 10^{-4}$	$2.47 \cdot 10^{-4}$	$2.92 \cdot 10^{-1}$	$7.76 \cdot 10^{-3}$	$1.57 \cdot 10^{-2}$
	MA2_3	$5.77 \cdot 10^{-4}$	$1.64 \cdot 10^{-4}$	$2.84 \cdot 10^{-1}$	$1.12 \cdot 10^{-3}$	$1.82 \cdot 10^{-2}$
	Mean	$\mathbf{5.18 \cdot 10^{-4}}$	$\mathbf{1.09 \cdot 10^{-4}}$	$\mathbf{2.10 \cdot 10^{-1}}$	$\mathbf{4.04 \cdot 10^{-3}}$	$\mathbf{1.02 \cdot 10^{-2}}$
	Mean*	$\mathbf{3.24 \cdot 10^{-4}}$	$\mathbf{1.20 \cdot 10^{-5}}$	$\mathbf{4.00 \cdot 10^{-2}}$	$\mathbf{3.64 \cdot 10^{-3}}$	$\mathbf{3.52 \cdot 10^{-3}}$
CG-promoter	MA1_1	$4.14 \cdot 10^{-5}$	$1.72 \cdot 10^{-4}$	4.16	$6.42 \cdot 10^{-3}$	$6.32 \cdot 10^{-3}$
	MA1_2	$2.92 \cdot 10^{-5}$	$1.33 \cdot 10^{-4}$	4.56	$8.46 \cdot 10^{-3}$	$5.82 \cdot 10^{-3}$
	MA1_3	$9.33 \cdot 10^{-5}$	$1.40 \cdot 10^{-3}$	15.0	$1.06 \cdot 10^{-3}$	$4.91 \cdot 10^{-3}$
	MA2_3	$4.28 \cdot 10^{-5}$	$6.45 \cdot 10^{-4}$	15.1	$1.47 \cdot 10^{-3}$	$6.18 \cdot 10^{-3}$
	Mean	$\mathbf{5.17 \cdot 10^{-5}}$	$\mathbf{5.88 \cdot 10^{-4}}$	**11.4**	$\mathbf{4.35 \cdot 10^{-3}}$	$\mathbf{5.81 \cdot 10^{-3}}$
CG-intergenic	MA1_1	$1.46 \cdot 10^{-4}$	$6.85 \cdot 10^{-5}$	$4.69 \cdot 10^{-1}$	$9.23 \cdot 10^{-3}$	$8.36 \cdot 10^{-3}$
	MA1_2	$6.13 \cdot 10^{-5}$	$6.36 \cdot 10^{-5}$	1.04	$1.15 \cdot 10^{-2}$	$8.18 \cdot 10^{-3}$
	MA1_3	$1.70 \cdot 10^{-4}$	$7.69 \cdot 10^{-4}$	4.52	$2.23 \cdot 10^{-3}$	$8.09 \cdot 10^{-3}$
	MA2_3	$8.35 \cdot 10^{-5}$	$4.01 \cdot 10^{-4}$	4.80	$1.96 \cdot 10^{-3}$	$9.45 \cdot 10^{-3}$
	Mean	$\mathbf{1.15 \cdot 10^{-4}}$	$\mathbf{3.25 \cdot 10^{-4}}$	**2.83**	$\mathbf{6.24 \cdot 10^{-3}}$	$\mathbf{8.52 \cdot 10^{-3}}$

Table S7: Forward epimutation rate, backward epimutation rate, expected number of epiheterozygots at the founder and estimated experimental error per experiment and per context. For CG-all the values were determined twice. One time using all consensus CGs for the analysis (CG-all†) and one time using a sample with the same annotation proportions as the TAIR 10 reference genome (CG-all). When indicated with *, the value is calculated taking only MA1_1 and MA1_2 into account.

Table S8

Accession	Tissue	# Ref CGs Cov >= 4	Accession	Tissue	# Ref CGs Cov >= 4	Accession	Tissue	# Ref CGs Cov >= 4
Not selected accessions			Ca_0	Leaf	2,279,403	Hs_0	Leaf	2,781,950
			Ba_1	Leaf	2,293,799	Nok_3	Bud	2,797,661
Rd_0	Leaf	689,987	Yo_0	Leaf	2,299,997	Seattle_0	Leaf	2,808,165
Ann_1	Leaf	866,895	Wt_5	Leaf	2,315,447	Bu_0	Leaf	2,828,071
Bik_1	Leaf	1,235,259	Bl_1	Leaf	2,317,892	Tscha_1	Bud	2,833,587
Boot_1	Leaf	1,730,121	Ragl_1	Bud	2,318,861	Abd_0	Leaf	2,846,882
Ei_2	Leaf	1,754,817	Baa_1	Leaf	2,344,622	Rome_1	Bud	2,897,355
Per_1	Leaf	1,759,888	Knox_18	Leaf	2,354,506	Br_0	Bud	2,901,811
Su_0	Leaf	1,788,179	Pu2_7	Bud	2,378,215	Westkar_4	Bud	2,906,675
			Kondara	Leaf	2,395,907	Ob_0	Bud	2,911,996
Selected accessions			Vind_1	Leaf	2,397,159	Lan_0	Leaf	2,940,569
			Cal_0	Leaf	2,429,694	Se_0	Bud	2,947,322
Si_0	Leaf	1,802,040	Da_1_12	Leaf	2,431,459	Sp_0	Bud	2,976,842
Cnt_1	Leaf	1,808,005	Kro_0	Leaf	2,435,736	Van_0	Bud	3,013,772
RRS_7	Leaf	1,859,948	Sei_0	Bud	2,437,088	Bsch_0	Leaf	3,021,106
An_1	Leaf	1,896,738	Ove_0	Bud	2,438,836	Np_0	Bud	3,040,823
Pi_0	Leaf	1,904,252	Jm_0	Leaf	2,447,422	Ema_1	Bud	3,055,389
Chi_0	Leaf	1,907,148	Wa_1	Leaf	2,464,725	Uk_1	Bud	3,088,578
Et_0	Leaf	1,913,880	In_0	Leaf	2,471,190	Fr_2	Bud	3,097,434
Bs_1	Leaf	1,929,437	**Kelsterbach_4**	**Leaf**	**2,473,192**	Old_1	Bud	3,104,895
Di_G	Leaf	1,932,294	Krot_0	Leaf	2,492,353	Wl_0	Bud	3,123,667
Stw_0	Leaf	1,978,991	Es_0	Leaf	2,494,428	Pog_0	Bud	3,125,162
Ak_1	Leaf	1,981,798	Ag_0	Bud	2,500,530	Neo_6	Bud	3,133,633
Anholt_1	Leaf	2,001,598	Dja_1	Leaf	2,503,046	Mh_0	Bud	3,141,404
Kas_1	Leaf	2,003,826	Sorbo	Bud	2,507,523	Kin_0	Bud	3,157,658
Ga_0	Leaf	2,018,069	Aa_0	Leaf	2,523,547	Rennes_1	Bud	3,163,011
Gr_1	Leaf	2,023,230	Ha_0	Leaf	2,524,505	Nc_1	Bud	3,178,632
Pt_0	Leaf	2,033,459	Chat_1	Leaf	2,545,889	**Cvi_0**	**Leaf**	**3,186,545**
Je_0	Leaf	2,054,801	Kyoto	Leaf	2,551,533	Pla_0	Bud	3,218,631
Gie_0	Leaf	2,075,285	Lm_2	Leaf	2,554,950	Ms_0	Bud	3,231,570
Uod_1	Bud	2,083,418	Col_0	Leaf	2,564,466	Litva	Bud	3,255,223
Ler_1	**Leaf**	**2,086,686**	Rubezhnoe_1	Bud	2,585,509	Zdr_1	Bud	3,280,094
Mc_0	Leaf	2,101,747	Pro_0	Bud	2,594,791	Sus_1	Bud	3,311,339
En_D	Leaf	2,104,639	Tamm_2	Bud	2,600,460	Ta_0	Bud	3,333,674
Altai_5	Leaf	2,116,306	Ra_0	Leaf	2,600,502	Li_2_1	Bud	3,334,684
Bla_1	Leaf	2,116,416	La_0	Leaf	2,600,689	Ws_2	Bud	3,364,964
Lp2_2	Leaf	2,117,646	Kl_5	Leaf	2,612,669	Sg_1	Leaf	3,366,980
Fi_0	Leaf	2,117,747	**Rmx_A02**	**Bud**	**2,614,208**	Mz_0	Bud	3,384,058
Appt_1	Leaf	2,122,766	Pu2_23	Bud	2,663,916	Co_1	Bud	3,391,122
Benk_1	Leaf	2,124,851	Er_0	Bud	2,672,750	Is_0	Bud	3,403,359
Com_1	Leaf	2,125,117	HR_5	Bud	2,681,750	Tol_0	Bud	3,445,852
Got_7	Leaf	2,136,512	Hey_1	Leaf	2,686,419	Gel_1	Leaf	3,450,055
Db_1	Leaf	2,148,635	Pna_17	Bud	2,703,910	Bor_4	Leaf	3,474,232
Est	Leaf	2,185,670	Amel_1	Leaf	2,721,759	Utrecht	Bud	3,524,232
Kz_9	Leaf	2,215,079	El_0	Leaf	2,734,257	Nw_0	Bud	3,607,743
Gy_0	Leaf	2,225,161	Tu_0	Bud	2,738,702	Ty_0	Bud	3,825,644
Gre_0	Leaf	2,240,201	Sq_8	Bud	2,745,427	Gifu_2	Leaf	4,301,803
Gu_0	Leaf	2,250,626	Anz_0	Leaf	2,751,450			
Kil_0	Leaf	2,260,851	Rld_1	Bud	2,760,088			

Table S8: **Selection of natural accessions and number of reference CGs with sufficient coverage:** Listed are the 140 accessions for which MethylC-seq data was available (Schmitz et al. 2013). Seven accessions with the lowest number of reference CGs with sufficient coverage (>= 4) were excluded from the analysis (top left, 5% of total). Light gray rectangles indicate accessions that were not used in the analysis of sequence divergence.