Gene Body Methylation in Plants: Mechanisms, Functions, and Important Implications for Understanding Evolutionary Processes

Aline M Muyle, Danelle K Seymour, Yuanda Lv, Bruno Huettel, Brandon S Gaut

▶ **To cite this version:**

Aline M Muyle, Danelle K Seymour, Yuanda Lv, Bruno Huettel, Brandon S Gaut. Gene Body Methylation in Plants: Mechanisms, Functions, and Important Implications for Understanding Evolutionary Processes. Genome Biology and Evolution, 2022, 14 (4), 10.1093/gbe/evac038 . hal-03862084

HAL Id: hal-03862084
https://hal.science/hal-03862084v1

Submitted on 20 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Gene-body methylation in plants: mechanisms, functions and important implications for understanding evolutionary processes

Review

Aline M. Muyle[1], Danelle K. Seymour[2], Yuanda Lv[3], Bruno Huettel[4], Brandon S. Gaut[1]

[1]Ecology and Evolutionary Biology, UC Irvine, Irvine.

[2]Botany & Plant Sciences, UC Riverside, Riverside, United States.

[3]Provincial Key Laboratory of Agrobiology, Institute of Crop Germplasm and Biotechnology, Jiangsu Academy of Agricultural Sciences, Nanjing, China.

[4]Max Planck Genome Centre Cologne, Max Planck Institute for Plant Breeding, Cologne, Germany

Abstract

Gene body methylation (gbM) is an epigenetic mark where gene exons are methylated in the CG context only, as opposed to CHG and CHH contexts (where H stands for A, C or T). CG methylation is transmitted transgenerationally in plants, opening the possibility that gbM may be shaped by adaptation. This presupposes, however, that gbM has a function that affects phenotype, which has been a topic of debate in the literature. Here we review our current knowledge of gbM in plants. We start by presenting the well elucidated mechanisms of plant gbM establishment and maintenance. We then review more controversial topics: the evolution of gbM and the potential selective pressures that act on it. Finally, we discuss the potential functions of gbM that may affect organismal phenotypes: gene expression stabilization and upregulation, inhibition of aberrant transcription (reverse and internal), prevention of aberrant intron retention and protection against TE insertions. To bolster the review of these topics, we include novel analyses to assess the effect of gbM on transcripts. Overall, a growing body of literature finds that gbM correlates with levels and patterns of gene expression. It is not clear, however, if this is a causal relationship. Altogether, functional work suggests that the effects of gbM, if any, must be relatively small, but there is nonetheless evidence that it is shaped by natural selection. We conclude by discussing the potential adaptive character of gbM and its implications for an updated view of the mechanisms of adaptation in plants.

Keywords: epigenetics, gene expression, transcription, DNA methylation, population epigenomics

Significance statement:

Gene body methylation (hereafter gbM) is a common phenomenon in plants and can affect up to 60% of the genes in some species. It has been controversial whether gbM has any function in plants but recent findings suggest it is under selection and correlated with fitness. Here we review the scientific literature, include novel analyses and discuss the potential role of gbM in rapid evolutionary change.

Introduction

Epigenetics is the study of changes in gene expression that can be inherited through cell divisions (either mitotic or meiotic) that are not due to modifications in the DNA sequence (Holliday 1994; Cavalli and Heard 2019). A longstanding question is whether epigenetics can play a role in adaptation (Charlesworth *et al.* 2017; Cavalli and Heard 2019; Boquete *et al.* 2021). Cavalli and Heard (2019) stated that "a direct demonstration that other molecules, in addition to DNA [sequence], carry substantial heritable information would represent an important conceptual change in evolutionary biology". Theoretically epigenetic marks may be the basis for this conceptual change. If epigenetic marks affect fitness, if they are inherited through generations, and if they epimutate over time, then they can be the target of selection and facilitate adaptation.

Cytosine methylation is one common epigenetic mark that is generally found in eukaryotes, including vertebrates, insects, fungi and plants (Zemach and Zilberman 2010; Schmitz *et al.* 2019). In some of these groups, cytosines are methylated only in a single context, when they are part of a CG dinucleotide. In plants, however, cytosine methylation occurs in three sequence contexts - CG, CHG and CHH (where H stands for A, T or C). Methylation marks in these three contexts are produced by different biochemical pathways and have different patterns of inheritance. For example, epimutation accumulation lines in *Arabidopsis thaliana* have demonstrated that genome-wide methylation divergence at CG dinucleotides increases over the course of >30 generations (van der Graaf *et al.* 2015), illustrating that plant CG DNA methylation is transmitted from generation to generation and epimutates over time (Yao *et al.* 2021). In contrast, CHH methylation is mostly erased by demethylation in the *A. thaliana* male germline and later reset during embryonic development (Calarco *et al.* 2012). Therefore, CHH methylation is only transmitted partially over, at most, one or a few generations (with the interesting exception of some asexual plants without meiosis; Boquete *et al.* 2021). The trans-generational inheritance of the third context - CHG methylation - remains unclear. Although CHG methylation is retained during gametogenesis (Calarco *et al.* 2012), epimutation accumulation lines in *A. thaliana* do not diverge for CHG methylation over generations (van der Graaf *et al.* 2015), suggesting that CHG methylation is not inherited at a genome-wide scale. It is possible, however, that some genomic sites inherit CHG methylation over a few generations, especially in some asexual species (Boquete *et al.* 2021). To summarize, of the three methylation contexts in plants, methylation in CG dinucleotides is the most prone to trans-generational inheritance and is therefore the best candidate for epigenetic adaptation.

To consider the possibility of epigenetic adaptation, it is also important to know where these marks reside in the genome. In flowering plants, patterns of DNA methylation vary among genomic regions. Methylation in all three contexts silences transposable elements (TEs) and prevents activity at regulatory elements (Luo *et al.* 2018; Schmitz *et al.* 2019). Both CHG and CHH genic methylation are associated with reduced expression levels, as is CG methylation in promoter regions (Zhang *et al.* 2006; Niederhuth *et al.* 2016). In contrast, the exons of some genes (~20% of *A. thaliana* genes; Takuno and Gaut 2012) are methylated only in the CG context, a phenomenon called gene body methylation (hereafter gbM). gbM is mostly found in moderately and constitutively expressed housekeeping genes (Zhang *et al.* 2006; Neri *et al.* 2017; Schmitz *et al.* 2019). However, since its initial discovery, the topic of gbM function been controversial (Zhang *et al.* 2006; Teixeira and Colot 2009; Bewick *et al.* 2017; Zilberman 2017). If it has no function, it is obviously unlikely to contribute to adaptive processes.

Here we review our current knowledge of gbM in plants, with the ultimate goal to critically evaluate whether it has a function and may be a target for natural selection. We start by presenting the mechanisms of plant gbM establishment and maintenance, because these mechanisms are crucial for understanding how selection could act on this epigenetic state. We then consider the evolution of gbM, specifically whether gbM is a neutral manifestation of epigenomic dynamics or whether there is evidence that it can be advantageous. Adaptive arguments presume that gbM has a phenotype on which natural selection can act. Some, but certainly not all, recent work has established a connection between gbM and gene expression, but questions about generality and mechanisms remain. To address these questions, we review functional analyses of mutants and also comparative epigenomic approaches that have studied hypothetical functions of gbM, namely its potential role in regulating and stabilizing expression, preventing aberrant transcription and improving the fidelity of intron splicing. Finally, we present a model synthesizing the prevalence, distribution and effect of gbM with its potential evolutionary significance.

GbM establishment and maintenance mechanisms

CG methylation is maintained during plant cell division by METHYLTRANSFERASE 1 (MET1), which adds a methyl group on the symmetrical CG dinucleotide of a complementary DNA strand (Kawashima and Berger 2014). Epimutation accumulation lines in *A. thaliana* have shown that the maintenance of CG methylation by MET1 is an inherently error-prone process, with the epimutation rate estimated to be ~10^{-3} per generation per haploid epigenome for the loss of CG

methylation in genes (van der Graaf *et al.* 2015). Without methylation maintenance mechanisms, CG methylation is quickly diluted and lost over cell divisions, as demonstrated by the absence of CG methylation in *A. thaliana met1* mutants (Cokus *et al.* 2008).

Studies in *Eutrema salsugineum*, a close relative of *A. thaliana* have recently clarified the mechanisms responsible for the establishment of gbM in plants (Bewick *et al.* 2016). *E. salsugineum* lacks both gbM and the *CHROMOMETHYLASE 3* (*CMT3*) gene. The link between *CMT3* loss and the absence of gbM was at first enigmatic because, until recently, CMT3 was not known to methylate CG sites. It is now known that the CMT3 protein is involved in a self-reinforcing feedback loop: CMT3 recognizes the histone mark H3K9me2 (histone H3 lysine 9 dimethylation) and then *de novo* methylates nearby cytosines predominantly in the CHG context but also occasionally in the CG context. CHG DNA methylation in turn leads to H3K9 methylation by SU(VAR) HOMOLOGUE 4 (SUVH4), leading to a positive feedback loop between CHG methylation and H3K9 methylation (Johnson *et al.* 2007). CHG methylation typically suppresses transcription, but in *A. thaliana* CHG methylation is removed in transcribed genes due to active demethylation of H3K9 by *INCREASED IN BONSAI METHYLATION 1 (IBM1)* (Saze *et al.* 2008; Miura *et al.* 2009).

The joint loss of *CMT3* and gbM evolved independently in two Brassicaceae species, corroborating their association. However, *cmt3* mutants in *A. thaliana* have shown that CMT3 does not affect the maintenance of gbM once it is established (Stroud *et al.* 2013), suggesting the action of CMT3 is limited to gbM establishment (Bewick *et al.* 2016; Niederhuth *et al.* 2016). Interestingly, transgenic reinsertion of *CMT3* into *E. salsugineum* re-established genic methylation in all three contexts in a subset of genes (Wendte *et al.* 2019). This subset of genes has been called ‘CHG-gain’ genes and tend to be orthologous to gbM genes in *A. thaliana* (Wendte *et al.* 2019). After the *CMT3* transgene was lost, CHG-gain genes only maintained methylation in the CG context, presumably due to maintenance by MET1 (Wendte *et al.* 2019). On average, CHG-gain genes are longer, contain more exons and exhibit a moderate -- but on average higher -- level of expression than non-CHG-gain genes (Wendte *et al.* 2019). CHG-gain genes are also enriched for CWG trinucleotides (CAG and CTG) as opposed to CCG trinucleotides, consistent with preferred substrate of CMT3 (Gouil and Baulcombe 2016; Stoddard *et al.* 2019). Finally, CHG-gain genes have a higher frequency of CHG cytosines compared to non-CHG-gain genes (Wendte *et al.* 2019).

The role of CMT3 in genic *de novo* methylation was recently confirmed in *A. thaliana* mutants that hyperexpress *CMT3* during late embryonic development (Papareddy *et al.* 2021). CMT3 hyperexpression induces embryonic hypermethylation predominantly in the CWG context, but hypermethylation is also found in other contexts, including CG dinucleotides. These findings confirm that CMT3 is sloppy and can methylate contexts other than CHG. Methylation changes caused by embryonic CMT3 hyperexpression were maintained over cell divisions and still observed in 3-week old plants, consistent with the model that CMT3-induced epimutations give rise to gbM that can be maintained by MET1 across cell divisions and generations. The same gene patterns were repeatedly observed in independent transgenic lines, confirming that CMT3 hypermethylation is not stochastic and tends to target a specific gene set (Papareddy *et al.* 2021). CMT3-induced hypermethylation was enriched in genes characterized by inaccessible chromatin marks and heterochromatin histone variants (Papareddy *et al.* 2021). Altogether, these observations lead to a model in which gbM establishment is caused by the recruitment of CMT3, the formation of a feedback loop that ultimately produces CHG, CHH and CG methylation, the eventual removal of CHG and CHH methylation, and the maintenance of the remaining CG methylation by MET1 (Figure 1).

gbM gene characteristics and evolution

gbM genes are typically defined statistically as being significantly more methylated than the genic average in the CG context and significantly less methylated than the genic average in the CHG and CHH contexts (Takuno and Gaut 2012). Once defined, the proportion of gbM genes varies greatly across species, with as many as ~60% of genes in *Mimulus guttatus* but 0% in *Marchantia polymorpha*, *Physcomitrella patens* and *E. salsugineum* (Takuno *et al.* 2016; Niederhuth *et al.* 2016; Niederhuth and Schmitz 2017). The lack of gbM in a few species has been used to argue that gbM is dispensable and thus has no function (Bewick *et al.* 2016, 2019). However, the loss of gbM in a few species does not imply that it is nonfunctional in all plants (Zilberman 2017).

A remarkable feature of gbM is that it is enriched over a conserved set of orthologs among species as distantly related as ferns and angiosperms (Takuno and Gaut 2013; Seymour *et al.* 2014; Takuno *et al.* 2016; Niederhuth *et al.* 2016; Seymour and Gaut 2019). Two alternative hypotheses can explain the remarkable conservation of gbM. The first is biased establishment of gbM in a subset of specific genes with inaccessible chromatin marks and heterochromatic (H3K9me2) histone variants (Wendte *et al.* 2019). If these biases are conserved across species, they could explain the distribution of gbM across both genes and species. This first hypothesis is neutral with respect to selection, because it

does not assume gbM has any effect on fitness. Instead, in this scenario, gbM is a consequence of CMT3 activity that is retained and transmitted over generations by MET1 (Teixeira and Colot 2009; Wendte *et al.* 2019; Papareddy *et al.* 2021).

An observation in favor of the neutral hypothesis is that gbM genes share many characteristics of the CHG-gain genes described previously. That is, the genes targeted by CMT3 are like gbM genes, in that they are generally characterized as being constitutively expressed at moderate levels and tend to be longer than unmethylated genes, with more exons and a higher frequency of CAG and CTG (as opposed to CCG) trinucleotides (Zhang *et al.* 2006; Lister *et al.* 2008; Takuno and Gaut 2012, 2013; Bewick *et al.* 2016, 2017; Takuno *et al.* 2016; Niederhuth *et al.* 2016). Moreover, CHG-gain genes in *E. salsugineum* tend to be orthologous to gbM genes in *A. thaliana* (Wendte *et al.* 2019). This observation suggests gbM establishment is biased towards specific genes, potentially explaining the conservation of gbM between orthologs (Wendte *et al.* 2019; Papareddy *et al.* 2021). The overlap between CHG-gain and gbM is not complete, because ~40% of CHG-gain genes in *E. salsugineum* are orthologous to a gbM gene in *A. thaliana* (Wendte *et al.* 2019). One explanation for the imperfect overlap between CHG-gain genes and gbM genes is because they were defined in different species – i.e., CHG-gain genes were defined in *E. salsugineum* and gbM genes in *A. thaliana.* Morever, these two species diverged 47 million years ago (Arias *et al.* 2014), which may be ample time for the targets of CMT3 to diverge. Finally, another plausible explanation is temporal. The CHG-gain genes in *E. salsugineum* were established experimentally over only a few generations; continuation of this experiment over a much longer timeframe could lead to the establishment of methylation within more genes, potentially increasing the 40% overlap.

An alternative hypothesis is that the conservation of gbM across genes and species is shaped in part by the action of natural selection (Zilberman 2017). For example, selection could potentially explain the difference in the distribution of CHG-gain and gbM genes. Under this scenario, specific subsets of genes are targeted for *de novo* establishment of gbM, but selection on or against gbM removes or maintains CG methylation in different gene sets. At least three observations support the hypothesis that some gbM is under selection. First, DNA methylation is mutagenic and elevates C to T substitutions (Bird 1980). Therefore, the conservation of gbM in a specific set of orthologous genes is surprising, especially because gbM genes are generally 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). This suggests the possibility that the mutagenic nature of methylation is compensated by an advantageous effect that maintains gbM in specific genes (Zilberman 2017). A second observation in favor of the selective hypothesis comes from the comparison of gbM

status in orthologous genes of eight grass species, where shifts in the gbM status of genes are almost exclusive to the tips of the phylogeny (i.e. in a single species) (Seymour and Gaut 2019). This pattern suggests that shifts in gbM are deleterious and generally not favored over evolutionary time (Table 1); however, it is also possible that the pattern is driven by epimutational biases.

The third observation is based on population genetic analyses, because selection acting on gbM can be explicitly measured using DNA methylation variation among natural populations of a species. Indeed, if gbM is advantageous within a given gene, an unmethylated allele will be disadvantageous and removed by selection. In such a gene, only a small proportion of individuals should be observed with an unmethylated allele. To infer the intensity of selection, Charlesworth and Jain (2014) constructed a population model that relies on the site frequency spectrum (SFS) of epigenetic states. In an inspired analysis, Vidalis *et al.* (2016) applied this model to the SFS of CG sites within all genes of a sample of 92 *A. thaliana* individuals. They did not detect a deviation from neutrality (Table 1), but this result comes with two important caveats. The first is that the test is unlikely to be powerful with a small sample, particularly if gbM has a small impact on fitness. Vidalis *et al.* (2016) used the sample of 92 individuals that was available at the time, but larger samples now exist. The second is ~~the important point~~ that CG methylation within genes is not limited to gbM genes, but can also be found in genes that are methylated in all three contexts (i.e, TE-like methylation; Kawakatsu *et al.* 2016). Methylation in all three contexts within a gene can be caused by a nearby TE insertion, is known to suppress expression and may be an indication of pseudogenization. Therefore, most genes with TE-like methylation are likely to be under different evolutionary pressures than gbM genes, such that analyzing both gbM and TE-like methylated genes together, as done by Vidalis *et al.* (2016), is likely to confound opposing selection pressures.

For all these reasons, we recently repeated the analyses of Vidalis *et al.* (2016) with the important difference that we separated gbM gene sets from any genes with TE-like methylation (Muyle *et al.* 2021). We also relied on larger datasets – i.e., two distinct subsets of 876 and 120 individuals that originated from different sources -- from the 1001 methylomes project in *A. thaliana* (Kawakatsu *et al.* 2016). To assess whether selection acts on the gbM state, we characterized the population frequency of methylation at the gene level to estimate the SFS of gene allelic states. Using the population genetic model of Charlesworth and Jain (2014), we inferred that genes with ancestral gbM in Brassicaceae were under significant selection to remain CG methylated in *A. thaliana* (Table 1; Muyle *et al.* 2021), based on the larger dataset. Conversely, ancestrally unmethylated genes in Brassicaceae were under selection to remain unmethylated in *A. thaliana*. We repeated the analyses on the smaller dataset and also on an SFS drawn at the level of individual cytosines. The former had similar trends as the larger

dataset, but without a significant effect of gbM, and the latter corroborated our gene-level analyses. That is, the overall impression is that CG sites within ancestrally gbM genes in Brassicaceae have been under selection to remain methylated in *A. thaliana*, while CG sites within ancestrally unmethylated genes were under selection to be unmethylated in *A. thaliana* (Muyle *et al.* 2021). Importantly, the results were also confirmed after splitting the gene sets into CHG-gain and non-CHG-gain genes, as characterized by Wendte *et al.* (2019), showing that biases in epimutation rates between gene sets were not completely responsible for the inferred selection acting on gbM. In other words, this control using CHG-gain genes shows that *cis* effects (either genetic or epigenetic) that locally influence epimutation rates do not explain the inferred selective pressures.

Like all evolutionary analyses, there are caveats to this analysis, too. First, it relies on a model that simplifies the evolutionary process and includes assumptions that do not strictly fit the study organism (e.g., the model assumes random mating but *A. thaliana* is self-fertilizing). Second, there is always the possibility that results are driven by sampling effects, including demographic history, although the use of two data sets and separate partitions of those datasets somewhat discounts that notion here. Third, and perhaps most importantly, it is difficult to disentangle genetic from epigenetic effects. Overall, this work suggests that gbM has a measurable effect on fitness. The estimated selection coefficients were small ($4N_es = 1.4$) but nonetheless similar to the magnitude of selection acting on codon usage that has been measured in *A. thaliana*, *A. lyrata* and *Capsella rubella* (Qiu et al. 2011).

One interesting feature of codon bias, a phenomenon widely accepted as a genomic feature that is under weak selection, is that it varies among species, with selection detectable in species with large historical population sizes but not detectable in small N_e species (Galtier *et al.* 2018). An overarching feature of gbM is that much of the experimental and comparative work on gbM has focused on *A. thaliana*. It is worth noting that this species may be atypical in at least three respects. First, two independent studies relying on different datasets and approaches have inferred that *A. thaliana* has lost gbM three times faster than gaining it (Takuno *et al.* 2017; Muyle *et al.* 2021) relative to closely related outcrossing species. Second, the recent shift of *A. thaliana* to an inbreeding mating system reduced its effective population size (Mattila *et al.* 2020), which is likely to have weakened the efficacy of selection on gbM in that species. Finally, methylation mutants usually have little phenotypic effect in *A. thaliana,* whereas they are often lethal in taxa with higher TE load, such as maize (Li *et al.* 2014). Together these observations suggest that *A. thaliana* may not be the best model for measuring the evolutionary effects of methylation and yet there is still some evidence that selection acts on gbM in that species, raising the possibility that the effects of gbM may be more pronounced in other species.

For this reason, we advocate that similar analyses are extended to other taxa when large methylation datasets become available.

Does gbM affect gene expression?

Given that there are some indications that gbM may be under weak selection, one naturally wonders what its function might be. One consistent hypothesis has been that gbM affects gene expression levels. This hypothesis first came from the observation that genic methylation levels across genes within *A. thaliana* are associated with expression levels: methylated genes tend to be intermediately to highly expressed, with lower expression variance among tissues (Zhang *et al.* 2006; Zilberman *et al.* 2007; Takuno and Gaut 2012). These patterns have been interpreted in two ways: either gbM might affect expression patterns (Figure 2.a) or, conversely, active transcription might drive gbM (Teixeira and Colot 2009). Many highly expressed genes do not have gbM in *A. thaliana* (Zhang *et al.* 2006; Zilberman *et al.* 2007), an observation that discounts the second hypothesis or at least suggests that the relationship is not completely straight-forward. Moreover, it is now known that CMT3 does not depend on gene expression to methylate genes but instead on inaccessible chromatin marks and heterochromatin histone variants (Wendte *et al.* 2019; Papareddy *et al.* 2021), although it remains possible that the initial recruitment of CMT3 requires or depends on gene expression.

One difficulty in assessing the effect of gbM on gene expression comes from the possible confusion between genetic and epigenetic effects. Indeed, variation in gene expression can be caused by numerous factors – e.g., by nearby single nucleotide polymorphisms (SNPs) in regulatory sequences, by a nearby TE insertion (genetic *cis* effects), by a change in a transcription factor (*trans* effects) or by a change in the gene DNA methylation level (epigenetic *cis* effects). Genome-wide association studies (GWAS) and epigenome-wide association studies (EWAS) have shown that DNA methylation variants associated with expression variation are often in linkage disequilibrium with nearby SNPs (Kawakatsu *et al.* 2016), making it difficult to disentangle the respective contribution of SNPs and methylation variation on gene expression. However, Meng et al. (2016) found a significant association between *cis*-methylation and gene expression in hundreds of genes across 135 *A. thaliana* accessions. Interestingly, gbM was positively correlated with gene expression, and the effect remained significant after controlling for SNPs. Overall, the number and magnitude of affected loci by DNA methylation was smaller than the effect of SNPs, and hence the authors concluded that DNA methylation has limited effects on expression variation (Table 1; Meng *et al.* 2016).

The association between gbM and expression was further tested experimentally in epigenetic recombinant inbred lines (epiRILs) obtained through 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 have a mosaic methylome, with some regions derived from the *met1* mutant that originally lacked gbM and other regions containing CG methylation derived from the wild-type (WT) parent. Bewick *et al.* (2016) inferred differentially expressed genes between the *met1* derived regions of epiRILs and their WT homolog. They found only 6 out of 3,471 genes that were gbM in WT plants and became differentially expressed when located in *met1* derived regions in epiRILs. On the other hand, they found significantly more genes (46 out of 3,124, p-value = 2.55×10^{-9}) that were unmethylated in WT plants and became differentially expressed when located in *met1* derived regions in epiRILs. Taken together these results suggest that gbM loss has little, if any, effect on gene expression (Table 1). However, if gbM has a small effect on expression, it is likely to have been missed by differential expression analyses that typically detect individual genes with twofold or more expression differences. More subtle effects may be statistically detectable only by approaches that summarize trends across multiple genes. More importantly, the *met1* mutant might be a poor system to study the association between gene methylation and expression level, because both methylated and unmethylated genes were upregulated in *met1* mutants using microarray data (Zilberman *et al.* 2007).

Another approach to test for associations between gbM and expression has been to use comparative and evolutionary, rather than experimental approaches. For example, several studies have compared expression between gbM-deprived *E. salsugineum* with its close relative *A. thaliana,* but the results have been controversial. In the first study, Bewick *et al.* (2016) estimated the expression of unmethylated *E. salsugineum* genes that are orthologous to gbM genes in *A. thaliana* and found no difference in expression between species (Table 1). Muyle and Gaut (2019) reanalyzed the data from Bewick *et al.* (2016) and used genes that were unmethylated in both *A. thaliana* and *E. salsugineum* as a negative control to measure the average difference in expression between the two species. When taking into account this species effect in a linear model, gbM loss in *E. salsugineum* was associated with a small but significant decrease in expression (Table 1). In a third study using the same data, Bewick *et al.* (2019) disagreed on the use of unmethylated genes as a negative control because they have been shown to have more variable expression levels over evolutionary time and again found no effect of gbM loss on gene expression.

Another effort compared gbM and unmethylated genes between *A. thaliana* and *A. lyrata* (Takuno *et al.*, 2017). Methylated genes were expressed at significantly higher levels, on average, and with less variation between species than non-CG methylated genes. The authors identified genes that

changed methylation status between *A. thaliana* and *Arabidopsis lyrata* to examine whether the shift in methylation correlated with gene expression. They found that genes that had gained gbM in one of the two species also tended to shift toward higher expression levels, but these results were not statistically significant (Table 1). However, genes that differed in gbM status between *A. thaliana* and *A. lyrata* exhibited significantly higher variance in expression between species than genes that were gbM in both species (Takuno *et al.* 2017), consistent with previous studies suggesting gbM modulates expression variability (Zilberman 2017). Another comparative study compared the methylomes of eight grass species and found that genes that were gbM in all eight species tended to have higher and less variable expression compared to genes that varied in their methylation state across species (Seymour and Gaut, 2019). Although the effect was very small, the results suggest a positive effect of gbM on expression level and expression stabilization (Figure 2.c). It is worth emphasizing, however, that this approach, like most comparative approaches, cannot determine causality. More recently, we used the 876 *A. thaliana* methylomes to study the association between gbM and gene expression within a species by comparing the methylation state of alleles both to their expression level and to the variability in expression across the larger data subset of 876 *A. thaliana* methylomes (Muyle *et al.* 2021). Across genes with polymorphic methylation states, the expression of gbM alleles was consistently and significantly higher than unmethylated alleles (Table 1). Taken across the entire genome, gbM alleles also had a significantly less variable expression level compared to unmethylated alleles of the same gene (Muyle *et al.* 2021). Although consistent across the thousands of genes in the dataset, the effect was quite small: on average, a methylated allele had ~1 more RNAseq read than an unmethylated allele. A weakness of this work is that it did not disentangle potential genetic effects from epigenetic effects; however, the gbM effect did remain consistent when models included a proxy for genetic variation, by including the number of CG dinucleotides in statistical analyses. Consistent with our *A. thaliana* results, work on the outcrossing crucifer *Capsella grandiflora* has revealed that the presence of gbM is a major predictor of cis-regulatory constraint (Steige *et al.* 2017). GbM lowers the probability of allele specific expression via cis-regulation, again suggesting a stabilizing effect of gbM on expression level.

As we have just reviewed, several studies have established an association between gbM and stable expression level (Figure 2.b-c), which complement the proposal that gbM has a homeostatic effect on expression (Zilberman 2017). This phenomenon has been further investigated by Horvath *et al.* (2019), who studied gene expression levels in *A. thaliana* roots via single-cell RNA-seq. They found no significant correlation between gbM and gene expression noise (as measured by variation in expression level among single cells). However, gbM was significantly positively correlated with gene

expression consistency, which they measured as the number of single cell RNA-seq replicates in which the gene was expressed (Figure 2.b). This effect remained after correcting for other genomic features such as gene expression, gene length, gene conservation and gene duplication status. Therefore, Horvath *et al.* (2019) found that gbM genes are more consistently expressed than unmethylated genes across cells of a tissue, which can be interpreted as implying that gbM is involved in the maintenance of a consistent gene expression (Figure 2.b). If this is true, the mechanism by which this happens remains unknown. One hypothesis comes from the anticorrelation observed between genome-wide distributions of the histone variant H2A.Z and DNA methylation in *A. thaliana* (Zilberman *et al.* 2008). H2A.Z is typically associated with transiently expressed response genes, such as immune response or environmental stimulus response genes (Coleman-Derr and Zilberman 2012). In *met1* mutants, the loss of DNA methylation was accompanied by a gain in H2A.Z deposition (Zilberman *et al.* 2008). However, in an *h2a.z* mutant, DNA methylation patterns were only minimally affected (Coleman-Derr and Zilberman 2012), suggesting that DNA methylation prevents H2A.Z incorporation but not the converse. Based on these observations, it has been proposed that gbM serves to stabilize transcription by preventing deposition of the histone variant H2A.Z (Coleman-Derr and Zilberman 2012).

Finally, Shahzad *et al.* (2021) have used quantitative trait loci (QTL) mapping to identify ~1000 genes for which the proportion of methylated CG sites significantly correlates with expression level across a sample of over 900 natural *A. thaliana* accessions. The variance in expression explained by CG methylation is modest for most genes, but for some genes it reaches levels comparable to the effect of SNPs on expression. gbM is mostly positively correlated with expression; in contrast, TE-like methylation (i.e., in all three contexts) is, as expected, negatively correlated with expression. In a clever extension to control for the effect of linked SNPs, Shahzad *et al.* (Shahzad *et al.* 2021) identified SNPs with significant effects on expression, using GWAS. They then repeated the analysis linking CG methylation and expression within nested sets of accessions that carry the same GWAS allele. For the vast majority of genes, this approach confirmed the significant positive correlation between gbM and expression level, either because there was no GWAS SNP or because at least one nested sample had a significant correlation. A second control analyzing haplogroups, which corrects for *cis* genetic variation, led separately to the same conclusion. They then studied gene expression in the *met1 A. thaliana* mutant without gbM, and they found as expected a reduced expression level in genes with these three characteristics: (i) genes with a significant positive correlation between CG methylation and expression level, (ii) genes that were methylated in the WT Col-0 accession (which is the accession used for the *met1* mutant) and (iii) genes for which the effect of gbM was not confounded by linked

genetic variants. The authors concluded that gbM is positively correlated with gene expression in hundreds of genes, independently of local genetic variants.

Although Shazad et al. (2021) did attempt to account for *trans* effects as well as *cis* effects, a shortcoming of most of the comparative studies referred to in this section is that they do not account for possible trans genetic effects on gene expression, which may result in overestimated *cis* epigenetic effects. Altogether, however, evolutionary and comparative studies tend to find small but detectable relationships between gbM and either gene expression levels and gene expression variation. These results contrast with many direct experimental measurements of gbM based on *A. thaliana* mutants. If the comparative conclusions are correct, they are important because they suggest that gbM has a phenotype that may be the target of natural selection.

Potential effects of gbM on internal and reverse transcription

To date, studies have been inconsistent as to whether gbM associates with gene expression (Table 1). When it does associate with expression it is also difficult to disentangle cause from effect. If, however, we assume there is a real relationship between gbM and gene expression, there remains an open question: what is the mechanism(s) by which gbM affects expression? One hypothesis is that gbM improves transcription through regulation of alternative promoters within gene bodies, thereby potentially preventing aberrant internal and/or antisense transcription (Figure 2.d; Tran *et al.* 2005; Maunakea *et al.* 2010). This hypothesis stems from the observation that CG methylation is typically depleted within active promoter regions (Feng *et al.* 2010) and also that genes with CG-methylated promoters are silenced (Niederhuth *et al.* 2016). Aberrant transcription, whether in the sense or antisense orientation, is expected to be deleterious because it is energetically costly and leads to the accumulation of both unnecessary transcripts and truncated proteins that can be toxic for the cell. Aberrant antisense transcription is expected to disturb gene expression because RNA-polymerases coming from both directions may collide. Moreover, the RNA-directed DNA methylation pathway can be activated by the pairing of sense and antisense transcripts into double stranded RNA (Tran *et al.* 2005), which may further prevent gene expression. Hence, if gbM prevents aberrant reverse transcription, it could explain the aforementioned association between gbM and gene expression.

However, the results of tests for this effect have been inconsistent (Table 1). Some of these tests have taken place in mammalian systems, because they too exhibit CG methylation within genes (Yi 2017), even though they mostly do not methylate in the CHG and CHH contexts. For example, Neri *et al.* (2017) studied mouse embryonic stem cells in *DNA methyl-transferase 3b* (*Dnmt3b)* mutants that

lack gbM and compared them to WT. To quantify internal transcription, the authors used a ratio of the number of RNA-seq reads that map to the third exon divided by the number of reads mapping to the first exon (hereafter exon3/exon1). This ratio is expected to increase when there is cryptic intragenic initiation of transcription. Neri *et al.* (2017) found that the loss of gbM was accompanied by higher exon3/exon1 ratios compared to WT mice, suggesting an increase of spurious internal transcription between exons 1 and 3 when gbM is lost (Table 1). The results were confirmed by a sequencing technique that allows characterization of the exact position of the 5' end of mRNAs by targeting the mRNA cap (DECAP-seq). However, Teissandier and Bourc'his (2017) performed similar analyses in *Dnmt* triple mutants on highly expressed genes, and they were unable to corroborate the findings of Neri et al. (2017) (Table 1). Results from Teissandier and Bourc'his (2017) suggest that the role of gbM in suppressing spurious transcription initiation may be specific to the lack of DNMT3B, but only while other DNMTs are still present.

Similar work has sought evidence for an effect of gbM on aberrant transcription in plants. For example, Bewick *et al.* (2016) compared *met1* derived regions of *A. thaliana* epiRILs with orthologous wild type regions. They quantified antisense transcription and found that gbM loss did not lead to an increase in differentially expressed antisense transcripts (Table 1). However, Choi *et al.* (2020) detected that the expression of antisense transcripts was activated in 938 genes in *h1,met1* double mutants compared to WT. The number of upregulated antisense transcripts was comparatively low in single mutants when compared to WT (145 and 34 for *met1* and *h1* respectively), suggesting redundancy in H1 and MET1 repression of antisense transcription. This finding demonstrates that, at least for some genes, gbM may repress antisense transcription in *A. thaliana* jointly with histone H1 (Table 1). This study also again exemplifies redundancy among DNA methylation, histone variants and histone marks. These different epigenetic marks are interdependent and play overlapping roles in the cell, complicating the characterization and inference of potential gbM effects.

More recently, Li *et al.* (2021) used RNA long reads sequenced by Oxford Nanopore Technology Direct Sequencing (ONT DRS) to characterize transcription start sites (TSSs) in *A. thaliana*. They found that the *met1-3* mutant, which lacks CG methylation, has significantly more unique TSSs compared to WT, and these unique TSSs occurred in regions where mutant methylation was lower than WT. These results suggest that gbM can prevent the initiation of aberrant transcription. The transcription termination site (TTS) was also affected by DNA methylation (Li *et al.* 2021). Indeed, the *met1-3* mutant had a higher number of unique TTS than WT, indicating that CG methylation also inhibits aberrant transcription termination. Altogether, this work suggests that gbM could ensure proper transcription of genes from start to end (Figure 2.d).

Here, we revisited this issue by analyzing Isoseq (PacBio RNA long read) data in maize and *A. thaliana* (Supplementary Materials and Methods). We included maize because this is (to our knowledge) the first such attempt to examine this question in a plant other than *A. thaliana.* We focus on Isoseq data because it can represent full-length mRNA, thanks to the selection of mRNAs that contain a 3' poly-A tail and, in some cases, a 5' cap (when sequencing is done with a cap-trap step). The *A. thaliana* Isoseq dataset we analyzed has a 5' cap, so that most Isoseq reads likely represent full length mRNAs (Supplementary Materials and Methods). Some of the *A. thaliana* dataset was publicly available (Cartolano et al. 2016) and we also generated new Isoseq data for this study; in both cases the data were generated from Col-0 inflorescences. In contrast, the maize Isoseq data, which was generated on pooled RNA extracted from six tissues at different developmental stages of the B73 inbred line (Wang *et al.* 2016), was not generated with a cap-trap step. With both the maize and *A. thaliana* datasets, we considered aberrant internal transcription to be reflected in Isoseq reads that begin after the start of exon 1 (Figure 2.d). For each gene, the proportion of full-length Isoseq reads with a "conventional" TSS (i.e., that begin prior to the start of exon 1) was computed and compared between gbM and unmethylated (UM) genes. gbM and UM genes were categorized from publicly available methylation data (see Supplemental Materials and Methods).

In maize, we found that gbM genes had a significantly higher proportion of conventional TSSs (average 0.81) compared to UM genes (average 0.78, Wilcoxon test *p-value* = 5.63×10^{-11} Figure 3.a); superficially this observation complies with the prediction that gbM genes have less aberrant transcription. However, gbM is known to be associated with more highly expressed and longer genes (Zhang *et al.* 2006; Takuno and Gaut 2012) and these covariates must be taken into account. Genes with higher expression had a significantly higher proportion of conventional TSSs (generalized linear model contrast estimate=0.042, Z-ratio=69.15, p-value$<2 \times 10^{-16}$, Supplementary Table S1), perhaps reflecting higher selective pressures to remove aberrant transcription for highly expressed genes. Longer genes had significantly fewer conventional TSSs (generalized linear model contrast estimate=-0.178, Z-ratio=-130.7, p-value$<2 \times 10^{-16}$ Supplementary Table S1), which could be attributable to the higher probability of a long gene harboring an aberrant internal promoter or an experimental artifact (i.e., longer genes may have a higher chance of having their mRNA not fully reversed transcribed during sequencing, leading to 5' truncated transcripts and wrongly inferred aberrant TSS). Notably, however, only 321 of 2059 detected non-conventional TSSs occurred in introns. After taking gene length and gene expression into account in a generalized linear model (Supplementary Materials and Methods, Equation 6), UM genes had a significantly higher proportion of conventional TSSs compared to gbM genes (generalized linear model contrast estimate=0.214, Z-ratio=33.4, p-value$<2 \times 10^{-16}$,

Supplementary Table S1). This result is not consistent with the expectation that gbM prevents aberrant TSS (Table 1).

The *A. thaliana* results complement the maize results in some ways but not others. Since the data were generated with a 5' cap, the overall proportion of conventional TSSs in the *A. thaliana* Isoseq dataset was higher than in maize (Figure 3.a). gbM genes had a lower proportion of conventional TSSs (mean 0.90) compared to UM genes (0.96). However, after taking gene length and gene expression into account in a generalized linear model, gbM genes did have significantly more conventional TSS compared to UM genes (p-value<$2x10^{-16}$, Supplementary Table S2). We conclude that the Isoseq data do reflect some advantage of gbM in terms of avoiding internal transcription start in WT *A. thaliana,* but not in maize (Table 1).

We explored these ideas further by turning to a different approach that relies on RNA-seq reads in gbM mutants. Because 5' cap Isoseq data are not available for methylation mutants, we inferred internal transcription starts using the RNA-seq coverage ratio of exon3/exon1, following the approach of Neri *et al.* (2017) (see Supplementary Materials and Methods). If gbM prevents aberrant internal transcription start, this ratio should increase in gbM mutants relative to WT. We therefore measured exon3/exon1 RNA-seq coverage in WT and gbM mutants. We performed this comparison for two datasets based on two different gbM mutants. In Dataset 1, RNA-seq data was generated on 13 day old seedlings for three replicates of WT controls that were compared to three replicates of *met1-3* mutants (Zhang *et al.* 2017). In Dataset 2, RNA-seq data was generated on leaf tissue for three replicates of WT controls (these differed from the controls within Dataset 1) that were then compared to five replicates of *met1,sdg7-8* triple mutants (Bewick *et al.* 2016). In WT plants, gbM genes had a lower exon3/exon1 coverage ratio compared to UM genes (Figure 3.b), suggesting superficially that gbM could prevent internal transcription start. This was observed consistently for WT plants from dataset 1 (mean exon3/exon1 coverage 0.662 in gbM genes, 1.919 in UM genes, one-sided wilcoxon test p-value < $2.2x10^{-16}$) and also from dataset 2 (mean exon3/exon1 coverage 1.013 in gbM genes, 3.684 in UM genes, one-sided wilcoxon test p-value = $1.3x10^{-11}$). However, this same difference between gbM genes (as defined in WT plants) and UM genes was also apparent in mutant plants that lacked gbM (Figure 3.b). That is, gbM genes had a lower exon3/exon1 ratio compared to UM genes in *met1-3* mutants (0.512 versus 1.876, one-sided wilcoxon test p-value < $2.2x10^{-16}$) and in *met1,sdg7-8* triple mutants (1.232 versus 1.566, one-sided wilcoxon text p-value < $2.2x10^{-16}$). These observations suggest that the fact that gbM genes have lower exon3/exon1 coverage in RNA-seq data is not due to their methylation state alone, because the same pattern is observed in mutants without gbM. While one must always be careful that mutants can be complex and may reflect other (unknown) effects, the data again provide

little support for the notion that gbM alone prevents aberrant internal transcription initiation in *A. thaliana* genes (Table 1). Interestingly, Choi *et al.* (2020) showed that gbM and H1 play a redundant role in inhibiting aberrant reverse transcription, and Martin *et al.* (2021) hypothesized that another epigenetic mark, CHH islands, may have some redundant function with gbM. These redundancies could explain why we did not detect any change in internal transcription start in *A. thaliana* gbM mutants, because these redundant epigenetic marks may have been functioning in gbM mutants, thus complicating inferences about gbM effects.

Another aspect of aberrant transcription, aside from internal transcription initiation, is reverse transcription. We again used the Isoseq data from *A. thaliana* to estimate the proportion of antisense full-length reads, which correspond to reverse transcription events (see Supplementary Materials and Methods). gbM genes had a significantly lower mean proportion of antisense reads (average 0.0046) compared to UM genes (average 0.016, Wilcoxon test *p-value* = 3.16×10^{-12}). This result held after accounting for gene length and gene expression in a generalized linear model (Supplementary Table S3). However, in maize Isoseq data, the gbM genes had significantly more antisense transcription compared to UM genes (Supplementary Table S4). We conclude that there is evidence that gbM prevents antisense transcription in *A. thaliana* based on Isoseq data from WT plants (Table 1), but no such evidence based on maize. Altogether, however, we believe there are enough compelling observations – both from animals and from *A. thaliana* plants (Choi *et al.* 2020) – to suggest that further dissection of this potential function may be worthwhile.

Assessing the effect of gbM on splicing fidelity

Another hypothesis is that gbM improves splicing fidelity and prevents aberrant intron retention (Figure 2.e), but this raises the question of how splicing fidelity may drive a relationship between gbM and gene expression. One possibility is that poor splicing in the absence of gbM leads to the retention of introns (Figure 2.e) that contain premature stop codons. Aberrant transcripts containing premature stop codons are typically sent to the nonsense-mediated mRNA decay for destruction (Causier *et al.* 2017), which might in turn lower gene expression. This suggests a potential relationship among gbM, splicing fidelity and gene expression.

The effect of gbM on splicing fidelity has been tested across various taxa, and the results have been - like studies of aberrant transcription - somewhat inconsistent. In honey-bee and mouse embryonic stem cells, for example, it is clear that alteration of DNA methylation impacts alternative splicing (Lev Maor *et al.* 2015). In honeybees, DNA methylation is predominantly on gene bodies and

in the CG context. A knock-down of the expression of *dnmt3*, which is required for de novo DNA methylation, decreased global genomic methylation level and caused widespread changes in alternative splicing in fat tissue (Li-Byarlay *et al.* 2013). In mouse embryonic stem cells, Yearim *et al.* (2015) constructed an experimental system in which differential DNA methylation could be limited to a single gene while all other cellular factors remained identical. Using this system, they demonstrated a direct causal relationship between DNA methylation and the recruitment of splicing factors. Patterns of methylation near splice sites have also been studied in maize, where CHG methylation of the splicing acceptor site is associated with a lower efficiency of splicing and CHH methylation does not correlate with splicing efficacy (Regulski *et al.* 2013). Surprisingly, however, the effect of CG methylation was not tested explicitly. Horvath *et al.* (Horvath *et al.* 2019) followed the maize work by measuring splicing fidelity in *A. thaliana*. They found that gbM was negatively correlated with the amount of RNA-seq reads that map to introns, suggesting that gbM genes tend to retain fewer introns in their mRNA compared to UM genes. Similarly, Li *et al.* (2021) recently used ONT DRS to characterize splicing in *A. thaliana*. They found that retained introns had significantly lower CG methylation levels around their splicing sites (both donor and acceptor sites) compared to spliced introns in WT and some CHG and CHH methylation mutants. This suggests that gbM facilitates splicing. However, Bewick *et al.* (2016) found no evidence for this splicing effect when they compared *met1* epiRILs to wild type plants. In fact, they found that WT gbM genes retained significantly fewer intron reads than UM genes after they lost gbM in the *met1* background. This work suggests that this intron effect is a property of the genes, rather than gbM per se.

Given contradictory results in the literature, we further tested the hypothesis that gbM prevents aberrant intron retention using *A. thaliana* Isoseq data (Supplementary Material and Methods). The proportion of full-length Isoseq reads that retained at least one intron was higher in gbM genes (mean 0.149) compared to UM genes (mean 0.106). This result remained significant after taking gene length and gene expression into account in a generalized linear model (Table 1, Supplementary Table S5). We also measured intron RNA-seq coverage in WT and mutant *A. thaliana* plants that lack gbM (Supplementary Material and Methods). Similarly to Horvath *et al.* (2019), we found that gbM genes had a lower intron read coverage compared to UM genes in WT plants (Figure 3.c) both in dataset 1 (mean gbM genes intron coverage 61.51 RPKM, versus 133.42 for UM genes, one-sided wilcoxon test p-value$<2.2\text{x}10^{-16}$) and in dataset 2 (mean gbM genes intron coverage 53.13 RPKM, versus 208.68 in UM genes, one-sided wilcoxon test p-value$<2.2\text{x}10^{-16}$). However, the difference between gbM genes and UM genes was also found in mutant plants that lack gbM (Figure 3.c). Indeed, gbM genes had a lower intron read coverage compared to UM genes in met1-3 mutants (59.04 versus 131.65 RPKM,

one-sided wilcoxon test p-value<$2.2x10^{-16}$) and in met1,sdg7-8 triple mutants (51.27 versus 132.97 RPKM, one-sided wilcoxon test p-value<$2.2x10^{-16}$). This again suggests that the fact that gbM genes have lower intron coverage in RNA-seq data is not due to their methylation state (Table 1). Another possibility is again that some other epigenetic mark plays a redundant role with gbM in preventing aberrant intron retention. In summary, there is not yet a clear consensus, or even a clear trend, as to whether gbM plays a role in splicing fidelity. However, we note again that most of the work in plants has focused on *A. thaliana,* and, as previously discussed, it may be helpful to extend these analyses to other species. The question of the potential role of gbM in splicing fidelity, like its role in aberrant transcription, remains unresolved and will benefit from broader and more comprehensive investigation.

A potential relationship between gbM and Transposable Element (TE) insertion

Another hypothesized function of gbM is that methylation protects against the insertion of some TEs (Figure 2.f). This hypothesis primarily stems from two studies in maize that focused on the Robertson's Mutator (Mu) transposons, which typically insert within or near genes and can be highly deleterious by disrupting gene function. In the first study, Liu *et al.* (2009) found that Mu transposons insert preferentially within unmethylated regions of the B73 genome. However, the methylation context could not be determined, which motivated Regulsky *et al.* (2013) to repeat the analyses with context-specific DNA methylation data. They found that Mu transposon insertion sites within genes were strongly depleted in CG methylated regions (Table 1), but these regions were not depleted in CHG nor CHH methylation relative to average gene methylation. This raises the possibility that gbM is beneficial because it deters transposon insertions. This hypothesis is also difficult to disentangle from from covariates, particularly the observation that gbM genes tend to be under stronger selective constraint than unmethylated genes, as measured by non-synonymous divergence (Takuno and Gaut 2012), although there is conflicting evidence that gbM genes do (Niederhuth *et al.*, 2016) or do not (Takuno and Gaut, 2013) evolve more rapidly in the Poaceae. Nonetheless, it is possible that the apparent difference in TE insertion reflects different strengths of selection on gbM vs. unmethylated genes, rather than a direct effect of gbM on TE insertion rate.

It would be insightful to repeat analyses of the effect of DNA methylation on TE insertion in other plant species and with different TE types to test for the potential broader relevance of this idea. If this phenomenon occurs across diverse species and TE types, it could partially explain the link between gbM and expression stabilization within and between species. Indeed, a genic TE insertion might prevent proper gene expression because silencing of the TE by methylation in the three contexts might

spread to the gene and affect expression (Choi and Lee, 2020). Also, if the TE is inserted within an exon, it might lead to the appearance of premature stop codons and the destruction of mRNA by the nonsense-mediated mRNA decay. This would lead to more expression variation among accessions of a species and also among species (Figure 2c).

Conclusions and future research directions

The molecular mechanisms leading to gbM establishment and maintenance in plants have been remarkably well elucidated (Figure 1). However, the function and potential importance of gbM remain debated, as illustrated in this review by the numerous studies that are inconsistent or in some cases contradictory (Table 1). The main sources for this persistent uncertainty come first from the difficulty in disentangling epigenetic from genetic effects and second from the complex system of redundancies and overlapping functions among gbM, histone variants and other epigenetic marks. These dependencies and redundancies undoubtedly complicate the interpretation of experimental mutants, as illustrated for example by contrasting results based on *met1* plants with or without the *h1* mutation (Table 1; Bewick *et al.* 2016; Choi *et al.* 2020).

Despite these difficulties, there has been substantive progress toward understanding the effect, dynamics and potential adaptive impact of gbM. In the last few years, several studies have concluded that gbM is associated with both the level and the variance of gene expression (Table 1). An interesting corollary of these observations is that many experimental efforts to measure this association based on *A. thaliana* mutants have yielded negative results (Table 1). While these experiments may reflect reality, our view is that experimental approaches have nonetheless suffered from a few common shortcomings. First, we suspect (but certainly do not know) that the reliance on *A. thaliana* is limiting; it is illogical to expect a strong experimental effect in a system that is studied in part because methylation mutants are viable and thus may not have a strong effect relative to other plant systems. Second, as noted above, it is not clear that all mutants are equivalent, because some mutants may be unsuitable for detecting specific effects due to dependencies and redundancies. Finally, it can be exceedingly difficult to identify subtle effects using short-term experimental approaches. Unfortunately, however, the inability to detect an effect is often incorrectly interpreted as an absence of effect.

In contrast, evolutionary and comparative approaches based on genetic diversity or species comparison have often, but not always, found an association between gbM and gene expression, particularly expression homeostasis (Table 1). These analyses also suffer from a number of potential drawbacks, including reliance on simplified models, discrete definitions of which genes are (or are not)

gbM, and an inability to disentangle causation from correlation. One potential reason for the ability to detect an effect using these approaches is that even very subtle effects can accrue over time and thus be detected by evolutionary contrasts. It is difficult to establish whether these associations are causal, due to all the complex reasons cited above – i.e., functional redundancies among epigenetic marks and difficulties in discriminating genetic from epigenetic effects. Nonetheless, the apparent association between gbM and expression is important, because it provides a potential phenotype on which selection can act. Although more investigation is needed to test whether gbM is shaped by selection, both phylogenetic and population genetic studies suggest that selection acts to maintain gbM status in some genes. Moreover, population variation in gbM has been shown to associate with fitness under water stress and selection for flowering time (Shahzad *et al.* 2021). These fitness effects appear to stem from a correlation between gbM and gene expression, as demonstrated by the fact that experimental modification of candidate gene expression affects the trait under study (Shahzad *et al.* 2021). Altogether, these results suggest that gbM may affect fitness and phenotype through an effect on gene expression, thereby potentially affecting species adaptation independently of genetic variation. We emphasize, however, that these effects are subtle, at best, and so there is still much to learn, even about the simple question as to whether and when gbM associates with gene expression (Table 1).

If there is selection on gbM itself – or on another epigenetic feature that correlates with gbM – then this fact may be the basis for an "important conceptual change in evolutionary biology" (Cavalli and Heard 2019), because selection on epigenetic modifications has the potential to affect the timeframe of mutation and thus, potentially, of adaptative change. As an example, Figure 4 illustrates the tempo of epigenetic versus genetic change. As we have noted, epigenetic change can be incredibly rapid. For example, CHH methylation is reset every generation, making it unlikely to be under direct selection because it is not transmitted transgenerationally. CHH methylation (and to some extent CHG methylation) is perhaps better described as tracking genetic (i.e., transposable element and CMT3) activity. In contrast, CG methylation is heritable and mutates approximately three and six orders of magnitude faster than gene duplication and nucleotides, respectively (Figure 4). It is worth noting, however, that the rate of change of an entire gene or allele from gbM to UM is substantially slower, because it probably requires numerous changes of individuals sites. The rate of change for an entire allele, based on our previous SFS analysis in *A. thaliana* populations, is ~3 x 10^{-7} per gene per year (Muyle *et al.* 2021), an estimate based on models that require assumptions about the effective population size but, if accurate, is still faster than the rate of nucleotide change. In theory, then, methylation variation may provide a rapid source of phenotypic novelty that could be subjected to natural selection on rapid – and perhaps even ecological - time-scales.

Authors contributions

Aline Muyle and Brandon Gaut designed the analyses. Aline Muyle analyzed the data. Aline Muyle and Brandon Gaut wrote the manuscript. Danelle Seymour, Yuanda Liu and Bruno Huettel shared data.

Acknowledgments

AM is supported by an EMBO Postdoctoral Fellowship ALTF 775-2017 and by HFSPO Fellowship LT000496/2018-L. BSG is supported by NSF grant IOS-1542703. We would like to thank Galen Martin for commenting on the manuscript.

Data availability

Arabidopsis thaliana Isoseq data newly sequenced here was deposited in NCBI under BioProject accession PRJNA754773.

References

Arias T., M. A. Beilstein, M. Tang, M. R. McKain, and J. C. Pires, 2014 Diversification times among Brassica (Brassicaceae) crops suggest hybrid formation after 20 million years of divergence. Am. J. Bot. 101: 86–91. https://doi.org/10.3732/ajb.1300312

Bewick A. J., L. Ji, C. E. Niederhuth, E.-M. Willing, B. T. Hofmeister, *et al.*, 2016 On the origin and evolutionary consequences of gene body DNA methylation. Proc. Natl. Acad. Sci. U.S.A. 113: 9111–9116. https://doi.org/10.1073/pnas.1604666113

Bewick A. J., C. E. Niederhuth, L. Ji, N. A. Rohr, P. T. Griffin, *et al.*, 2017 The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. Genome Biol. 18: 65. https://doi.org/10.1186/s13059-017-1195-1

Bewick A. J., Y. Zhang, J. M. Wendte, X. Zhang, and R. J. Schmitz, 2019 Evolutionary and Experimental Loss of Gene Body Methylation and Its Consequence to Gene Expression. G3 (Bethesda) 9: 2441–2445. https://doi.org/10.1534/g3.119.400365

Bird A. P., 1980 DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 8: 1499–1504. https://doi.org/10.1093/nar/8.7.1499

Boquete M. T., A. Muyle, and C. Alonso, 2021 Plant epigenetics: phenotypic and functional diversity beyond the DNA sequence. Am J Bot 108: 553–558. https://doi.org/10.1002/ajb2.1645

Calarco J. P., F. Borges, M. T. A. Donoghue, F. Van Ex, P. E. Jullien, *et al.*, 2012 Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. Cell 151: 194–205. https://doi.org/10.1016/j.cell.2012.09.001

Causier B., Z. Li, R. De Smet, J. P. B. Lloyd, Y. Van de Peer, *et al.*, 2017 Conservation of Nonsense-Mediated mRNA Decay Complex Components Throughout Eukaryotic Evolution. Sci Rep 7: 16692. https://doi.org/10.1038/s41598-017-16942-w

Cavalli G., and E. Heard, 2019 Advances in epigenetics link genetics to the environment and disease. Nature 571: 489–499. https://doi.org/10.1038/s41586-019-1411-0

Charlesworth B., and K. Jain, 2014 Purifying selection, drift, and reversible mutation with arbitrarily high mutation rates. Genetics 198: 1587–1602. https://doi.org/10.1534/genetics.114.167973

Charlesworth D., N. H. Barton, and B. Charlesworth, 2017 The sources of adaptive variation. Proc Biol Sci 284: 20162864. https://doi.org/10.1098/rspb.2016.2864

Choi J., D. B. Lyons, M. Y. Kim, J. D. Moore, and D. Zilberman, 2020 DNA Methylation and Histone H1 Jointly Repress Transposable Elements and Aberrant Intragenic Transcripts. Mol. Cell 77: 310-323.e7. https://doi.org/10.1016/j.molcel.2019.10.011

Choi, J.Y, Y.C.G. Lee, 2020. Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. Plos Genetics. https://doi.org/10.1371/journal.pgen.1008872.

Cokus S. J., S. Feng, X. Zhang, Z. Chen, B. Merriman, *et al.*, 2008 Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452: 215–219. https://doi.org/10.1038/nature06745

Coleman-Derr D., and D. Zilberman, 2012 DNA methylation, H2A.Z, and the regulation of constitutive expression. Cold Spring Harb. Symp. Quant. Biol. 77: 147–154. https://doi.org/10.1101/sqb.2012.77.014944

Cartolano M, Huettel B, Hartwig B, Reinhardt R, Schneeberger K. 2016. cDNA Library Enrichment of Full Length Transcripts for SMRT Long Read Sequencing. PLoS ONE. 11:e0157779. doi: 10.1371/journal.pone.0157779.

Feng S., S. J. Cokus, X. Zhang, P.-Y. Chen, M. Bostick, *et al.*, 2010 Conservation and divergence of methylation patterning in plants and animals. PNAS 107: 8689–8694. https://doi.org/10.1073/pnas.1002720107

Galtier N., C. Roux, M. Rousselle, J. Romiguier, E. Figuet, *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. https://doi.org/10.1093/molbev/msy015

Gaut B., L. Yang, S. Takuno, and L. E. Eguiarte, 2011 The Patterns and Causes of Variation in Plant Nucleotide Substitution Rates. Annual Review of Ecology, Evolution, and Systematics 42: 245–266. https://doi.org/10.1146/annurev-ecolsys-102710-145119

Gore A. V., K. A. Tomins, J. Iben, L. Ma, D. Castranova, *et al.*, 2018 An epigenetic mechanism for cavefish eye degeneration. Nature Ecology & Evolution 2: 1155–1160. https://doi.org/10.1038/s41559-018-0569-4

Gouil Q., and D. C. Baulcombe, 2016 DNA Methylation Signatures of the Plant Chromomethyltransferases. PLoS Genet 12: e1006526. https://doi.org/10.1371/journal.pgen.1006526

Graaf A. van der, R. Wardenaar, D. A. Neumann, A. Taudt, R. G. Shaw, *et al.*, 2015 Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. Proc. Natl. Acad. Sci. U.S.A. 112: 6676–6681. https://doi.org/10.1073/pnas.1424254112

Holliday R., 1994 Epigenetics: an overview. Dev Genet 15: 453–457. https://doi.org/10.1002/dvg.1020150602

Horvath R., B. Laenen, S. Takuno, and T. Slotte, 2019 Single-cell expression noise and gene-body methylation in Arabidopsis thaliana. Heredity (Edinb) 123: 81–91. https://doi.org/10.1038/s41437-018-0181-z

Jelesko J. G., K. Carter, W. Thompson, Y. Kinoshita, and W. Gruissem, 2004 Meiotic recombination between paralogous RBCSB genes on sister chromatids of Arabidopsis thaliana. Genetics 166: 947–957. https://doi.org/10.1534/genetics.166.2.947

Johnson L. M., M. Bostick, X. Zhang, E. Kraft, I. Henderson, *et al.*, 2007 The SRA methyl-cytosine-binding domain links DNA and histone methylation. Curr Biol 17: 379–384. https://doi.org/10.1016/j.cub.2007.01.009

Kawakatsu T., S.-S. C. Huang, F. Jupe, E. Sasaki, R. J. Schmitz, *et al.*, 2016 Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell 166: 492–505. https://doi.org/10.1016/j.cell.2016.06.044

Kawashima T., and F. Berger, 2014 Epigenetic reprogramming in plant sexual reproduction. Nat. Rev. Genet. 15: 613–624. https://doi.org/10.1038/nrg3685

Lev Maor G., A. Yearim, and G. Ast, 2015 The alternative role of DNA methylation in splicing regulation. Trends Genet 31: 274–280. https://doi.org/10.1016/j.tig.2015.03.002

Li Q., S. R. Eichten, P. J. Hermanson, V. M. Zaunbrecher, J. Song, *et al.*, 2014 Genetic perturbation of the maize methylome. Plant Cell 26: 4602–4616. https://doi.org/10.1105/tpc.114.133140

Li Q., S. Chen, A. W.-S. Leung, Y. Liu, Y. Xin, *et al.*, 2021 *DNA methylation affects pre-mRNA transcriptional initiation and processing in Arabidopsis*.

Li-Byarlay H., Y. Li, H. Stroud, S. Feng, T. C. Newman, *et al.*, 2013 RNA interference knockdown of DNA methyl-transferase 3 affects gene alternative splicing in the honey bee. PNAS 110: 12750–12755. https://doi.org/10.1073/pnas.1310735110

Lister R., R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, *et al.*, 2008 Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133: 523–536. https://doi.org/10.1016/j.cell.2008.03.029

Liu S., C.-T. Yeh, T. Ji, K. Ying, H. Wu, *et al.*, 2009 Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet 5: e1000733. https://doi.org/10.1371/journal.pgen.1000733

Luo C., P. Hajkova, and J. R. Ecker, 2018 Dynamic DNA methylation: In the right place at the right time. Science 361: 1336–1340. https://doi.org/10.1126/science.aat6806

Martin G. T., D. K. Seymour, and B. S. Gaut, 2021 CHH Methylation Islands: A Nonconserved Feature of Grass Genomes That Is Positively Associated with Transposable Elements but Negatively Associated with Gene-Body Methylation. Genome Biol Evol 13: evab144. https://doi.org/10.1093/gbe/evab144

Mattila T. M., B. Laenen, and T. Slotte, 2020 Population Genomics of Transitions to Selfing in Brassicaceae Model Systems. Methods Mol. Biol. 2090: 269–287. https://doi.org/10.1007/978-1-0716-0199-0_11

Maunakea A. K., R. P. Nagarajan, M. Bilenky, T. J. Ballinger, C. D'Souza, *et al.*, 2010 Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466: 253–257. https://doi.org/10.1038/nature09165

Meng D., M. Dubin, P. Zhang, E. J. Osborne, O. Stegle, *et al.*, 2016 Limited Contribution of DNA Methylation Variation to Expression Regulation in Arabidopsis thaliana. PLoS Genet 12: e1006141. https://doi.org/10.1371/journal.pgen.1006141

Miura A., M. Nakamura, S. Inagaki, A. Kobayashi, H. Saze, *et al.*, 2009 An Arabidopsis jmjC domain protein protects transcribed genes from DNA methylation at CHG sites. EMBO J. 28: 1078–1086. https://doi.org/10.1038/emboj.2009.59

Muyle A., and B. S. Gaut, 2019 Loss of Gene Body Methylation in Eutrema salsugineum Is Associated with Reduced Gene Expression. Mol. Biol. Evol. 36: 155–158. https://doi.org/10.1093/molbev/msy204

Muyle A., J. Ross-Ibarra, D. K. Seymour, and B. S. Gaut, 2021 Gene body methylation is under selection in Arabidopsis thaliana. Genetics 218: iyab061. https://doi.org/10.1093/genetics/iyab061

Neri F., S. Rapelli, A. Krepelova, D. Incarnato, C. Parlato, *et al.*, 2017 Intragenic DNA methylation prevents spurious transcription initiation. Nature 543: 72–77. https://doi.org/10.1038/nature21373

Niederhuth C. E., A. J. Bewick, L. Ji, M. S. Alabady, K. D. Kim, *et al.*, 2016 Widespread natural variation of DNA methylation within angiosperms. Genome Biology 17: 194. https://doi.org/10.1186/s13059-016-1059-0

Niederhuth C. E., and R. J. Schmitz, 2017 Putting DNA methylation in context: from genomes to gene expression in plants. Biochim Biophys Acta Gene Regul Mech 1860: 149–156. https://doi.org/10.1016/j.bbagrm.2016.08.009

Ossowski S., K. Schneeberger, J. I. Lucas-Lledó, N. Warthmann, R. M. Clark, *et al.*, 2010 The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science 327: 92–94. https://doi.org/10.1126/science.1180677

Papareddy R. K., K. Páldi, A. D. Smolka, P. Hüther, C. Becker, *et al.*, 2021 Repression of CHROMOMETHYLASE 3 prevents epigenetic collateral damage in Arabidopsis. Elife 10: e69396. https://doi.org/10.7554/eLife.69396

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.

Regulski M., Z. Lu, J. Kendall, M. T. A. Donoghue, J. Reinders, *et al.*, 2013 The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. Genome Res. 23: 1651–1662. https://doi.org/10.1101/gr.153510.112

Reinders J., B. B. H. Wulff, M. Mirouze, A. Marí-Ordóñez, M. Dapp, *et al.*, 2009 Compromised stability of DNA methylation and transposon immobilization in mosaic Arabidopsis epigenomes. Genes Dev. 23: 939–950. https://doi.org/10.1101/gad.524609

Saze H., A. Shiraishi, A. Miura, and T. Kakutani, 2008 Control of genic DNA methylation by a jmjC domain-containing protein in Arabidopsis thaliana. Science 319: 462–465. https://doi.org/10.1126/science.1150987

Schmitz R. J., Z. A. Lewis, and M. G. Goll, 2019 DNA Methylation: Shared and Divergent Features across Eukaryotes. Trends Genet. 35: 818–827. https://doi.org/10.1016/j.tig.2019.07.007

Seymour D. K., D. Koenig, J. Hagmann, C. Becker, and D. Weigel, 2014 Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. PLoS Genet. 10: e1004785. https://doi.org/10.1371/journal.pgen.1004785

Seymour D. K., and B. S. Gaut, 2019 Phylogenetic shifts in gene body methylation correlate with gene expression and reflect trait conservation. Mol. Biol. Evol. https://doi.org/10.1093/molbev/msz195

Shahzad Z., J. D. Moore, and D. Zilberman, 2021 Gene body methylation mediates epigenetic inheritance of plant traits. bioRxiv 2021.03.15.435374. https://doi.org/10.1101/2021.03.15.435374

Steige K. A., B. Laenen, J. Reimegård, D. G. Scofield, and T. Slotte, 2017 Genomic analysis reveals major determinants of cis-regulatory variation in Capsella grandiflora. Proc. Natl. Acad. Sci. U.S.A. 114: 1087–1092. https://doi.org/10.1073/pnas.1612561114

Stoddard C. I., S. Feng, M. G. Campbell, W. Liu, H. Wang, *et al.*, 2019 A Nucleosome Bridging Mechanism for Activation of a Maintenance DNA Methyltransferase. Mol Cell 73: 73-83.e6. https://doi.org/10.1016/j.molcel.2018.10.006

Stroud H., M. V. C. Greenberg, S. Feng, Y. V. Bernatavichute, and S. E. Jacobsen, 2013 Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell 152: 352–364. https://doi.org/10.1016/j.cell.2012.10.054

Takuno S., and B. S. Gaut, 2012 Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly. Mol. Biol. Evol. 29: 219–227. https://doi.org/10.1093/molbev/msr188

Takuno S., and B. S. Gaut, 2013 Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. Proc. Natl. Acad. Sci. U.S.A. 110: 1797–1802. https://doi.org/10.1073/pnas.1215380110

Takuno S., J.-H. Ran, and B. S. Gaut, 2016 Evolutionary patterns of genic DNA methylation vary across land plants. Nat Plants 2: 15222. https://doi.org/10.1038/nplants.2015.222

Takuno S., D. K. Seymour, and B. S. Gaut, 2017 The Evolutionary Dynamics of Orthologs That Shift in Gene Body Methylation between Arabidopsis Species. Mol. Biol. Evol. 34: 1479–1491. https://doi.org/10.1093/molbev/msx099

Teissandier A., and D. Bourc'his, 2017 Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. EMBO J. 36: 1471–1473. https://doi.org/10.15252/embj.201796812

Teixeira F. K., and V. Colot, 2009 Gene body DNA methylation in plants: a means to an end or an end to a means? EMBO J. 28: 997–998. https://doi.org/10.1038/emboj.2009.87

Tran R. K., J. G. Henikoff, D. Zilberman, R. F. Ditt, S. E. Jacobsen, *et al.*, 2005 DNA methylation profiling identifies CG methylation clusters in Arabidopsis genes. Curr. Biol. 15: 154–159. https://doi.org/10.1016/j.cub.2005.01.008

Vidalis A., D. Živković, R. Wardenaar, D. Roquis, A. Tellier, *et al.*, 2016 Methylome evolution in plants. Genome Biol. 17: 264. https://doi.org/10.1186/s13059-016-1127-5

Wang B., Tseng E., Regulski M., Clark T., Hon T., *et al.*, 2016. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun. 7:11708. doi: 10.1038/ncomms11708.

Wendte J. M., Y. Zhang, L. Ji, X. Shi, R. R. Hazarika, *et al.*, 2019 Epimutations are associated with CHROMOMETHYLASE 3-induced de novo DNA methylation. Elife 8. https://doi.org/10.7554/eLife.47891

Yao N., R. J. Schmitz, and F. Johannes, 2021 Epimutations Define a Fast-Ticking Molecular Clock in Plants. Trends Genet 37: 699–710. https://doi.org/10.1016/j.tig.2021.04.010

Yearim A., S. Gelfman, R. Shayevitch, S. Melcer, O. Glaich, *et al.*, 2015 HP1 Is Involved in Regulating the Global Impact of DNA Methylation on Alternative Splicing. Cell Reports 10: 1122–1134. https://doi.org/10.1016/j.celrep.2015.01.038

Yi S. V., 2017 Insights into Epigenome Evolution from Animal and Plant Methylomes. Genome Biology and Evolution 9: 3189–3201. https://doi.org/10.1093/gbe/evx203

Zemach A., and D. Zilberman, 2010 Evolution of eukaryotic DNA methylation and the pursuit of safer sex. Curr Biol 20: R780-785. https://doi.org/10.1016/j.cub.2010.07.007

Zhang X., J. Yazaki, A. Sundaresan, S. Cokus, S. W.-L. Chan, *et al.*, 2006 Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell 126: 1189–1201. https://doi.org/10.1016/j.cell.2006.08.003

Zhang R., C. P. G. Calixto, Y. Marquez, P. Venhuizen, N. A. Tzioutziou, *et al.*, 2017 A high quality Arabidopsis transcriptome for accurate transcript-level analysis of alternative splicing. Nucleic Acids Res 45: 5061–5073. https://doi.org/10.1093/nar/gkx267

Zilberman D., M. Gehring, R. K. Tran, T. Ballinger, and S. Henikoff, 2007 Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat. Genet. 39: 61–69. https://doi.org/10.1038/ng1929

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

Zilberman D., 2017 An evolutionary case for functional gene body methylation in plants and animals. Genome Biol. 18: 87. https://doi.org/10.1186/s13059-017-1230-2

FIGURE LEGENDS:

Figure 1: The establishment and maintenance of gbM in plants. The DNA is represented as a line coiled around nucleosomes. Red dots indicate methylated H3K9 tails. CG, CHG and CHH DNA methylation are drawn as black, gray and white lollipops, respectively. **a-b)** CMT3 induces *de novo* methylation at CHG sites of genes associated with inaccessible chromatin marks and heterochromatin histone variants (Papareddy *et al.* 2021). **b-c)** The CHG-H3K9me2 self-reinforcing feedback loop is then established. **c-d)** CMT3 preferentially *de novo* methylates CWG sites but to a lesser extent also methylates other contexts, such as CG. **d-e)** Demethylation of H3K9 by IBM1 is coupled to gene transcription. **f)** After a few cell divisions, only CG methylation (mCG) remains due to MET1 maintenance.

Figure 2: Potential gbM functions and evolutionary consequences. Unmethylated genes, represented on the left column, are compared to gbM genes on the right. TSS stands for transcription start site, TTS for transcription termination site and TE for transposable element. **(a)** gbM is hypothesized to upregulate gene expression. The number of mRNA molecules, represented by wavy lines, illustrates the gene expression level. **(b)** gbM genes may stabilize gene expression by triggering consistent expression levels among the cells of a tissue. **(c)** gbM may stabilize gene expression, as seen by the more constant and conserved expression levels observed among species. **(d)** gbM could prevent aberrant internal and reverse transcription by silencing alternative promoters within genes. gbM might also inhibit aberrant TTS. These hypotheses are coherent with the typical depletion of CG methylation observed around the TSS and TTS of genes. **(e)** gbM is hypothesized to facilitate correct splicing and prevent aberrant intron retention. **(f)** Some TEs preferentially insert into genes, however, gbM may protect against deleterious insertions within genes.

Figure 3: Novel analyses to assess the effect of gbM on transcripts. a) Proportion of full-length Isoseq reads with conventional transcription start site (TSS) in gbM and UM genes in maize and *A. thaliana*. Isoseq reads that started after the start of exon 1 were considered as non-conventional. **b)** RNA-seq read coverage ratio between exon 3 and exon 1 for gbM and UM genes in *A. thaliana*. Internal transcription start happening between exon 1 and exon 3 is expected to increase the ratio of exon 3 to exon 1 coverage. **c)** RNA-seq read coverage of introns (in RPKM) for gbM and UM genes in *A. thaliana*. Pools of gbM genes are drawn in red and unmethylated genes (UM) in turquoise. In

Dataset 1, WT controls were compared to *met1-3* mutants. In Dataset 2, WT controls were compared to *met1,sdg7-8* triple mutants. The boxplots show the median, the hinges are the first and third quartiles (the 25th and 75th percentiles) and the whiskers extend from the hinge to the largest or smallest value no further than 1.5 times the interquartile range (distance between the first and third quartiles).

Figure 4: Tempo of epigenetic versus genetic change. The rates of change come from a series of sources (Jelesko *et al.* 2004; Ossowski *et al.* 2010; Gaut *et al.* 2011; van der Graaf *et al.* 2015).

Table 1: Review of the literature on the possible functions of gbM and selective pressures that act on it. Many studies contradict one another, suggesting that if gbM indeed has a function, its effect must be subtle.

Reference	Species and genotypes	Data type	gbM upregulates gene expression	gbM stabilizes gene expression	gbM prevents internal transcription start	gbM prevents aberrant transcription termination	gbM prevents antisens transcription	gbM prevents intron retention	gbM prevents TE insertion	gbM is under selection
Regulsky *et al.* 2013	Maize B73	BS-seq							yes (Mu element)	
Bewick *et al.* 2016	*Eutrema salsugineum* WT, *Arabidopsis thaliana* WT + *met1* epiRIL + *met1,sdg7,sdg8*	BS-seq, RNA-seq	no				no	no		
Vidalis *et al.* 2016	92 WT *A. thaliana*	BS-seq								no
Meng *et al.* 2016	135 *A. thaliana* wild accessions	BS-seq, RNA-seq	yes (for a few genes)							
Takuno *et al.* 2017	*A. thaliana*, *A. lyrata*	BS-seq, RNA-seq	trend	yes						
Steige *et al.* 2017	*Capsella grandiflora*	BS-seq, RNA-seq		yes						
Neri *et al.* 2017	mouse embryonic stem cells WT + *Dnmt3b* + *SetD2* knockdown + DNMT3B rescue	Dnmt3b ChIP-seq, BS-seq, RNA-seq, Pol II ChIP-seq, CAPIP-seq, DECAP-seq			yes					
Teissandier and Bourc'his 2017	mouse embryonic stem cells WT + *Dnmt- tKO* (triple mutant) + chemical inhibition of methylation	RNA-seq			no					
Muyle and Gaut 2019	*E. salsugineum*, *A. thaliana*	BS-seq, RNA-seq	yes							
Bewick *et al.* 2019	*E. salsugineum*, *A. thaliana* WT + *met1* epiRIL	BS-seq, RNA-seq	no							

Horvath *et al.* 2019	*A. thaliana* WT	BS-seq, Single-cell RNA-seq		yes				yes		
Seymour and Gaut 2019	8 species of the grass family (Poaceae)	BS-seq, RNA-seq	yes	yes						yes
Choi *et al.* 2020	*A. thaliana* WT + *met1* + *h1* + *h1,met1*	BS-seq, RNA-seq, decap-seq					yes, jointly with histone H1			
Muyle *et al.* 2021	1001 WT *A. thaliana*	BS-seq, RNA-seq	yes	yes						yes
Shahzad *et al.* 2021	1001 WT *A. thaliana*, *A. thaliana* mutant collection	BS-seq, RNA-seq	yes							
Li *et al.* 2021	WT *A. thaliana* and *met1* mutant.	BS-seq, ONT DRS			yes	yes		yes		
This manuscript	Maize WT, *A. thaliana* WT + *met1* + *met1,sdg7,sdg8*	BS-seq, RNA-seq, Isoseq			yes in WT *A. thaliana* Isoseq data but not in maize nor in RNA-seq data.		yes in WT *A. thaliana* Isoseq data but not in maize.	no		

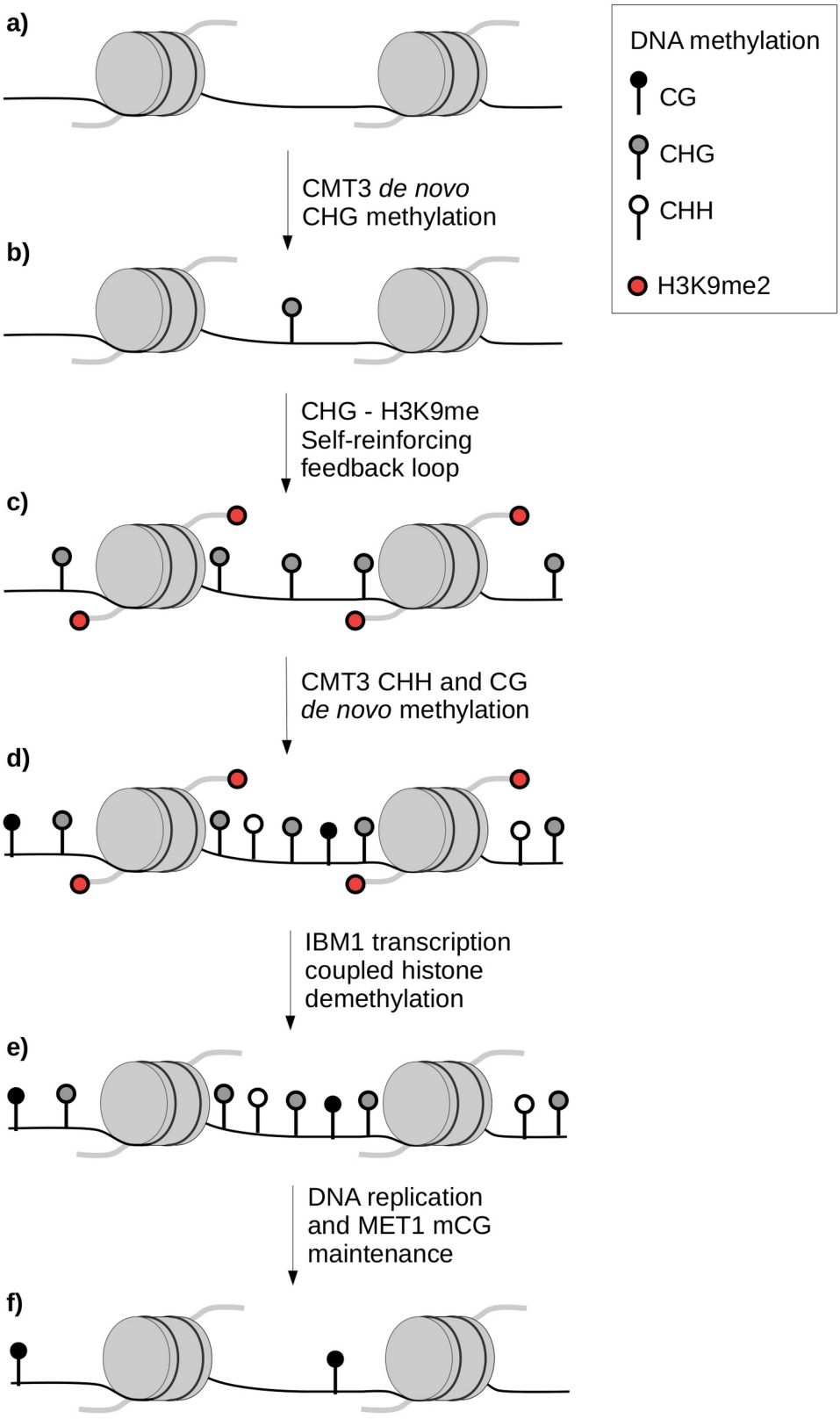
a)
CMT3 *de novo*
CHG methylation
b)
CHG - H3K9me
Self-reinforcing
feedback loop
c)
CMT3 CHH and CG
de novo methylation
d)
IBM1 transcription
coupled histone
demethylation
e)
DNA replication
and MET1 mCG
maintenance
f)
DNA methylation
CG
CHG
CHH
H3K9me2

Fig 1

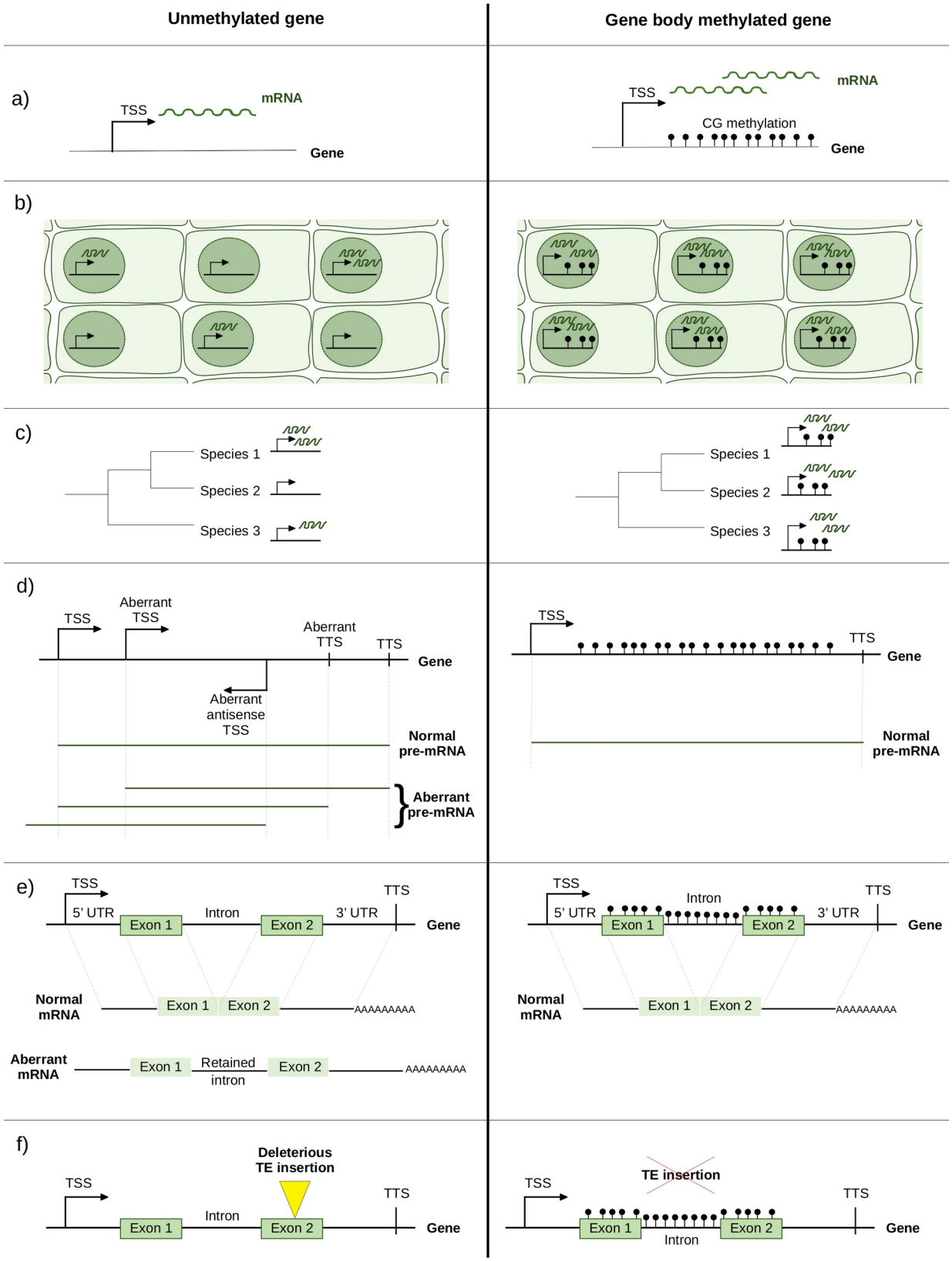
Unmethylated gene
Gene body methylated gene
a)
TSS
mRNA
Gene
CG methylation
b)
c)
Species 1
Species 2
Species 3
d)
Aberrant TSS
Aberrant TTS
TTS
Aberrant antisense TSS
Normal pre-mRNA
Aberrant pre-mRNA
e)
5' UTR
Exon 1
Intron
Exon 2
3' UTR
Normal mRNA
AAAAAAAAA
Aberrant mRNA
Retained intron
f)
Deleterious TE insertion
TE insertion

Fig 2

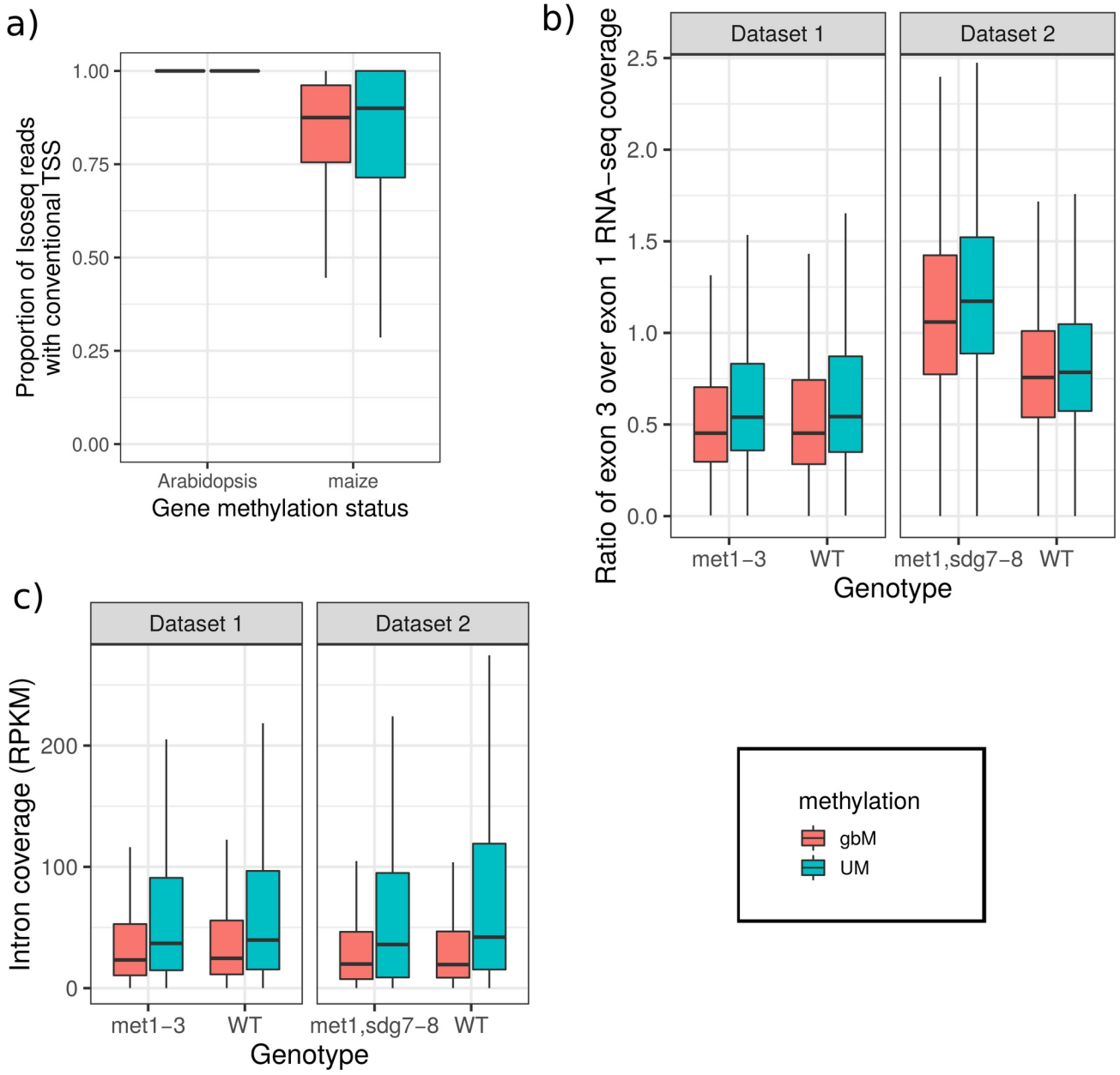
a)
Proportion of Isoseq reads with conventional TSS
1.00
0.75
0.50
0.25
0.00
Arabidopsis
maize
Gene methylation status
b)
Dataset 1
Dataset 2
Ratio of exon 3 over exon 1 RNA−seq coverage
2.5
2.0
1.5
1.0
0.5
0.0
met1−3
WT
met1,sdg7−8
WT
Genotype
c)
Dataset 1
Dataset 2
Intron coverage (RPKM)
200
100
0
met1−3
WT
met1,sdg7−8
WT
Genotype
methylation
gbM
UM

Fig 3

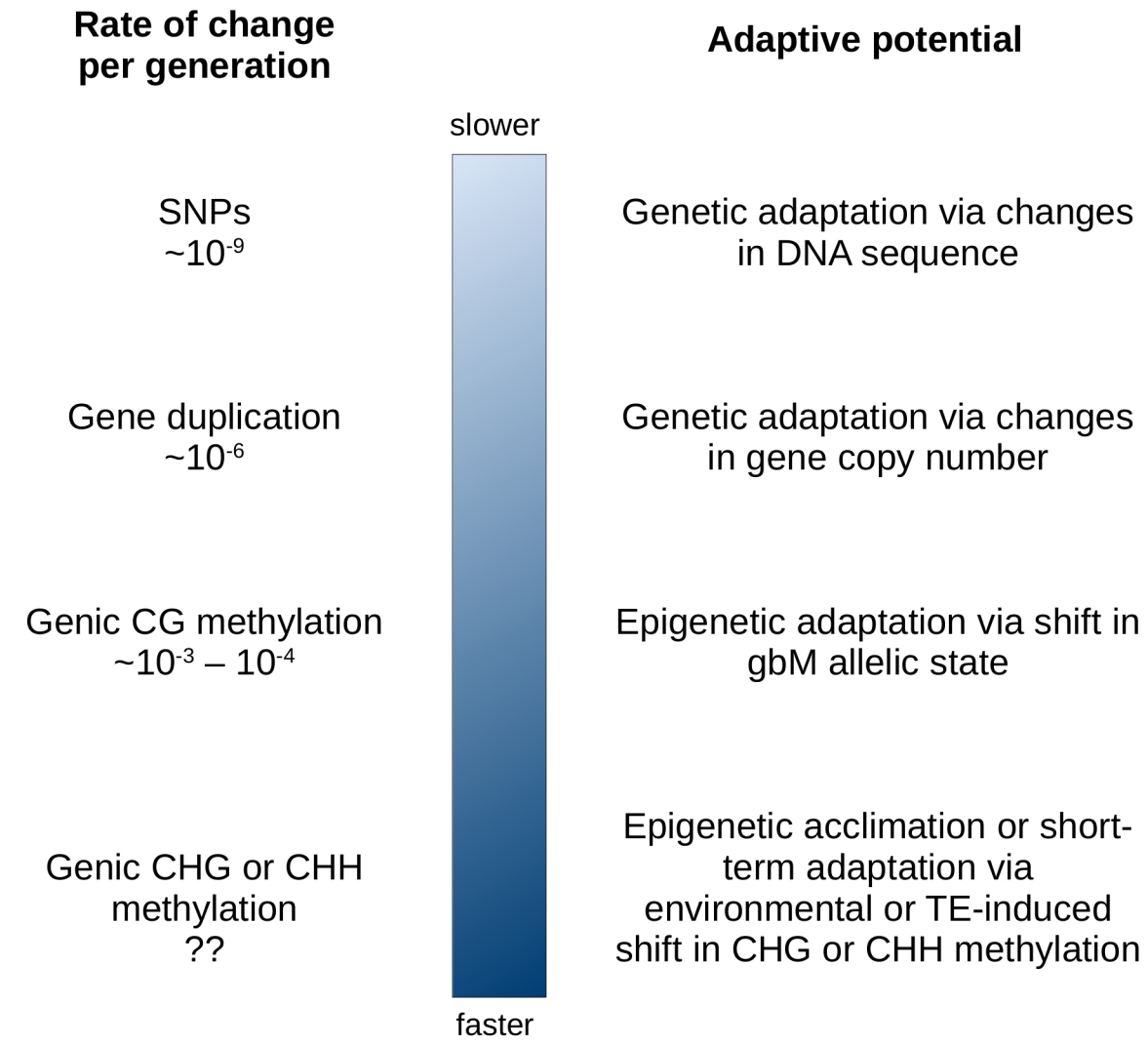
Rate of change per generation
Adaptive potential
slower
faster
SNPs
~10-9
Genetic adaptation via changes in DNA sequence
Gene duplication
~10-6
Genetic adaptation via changes in gene copy number
Genic CG methylation
~10-3 – 10-4
Epigenetic adaptation via shift in gbM allelic state
Genic CHG or CHH methylation
??
Epigenetic acclimation or short-term adaptation via environmental or TE-induced shift in CHG or CHH methylation

Fig 4