OXFORD GENETICS

GENETICS, 2021, 218(2), iyab061
DOI: 10.1093/genetics/iyab061
Advance Access Publication Date: 19 April 2021
Investigation
Highlighted Article

Gene body methylation is under selection in *Arabidopsis thaliana*

Aline Muyle [1,*] Jeffrey Ross-Ibarra,[2] Danelle K. Seymour,[3] and Brandon S. Gaut[1]

[1]Ecology and Evolutionary Biology, University of California, Irvine, Irvine, CA 92697-2525, USA,
[2]Evolution and Ecology, Center for Population Biology and Genome Center, University of California, Davis, Davis, CA 95616, USA, and
[3]Botany & Plant Sciences, University of California, Riverside, Riverside, CA 92521, USA

*Corresponding author: Ecology and Evolutionary Biology, University of California, Irvine, 321 Steinhaus Hall, Irvine, CA 92697-2525, USA. Email: amuyle@uci.edu

Abstract

In plants, mammals and insects, some genes are methylated in the CG dinucleotide context, a phenomenon called gene body methylation (gbM). It has been controversial whether this phenomenon has any functional role. Here, we took advantage of the availability of 876 leaf methylomes in *Arabidopsis thaliana* to characterize the population frequency of methylation at the gene level and to estimate the site-frequency spectrum of allelic states. Using a population genetics model specifically designed for epigenetic data, we found that genes with ancestral gbM are under significant selection to remain methylated. Conversely, ancestrally unmethylated genes were under selection to remain unmethylated. Repeating the analyses at the level of individual cytosines confirmed these results. Estimated selection coefficients were small, on the order of $4N_e s = 1.4$, which is similar to the magnitude of selection acting on codon usage. We also estimated that *A. thaliana* is losing gbM threefold more rapidly than gaining it, which could be due to a recent reduction in the efficacy of selection after a switch to selfing. Finally, we investigated the potential function of gbM through its link with gene expression. Across genes with polymorphic methylation states, the expression of gene body methylated alleles was consistently and significantly higher than unmethylated alleles. Although it is difficult to disentangle genetic from epigenetic effects, our work suggests that gbM has a small but measurable effect on fitness, perhaps due to its association to a phenotype-like gene expression.

Keywords: Site frequency spectrum; selection efficacy; DNA methylation; gene expression

Introduction

Cytosine DNA methylation is a type of epigenetic mark in which a methyl group is added to the 5th carbon of cytosines. In plants, it can occur in three sequence contexts—CG, CHG, and CHH (where H stands for A, T, or C)—but levels and patterns of DNA methylation vary among genomic regions. In flowering plants, methylation in all three contexts has a well-established repressive function on transposable elements (TEs) and regulatory elements (Luo *et al.* 2018; Schmitz *et al.* 2019). Both CHG and CHH methylation within genes are associated with reduced expression levels in angiosperms, together with CG methylation in the promoter region (Niederhuth *et al.* 2016). In contrast, exons in plants, insects and mammals are sometimes methylated only in the CG context; this gene body methylation (gbM) can be found within moderately and constitutively expressed housekeeping genes (Zhang *et al.* 2006; Neri *et al.* 2017; Schmitz *et al.* 2019) and is linked to active transcription in plants (Zhang *et al.* 2006; Zilberman *et al.* 2007; Cokus *et al.* 2008; Lister *et al.* 2008). However, it is not yet clear if gbM has a function, because the study of mutants deprived of gbM has failed to reveal a clear effect on phenotype (Teixeira and Colot 2009; Bewick and Schmitz 2017; Zilberman 2017).

The mechanisms responsible for the establishment of gbM in plants have recently been clarified, due in large part to studies in *Eutrema salsugineum*, a close relative of *Arabidopsis thaliana* that lacks both gbM and the *CHROMOMETHYLASE 3* (*CMT3*) gene (Bewick *et al.* 2016). The CMT3 protein has previously been shown to be involved in a self-reinforcing feedback loop: the histone mark H3K9me2 is recognized by CMT3 which then *de novo* methylates nearby cytosines in the CHG context and in turn leads to H3K9 methylation (Kawashima and Berger 2014). The deposition of CHG methylation typically suppresses transcription, but it is removed within transcribed genic regions by INCREASED IN BONSAI METHYLATION 1 (IBM1) (Saze *et al.* 2008; Miura *et al.* 2009).

The critical role of CMT3 in gbM establishment is supported by the facts that not just one but two Brassicaceae species have independently lost *CMT3* and both lack gbM (Bewick *et al.* 2016; Niederhuth *et al.* 2016). Moreover, transgenic reinsertion of *CMT3* into *E. salsugineum* re-establishes genic methylation in all three contexts in a subset of genes that tend to be orthologous to gbM genes in *A. thaliana* (Wendte *et al.* 2019). This subset of genes has been called "CHG-gain" genes (Wendte *et al.* 2019), and remarkably, these genes remained methylated only in the CG context following the loss of the *CMT3* transgene (Wendte *et al.* 2019). It remains unclear how CMT3 (and/or H3K9me2) is directed to a specific subset of genes for *de novo* DNA methylation and how these CHG-gain genes also become *de novo* methylated in the

Received: March 09, 2021. **Accepted:** April 07, 2021

CG and CHH contexts (Wendte et al. 2019), but *cmt3* mutants in *A. thaliana* clearly demonstrate that CMT3 does not affect the maintenance of gbM once it is established (Stroud et al. 2013). Once CG methylation is established, it is maintained by METHYLTRANSFERASE 1 (MET1), which adds a methyl group on the symmetrical CG dinucleotide of a complementary DNA strand during cell division (Kawashima and Berger 2014). Maintenance by MET1 is an inherently error-prone process, as illustrated by epimutation accumulation in *A. thaliana* (Becker et al. 2011; Schmitz et al. 2011; van der Graaf et al. 2015). The accumulation of these epimutations over time illustrates that CG methylation is heritable.

Although gbM is widespread across species and relatively common within a genome—it is found for example in ~20% of *A. thaliana* genes (Takuno and Gaut 2012)—it remains unclear whether gbM is functionally relevant (Teixeira and Colot 2009; Bewick and Schmitz 2017; Zilberman 2017). The question of its potential function has focused on three interrelated hypotheses. The first is that gbM affects gene expression. This hypothesis is supported by the fact that gbM genes exhibit a positive correlation between methylation and expression levels across genes (Zhang et al. 2006; Zilberman et al. 2007; Takuno and Gaut 2012), suggesting either that gbM might cause higher expression or, conversely, that active transcription drives gbM (Teixeira and Colot 2009). However, further tests of this association have led to contradictory results. For example, not all highly expressed genes have gbM in *A. thaliana* (Zhang et al. 2006; Zilberman et al. 2007), illustrating that any association is not absolute. The association has also been tested experimentally in epigenetic recombinant inbred lines (epiRILs) that were developed from the cross of a *met1* mutant and wild-type (WT) *A. thaliana*, followed by eight generations of inbreeding (Reinders et al. 2009). The resulting epiRILs had a mosaic methylome, with regions that have normal CG methylation derived from the WT parent and other regions derived from the *met1* mutant that originally lacked gbM. Analysis of gene expression in these lines detected no significant changes in the *met1* derived regions of epiRILs compared to orthologous WT regions (Bewick et al. 2016). Moreover, the epiRILs did not reestablish the original pattern of gbM after eight generations of epimutations (Bewick et al. 2016), suggesting that expression was not sufficient to drive gbM reestablishment, at least not within a few generations. However, Zilberman et al. (2007) found that both methylated and unmethylated genes were upregulated in *met1* mutants using microarray data, suggesting *met1* methylation mutants may have unanticipated global expression effects that make them a poor system for studying the association between gbM and expression.

Another approach to test for associations between gbM and expression has been comparative genomics, which has the advantage of integrating effects over evolutionary time. Here, again the results have been inconsistent. For example, Bewick et al. (2016) and Bewick et al. (2019) found no effect of the loss of gbM on gene expression in *E. salsugineum* compared to *A. thaliana*. In contrast, Muyle and Gaut (2019) found a small but significant decrease in expression associated with genes that lost gbM in *E. salsugineum*, based on a reanalysis of the data from Bewick et al. (2016). In another effort, Takuno et al. (2017) identified genes that changed methylation status between *A. thaliana* and *Arabidopsis lyrata*. They found a trend: genes that had gained gbM between species tended to also shift toward higher expression levels. Finally, Seymour and Gaut (2019) studied eight grass species and found that genes that were gbM in all eight species tended to have higher and less variable expression, although the effect is small. This last observation is consistent with previous observations that gbM is associated with less variable gene expression both within and between species (Zilberman et al. 2008; Coleman-Derr and Zilberman 2012; Steige et al. 2017; Takuno et al. 2017; Horvath et al. 2019; Seymour and Gaut 2019), suggesting it has a homeostatic effect on expression (Zilberman 2017).

In addition to a potential—but unresolved—association with gene expression, a second hypothesis of gbM function is that it prevents aberrant internal and/or antisense transcription (Tran et al. 2005; Maunakea et al. 2010). Here, again the evidence is unclear because studies comparing gbM mutants to WT mouse embryonic stem cells have been contradictory (Neri et al. 2017; Teissandier and Bourc'his 2017). In plants, Bewick et al. (2016) found no evidence that gbM prevents antisense transcription in *met1* derived regions of *A. thaliana* epiRILs compared to orthologous WT regions. However, Choi et al. (2020) has shown that gbM and histone H1 jointly suppress antisense transcription in a comparison of *met1, h1* double mutants to WT *A. thaliana*.

The third hypothesis is that gbM improves splicing fidelity and prevents intron retention. There is some evidence for this hypothesis, because the alteration of DNA methylation impacts alternative splicing in honey bee and mouse embryonic stem cells (Li-Byarlay et al. 2013; Yearim et al. 2015). Horvath et al. (2019) has found evidence to support this hypothesis by comparing gbM genes to unmethylated genes in *A. thaliana*, but Bewick et al. (2016) found no evidence for this effect by comparing *met1* epiRILs to WT plants. Overall, the contradictory findings regarding the possible function of gbM suggest that its effects, if any, must be relatively small.

While assays of the functional relevance of gbM have provided mixed results, evolutionary patterns of gbM have provided consistent but indirect evidence of its potential importance. Across plant species, gbM genes are generally longer, enriched for housekeeping and other important functions and evolve more slowly than unmethylated genes (Takuno and Gaut 2012, 2013; Takuno et al. 2017; Seymour and Gaut 2019). Moreover, comparative analyses have shown that gbM is conserved for orthologous genes between species as distantly related as ferns and angiosperms (Takuno and Gaut 2013; Seymour et al. 2014; Niederhuth et al. 2016; Takuno et al. 2016; Seymour and Gaut 2019). This last characteristic of gbM is surprising because DNA methylation is mutagenic and elevates C to T substitutions (Bird 1980). Hence, the conservation of gbM over millions of years suggests that the mutagenic feature of methylation is counterbalanced by an advantageous effect that acts to maintain gbM in specific genes (Zilberman 2017). However, another possible explanation for the strong conservation of gbM within a specific set of genes is that *de novo* methylation biases, such as those that target the CHG-gain genes of *E. salsugineum* (Wendte et al. 2019), have been conserved across species over vast periods of evolutionary time.

Clearly several questions about gbM function and evolution remain unresolved. Here, we move away from experiments and comparative studies and employ population genetic approaches to study gbM. Thus far, the tools of population genetics have been applied to epigenetic phenomena in only a handful of studies. For example, van der Graaf et al. (2015) found similar epimutation rates between *A. thaliana* populations and >31 generations of epimutation accumulation lines, suggesting that selection has not impacted global patterns of CG methylation diversity in that species. They nonetheless argued, based on the rate of epimutation events, that selection of epiallelic states could be an important process. Wang and Fan (2014) developed a modification of Tajima's D for application to methylation data and used it to

demonstrate that new genes have an excess of rare epialleles, which they interpreted was consistent with directional selection on an epigenetic state. Two other studies have used site frequency spectra (SFS) to test for selection on methylation data. In the first, Vidalis *et al.* (2016) estimated the SFS of cytosine sites within genes of a sample of 92 *A. thaliana* individuals, but they did not detect a deviation from neutrality. More recently, studies have hinted at selection on methylation, because an SFS analysis at the level of 100 bp regions detected weak but significant selection on methylation levels (Xu *et al.* 2020) and because germline promoter methylation was inferred to be deleterious in humans (Boukas *et al.* 2020).

Here, we extend the SFS approach to data from the 1001 methylomes project in *A. thaliana* (Kawakatsu *et al.* 2016), to test two features of gbM. The first is whether there is evidence that gbM is subject to selection. To do so, we focus on the methylation state of genes, rather than individual sites. We focus on genes because previous work has shown that methylation is evolutionary conserved at the level of genes and not within individual sites, suggesting that the methylation state of a gene region could be the unit under selection (Takuno and Gaut 2013). However, we do not rely solely on gene-level analyses but also analyze the SFS of individual cytosines, as did Vidalis *et al.* (2016), but with a much larger dataset. The second is that we provide an intraspecific test of the association of gbM and gene expression by comparing the methylation state of alleles to their level and variability in expression. By harnessing the power of an extensive *A. thaliana* dataset, we uncover new information on the evolutionary forces that may act on epigenetic phenomena and the potential functional significance of gbM.

Materials and methods

Datasets

Methylation and expression files for *A. thaliana* were retrieved from data banks indicated in Supplementary Table S1. The files consisted of tables with one line per cytosine showing the number of methylated and unmethylated bisulfite sequencing (BS-seq) reads for each methylome, and tables with one line per gene showing the number of reads mapping for each transcriptome. The dataset included a total 1211 samples sequenced by BS-seq and 1195 by RNA-seq (see Supplementary Table S1 for a synthesis). More precisely, 927 *A. thaliana* were grown at 22°C and their methylomes were sequenced by BS-seq at the SALK Institute (Kawakatsu *et al.* 2016), of which 876 came from leaves and 51 from flower buds (with only a partial overlap in accessions between the two tissues). One hundred and forty-four of these samples had their leaf transcriptome profiled with the SOLiD system (Schmitz *et al.*, 2013), and 728 samples had their leaf transcriptome sequenced by Illumina RNA-seq (Kawakatsu *et al.* 2016). Another set of accessions, most of which were from Sweden, had their leaf BS-seq data generated at the Gregor Mendel Institute (GMI) (Dubin *et al.* 2015). These included 152 accessions that were grown at 10°C and another 120 accessions grown at 16°C. Some accessions had replicates sequenced, resulting in a total of 284 methylomes from GMI. These had corresponding leaf transcriptome data from 160 accessions grown at 10°C and from 163 accessions grown at 16°C, for a total of 323 samples sequenced by Illumina RNA-seq (Dubin *et al.* 2015). While the total dataset was 1211 accessions, we detected a strong Institute-of-origin effect in the data (see *Results*). We therefore opted to treat data from the two institutes separately and focused our analyses on leaf data from the Salk Institute (the "Salk dataset" of 876 accessions, Supplementary Table S1) and also analyzed a dataset from GMI (the "GMI dataset") consisting of the 120 accessions grown at 16°C.

For outgroup data, we retrieved *A. lyrata* MN47 and *Capsella rubella* MTE ~10-day-old seedling shoot methylation files (Seymour *et al.* 2014). For each species, two replicates grown at 23°C were used.

Inference of cytosine methylation

Cytosine methylation calls were already in the downloaded files from the Salk institute, and these calls were based on the method of Kawakatsu *et al.* (2016). For *A. thaliana* data from GMI (273 samples plus 11 replicates) as well as for *A. lyrata* and *C. rubella* data, we inferred cytosine methylation using the same method. Briefly, methylation was inferred for each site by performing a binomial test on the number of methylated and unmethylated reads, while taking into account the no-conversion rate (Lister *et al.* 2008). For the GMI data, the average no-conversion rate of 0.0041 was used for all samples (Dubin *et al.* 2015). *P*-values were corrected for multiple tests using Benjamini and Hochberg correction. Sites with ≤ 2 reads were considered as unmethylated, and sites with a corrected *P*-value under 0.001 were considered to be methylated.

Inference of gene body methylation

For each gene, the methylation state was inferred using data from coding sequences (CDS), which included exons but excluded both untranslated terminal regions and introns. We used the annotation of the longest transcript to define the CDS. For each accession separately, we computed an expected methylation rate for each context (CG, CHG, and CHH) across all CDSs annotated in the genome, and we used binomial tests to assess whether gene CDSs had a significantly higher proportion of methylated cytosines than the genome-wide background level of CDS methylation (Takuno and Gaut 2012). This was performed for each accession and cytosine context separately. *P*-values were corrected for multiple tests using the Benjamini and Hochberg correction for each accession separately.

Given the binomial results, a gene within an accession was inferred to be **gene body methylated (gbM)** if: (i) it had ≥ 20 CG sites, (ii) CG methylation was significantly higher than the background (one-sided binomial $P \leq 0.05$), and (iii) CHG and CHH methylation were not significantly higher than the background (one-sided $P > 0.05$). Similarly, a gene was inferred to be **CHG methylated** if it had ≥ 20 CHG sites, if CHG methylation was higher than the background (one-sided $P \leq 0.05$) and CHH methylation was not significantly higher than the background (one-sided $P > 0.05$). CHG methylated genes also tended to be CG methylated, but CG methylation was not required in our categorization. A gene was inferred to be **CHH methylated** if it had ≥ 20 CHH sites and if CHH methylation was higher than the background (one-sided $P \leq 0.05$). CHH methylated genes also tend to be CG and CHG methylated. Finally, a gene was inferred to be **unmethylated (UM)** if it had ≥ 20 CG sites and if CG, CHG, and CHH methylation were not significantly higher than the background (one-sided $P > 0.05$). In any other case, the gene methylation state was not inferred. Altogether, by applying this approach, we identified the frequency of methylation states across alleles among 1211 accessions and for ~27,000 genes. Note, however, that we focused our analyses on the subset of accessions that we called the Salk and GMI datasets (Supplementary Table S1).

Inference of ancestral methylation state

For each gene and each dataset, the ancestral methylation state in *A. thaliana* was inferred using methylation data from *A. lyrata* and *C. rubella*. To this end, we used the CoGe tool SynMap3D (Lyons and Freeling 2008) to infer orthologous syntelogs among *A. thaliana*, *A. lyrata* and *C. rubella*. We differentiated between orthologs and out-paralogs (paralogs caused by duplications that predate speciation) using pairwise dS values between syntelogs. Based on the distribution of dS values (Supplementary Figure S1), log10(dS) values were filtered to be lower than −0.39 for all species pairwise comparisons, which is equivalent to dS values lower than 0.407. After this filtering, 14,718 orthologous syntelogs were identified among the three species.

Two shoot replicates grown at 23°C were available for each outgroup species (*A. lyrata* and *C. rubella*) (Seymour *et al.* 2014). For every gene, the ancestral methylation state was inferred as the shared state between the two outgroups and their replicates. If the two replicates of a species had different methylation states for a gene, or if the gene had different methylation states between *A. lyrata* and *C. rubella*, we classified the gene as having an ambiguous ancestral state.

Inference of genes undergoing CG methylation epimutations

We also investigated the set of CHG-gain genes from *E. salsugineum CMT3* overexpressing transgenic lines by retrieving the list of 8704 CHG-gain genes from Wendte *et al.* (2019). The best blast hit—as provided in the genome reference—was used to infer the ortholog in *A. thaliana* for 8025 of these CHG-gain genes.

Site frequency spectrum

Most of our analyses were done at the genic level, so that the SFS was based on gene allelic states (*i.e.*, epialleles). The unfolded SFS was drawn for two gene methylation states, gbM and UM. mCHG and mCHH states were excluded from the SFS. For the Salk dataset, which consisted of 876 leaf methylomes, we only included genes in the SFS when they had ≥600 accessions with a UM or gbM methylation state; for the GMI dataset, the corresponding number was ≥80 of the 120 accessions. The distribution of the proportion of mCHG and mCHH accessions across all genes in the Salk dataset after applying this filter is shown in Supplementary Figure S2. For both datasets, genes that had >70% of accessions with mCHG or mCHH methylation state were discarded as possibly being pseudogenes or misannotated TEs.

The number of accessions with an inferred methylation state n varied among genes due to missing data, so that the site frequency spectrum sample size varied among genes. To cope with this missing data, we defined n', the minimum required number of accessions with characterized methylation, and applied a hypergeometric projection of the observed SFS into a subsample of size n'=600 for the Salk dataset and n'=80 for the GMI dataset. This is a mathematical transformation that down-samples all genes to have the same sample size. Genes sampled in less than n' accessions were discarded. Given the frequency k of the derived allele in the original sample of size n, the probability that i copies are observed in the reduced sample of size n' is (Hernandez *et al.* 2007):

$$P\left(\frac{i}{n'}\middle|\frac{k}{n}\right)=\frac{C_k^i C_{n-k}^{n'-i}}{C_n^{n'}} \quad (1)$$

Estimation of selection using the site frequency spectrum

Given the SFS, we estimated the strength of selection acting on methylation variants using the model of Charlesworth and Jain (2014). The model was designed to characterize the evolutionary forces acting on epigenetic markers, which evolve at much higher rates than DNA sequences when single sites are considered (Becker *et al.* 2011; Schmitz *et al.* 2011). We adapted the model for application to our biological question of whether selection acts on gene methylation states. Genes, which were the predominant unit considered here, can either be gbM or UM, with μ the mutation rate from UM to gbM and ν the mutation rate from gbM to UM.

The model assumes a randomly mating diploid population of constant effective population size N_e which is at mutation-selection equilibrium. The model further assumes that alleles are semi-dominant and that sites are independent. We estimated N_e using available polymorphism measures in *A. thaliana* (Alonso-Blanco *et al.* 2016): 10,707,430 total SNPs were detected in 1135 genomes of size 135 Mb, resulting in a Watterson theta θ_w = 0.00955 (Charlesworth and Charlesworth 2010). Using a mutation rate $\mu = 7 \times 10^{-9}$ (Ossowski *et al.* 2010) and θ_w=$4N_e\mu$, we estimated $N_e \approx$ 341,000. These values were similar to previous diversity measurements in *A. thaliana*, where intronic θ_w was estimated to be 0.0082 (Nordborg *et al.* 2005), but the actual θ_w may be higher due to biases in its estimation (Korunes and Samuk 2020).

If the UM state is advantageous over the gbM state, the probability that a sample of n individuals segregates for k UM variants and $(n\text{-}k)$ gbM variants at a given gene is (Charlesworth and Jain 2014):

$$p(k)=\binom{n}{k}\frac{F_1(\beta+k,\alpha+\beta+n,\gamma)(\beta)_k(\alpha)_{n-k}}{F_1(\beta,\alpha+\beta+n,\gamma)(\alpha+\beta)_n} \quad (2)$$

Where F_1 is the confluent hypergeometric function, $(x)_n$ is Pochhammer's symbol, $\alpha = 4N_e\mu$, $\beta = 4N_e\nu$ and $\gamma = 4N_e s_{UM}$ with s_{UM} the selective advantage of the UM methylation state over gbM. The model can easily be adapted to a case where the gbM state is advantageous over the UM state by switching α and β in Equation (2) and defining s_{gbM} the selective advantage of the gbM state.

The likelihood of the model is:

$$L=\prod_{k=0}^{n} p(k)^{d_k} \quad (3)$$

Where d_k is he number of genes observed with k UM accessions and $(n\text{-}k)$ gbM accessions.

Parameters of the model μ, ν, and s_{UM} (or s_{gbM}) were estimated using a Markov Chain Monte Carlo (mcmc) random walk with 100,000 generations as in Xu *et al.* (2020). The first 25% of mcmc generations were removed as burn-in. Parameters were sampled every 100 generations, providing around 750 samples for the posterior distributions of parameters (see Supplementary Figure S3 for an example of mcmc run diagnostics). The lambda parameters for scale proposal distribution were adjusted to obtain parameter acceptance rates between 20 and 70%. Both segregating and fixed sites of the SFS were used in the model. Final parameter values were obtained from the mean of the posterior distribution

and the credible interval from the 95% margins of the posterior distribution. We ran the algorithm three times with random starting points to ensure that the global maximum was found. In order to infer whether selection acting on the UM or the gbM state was significant, we compared the inferred value of $4.N_e.s$ to 1.0. If 1.0 was lower than $4.N_e.s$ and outside of its credibility interval (hereafter abbreviated $4.N_e.s \gg 1$), then we inferred that there was significant selection on methylation state, as values of $4.N_e.s$ lower than 1.0 are typically interpreted as cases of neutral evolution (Charlesworth and Charlesworth 2010). For each run, the expected SFS (using inferred parameter values from the best model) was compared to the observed SFS using a Pearson's χ^2 test in R.

SFS analysis on individual cytosines

The previous two sections used genes as the unit to draw the SFS and their methylation state as epialleles. In an attempt to generalize our results, we also analyzed the SFS of individual cytosines at CG sites for the Salk dataset. To this end, we (i) isolated CG cytosines with 3 or more read coverage within the CDS of genes; (ii) separated cytosines inside ancestrally gbM and ancestrally UM genes; and (iii) excluded the cytosines of accessions where the CDS had either mCHG or mCHH methylation state (as inferred at the gene level). CG cytosines were assigned to be either unmethylated or methylated as indicated in the Salk Institute methylation files. After inferring the SFS for individual cytosines within different genic sets, we applied the same mcmc approach as described above to estimate selection coefficients (s_{mC} for advantageous methylated cytosines and s_C for advantageous unmethylated cytosines). The state of the individual cytosines in the outgroup was not considered here because we analyzed separately cytosines that fell within ancestrally gbM and ancestrally UM genes.

Statistical study of the link between gbM and gene expression level

We measured the effect of gbM on expression level using the Salk dataset, which consisted of 679 accessions with both leaf methylation data and leaf expression data in the form of raw RNA-seq read counts. This number of accessions differed from the previous 876 Salk accessions used for the SFS analysis due to missing leaf expression data for some accessions. We constructed a linear model with mixed effects (Equation 4) to examine the data, which was run with the R package lme4 (Bates *et al.* 2015). We did not normalize gene expression and used raw read numbers, but the results were equivalent when using normalized read numbers as provided in GEO expression files. The aim of the model was to test, within each gene, for an association between a change in gene methylation state and gene expression across *A. thaliana* samples. To account for expression variability among genes, the model incorporated a random gene effect (see Equation 4). The random gene effect captures variability in gene expression due to average differences among genes. We also defined a fixed effect called gene methylation state (Equation 4), which consists of the states described above (*e.g.*, gbM, mCHG, mCHH, and UM) and applies to each gene epiallelic state within each accession. Significance for the fixed effect was determined by comparing the fit of the full model to a nested model without the fixed effect, using the anova function in R. Expression level was measured as raw read counts and log transformed. The R package lsmeans (Lenth 2016) was used to estimate pairwise differences between each pair of methylation states (*i.e.*, gbM *vs* UM, UM *vs* mCHG, and so on).

Our linear model can be expressed as:

$$\log(\text{Gene Expression} + 1) \sim \text{gene methylation state} + (1|\text{Gene}) \quad (4)$$

We also developed two linear mixed-effects models to investigate the potential relationship between genetic and epigenetic states of alleles. The model included the number of CG dinucleotides (#CG) and the epiallelic methylation state, as fixed effects, and the random gene effect:

$$\text{CG} \sim \text{gene methylation state} + (1|\text{Gene}) \quad (5)$$

$$\log(\text{Gene Expression} + 1) \sim \text{CG} + \text{gene methylation state} + (1|\text{Gene}) \quad (6)$$

Data availability

All data used in this manuscript were previously published (GEO accessions GSE43857, GSE80744, GSE54292, GSE43858, GSE54680). Supplementary material is available at figshare: https://doi.org/10.25386/genetics.14390681 (last accessed 29 April 2021).

Results

Detection of selection acting on gene methylation level

We used publicly available methylation datasets (Supplementary Table S1) to infer the methylation state of genes in *A. thaliana* accessions and two closely related outgroups. We first recognized that BS-seq of *A. thaliana* methylomes was carried out by two research Institutes (Salk and GMI), and so we compared methylation patterns and levels between their data. We found that the global rate of CHH methylation was significantly higher in accessions sequenced by GMI (2.3%) compared to the Salk Institute (0.28%, Supplementary Figure S4), regardless of the geographic origin of accessions (Supplementary Figure S5). This heterogeneity in the raw data had the potential to impact downstream gene methylation inferences (Supplementary Figure S6). We therefore focused most of our analyses on a single source—*i.e.*, the Salk data—because it had the highest number of samples. We also focused on methylome data from only a single tissue (leaf), leading to total analysis sample of 876 accessions in the Salk dataset (Supplementary Table S1).

Given the data, we inferred the gbM status for each gene in each accession separately to calculate the unfolded SFS for two gene methylation states—gbM and UM—after using a mathematical transformation to downsample all genes to a sample size of 600 accessions (see *Materials and Methods* for details). Altogether, we plotted the SFS based on 22,609 genes and found that, after the hypergeometric transformation, many genes were fixed for the UM methylation state in *A. thaliana* wild populations (10,090 genes), but there was also a subset of 652 genes fixed for gbM alleles (Figure 1A). Given the inferred SFS, we applied the model of Charlesworth and Jain (2014) to infer the selection coefficient, which is s_{UM} in a model where the UM state is advantageous and s_{gbM} in a model for which the gbM state is advantageous. Based on all 22,609 genes together (ancestrally UM, ancestrally gbM and genes with ambiguous or missing ancestral states), we found that the model that best fit the data was one without selection on gene methylation state, with estimated values of $4.N_e.s$ close to zero (Table 1). Note that the expected SFS based on fixing s to zero and estimating other parameters fit the observed SFS quite

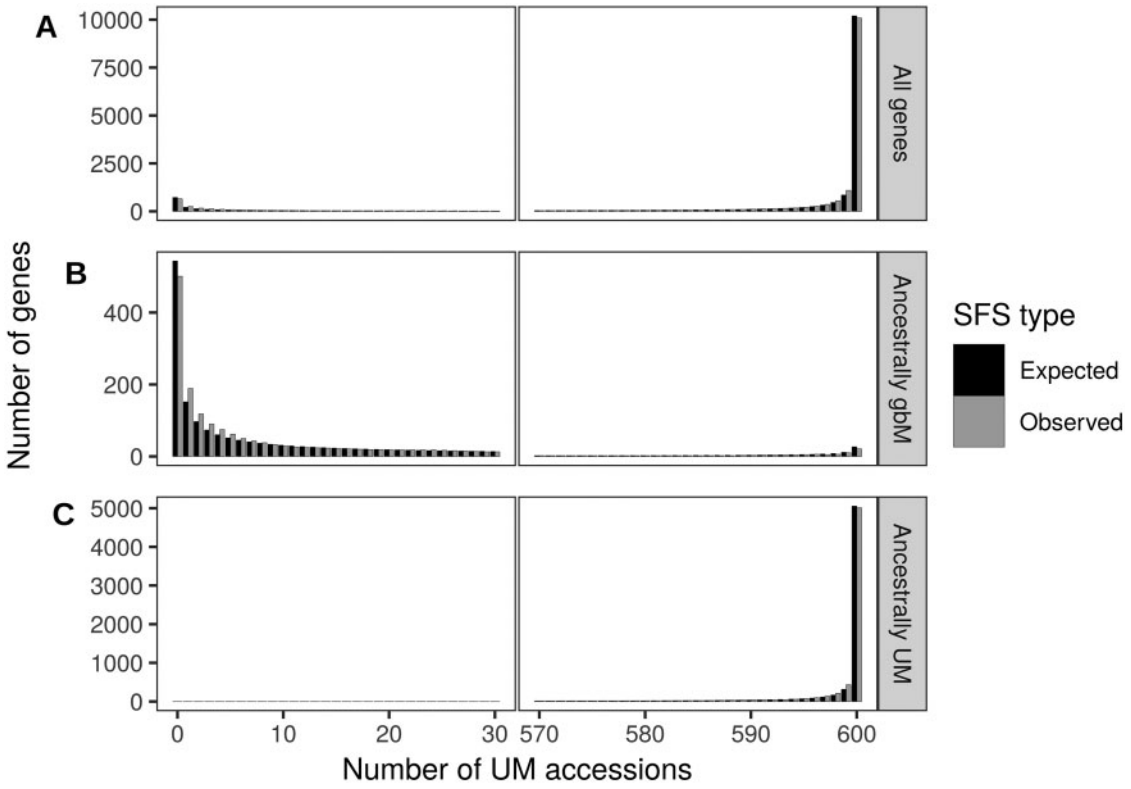

Figure 1 Expected and observed SFS of gbM based on the Salk dataset of 876 accessions at the gene level. The x-axis provides the number of UM accessions, out of a sample of 600. For these data, accessions that are not UM are gbM, meaning that genes with 600 UM accessions are fixed for the UM state in *A. thaliana* and genes with 0 UM accessions are fixed for the gbM state. The number of genes is provided on the y-axis. (A) All genes (22,609 genes with either UM, gbM, missing or ambiguous ancestral state), (B) Ancestrally gbM genes (3,243 genes), and (C) Ancestrally UM genes (7,626 genes). For visualization purposes, a gap was introduced in the x-axis from 30 to 570. The expected SFS were drawn using the parameters estimated by the mcmc, using the inferred selection pressure shown in Table 1. All three expected SFS fit the observed SFS well and did not differ significantly from the observed distribution (Pearson's χ^2 test $P > 0.4$).

Table 1 Estimation of selection acting on gene methylation state on the Salk dataset

		Gene number	Mean $4.N_e.s$	$4.N_e.s$ credible interval	Conclusion
All genes		22,609	$4.N_e.s_{UM} = 0.064$	0.011–0.127	No selection on gene methylation ($4.N_e.s \ll 1$)
Ancestrally gbM genes		3,243	$4.N_e.s_{gbM} = 1.637$	1.359–1.896	gbM state advantageous ($4.N_e.s_{gbM} \gg 1$)
Ancestrally UM genes		7,626	$4.N_e.s_{UM} = 1.896$	1.514–2.237	UM state advantageous ($4.N_e.s_{UM} \gg 1$)
CHG-gain orthologs		7,534	$4.N_e.s_{gbM} = 0.286$	0.176–0.393	No selection on gene methylation ($4.N_e.s \ll 1$)
No-CHG-gain orthologs		15,093	$4.N_e.s_{UM} = 0.355$	0.246–0.468	No selection on gene methylation ($4.N_e.s \ll 1$)
Ancestrally gbM	**CHG-gain orthologs**	1,984	$4.N_e.s_{gbM} = 1.623$	1.337–1.910	gbM state advantageous ($4.N_e.s_{gbM} \gg 1$)
	no-CHG-gain orthologs	1,250	$4.N_e.s_{gbM} = 1.637$	1.266–2.005	gbM state advantageous ($4.N_e.s_{gbM} \gg 1$)
Ancestrally UM	**CHG-gain orthologs**	2,184	$4.N_e.s_{UM} = 3.437$	2.905–3.915	UM state advantageous ($4.N_e.s_{UM} \gg 1$)
	no-CHG-gain orthologs	5,349	$4.N_e.s_{UM} = 3.178$	2.728–3.601	UM state advantageous ($4.N_e.s_{UM} \gg 1$)

Parameters of the SFS model were estimated using an mcmc approach (see *Materials and Methods* for details). The estimated mean selection efficacy (either $4.N_e.s_{UM}$ or $4.N_e.s_{gbM}$, depending on which is higher) is shown and its 95% credible interval. If $4.N_e.s > 1.0$ and if 1.0 is not included in the credible interval, then significant selection is inferred to act on methylation state (either the UM or the gbM state is advantageous). Details on inferred values of other parameters of the model can be found in Supplementary Table S2. These results come from one mcmc run and are equivalent to results obtained from two independent runs with random parameter initiation values (Supplementary Table S2).

well (Figure 1A), with no significant difference between them (Pearson's χ^2 test $P = 0.406$).

Our SFS illustrates that gbM is bi-modal, which is consistent with the fact that gbM is associated with a finite but conserved set of orthologous genes across angiosperms (Takuno *et al.* 2016). It seems reasonable to presume, then, that genes that are evolutionary conserved as gbM may be under different selection regimes than those that are evolutionary conserved as UM. Accordingly, we repeated analyses after splitting ancestrally gbM and UM genes (Figure 1, B and C). To infer the ancestral state, we used two outgroups (*A. lyrata* and *C. rubella*), identified syntelogs for 14,718 genes among the three species, and then inferred the ancestral methylation state by parsimony when both outgroups and their replicates had the same methylation state. After excluding 3699 genes for either missing methylation state inferences or for having ambiguous ancestral state, we applied the

model to a set of 3243 genes that were inferred to be ancestrally gbM, estimating a small selection coefficient ($s_{gbM} = 1.2 \times 10^{-6}$) but significant selection ($4.N_e.s_{gbM} \gg 1$, Table 1). This result implies that there is weak but detectable selection to maintain methylated alleles within genes that are ancestrally gbM. In contrast, 7626 ancestrally UM genes were estimated to be under selection to maintain UM alleles in *A. thaliana* wild populations ($s_{UM} = 1.39 \times 10^{-6}$, $4.N_e.s_{UM} \gg 1$, Table 1).

Additional analyses confirm selection on gene methylation states

We have mentioned both that there is no gbM in *E. salsugineum* due to the loss of *CMT3* (Bewick *et al.* 2016) and that complementation of *E. salsugineum* with a functional copy of *A. thaliana CMT3* leads to the accumulation of DNA methylation in "CHG-gain" genes (Wendte *et al.* 2019). These results suggest that DNA methylation epimutation rates are not homogeneous among genes, which could be problematic for our model. We therefore repeated the previous analyses separately for CHG-gain and for no-CHG-gain genes in *A. thaliana*, based on identifying the 8,025 orthologs of CHG-gain genes from *E. salsugineum* (see *Materials and Methods*). After excluding genes with <600 accessions with a gbM or UM methylation state, we found that 7,534 CHG-gain genes were under no selection on gene methylation state ($4.N_e.s \ll 1$, Table 1). Similarly, the set of 15,093 no-CHG-gain genes were estimated to be under no selection to retain UM nor gbM status ($4.N_e.s \ll 1$, Table 1). The model also estimates epimutation rates; consistent with implications of the *E. salsugineum* experiment (Wendte *et al.* 2019), we found that the mutation rate μ from the UM state to the gbM state was ~1.79 times higher in CHG-gain compared to no-CHG-gain genes. In contrast, the epimutation rate ν from gbM to UM was ~0.79 times lower in CHG-gain compared to no-CHG-gain genes (Supplementary Table S2).

Recognizing that CHG-gain and no-CHG-gain genes have different epimutation rates, we repeated the analyses after splitting CHG-gain and no-CHG-gain genes into ancestrally gbM and ancestrally UM genes. Ancestrally gbM genes were under significant selection to remain gbM ($4.N_e.s_{gbM} \gg 1$), regardless of whether they were targeted by additional methylation epimutations (CHG-gain) or not (no-CHG-gain, Table 1). Conversely, ancestrally UM genes were under significant selection to remain UM ($4.N_e.s_{UM} \gg 1$), regardless of their epimutational bias (Table 1). Our results were therefore confirmed after splitting the dataset into sets of genes that differ in epimutation rates.

Thus far, all of our results have been based on the Salk dataset. To confirm these results with an independent dataset, we ran the same gene-level SFS analyses on the GMI dataset (*i.e.*, for 120 accessions grown at 16°C; Supplementary Table S1). The observed SFS fit well to the expected SFS drawn from the parameter values estimated with the mcmc (Pearson's χ^2 test $P > 0.3$, Supplementary Figure S7). This dataset was expected to be less statistically powerful due to the lower number of accessions and because fewer genes were included in the SFS after filtering for having ≥80 accessions with UM or gbM methylation state (Table 2). We could nonetheless confirm some of our previous conclusions based on the Salk dataset (Table 2). Ancestrally UM genes are under selection to remain UM and all genes taken together were under no selection for methylation state. However, unlike the Salk dataset, ancestrally gbM genes did not yield evidence for significant selection on the gbM state, perhaps because the datasets were smaller (Table 2). These results for the GMI dataset were consistent across three independent mcmc runs (Supplementary Table S3).

Finally, we recognized that tests of allelic states may violate one of the assumptions of Charlesworth and Jain (2014), which is high mutation rates. Hence, as a final analysis, we ran the SFS analysis on individual cytosines rather than entire genes, again using the Salk dataset (see *Materials and Methods*). We analyzed separately cytosines within ancestrally gbM genes (361,038 cytosines) and cytosines within ancestrally UM genes (602,890 cytosines) and plotted their SFS (Figure 2). The model inferred that cytosines inside ancestrally gbM genes were under significant selection to be methylated ($4.N_e.s_{mC} = 2.101$, with credible interval 2.073–2.114, hence $4.N_e.s_{mC} \gg 1$, Supplementary Table S4). The inferred selection coefficient acting on individual cytosines ($s_{mC} = 1.54 \times 10^{-6}$, Supplementary Table S4) was similar to the one estimated on entire genes ($s_{gbM} = 1.2 \times 10^{-6}$, Table 1). Conversely, cytosines inside ancestrally UM genes were under significant selection to be unmethylated ($4.N_e.s_C = 2.51$, with credible interval 2.482–2.537, hence $4.N_e.s_C \gg 1$, Supplementary Table S4). These results were also consistent across three independent mcmc runs (Supplementary Table S4). Overall, our results provide evidence that the methylation state of alleles is associated with natural selection within genes.

Effect of gene methylation state on gene expression level in *A. thaliana* wild populations

E. salsugineum lacks gbM due to the loss of *CMT3* (Bewick *et al.* 2016), but there has been some debate about the effects of this gbM loss on gene expression. Bewick *et al.* (2016) found no effect, a result upheld by later analyses (Bewick *et al.* 2019). However, using different statistical approaches, Muyle and Gaut (2019) found evidence for a small but significant decrease in expression level for *E. salsugineum* genes that had lost gbM relative to the same gbM genes in *A. thaliana*. We further investigated the possible association between gbM and gene expression by analyzing expression levels. To make this assessment, we focused on accessions within the Salk dataset that had both RNAs-eq data and methylation data from leaves. The dataset consisted of 679 accessions (Supplementary Table S1) and 23,261 genes with polymorphic methylation states (*i.e.*, genes fixed for a given methylation state were removed from consideration). The availability of these data

Table 2 Estimation of selection acting on gene methylation state on the GMI dataset

	Gene number	Mean $4.N_e.s$	$4.N_e.s$ credible interval	Conclusion
All genes	15,720	$4.N_e.s_{gbM} = 0.199$	0.052–0.378	No selection on methylation ($4.N_e.s \ll 1$)
Ancestrally gbM genes	1,383	$4.N_e.s_{gbM} = 0.866$	0.368–1.341	Nonsignificant selection on gbM state ($4.N_e.s_{gbM} \approx 1$)
Ancestrally UM genes	6,078	$4.N_e.s_{UM} = 3.751$	3.274–4.242	UM state advantageous ($4.N_e.s_{UM} \gg 1$)

Parameters of the SFS model were estimated using an mcmc approach (see *Materials and Methods* for details). The estimated mean selection efficacy (either $4.N_e.s_{UM}$ or $4.N_e.s_{gbM}$, depending on which is higher) is shown with its 95% credible interval. If $4.N_e.s > 1.0$ and if 1.0 is not included in the credible interval, then significant selection is inferred to act on methylation state (the UM or the gbM state is advantageous). Details on inferred values of other parameters of the model can be found in Supplementary Table S3. These results come from one mcmc run and are equivalent to results obtained from two independent runs with random parameter initiation values (Supplementary Table S3).

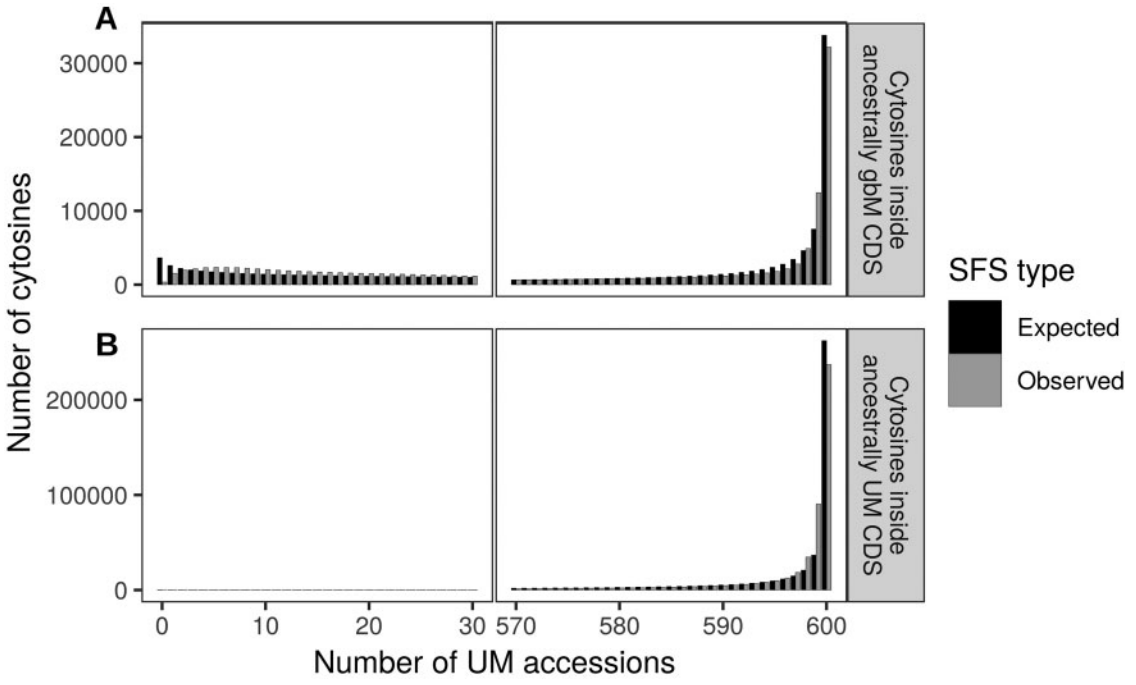

Figure 2 Expected and observed SFS of individual cytosine methylation in the Salk dataset. The x-axis provides the number of unmethylated (UM) accessions out of a sample of 600. For these data, accessions that are not UM are methylated, meaning that cytosines with 600 UM accessions are fixed for the UM state in *A. thaliana* and cytosines with 0 UM accessions are fixed for the methylated state. The number of cytosines is provided on the *y*-axis. (A) Cytosines within ancestrally gbM genes (361,038 cytosines) and (B) Cytosines within ancestrally UM genes (602,890 cytosines). The expected SFS were drawn using the parameters estimated by the mcmc, using the best model in Supplementary Table S4. Both expected SFS did not differ significantly from the observed SFS (Pearson's χ^2 test $P > 0.3$).

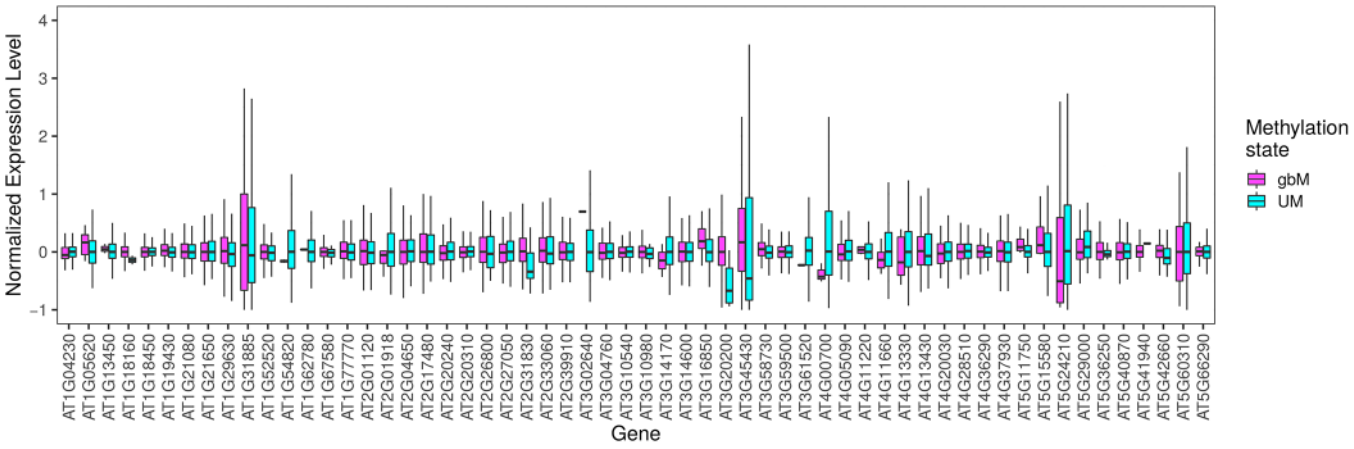

Figure 3 Normalized expression levels in a random set of genes with both UM and gbM epialleles. To make this figure, a random set of 100 genes was selected and only genes with both UM and gbM epialleles were retained. Genes with median read number under 10 were also excluded, leaving 57 genes. For each gene, the box plots indicate normalized expression levels for gbM and UM epialleles, with the lines representing the medians and the width of the boxplot the 1st and 3rd quartile. The figure shows that accessions that have a gbM epiallele tend to have a higher median normalized expression compared to accessions that have a UM epiallele, because 34 out of 57 genes in the represented random gene set have higher median expression levels for gbM epialleles. Although the differences are small and some genes show the opposite pattern, the overall effect is significant in a linear model across genes (see text for details). For each accession, normalized expression level was computed as follows: (accession normalized read number—median gene normalized read number)/median gene normalized read number.

permitted a test of whether epiallelic methylation states are associated with differences in expression.

We analyzed the data using a linear model with mixed effects (Equation 4, see *Materials and Methods*). The model was written to measure within gene expression variation, and then test for a significant effect of methylation state across all genes. This approach is possible due to polymorphisms in gene epiallelic states (or epialleles) among accessions. Treating genes as random effects and epiallele state (gbM, UM, mCHG, or mCHH) as a fixed effect, we found that epiallele methylation state had a significant effect on gene expression level ($\chi^2 = 19{,}300$, $P < 2.2 \times 10^{-16}$ when comparing a linear model with and without gene methylation state effect). We repeated these analyses between pairs of methylation states, comparing along an expected hierarchy of expression levels defined as gbM > UM > mCHG > mCHH. Our results confirmed that expected hierarchy, because gene expression in

Table 3 Pairwise comparison of the effect of gene methylation state on gene expression level in the *A. thaliana* Salk dataset

Contrast	Estimate	Standard error	z-ratio	*P*-value
gbM—UM	0.0563	0.0011	51.56	$<2.2 \times 10^{-16}$
UM—mCHG	0.1093	0.0033	33.13	$<2.2 \times 10^{-16}$
mCHG—mCHH	0.2275	0.00359	63.33	$<2.2 \times 10^{-16}$

A generalized linear model with mixed effects was used to estimate the effect of gene methylation state on gene expression (see *Materials and Methods* for details). Gene expression was measured as raw read counts and log transformed, but the results were equivalent when performed on normalized read counts. The table shows the average differences in log expression levels between pairs of methylation states (estimates) and their associated standard error, *t*-ratio and *P*-value after correction for multiple tests. For example, the gbM state is consistently associated with a 0.0563 higher log read count compared to the UM state, on average across all genes and accessions.

accessions that had the gbM epiallelic state was significantly higher than for accessions that had the UM epiallelic state for that same gene, globally across all genes (Figure 3). Similarly, we found that UM alleles had higher expression than mCHG alleles and that mCHG alleles were more highly expressed than mCHH alleles (Table 3). Altogether, these results show that within a gene, an accession with the gbM epiallelic state is consistently associated with the highest gene expression level, while the mCHH state is consistently associated with the lowest gene expression level. Our results are consistent with previous studies that associate CHG and CHH methylation with a reduced gene expression level in angiosperms (Niederhuth *et al.* 2016). However, the estimated differences in expression levels were very small (0.0563 log read count difference on average between gbM and UM methylation states, Table 3, which is equivalent to 1.058 raw read count difference on average). It is important to emphasize that linear models can detect small mean differences as significant so long as those differences are prevalent across the entire dataset.

In order to further test the robustness of our results, we performed two additional analyses. First, to confirm that the results were not an artifact of the linear model, we reran the model after randomly permuting methylation states without replacement among accessions and genes. These permutations removed associations between methylation states of an allele and their expression, and hence we did not expect to detect significant effects with permuted data. We ran the model on 1001 permuted datasets. As expected, the correlation between gene methylation state and expression was significant at $\alpha = 0.05$ only ~5.0% of the time, because we detected significance in 54 of 1001 permutations (Supplementary Figure S8). What is more, the *P*-value obtained from the real dataset was more than eight orders of magnitude lower than the lowest *P*-value obtained on any of the permuted datasets. These permutation results illustrate both that the model is well-behaved and that the observed data are strongly unexpected under a null hypothesis in which methylation and expression are not linked. Second, we compared expression levels between UM and gbM alleles within single genes for the 11,613 genes that had at least one UM accession and one gbM accession. We found accessions with gbM epialleles had a higher median expression level than accessions with the UM state for 6122 out of 11,613 (or 52.72% of genes). This proportion represents a significant deviation (binomial test, two-sided, $P = 5 \times 10^{-9}$) from the expected value of 50% under the null hypothesis that epiallelic state does not affect expression.

These data also present the opportunity to test the homeostasis hypothesis, which posits that gbM acts to stabilize gene expression (Zilberman 2017). Under this hypothesis, we expect the coefficient of variation to be lower across gbM epialleles than for UM epialleles. To test the hypothesis, we focused on the 10,327 genes that harbored at least two accessions of both UM and gbM epiallelic states in the population and then controlled for sample size differences by randomly sampling the same number of accessions for both epialleles within each gene. We found that the gbM state had a significantly lower coefficient of variation than the UM state (one-sided Wilcoxon paired signed rank test $P = 4.77 \times 10^{-6}$). Alternatively, we simply counted the number of genes that had a higher coefficient of variation for the UM state compared to the gbM state; 5357 (or 61.56%) genes had more variable expression among UM epialleles, representing a highly significant deviation from the null expectation of 50% (binomial test, two-sided, $P = 1.5 \times 10^{-4}$).

Discussion

We have utilized the dataset of 1001 *A. thaliana* methylomes (Kawakatsu *et al.* 2016) to examine features of the population genomics of epiallelic states, with a focus on gbM. Our analyses reveal two main observations. The first is that the SFS of allelic states yields information about selection. We find that all genes taken together are not under selection for methylation state. However, we also find that genes with ancestral gbM are under selection to maintain gbM in *A. thaliana*. The second observation is that analysis of this extensive dataset reveals an association between epiallelic state and patterns of expression, in terms of both expression level and stability. Both observations have broad relevance but also have caveats that must be considered.

gbM is under selection in *A. thaliana*

This study relies on 876 leaf methylomes in *A. thaliana* to investigate whether gbM is under selection and also on an SFS approach to detect selection that is specifically designed for epigenomic data (Charlesworth and Jain 2014). Our work offers the first evidence that ancestrally gbM genes are subject to natural selection to remain gbM. As selection only acts on traits that impact fitness, these results suggest that gbM has a function, at least in *A. thaliana* and maybe in other organisms that have similar gene methylation patterns (*i.e.*, plants, mammals, and insects).

The estimated selection coefficient for the advantage of gbM, s_{gbM}, is small (1.2×10^{-6} on average, based on the Salk dataset, Supplementary Table S2), resulting in $\gamma = 4N_e s_{gbM} = 1.64$. Values of $4N_e s < 1.0$ are typically considered neutral (Charlesworth and Charlesworth 2010). These inferred values of selection coefficients acting on gbM are similar to values estimated for codon usage bias, a phenomenon known to be under weak but significant selection in species with large enough *Ne* (Galtier *et al.* 2018). For example, leucine, valine, isoleucine, and arginine have estimated γ values for selection on codon usage between 1 and 2 in *A. lyrata* (Qiu *et al.* 2011). GbM is therefore a trait that seems to have a weak impact on fitness, but natural selection may be substantial enough to maintain it in a subset of important genes through time, just like for codon bias.

Our inferences are of course subject to caveats. The first set of caveats is related to our application of the model of Charlesworth and Jain (2014), which assumes uniform epimutation rates across the entire genome. To help address this limitation of the model, we investigated ancestrally UM and gbM genes separately, and we also separated the set of CHG-gain genes identified in *E. salsugineum* from the no-CHG-gain genes. We separated the latter two because CHG-gain genes may have elevated rates of *de novo* epimutation (Wendte *et al.* 2019); indeed, we estimate that CHG-gain genes have 1.79-fold higher mutation rates from the UM to gbM

state. However, we inferred similar trends from all subsets of the Salk dataset, including estimates of $4.N_e.s_{gbM} \gg 1$, *i.e.*, selection to retain methylation for ancestrally gbM genes (Table 1). These results suggest that heterogeneity in epimutation rates across genes are unlikely to drive our results, although we advocate for future investigations into CMT3 targeted genes in *A. thaliana*, because their dynamics could differ from *E. salsugineum* given the ~47 million year divergence between species (Arias *et al.* 2014).

Additional limitations of the model include assumptions about outcrossing, semi-dominance, absence of population structure, independence among sites, demographic equilibrium, and mutation-selection balance (Charlesworth and Jain 2014). Clearly, the first three of these assumptions are violated by our study organism (*A. thaliana*), which is predominantly selfing with strongly structured populations. However, we treated each individual as haploid (in the sense that we did not separate the two alleles of a gene), which reduces to sampling one allele per individual from an outcrossed population. We nonetheless find that the model fits the data well despite these limitations (Figures 1 and 2; Supplementary Figure S7).

Another set of limitations surround the data and our treatment of the data. For example, we used data from 10-day shoots—which are a mix of leaves, stems, and leaf buds—to infer ancestral states of *A. thaliana* leaves. We therefore assume that shoots adequately reflect methylation states in leaves, an assumption that appears to be reasonable for genic methylation states across tissues of *Brachypodium distachyon* (Roessler *et al.* 2016). We also treat complete CDS regions as an epiallelic state—*e.g.*, gbM, UM—and inferred the SFS of those states. We chose to focus this study on this gene-level approach based on the study of Takuno and Gaut (2013). This study found that an ortholog that is gbM in one species is highly likely to be gbM in another species, even when the two species in question (in this case, rice and *B. distachyon*) diverged ~50 my ago. The remarkable feature of this observation is that the methylation of orthologs was conserved but the methylation of individual nucleotides was not. In other words, this and subsequent studies have suggested that gbM is a property of genes, not nucleotide sites nor DMRs (differentially methylated regions), which are an amorphous and often statistically problematic concept (Roessler *et al.* 2016). Our focus on genes is also justified from observations about CHG-gain genes in *E. salsuginuem*, which show that epimutation by CMT3 locally spreads over entire genes, rather than methylating scattered isolated cytosines (Wendte *et al.* 2019).

We nonetheless repeated the SFS analyses at the level of individual cytosines located within CDS. Unlike Vidalis *et al.* (2016), we analyzed separately cytosines within ancestrally gbM genes and cytosines within ancestrally UM genes. This may be one reason why our conclusions differ from Vidalis *et al.* (2016), because they analyzed all genes together and they also had a smaller dataset with presumably less statistical power. Our results on individual cytosines confirm our results at the gene level: cytosines within ancestrally gbM genes are under selection to remain methylated ($4.N_e.s_{mC} \gg 1$). We used the ancestral gbM status of the entire gene to study individual cytosines, rather than the ancestral state of each individual cytosine, because—as mentioned above—methylation of orthologs is conserved over time, but the state of individual cytosines is not (Takuno and Gaut 2013). The inferred selection coefficients are also similar between the individual cytosines and the gene-level analysis (Supplementary Tables S1 and S3).

Interestingly, our results on ancestrally UM genes were opposite to those based on ancestrally gbM genes, both at the gene and cytosine levels, because the former were inferred to be under selection to be unmethylated (Supplementary Tables S2 and S4). We obtained a similar result on the GMI dataset when running the SFS at the gene level (Supplementary Table S3). However, selection on ancestrally gbM genes was not significant in the GMI dataset at the gene level, although the trends were in the same direction as the Salk dataset—*i.e.*, toward an advantage of the methylated state in ancestrally gbM genes. These results suggest, first, that the selection on gbM state varies among genes and, second, that gbM might be associated with a selective trade-off (Kiefer *et al.* 2019). That is, gbM is advantageous for some genes, but could also be deleterious, perhaps due to the increased mutation rate on methylated cytosines, energetic costs, or effects on chromatin structure. We therefore argue that the advantages of gbM outweigh its putatively deleterious mutagenic effects in a subset of genes, but in other genes gbM offers either no advantage or the advantage is not strong enough to compensate for higher mutation rates. In other words, gbM could be under stabilizing selection, with varying optima across genes.

Finally, we focus on the use of *A. thaliana* as a study organism for studies of methylation. *A. thaliana* has been used as a model system for good reason; without its genetic tools, the pathways and mechanisms of cytosine methylation in plants would not be nearly as well understood (Law and Jacobsen 2010). Similarly, the fact that it is selfing with a small genome size makes it ideal for some applications such as population genomics and epigenomics (Alonso-Blanco *et al.* 2016; Kawakatsu *et al.* 2016), leading to the generation of unique datasets like the one we have analyzed here. However, *A. thaliana* may not be the ideal model to study methylation mutants precisely because those mutants have less phenotypic effect in *A. thaliana* than in some other plants—for example, methylation mutants are often lethal in maize (Li *et al.* 2014). Consistent with this conjecture, a previous study comparing *A. thaliana* and *A. lyrata* gene methylation states has inferred that *A. thaliana* has lost gbM three times faster than gaining it (Takuno *et al.* 2017). Our estimated values of epimutation rates on all genes (Supplementary Table S2) from gbM to UM ($\nu = 2.09 \times 10^{-7}$) and from UM to gbM ($\mu = 6.2 \times 10^{-8}$) exactly reiterate this threefold difference ($\nu/\mu = 3.37$). Thus, the growing consensus is that *A. thaliana* is losing gbM through time. We hypothesize that one reason for this is the recent shift of *A. thaliana* to an inbreeding mating system, which has reduced its effective population size (Mattila *et al.* 2020) and likely led to weaker selection on epigenetic states. The overarching—and more important—point is that *A. thaliana* is likely to be a poor model to study the evolutionary forces that act on gbM, and yet our study nonetheless detects a significant selective effect.

gbM is associated with gene expression

As we noted in the Introduction, the question of gbM function has been raised in many studies, and gene expression has been used as the proxy for function in most of these studies. The field has thus focused on a relatively simple question: Is gbM associated with gene expression? Unfortunately, the outcome of these studies has been inconsistent, owing to a wide variety of reasons that may include that: (i) the effect of gbM on expression is minor; (ii) some studies are underpowered to detect such an effect, particularly over short temporal scales, (iii) researchers disagree on statistical approaches, particularly whether UM genes can be utilized as a control comparison to gbM genes (Bewick *et al.* 2019; Muyle and Gaut 2019); and (iv) independent epigenetic marks may have redundant functions that hide the effects of gbM loss in methylation mutants (Choi *et al.* 2020).

Our work here has, however, taken a unique approach, which is to examine the association of intraspecific variation in epialleles and expression levels across genes. This approach makes it possible to test (both within and across genes) whether a change in methylation state within the population associates with differences in expression level. To our knowledge, this is the first study to integrate intraspecific variation in gbM state with expression level in WT plants. Our linear model consistently identified an effect of methylation state on expression, whether we investigated all of the defined states or compared pairs of states (*e.g.*, gbM *vs* UM; Table 3). The power of this approach undoubtedly comes from the extensive data generated by the 1001 methylome consortium, because the size of the estimated effect is small. In real terms, the difference between a gbM allele and a UM allele is about 1 raw sequence read, averaged over the entire dataset. Nonetheless, it is clear that this result is not an artifact of the approach, because we permuted the data and found that the observed results are far more extreme (by 1000-fold) than the permuted data. In short, the evidence for the effect is strong, even though it is small. This adds to a growing number of experimental and comparative genomic approaches that point consistently to some association between gbM and expression (Zilberman *et al.* 2007, 2008; Coleman-Derr and Zilberman 2012; Steige *et al.* 2017; Takuno *et al.* 2017; Horvath *et al.* 2019; Seymour and Gaut 2019). We also show that the variation in gene expression among accessions is lower for the gbM compared to the UM epiallelic state. This is in agreement with other studies that suggest that gbM stabilizes gene expression (Zilberman *et al.* 2008; Coleman-Derr and Zilberman 2012; Steige *et al.* 2017; Takuno *et al.* 2017; Horvath *et al.* 2019; Seymour and Gaut 2019).

Our results point to selection on gbM perhaps, in part, due to its association with gene expression. But there remain two difficult questions. The first is whether selection is on gbM itself—*i.e.*, the epigenetic states directly—or on associated factors, such as chromatin factors or even underlying sequence features that may contribute to gbM in some unknown way. Unfortunately, we find no convincing method to discriminate among an associated *vs* a direct effect of gbM, and we must thus be careful to conclude that selection acts directly on the epigenetic state. However, to investigate this question, we ran a linear model with mixed effects to study the association between the number of CG dinucleotides (#CG) and gene methylation states in the Salk Institute data (see *Materials and Methods*, Equation 5 for details). There was a significant correlation between #CG and methylation states ($\chi^2 = 17{,}262$ and $P < 2.2 \times 10^{-16}$ when comparing a linear model with and without gene methylation state effect). This model also demonstrates that gbM epialleles have more CG dinucleotides than UM epialleles (linear model pairwise contrast estimate = 1.338, $P < 0.0001$). Surprisingly, however, when including both methylation state and #CG in a linear model to explain expression variation (see *Materials and Methods* Equation 6), the methylation state remains the main influence on gene expression. Moreover, accessions with higher #CG are significantly less expressed (linear model estimate -5.77×10^{-3}, $P < 2.2 \times 10^{-16}$), which opposes the effect of gbM on expression. Together, these analyses illustrate that the epiallelic state is not independent of the underlying sequence, as measured by #CG, but it also hints that epigenetic state contributes to phenotype in a way that is not easily explained by variation in the number of CG dinucleotides alone.

The second difficult question is function: what does gbM actually do? We cannot yet answer this question, especially given the inconsistent evidence from a variety of organisms and experiments (Zilberman *et al.* 2007, 2008; Coleman-Derr and Zilberman 2012; Li-Byarlay *et al.* 2013; Yearim *et al.* 2015; Bewick *et al.* 2016, 2019; Neri *et al.* 2017; Steige *et al.* 2017; Takuno *et al.* 2017; Teissandier and Bourc'his 2017; Horvath *et al.* 2019; Muyle and Gaut 2019; Seymour and Gaut 2019; Choi *et al.* 2020). We note, however, that histone H1 was recently shown to have a similar effect to DNA methylation in TEs and genes (Choi *et al.* 2020). In that study, expression of antisense transcripts was activated in 710 genes following methylation loss in *h1, met1* double mutants, at a level that was not positively correlated to sense transcription changes. This finding demonstrates that, at least for some genes, gbM can repress antisense transcription in *A. thaliana* jointly with H1. We hypothesize that the inhibition of antisense transcription requires a threshold of cytosine methylation, which we have captured by studying the methylation state for the entire CDS. Even if this is true, there are still unanswered mechanistic questions about how the effect of gbM on anti-sense transcription affects the level and stability of expression.

Acknowledgments

The authors would like to thank Mike May for help developing the mcmc approach used here.
A.M. and B.S.G. conceived the project. A.M. ran the analyses. J.R.I. provided the mcmc R code and critical ideas. D.K.S. shared data. A.M. and B.S.G. wrote the manuscript with input from all authors.

Funding

A.M. is supported by the €European Molecular Biology Organization (EMBO) Postdoctoral Fellowship ALTF 775-2017 and by the Human Frontier Science Program (HFSP) Fellowship LT000496/2018-L. B.S.G. is supported by the National Science Foundation (NSF) grant IOS-1542703. J.R.I. is supported by NSF grant 1546719 and the €United States Department of Agriculture (USDA) Hatch project CA-D-PLS-2066-H.

Conflicts of interest

None declared.

Literature cited

Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, *et al.* 2016. 1135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell. 166:481–491.

Arias T, Beilstein MA, Tang M, McKain MR, Pires JC. 2014. Diversification times among Brassica (Brassicaceae) crops suggest hybrid formation after 20 million years of divergence. Am J Bot. 101:86–91.

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. J Stat Soft. 67: 1. 10.18637/jss.v067.i01 (last accessed 29 April 2021).

Becker C, Hagmann J, Müller J, Koenig D, Stegle O, *et al.* 2011. Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. Nature. 480:245–249.

Bewick AJ, Ji L, Niederhuth CE, Willing E-M, Hofmeister BT, *et al.* 2016. On the origin and evolutionary consequences of gene body DNA methylation. Proc Natl Acad Sci USA. 113:9111–9116.

Bewick AJ, Schmitz RJ. 2017. Gene body DNA methylation in plants. Curr Opin Plant Biol. 36:103–110.

Bewick AJ, Zhang Y, Wendte JM, Zhang X, Schmitz RJ. 2019. Evolutionary and experimental loss of gene body methylation and its consequence to gene expression. G3 (Bethesda). 9: 2441–2445.

Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucl Acids Res. 8:1499–1504.

Boukas L, Bjornsson HT, Hansen KD. 2020. Purifying selection acts on germline methylation to modify the CpG mutation rate at promoters. bioRxiv 2020.07.04.187880.

Charlesworth B, Charlesworth D. 2010. Elements of Evolutionary Genetics. Greenwood Village, Colorado, USA: Roberts and Company Publishers.

Charlesworth B, Jain K. 2014. Purifying selection, drift, and reversible mutation with arbitrarily high mutation rates. Genetics. 198: 1587–1602.

Choi J, Lyons DB, Kim MY, Moore JD, Zilberman D. 2020. DNA Methylation and histone H1 jointly repress transposable elements and aberrant intragenic transcripts. Mol Cell. 77: 310–323.e7.

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, *et al.* 2008. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature. 452:215–219.

Coleman-Derr D, Zilberman D. 2012. DNA methylation, H2A.Z, and the regulation of constitutive expression. Cold Spring Harb Symp Quant Biol. 77:147–154.

Dubin MJ, Zhang P, Meng D, Remigereau M-S, Osborne EJ, *et al.* 2015. DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. Elife. 4:e05255.

Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, *et al.* 2018. Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. Mol Biol Evol. 35:1092–1103.

Hernandez RD, Williamson SH, Zhu L, Bustamante CD. 2007. Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. Mol Biol Evol. 24:2196–2202.

Horvath R, Laenen B, Takuno S, Slotte T. 2019. Single-cell expression noise and gene-body methylation in *Arabidopsis thaliana*. Heredity (Edinb) 123:81–91.

Kawakatsu T, Huang S-SC, Jupe F, Sasaki E, Schmitz RJ, *et al.* 2016. Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. Cell. 166:492–505.

Kawashima T, Berger F. 2014. Epigenetic reprogramming in plant sexual reproduction. Nat Rev Genet. 15:613–624.

Kiefer C, Willing E-M, Jiao W-B, Sun H, Piednoël M, *et al.* 2019. Interspecies association mapping links reduced CG to TG substitution rates to the loss of gene-body methylation. Nat Plants. 5: 846–855.

Korunes KL, Samuk K. 2020. pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. bioRxiv 2020.06.27.175091.

Law JA, Jacobsen SE. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 11:204–220.

Lenth RV. 2016. Least-squares means: the R package lsmeans. J Stat Soft. 69:1–33.

Li Q, Eichten SR, Hermanson PJ, Zaunbrecher VM, Song J, *et al.* 2014. Genetic perturbation of the maize methylome. Plant Cell. 26: 4602–4616.

Li-Byarlay H, Li Y, Stroud H, Feng S, Newman TC, *et al.* 2013. RNA interference knockdown of DNA methyl-transferase 3 affects gene alternative splicing in the honey bee. Proc Natl Acad Sci USA. 110: 12750–12755.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, *et al.* 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 133: 523–536.

Luo C, Hajkova P, Ecker JR. 2018. Dynamic DNA methylation: in the right place at the right time. Science. 361:1336–1340.

Lyons E, Freeling M. 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. 53: 661–673.

Mattila TM, Laenen B, Slotte T. 2020. Population genomics of transitions to selfing in brassicaceae model systems. Methods Mol Biol. 2090:269–287.

Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, *et al.* 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 466:253–257.

Miura A, Nakamura M, Inagaki S, Kobayashi A, Saze H, *et al.* 2009. An Arabidopsis jmjC domain protein protects transcribed genes from DNA methylation at CHG sites. EMBO J. 28: 1078–1086.

Muyle A, Gaut BS. 2019. Loss of gene body methylation in *Eutrema salsugineum* is associated with reduced gene expression. Mol Biol Evol. 36:155–158.

Neri F, Rapelli S, Krepelova A, Incarnato D, Parlato C, *et al.* 2017. Intragenic DNA methylation prevents spurious transcription initiation. Nature. 543:72–77.

Niederhuth CE, Bewick AJ, Ji L, Alabady MS, Kim KD, *et al.* 2016. Widespread natural variation of DNA methylation within angiosperms. Genome Biol. 17:194.

Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, *et al.* 2005. The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. 3:e196.

Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, *et al.* 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science. 327:92–94.

Qiu S, Zeng K, Slotte T, Wright S, Charlesworth D. 2011. Reduced efficacy of natural selection on codon usage bias in selfing *Arabidopsis* and *Capsella* species. Genome Biol Evol. 3:868–880.

Reinders J, Wulff BBH, Mirouze M, Marí-Ordóñez A, Dapp M, *et al.* 2009. Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes. Genes Dev. 23:939–950.

Roessler K, Takuno S, Gaut BS. 2016. CG methylation covaries with differential gene expression between leaf and floral bud tissues of *Brachypodium distachyon*. PLoS One. 11:e0150002.

Saze H, Shiraishi A, Miura A, Kakutani T. 2008. Control of genic DNA methylation by a jmjC domain-containing protein in *Arabidopsis thaliana*. Science. 319:462–465.

Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, *et al.* 2013. Patterns of population epigenomic diversity. Nature. **495**: 193–198.

Schmitz RJ, Lewis ZA, Goll MG. 2019. DNA methylation: shared and divergent features across eukaryotes. Trends Genet. 35:818–827.

Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, *et al.* 2011. Transgenerational epigenetic instability is a source of novel methylation variants. Science. 334:369–373.

Seymour DK, Gaut BS. 2019. Phylogenetic shifts in gene body methylation correlate with gene expression and reflect trait conservation. Mol Biol Evol. 37:31–43.

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: e1004785.

Steige KA, Laenen B, Reimegård J, Scofield DG, Slotte T. 2017. Genomic analysis reveals major determinants of cis-regulatory

variation in Capsella grandiflora. Proc Natl Acad Sci USA. 114: 1087–1092.

Stroud H, Greenberg MVC, Feng S, Bernatavichute YV, Jacobsen SE. 2013. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell. 152:352–364.

Takuno S, Gaut BS. 2012. Body-methylated genes in *Arabidopsis thaliana* are functionally important and evolve slowly. Mol Biol Evol. 29:219–227.

Takuno S, Gaut BS. 2013. Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. Proc Natl Acad Sci USA. 110:1797–1802.

Takuno S, Ran J-H, Gaut BS. 2016. Evolutionary patterns of genic DNA methylation vary across land plants. Nat Plants. 2:15222.

Takuno S, Seymour DK, Gaut BS. 2017. The evolutionary dynamics of orthologs that shift in gene body methylation between Arabidopsis species. Mol Biol Evol. 34:1479–1491.

Teissandier A, Bourc'his D. 2017. Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. EMBO J. 36:1471–1473.

Teixeira FK, Colot V. 2009. Gene body DNA methylation in plants: a means to an end or an end to a means? EMBO J. 28:997–998.

Tran RK, Henikoff JG, Zilberman D, Ditt RF, Jacobsen SE, *et al.* 2005. DNA methylation profiling identifies CG methylation clusters in Arabidopsis genes. Curr Biol. 15:154–159.

van der Graaf A, Wardenaar R, Neumann DA, Taudt A, Shaw RG, *et al.* 2015. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. Proc Natl Acad Sci USA. 112: 6676–6681.

Vidalis A, Živković D, Wardenaar R, Roquis D, Tellier A, *et al.* 2016. Methylome evolution in plants. Genome Biol. 17:264.

Wang J, Fan C. 2014. A neutrality test for detecting selection on DNA methylation using single methylation polymorphism frequency spectrum. Genome Biol Evol. 7:154–171.

Wendte JM, Zhang Y, Ji L, Shi X, Hazarika RR, *et al.* 2019. Epimutations are associated with CHROMOMETHYLASE 3-induced *de novo* DNA methylation. Elife. 8:e47891.

Xu G, Lyu J, Li Q, Liu H, Wang D, *et al.* 2020. Adaptive evolution of DNA methylation reshaped gene regulation in maize. bioRxiv 2020.03.13.991117.

Yearim A, Gelfman S, Shayevitch R, Melcer S, Glaich O, *et al.* 2015. HP1 is involved in regulating the global impact of DNA methylation on alternative splicing. Cell Reports. 10:1122–1134.

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW-L, *et al.* 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in Arabidopsis. Cell. 126:1189–1201.

Zilberman D. 2017. An evolutionary case for functional gene body methylation in plants and animals. Genome Biol. 18:87.

Zilberman D, Coleman-Derr D, Ballinger T, Henikoff S. 2008. Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. Nature. 456:125–129.

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. 2007. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet. 39:61–69.

Communicating editor: S. Wright