

ORIGINATION, DISTRIBUTION, AND IMPLICATION OF SPONTANEOUS EPIALLELES IN PLANTS

by

YINWEN ZHANG

(Under the Direction of Robert J. Schmitz)

ABSTRACT

As one of the major chromatin modifications, DNA methylation plays an important role in chromosome stability, transposable element (TE) deactivation and gene expression regulation. Although the majority of DNA methylation is occupied on repeats and TEs in heterochromatic regions, many coding genes also contain DNA methylation. In flowering plants, many genes contain exclusively CG methylation in the transcribed region, which is referred to as gene body DNA methylation (gbM). gbM is prevalent throughout angiosperms and is a major driver of methylation variations. Curiously, the loss of gbM was found in *E. salsugineum*, a member from the Brassicaceae family, and it was accompanied by a loss of the DNA methyltransferase CHROMOMETHYLASE 3 (CMT3). Those findings supported a hypothesis that CMT3 is required for the origination of gbM.

My thesis was focused on investigating the mechanistic origin of gbM. Firstly, we tested what happens when the *CMT3* gene of *A. thaliana* was transferred into *E. salsugineum*. I showed that *de novo* CHG methylation accumulated on genes over generations and was subsequently followed by the establishment of CG methylation. Additionally, by tracing CG methylation changes for six generations, I showed that the *de novo* CG methylation that accumulated in genes could be inherited even following silencing of the *CMT3* transgene.

Next, I investigated the methylomes of 725 *A. thaliana* accessions. I found that DNA methylation is highly variable on both gene and heterochromatic regions among accessions. The number of genes possessing DNA methylation is inversely correlated with the abundance of methylation in heterochromatin. I trained a machine learning model that could utilize genic features to accurately predict gbM status, with gene length and CWG frequency being the most

important predictors. Importantly, I found that the most conserved gbM genes were more likely to shift to transposon-like methylation, which affects transcription and could lead to phenotypic consequences.

Finally, I analyzed the methylomes for a variety of mutant lines to test a model of how heterochromatin methylation affects gbM. The results showed that ectopic hypermethylation on genes is due to feedback regulation of pathways involved in maintaining DNA methylation at heterochromatin. Using methylomes of *ddm1* epigenetic recombinant inbred lines (epiRILs), I showed that hypomethylated chromosomes lead to ectopic DNA methylation on genes. Epiallelic variation in euchromatin is associated with QTL primarily located in heterochromatin.

Together, those studies investigated DNA methylomes for a variety of experimental mutant lines and natural populations, providing mechanistic investigation into spontaneous epiallele formation, which is associated with feedback regulation of pathways involved in maintenance of DNA methylation in heterochromatin. These findings expand our knowledge on the origination, distribution, and implication of spontaneous epialleles.

INDEX WORDS: Epigenetics, Epiallele, DNA methylation, Epigenomics, gbM,
Heterochromatin, *Arabidopsis thaliana*

ORIGINATION, DISTRIBUTION, AND IMPLICATION OF SPONTANEOUS EPIALLELES
IN PLANTS

by

YINWEN ZHANG

BS, Northwest A&F University, China, 2010

MS, Northwest A&F University, China, 2013

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial
Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2021

© 2021

Yinwen Zhang

All Rights Reserved

ORIGINATION, DISTRIBUTION, AND IMPLICATION OF SPONTANEOUS EPIALLELES
IN PLANTS

by

YINWEN ZHANG

Major Professor:	Robert J. Schmitz
Committee:	Zachary A. Lewis
	James H. Leebens-Mack
	Casey M. Bergman

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
December 2021

ACKNOWLEDGEMENTS

I always thought it is my best luck to have Bob Schmitz as my Ph.D. advisor at UGA. He is the person who provided me with systematical training in science and opportunities to work with scientists from various disciplines. He is a great model who set an example of what a great scientist should be. During this journey, he is the person who always provide me with great ideas that promote my research. He is also the person who always encourages my thoughts and work, which means a lot to me. He is the great leader, mentor, and friend of my entire graduate career.

Many thanks to the members of the Schmitz lab for the various roles they have played in this journey. I am especially indebted to Jered Michael Wendte and Hosung Jang who contribute a great many thoughts and experimental efforts that make this work possible. I would also like to thank Adam Bewick, and Lexiang Ji for always being available when I had questions or needed advice. And many thanks to William Jordan for continually being willing to edit and improve my writing.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iv
LIST OF FIGURES	vii
CHAPTER	
1 A REVIEW OF DNA METHYLATION AND EPIALLELES IN PLANTS	1
OVERVIEW OF DNA METHYLATION	1
EPIMUTATIONS IN PLANTS.....	6
EPIGENETIC MUTANT LINES.....	9
TRADEOFF BETWEEN CHROMATIN MAINTENANCE AND GENIC	
EPIMUTATION	13
DISSERTATION OUTLINE.....	15
2 EPIMUTATIONS ARE ASSOCIATED WITH CHROMOMETHYLASE 3-	
INDUCED DE NOVO DNA METHYLATION	16
ABSTRACT.....	17
INTRODUCTION	17
RESULTS	21
DISCUSSION	34
MATERIALS AND METHODS.....	36
3 NATURAL VARIATION IN DNA METHYLATION HOMEOSTASIS AND THE	
EMERGENCE OF EPIALLELES.....	67

ABSTRACT.....	68
INTRODUCTION.....	69
RESULTS	74
DISCUSSION	84
MATERIALS AND METHODS.....	86
4 HETEROCHROMATIN IS A QUANTITATIVE TRAIT ASSOCIATED WITH SPONTANEOUS EPIALLELE FORMATION.....	110
ABSTRACT.....	111
INTRODUCTION	111
RESULTS	116
DISCUSSION.....	123
MATERIALS AND METHODS.....	127
REFERENCES	150

LIST OF FIGURES

	Page
Figure 2.1: Expression of <i>AtCMT3</i> in <i>E. salsugineum</i> results in increased CHG methylation	47
Figure 2.2: <i>AtCMT3</i> expression results in mCHG over repeats and a subset of gene bodies.....	48
Figure 2.3: Gains in CHG methylation do not alter H3K9me2 levels or distribution	49
Figure 2.4: Genes that gain CHG methylation have <i>A. thaliana</i> gbM gene characteristics	51
Figure 2.5: mCHG over gene bodies is associated with gains in non-mCHG.....	52
Figure 2.6: Hypothetical model for <i>CMT3</i> establishment of gbM	53
Figure S2.1: Diagram of all experimental <i>E. salsugineum</i> lines analyzed in this study with associated data collected	54
Figure S2.2: CHG methylation increases in <i>AtCMT3</i> -expressing <i>E. salsugineum</i> lines	55
Figure S2.3: CG, CHG, and CHH DMRs co-localize	56
Figure S2.4: <i>AtCMT3</i> expression results in gains in CHG methylation over similar regions between lineages and across generations	57
Figure S2.5: Genome wide percent CHG methylation is correlated with <i>AtCMT3</i> transgene expression levels.	58
Figure S2.6: <i>AtCMT3</i> expression results in gains in CHG methylation over similar genes between lineages and across generations	59
Figure S2.7: The number of genes gaining a minimum of 5% CHG methylation (CHG-gain genes) is correlated with <i>AtCMT3</i> transgene expression levels.....	60
Figure S2.8: Gains in CHG methylation do not alter H3K9me2 levels or distribution.....	61

Figure S2.9: Gains in CHG methylation do not alter H3K9me2 levels or distribution.....	62
Figure S2.10: Changes in expression are not related to CHG methylation levels.....	63
Figure S2.11: mCHG over gene bodies is associated with gains in non-mCHG	64
Figure S2.12: AtCMT3-induced genic CG methylation is maintained at higher levels than background following loss of AtCMT3 expression.....	65
Figure 3.1: Genic DNA methylation patterns represent a continuum of epigenetic states.....	97
Figure 3.2: Distribution of genic DNA methylation patterns within <i>A. thaliana</i>	98
Figure 3.3: Genetic and epigenetic factors associated with gbM gene number.....	99
Figure 3.4: Genic features are predictive of gbM status and correlate with gbM frequency	100
Figure 3.5: Genes with high gbM frequencies are preferentially methylated by CMT3.....	102
Figure 3.6: GbM genes are more susceptible to epiallelic shifts that lead to transcriptional silencing	103
Figure S3.1: GWAS analysis for 620 accessions and corresponding QQ plots	105
Figure S3.2: Correlations of gbM gene number with various genomic features	106
Figure S3.3: Gene location comparison between gbM and UM genes for nature accessions	107
Figure S3.4: Gene methylation status prediction model construction and algorithm selection...	108
Figure S3.5: gbM and teM-status are on a continuous spectrum.....	109
Figure 4.1: Ectopic CHG methylation of gbM genes increases over generations in <i>ibm1</i> and <i>ddm1</i> mutants	136
Figure 4.2: Hypomethylated chromosomes from the <i>ddm1</i> parent led to ectopic DNA methylation in gbM genes.....	137
Figure 4.3: Global hypomethylation leads to ectopic activity of CMT3 at genes present within wild-type and <i>ddm1</i> haplotypes	138

Figure 4.4: Methylation QTL for genes with ectopic mCHG are especially enriched in the pericentromeric regions	139
Figure 4.5: Mutations in <i>ibm1</i> lead to ectopic methylation preferentially in genes that previously possessed gbM	140
Figure 4.6: Loci that are susceptible to spontaneous epiallele formation have a greater contact frequency with H3K9me2 region.....	141
Figure S4.1: Ectopic CHH methylation on genes are mainly at CWA sites in <i>ibm1</i> and <i>ddm1</i> mutants	142
Figure S4.2: Gradual increase of non-CG methylation at gbM genes in <i>ibm1</i> and <i>ddm1</i> mutants over generations	143
Figure S4.3: DNA methylation patterns at gbM, UM and teM genes for <i>ddm1</i> epiRILs	144
Figure S4.4: Ectopic non-CG methylation was not influenced by intronic methylation of IBM1 in <i>ddm1</i> epiRILs.....	145
Figure S4.5: Except chromosome 2, no obvious large region in <i>ibm1</i> ; <i>met1</i> epiRIL-12 is derived from the original <i>met1</i> -3 parent	146
Figure S4.6: Ectopic mCHG-gain genes in <i>met1</i> mutant are enriched in gbM genes	147
Figure S4.7: Total size and intersections of ectopic mCHG-gain genes among four datasets in this study	148
Figure S4.8: Loci that are susceptible to spontaneous epiallele formation in <i>ddm1</i> mutant have a greater contact frequency with H3K9me2 region	149

CHAPTER 1

A REVIEW OF DNA METHYLATION AND EPIALLELES IN PLANTS

OVERVIEW OF DNA METHYLATION

DNA methylation is an important chromatin modification, and plays important roles in gene regulation, repeat sequence repression, and maintenance of chromatin structure [1-4]. In mammals, DNA methylation is restricted to the symmetric CG context [5]. There are three DNA methyltransferases: DNMT1, DNMT3A, and DNMT3B. DNA (cytosine-5)-methyltransferase (DNMT1) is recruited by the E3 ubiquitin-protein ligase (UHRF1) at hemi-methylated DNA throughout S phase of DNA replication and maintains DNA methylation during cell divisions, whereas DNMT3A and DNMT3B are responsible for establishing *de novo* DNA methylation [6-8]. In plants, DNA methylation occurs in the context of CG, CHG and CHH (H= A, C, or T)[9, 10]. METHYLTRANSFERASE 1 (MET1), the homolog of DNMT1, recognizes hemi-methylated CG dinucleotides and methylates the unmethylated cytosine in the new strand in a symmetrical way [11]. MET1 is recruited by VARIANT IN METHYLATION 1 (VIM1), which is a UHRF1 ortholog, during DNA replication [12]. However, in plants there is also abundant non-CG methylation. Maintenance of symmetrical CHG is catalyzed by the DNA methyltransferase CHROMOMETHYLASE 3 (CMT3)[13, 14]. Asymmetrical CHH can be methylated by both CMT2, and the DNA methyltransferase DOMAINS REARRANGED METHYLASE 2 (DRM2), which catalyzes *de novo* DNA methylation in a sequence-

independent manner and is regulated by the RNA directed DNA methylation pathway (RdDM) [15, 16].

DNA methylation is highly correlated with the histone modifications in regulating gene expression and suppressing transposon activity. Methylation of histone 3 at lysine 4 (H3K4) and H3K9 are among the most highly conserved histone modifications that correlate with gene activation and gene silencing, respectively [6, 7, 17]. In plants, non-CG methylation catalyzed by CMT2 and CMT3 is dependent on H3K9me2 modifications. The protein structure of the maize CMT3 demonstrated that its bromo-adjacent homology (BAH) and chromo domains bind to H3K9me2. The absence of this interaction leads to a complete loss of CMT3 activity [18]. Three histone methyltransferases are responsible for the majority of H3K9 methylation (SUPPRESSOR OF VARIATION 3-9 HOMOLOGUE PROTEIN 4 (SUVH4), SUVH5, SUVH6, also known as KYP) [19-22]. The SRA domain of KYP can directly bind to methylated CHG and carry out H3K9 methylation [23, 24]. The SUVH4, SUVH5 and SUVH6 triple mutant abolishes H3K9me2 and substantially reduce CHG methylation [25]. In return, elimination of all non-CG methylation occurs in mutants lacking *drm1*, *drm2*, *cmt2* and *cmt3* and also results in a total loss of all H3K9 methylation [14]. These observations support the existence of a feedback loop between histone H3K9 methylation and non-CG methylation.

On the other hand, methylation of H3K4, which is associated with active gene transcription, appears to be mutually exclusive with DNA methylation [26]. A JmjC protein (JMJ14) shows an example of the mutual exclusion. JMJ14 acts downstream from RdDM components to demethylate H3K4me3 and to promote DNA methylation [27]. Except for H3K9 and H3K4 methylation, there are a variety of histone modifications and histone variants that are important for controlling chromatin structure and function, but the function and crosstalk

between them remains largely unknown [28, 29]. Therefore, the computational integration of high throughput epigenomic profiling has been constructed for several model plants. These data have led to proposals that plant genomes can be divided into ~30 chromatin states such as promoter-associated, transcription-associated, active intergenic, large-scale repressed and repeat-associated states, to interpret the possible function for a variety of epigenomic environments [30, 31].

DNA methylation is found throughout the tree-of-life, yet the manner in which it is utilized by genomes and the mechanistic of action varies depending on the species. Both prokaryotes and eukaryotes can have DNA methylation in their genomes, but they have different preferences to the location where a methyl group occurs in the DNA sequence. 5-methylcytosine 5mC occurs at the C5-position of cytosine and is the dominant methylation type in eukaryotic genomes, whereas 6mA and 4mC constitute the majority of DNA base modifications in prokaryotes [2, 32-34]. 5-methylcytosine is most studied base modification and is found to be widespread in eukaryotes including plants, fungi, and animals [6]. However, there numerous major model systems that do not require DNA methylation such as *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Caenorhabditis elegans* and *Drosophila melanogaster* [35-39]. DNA methylation patterns vary across organisms. In mammalian genomes, methylation is ubiquitous, not specific to TEs and genic regions. Approximately 60–80% of CpGs are generally methylated, and less than 10% of the CpGs occur in CG-dense regions that are unmethylated, termed CpG islands, which are prevalent at transcription start sites of active genes [40]. In plants, methylation occurs in both gene and repeat regions, but the overall methylation level and distribution in CG and non-CG sites are highly variable, which was found to be correlated to genome size and repeat number [41]; In fungi, DNA methylation in *Neurospora crassa* and

Tuber melanosporum has been found exclusively in transposons whereas DNA methylation in insects is restricted to the transcribed regions of genes [42-44]. With increasing comparative epigenomic studies, it was found that samples within the same lineage, even within species, can have significant methylation variation [41, 44, 45]. It is proposed that, evolution of underlying methylation enzymatic machinery may directly link to the diverse patterns of DNA methylation found throughout eukaryotes [44, 46].

The inheritance of DNA methylation as a result of reproduction is also different between plants and mammals [47]. In mammals, genome-wide epigenetic reprogramming occurs at two distinct stages: in primordial germ cells and in the early embryo beginning in the zygote immediately after fertilization and extending to blastocyst stage [48-50]. This reprogramming entails erasure of DNA methylation, which is reset by DNMTs during embryo development to reduce the risk of perpetuating deleterious epigenetic alleles from parent to progeny [51]. However, although large-scale reprogramming occurs in non-germ line reproductive cell, plants are not shown to undergo genome-wide waves of demethylation in germ cells. DNA methylation is stably maintained to keep transposons and repetitive DNA elements under tight regulation in reproductive cells [47, 52]. The faithful propagation of DNA methylation over generations in plants makes them ideal models for studies of epigenetics, since traditional approaches of constructing genetic populations can be applied to study the inheritance of methylation variants and their effects on phenotypes [53].

Although, DNA methylation plays important roles in chromatin regulation, it comes at a cost. DNA methylation itself is an endogenous mutagen. Methylated cytosines can spontaneously deaminate compared to unmethylated cytosines and the repair efficiency for resulting T-G mismatches is relatively low [4, 54-57]. CpG to TpG transitions can dominate substitution

profiles between species and genomes where CpG methylation is common [58, 59]. For example, the frequency of CpG dinucleotides in human genomes is less than one-fifth of the expected frequency [60]. Recent analyses have linked elevated mutation rates directly to DNA methylation by integrating base-resolution methylation maps with sequence variation or SNP data [61-63]. In addition, genomic regions with different methylation levels can cause different mutation rates [63]. Analysis on DNA methylation data in humans, *Arabidopsis thaliana*, and *Oryza sativa* shows that DNA methylation states affect mutation rate not only at the focal cytosine but also at neighboring nucleotides [64]. Importantly, methylation induced mutation is related to cancer. CpG sites in the coding regions of tumor suppressor genes are strong mutation hotspots leading to cancer [65]. For example, the methylated CpG sites in the *p53* coding region contribute to 50% of all inactivating mutations in colon cancer and 25% in cancers in general [66].

Understanding of the function, distribution and regulatory mechanism of DNA methylation is fundamental for disease treatment and crop improvement, which relies on technological advances in detection and in the analysis of DNA methylomes. Methodologies to explore the patterns of DNA methylation in genomes have evolved rapidly. Previously, DNA methylation within genome were detected by methylation-sensitive restriction digestion methods and later immunoprecipitation-based assays using 5-methylcytosine (5mC) specific antibodies coupled with microarrays or high-throughput sequencing (MeDIP-chip or MeDIP-seq), which resulted in relatively low-resolution DNA methylation maps [67, 68]. With the emergence of sodium bisulfite conversion combined with high-throughput sequencing (WGBS), whole genome-wide DNA methylation can be determined at single-base resolution [9, 69]. The principle of this technique includes: first, treatment of DNA with sodium bisulfite converts

unmethylated cytosine (C) to uracil (U); thereafter, PCR amplification converts U to thymine (T), while the methylated cytosine is protected and thus not converted. After PCR and high-resolution sequencing, raw reads of the converted DNA are then aligned to a reference assembly to identify the methylation status of every cytosine in the genome at a single base resolution [9, 69]. WGBS has led to a number of comparative epigenomic studies, which has expanded our knowledge of DNA methylation patterns in different species, populations and cell types [5, 41, 70, 71]. The technology has even advanced to discover single base resolution DNA methylation at single cell resolution [72-74]. With the development of high-throughput sequencing, the size and scope of ‘omics’ data sets including DNA methylomes, have increased rapidly. As a bioinformatics researcher, how to combine varied ‘omics’ dataset together to unravel the intricate relationship between genotype, epigenotype and phenotype is a challenging but essential task.

EPIMUTATIONS IN PLANTS

DNA methylation has been widely characterized both within and between plant species and found to be highly variable in its genome-wide levels and distribution [41, 71, 75]. These variants are referred to as epigenetic alleles (epialleles) and they can be transmitted across generations [53, 76]. In plant genomes, DNA methylation variations mainly exist in two groups. One group includes transposons, methylation variance is dependent on variation of repeat sequence or transposon insertions. This group is strongly targeted by the RdDM pathway. A second group contains hypervariable genic DNA methylation which accumulates with cell division time (and generations) and is more variable among accessions of the same species [71, 77].

In natural populations, it is challenging to study the epigenetic nature of DNA methylation variation, as there is abundant genetic variation that can induce it. Models have been proposed to conceptualize the dependence of epialleles on genetic variation, which has led to the classification of three distinct groups [78]. The first class is called ‘obligate epialleles’, which indicates a complete dependency on genetic variant; in other words, the epigenotype is an obligatory phenotype of the genotypes. The second class is called ‘facilitated epialleles’ and they can occur, in a probabilistic manner, only in the presence of a genetic variant such as a nearby transposon insertion. The third class is called pure epialleles (spontaneous epimutation), which are independent from the genotype. It has been proposed that spontaneous epimutation results from inefficiencies in transferring methylation information across cell divisions/generations [79], because the sites of epiallelic variation overlaps significantly with loci that are also affected in mutants in DNA methyltransferases or DNA demethylases [9, 80, 81]. Errors are generated from inefficiency, deficiency, or perturbation of the methylation maintenance system and as a result change arising from these stochastic events can generate epialleles at some finite frequency regardless of the genotype.

The majority of epialleles identified in *Arabidopsis thaliana* are either obligate or facilitated epialleles, as their formation has been induced using mutations in machinery required for maintenance of DNA methylation or they were recovered from mutagenesis experiments [82-87]. The identification of naturally occurring meiotically pure epialleles has remained elusive [79, 88]. With the accumulation of DNA methylome datasets, more natural epigenetic variants that are determined by genetic variation have been captured by GWAS and linkage mapping approaches, which combine sequence information with methylome data [12, 71, 75].

Although rare, there are significant examples of spontaneous epialleles that cause heritable phenotypic variation, such as peloric (*Linaria vulgaris*) [89], color less non-ripening (*Solanum lycopersicum*) [90] and B' (*Zea mays*) [91] epialleles. However, there is still a lack of knowledge about the impact of spontaneous epialleles on phenotypic variation in plants [92]. Therefore, there is a lot of focus on the variable methylation that occurs in genes and on identifying the epigenetic basis of plant traits. Comparative studies between *Arabidopsis thaliana* accessions found that DNA methylation variation immediately upstream of genes was strongly correlated with transcriptional variation, whereas variation of gene body methylation had no significant impact on gene expression levels [71, 93]. Studies on *Arabidopsis thaliana* and rice mutation accumulation (MA) lines also show the association between a small subset of differentially methylated regions (DMRs) and transcriptional changes of nearby genes [88, 94]. It has been hypothesized that epimutations lead to functional divergence by fine tuning gene expression. However, attempts to link such epimutation-induced gene expression changes with agricultural and evolutionary relevant plant traits remains challenging [92].

Changes in cytosine methylation are inherited over generations in plants [53]. Epimutations that contribute to phenotypes, might undergo pressure from natural selection and then spread or become extinct from a population at certain rate. This has led to the studies whereby principles of evolutionary/population genetics are applied to epimutations. The main goals of epimutation spectrum research includes studying origins, distributions, and inheritance of epimutations, and exploring how epimutations interact with phenotypic changes.

The rate of spontaneous epimutation has been well documented by tracing changes in DNA methylation using *Arabidopsis thaliana* MA lines that were generated from one parent by single seed descent growth for 30 generations [95]. These studies estimated that the rate for

single site spontaneous methylation variation was at least four orders of magnitude greater than genetic mutations. Furthermore, some captured spontaneous mutations resulted in significant transcriptional variation of the affected locus [79, 88]. However, the majority of the identified epimutations are neutral, under no or low selection pressure. Although they undergo a higher epimutation rate than nucleotide substitutions, once formed they are stably inherited; that is, they accumulate over generational time. Estimates of the stability show that once new DNA methylation changes form, they are inherited over multiple generations with 99.998% accuracy [96]. Due to the stability and accumulation of this heritable variation, DNA methylation and its variants can be used as a molecular clock to infer evolutionary history of populations over short time scales (up to a few thousand years) [97].

In addition to genetic variation, environmental factors also contribute to epimutations. Multi-generational environmental stress is known to elicit extensive methylation changes. MA lineages of *Arabidopsis thaliana* under salt-stress show a 45% increase in stochastic CG methylation changes after 10 generations compared with non-stressed control lines [98].

EPIGENETIC MUTANT LINES

Although extensive methylation variation was identified between natural accessions and between species, it is challenging to unravel the impact of epimutation on phenotypic diversity due to the presence of so much segregating genetic variation. One strategy to better understand the impact of epialleles, while minimizing genetic variation, is to use mutants in machinery required for the maintenance of DNA methylation. This approach allows the tracing of the distribution and changes to DNA methylation upon induction of mutation at the target locus and helps in our understanding of the formation and recovery of epimutation. This provides insight

into what happens if spontaneous mutations occur in methylation related enzymes in nature [76, 92].

There are a variety of mutants that have been essential to our understanding of the potential for disruption of DNA methylation to impact phenotype. These include *met1*, *ddm1* and *ibm1*. Loss of *met1* leads to an erasure of CG methylation and indirect losses of non-CG methylation in certain regions of the genome [11]. It also results in the suppression of DNA demethylation activities, altered RNA directed de novo DNA methylation, and the redistribution of other repressive histone modifications such as histone H3 demethylation in Lys 9 and trimethylation in Lys 27, creating further epigenetic variation [99, 100]. *met1* mutants display a few phenotypic and developmental abnormalities including late flowering as a result of demethylation on the promotor of flowering repressor FWA (FLOWERING WAGENINGEN) [1, 11, 101].

DECREASE IN DNA METHYLATION 1 (DDM1), which encodes a SWI–SNF2-like protein that is related to chromatin remodelers in mice and humans [102]. *ddm1* mutants result in a genome-wide depletion of DNA methylation at all cytosines independent of their sequence context. It's strongest affects are found in repeat regions of the genome [103, 104], which leads to strong transcriptional activation of repeats and TEs [105]. *ddm1* plants exhibit only mild phenotypic alterations, except after repeated self-crossing, in which case the severity and the number of aberrant phenotypes increases [106]. Although *ddm1* mutant plants exhibit a ~70% reduction of DNA methylation overall, some loci, such as *BONSAI*, can be hypermethylated in *ddm1* mutants after several generations, leading to developmental abnormalities [86]. This reveals that there are feedback mechanisms that are important to local control of DNA methylation based on global levels of DNA methylation.

INCREASE IN BONSAI METHYLATION 1 (IBM1) encodes a protein with histone H3mK9 demethylation activity. It preferentially demethylates H3mK9 at transcribed regions of the genome to protect them from silencing by nearby heterochromatin. *Ibm1* mutants possess ectopic accumulation of H3K9me and CHG methylation on thousands of genes which lead to multiple developmental phenotypes, including flower and seed development [107-109].

Genetic analysis has been useful to further define the mechanistic relationship between the genes identified that are important to maintenance of DNA methylation in *Arabidopsis thaliana*. For example, in *ibm1; cmt3* or *ibm1; kyp* double mutants, ectopic genic DNA methylation and aberrant phenotypes in *ibm1* mutants were suppressed by the *cmt3* or *kyp* mutation [107, 109], suggesting the two enzymes take opposite role in the establishment of DNA methylation on genes. An extensive study of 86 silencing mutants involved in DNA methylation has helped reveal both the specificity of each type of methylation as well as the degree of interaction amongst pathways [25]. Collectively, mutant analysis has contributed in a major way to our understanding of how DNA methylation is established and maintained in plant genomes.

The relatively stable transmission of epigenetic variation in *ddm1* and *met1* mutants provided the basis for creation of epigenetic recombinant inbred lines (epiRILs) [110, 111]. As RIL populations are used for mapping causal genetic variants for traits that are variable between two parental line, epiRIL populations can build the association between epigenetic variants and phenotypes with reduced genetic variability [76]. My thesis research took advantage of two independent epiRIL populations (*met1* epiRILs [111] and *ddm1* epiRILs [110]). Both epiRILs were created using mutants crossed with wild-type Col-0 plants to generate F1 hybrids. Then only F2 plants with homozygous *DDM1* (or backcross hybrid to wild type) or *MET1* were propagated by self-crossing and single-seed descent. Because of segregation and recombination,

each epiRIL line within population has a unique combination of genome fragments from either mutant or wild-type parental line, leading to epigenetic variation among samples without changes to the DNA sequence.

met1 epiRILs show extensive phenotypic variation within and between lines [111]. Some traits appear are stably inherited, including flowering time, plant growth, and stress responses. For example, epiRIL that had a *met1*-like *fwa* epialleles show stable inheritance of a late-flowering phenotype. Microarray-based and bisulfite sequencing approaches show that many of the epialleles reverted to wild-type status over generations whereas many genic regions that lost CG methylation remained unmethylated over time [25, 111]. In my research, we take advantage of the persistent hypomethylated region in *met1* epiRILs, and further cross a *met1* epiRIL line to an *ibm1* mutant to observe how spontaneous epimutation is distributed on those CG hypomethylation regions.

ddm1 epiRILs also show heritable variation for phenotypes such as flowering time and plant height, which is due to the stable inheritance and segregation of parental epialleles and/or DNA insertion variants [110]. *ddm1* mutants lead to a significant decrease of DNA methylation of all cytosine sequence contexts, but the majority of them efficiently regain DNA methylation within two to five generations following restoration of DDM1 function. This is likely due to recruitment of the RNA-directed DNA methylation machinery [112]. However, some hypomethylated epialleles are stably inherited over generations [110]. Those stably inherited epialleles can be used as markers to build a linkage map and to find QTL for interested phenotypic traits [113, 114], which was also strategy we used in my research.

TRADEOFF BETWEEN CHROMATIN MAINTENANCE AND GENIC EPIMUTATION

In flowering plants, although the majority of cytosine methylation in all contexts occurs at TE and repeat sequences within heterochromatin, many coding genes also contain DNA methylation [9, 41, 69]. In *Arabidopsis thaliana*, a small subset of genes (~ 4%) have TE-like methylation patterns, which appears to be associated with transcriptional repression. Around 17% of genes contain exclusively mCG in the transcribed region, which is referred to as gbM [115, 116]. Unlike DNA methylation at TEs, gbM shows no convincing evidence thus far that it is associated with gene repression [41, 71, 116]. Alternatively, genes with gbM tend to be housekeeping genes with moderate to high levels of gene expression [117, 118]. Maintenance of gbM requires MET1, as in a *met1* mutant, gbM is erased from genes [9, 69]. However, complementation of *met1* mutants with wild-type *MET1* fails to restore gbM at least during generations of crossing lines [111, 116], indicating the establishment of gbM may involve methylation activity of other pathways and longer generation times.

There is significant feedback between DNA methylation variation between local and global regions of the genome. In *ibm1* and *ddm1* mutants thousands of transcribed genes ectopically gain mCHG and H3K9me2 [108, 119], which are normally limited to heterochromatin where CMT3 and KYP form a self-reinforcing loop for efficient maintenance of mCHG and H3K9me2 [23]. However ectopic non-CG methylation observed on transcribed genes suggests that the self-reinforcing loop can occur at genes when the normal balance demarcating heterochromatin and euchromatin is disturbed. Therefore, it leads to a model whereby the self-reinforcing machinery is likely involved in the establishment of gbM, as gbM-containing genes are preferred targets for hypermethylation in the *ibm1* mutant [108].

This hypothesis is further supported by results from an epigenomics study of natural variation in plants. The study found that *E. salsugineum*, a species in the same Brassicaceae family with *Arabidopsis thaliana*, has lost gbM [41, 116], but still possessed high levels of mCG in repetitive sequence of *E. salsugineum* suggesting that the loss of gbM was not caused by MET1. Instead, a natural deletion of *CMT3* and a low mCHG level were found in *E. salsugineum* [116].

These studies provided the foundation for my thesis, which led to the hypothesis that *CMT3* is involved in the establishment of gbM and that there exists trade-off between maintenance of DNA methylation at heterochromatin and the origins of spontaneous epimutation. The underlying logic is that gbM evolved from spontaneous epimutations triggered by perturbation or constraints of pathways that maintain heterochromatin. In other words, epimutations are an inevitable cost of the heterochromatin maintenance process.

Heterochromatin maintenance is important for transcriptional silencing of repeats and genome stability. Deficiency in maintenance of heterochromatin leads to numerous abnormalities. In humans, global loss of DNA methylation is accompanied by ectopic methylation accumulation in CpG islands that is identified in many cancer types [120]. Similarly, global decrease of DNA methylation in *ddm1* also leads to ectopic genic methylation that results in developmental defects [106, 119]. This evidence suggests feedback regulations exist in epigenome maintenance, which is sensitive to the changes of genome-wide levels of DNA methylation, resulting in its redistribution.

The *CMT3* model in establishment of gbM might also reflect some level of feedback regulation in coordination of epigenomic states between chromosomes. Moreover, in this model, although starting as epimutations from self-reinforcing machinery, gbM is not eliminated from

natural selection and is instead maintained passively as exclusive mCG has reduced deleterious consequences on gene silencing [116, 121, 122]. Therefore, mechanistic and epigenomic comparative studies of how this process is established and how it evolves can help us to understand the formation, inheritance, and implications of epimutations from an evolutionary perspective.

DISSERTATION OUTLINE

My dissertation presents three interconnected studies on this process. First, I identified the mechanistic basis for gbM establishment by expressing the *Arabidopsis thaliana* *CMT3* gene in *E. salsugineum*, which naturally lacks *CMT3*. I traced the distribution and inheritance pattern of DNA methylation over six generations of transgenic line growth. The experiment shows the accumulation and stable inheritance of CG methylation following *de novo* CHG methylation induced from expressed *CMT3* (Chapter 2). Next, I conducted a comparative analysis of DNA methylomes among 725 *Arabidopsis thaliana* natural accessions, providing genetic and epigenetic evidence to support the association between maintenance of heterochromatin DNA methylation and gbM. Importantly, I also explored possible phenotypic and evolutionary implications of gbM (Chapter 3). Finally, I analyzed methylomes of a variety of mutant lines and a *ddm1* epiRIL population, providing mechanistic investigation into feedback regulation of pathways that are involved in heterochromatin maintenance and show they are associated with spontaneous epiallele formation (Chapter 4).

CHAPTER 2

EPIMUTATIONS ARE ASSOCIATED WITH CHROMOMETHYLASE 3-INDUCED DE
NOVO DNA METHYLATION ¹

¹ Yinwen Zhang, Jered M. Wendte, Lexiang Ji, Xiuling Shi, Rashmi R. Hazarika, Yadollah Shahrari, Frank Johannes, Robert J. Schmitz. *Elife*, 2019, 8: e47891

Reprinted here with permission of the publisher.

Yinwen Zhang and Jered M Wendte are equal contributors

ABSTRACT

In many plant species, a subset of transcribed genes are characterized by strictly CG-context DNA methylation, referred to as gene body methylation (gbM). The mechanisms that establish gbM are unclear, yet flowering plant species naturally without gbM lack the DNA methyltransferase, CMT3, which maintains CHG (H=A, C, or T) and not CG methylation at constitutive heterochromatin. Here, we identify the mechanistic basis for gbM establishment by expressing CMT3 in a species naturally lacking CMT3. CMT3 expression reconstituted gbM through a progression of de novo CHG methylation on expressed genes, followed by the accumulation of CG methylation that could be inherited even following loss of the CMT3 transgene. Thus, gbM likely originates from the simultaneous targeting of loci by pathways that promote euchromatin and heterochromatin, which primes genes for the formation of stably inherited epimutations in the form of CG DNA methylation.

INTRODUCTION

Heritable gains or losses of DNA methylation, or epimutations, can have important phenotypic consequences. Examples include peloric mutants of toadflax, where differential methylation of a single transcription factor can change flowers from bilaterally to radially symmetrical and the colorless non-ripening locus of tomato, where differential methylation affects fruit ripening [89, 90]. Despite the potential implications for phenotypic change, little is known about how epimutations form.

One form of genic DNA methylation that may provide clues to the mechanisms of epimutation is gene body methylation (gbM), which is found on a subset of expressed genes in many eukaryotic genomes, including most flowering plants [9, 10, 41, 46, 69, 115, 117, 121,

123-130]. GbM has been functionally implicated in certain processes, including transcriptional regulation, transcript processing, and suppression of intragenic transcripts [117, 121, 126, 131-135]. However, natural and experimental losses of gbM have also been documented with no obvious effects on expression or chromatin structure [10, 41, 116, 136]. One hypothesis that may reconcile these seemingly conflicting results explains the formation of gbM as a passive process resulting from transient localization of the proteins that maintain heterochromatin to genic space (i.e. euchromatin) [116, 121, 122, 137, 138]. Here, gbM is proposed to arise passively, but, once established, is maintained due to the preferential recruitment of maintenance methyltransferases to previously methylated sites. Under this line of reasoning, gbM may have functional consequences in some cases, but can also be non-functional, which could explain experimental findings. A better understanding of the mechanistic basis for the establishment of gbM will provide important insights into its possible functions, the mechanisms underlying epimutations, as well as to factors that might influence DNA methyltransferase mistargeting in disease states [122].

In plants, most knowledge of DNA methylation is derived from studies of *Arabidopsis thaliana*, where cytosines in different sequence contexts are the preferred substrates of distinct pathways. The RNA directed DNA methylation (RdDM) pathway acts to de novo methylate cytosines in all sequence contexts [139]. In RdDM, DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) is targeted to chromatin via non-coding RNAs produced by two plant-specific multi-subunit RNA polymerases [140-142]. Following de novo establishment, methylation of symmetrical CG cytosines is maintained by METHYLTRANSFERASE 1 (MET1), which is recruited to hemi-methylated CG sites to methylate the complementary, unmethylated CG sites [1, 12, 101, 143]. Symmetrical CHG (H =

A, T, or C) cytosine methylation is maintained by CHROMOMETHYLASE 3 (CMT3) [13, 144, 145]. Among CHG sites, CMT3 shows a marked preference for CWG (W = A or T) cytosines relative to CCG cytosines [146, 147]. An additional CHROMOMETHYLASE, CMT2, targets the methylation of CHH cytosines [14, 104]. In *A. thaliana*, CHH sites targeted by CMT2 can be distinguished from those targeted by RdDM, as they show an enrichment in CWA methylation relative to other CHH contexts, in contrast to RdDM target regions which show no preferred site enrichment [146].

CMT2 and CMT3 participate in a self-reinforcing feedback loop with an additional heterochromatin modification, histone H3 lysine 9 di-methylation (H3K9me₂) [14, 17, 18, 147]. CMT2 and CMT3 are both thought to depend on direct physical binding to H3K9me₂ for targeting to chromatin and methylation [14, 18, 147]. DNA methylation can be physically bound by the H3K9 methyltransferases, which reinforces the co-localization of H3K9me₂ and CMT-dependent DNA methylation [17, 23, 24, 148, 149]. This co-dependency results in losses of CMT-dependent DNA methylation in H3K9 methyltransferase mutants, as well as losses of H3K9me₂ in *cmt* mutants [14, 17, 19, 25, 150-153].

GbM is restricted to CG context cytosines [9, 41, 69, 115, 121, 125], and therefore dependent on MET1 [9, 69]. However, complementation of *met1* mutants with wild-type MET1 fails to restore gbM, presumably because hemi-methylated CG cytosines required for MET1 recruitment have been lost [111, 116]. Thus, the mechanisms required for the establishment of gbM are unclear.

Recent comparative analyses have identified Angiosperm plant species lacking gbM [41, 116]. Concurrent with the loss of gbM is the loss of the gene encoding CMT3, which has led to the hypothesis that CMT3 is required for the initial establishment of gbM [46, 116, 121]. The

lack of immediate restoration of gbM in MET1 complemented met1 mutants has led to the proposition that CMT3 is only rarely localized to gene bodies such that gbM only accumulates slowly over evolutionary time scales [116, 121, 122, 137]. Direct experimental testing of this model is difficult, and it is also currently unclear whether CMT3 can be active at regions with no pre-existing DNA or H3K9 methylation. Furthermore, the favored activity of ZMET2, the maize ortholog of CMT3, is a maintenance methyltransferase of hemi-methylated cytosines in the CHG sequence context [147], and, thus, a de novo activity that results in methylation of CG cytosines in vivo is unprecedented.

To gain insights into this process and provide a direct test of a role for CMT3 in gbM, we heterologously expressed *A. thaliana* CMT3 (AtCMT3) in a plant species that has lost both CMT3 and gbM, *Eutrema salsugineum* [116]. Expression of AtCMT3 in *E. salsugineum* resulted in gains of CHG methylation over repetitive sequences characterized by the presence of H3K9me2, as predicted based on the known mechanism of CMT3 targeting. However, AtCMT3 expressing lines also exhibited ectopic CHG methylation over a subset of genes in an AtCMT3-expression dependent manner. Genes that gained CHG methylation were orthologs of *A. thaliana* gbM genes and had no prior CHG methylation or detectable H3K9 methylation, suggesting de novo methylation activity of CMT3 in vivo. Unexpectedly, gains of CHG methylation did not result in stable accumulation of H3K9 methylation over gene bodies or expression changes, showing that the genic CHG methylation was uncoupled from transcriptional silencing, similar to gbM, likely due to transcription-coupled de-methylases. Gains in CHG methylation were also associated with gains in CG and CHH methylation, and removal of the transgene via crossing to non-transgenic parents, or progressive AtCMT3 transgene silencing over six generations of propagation, revealed that ectopic genic CG methylation was preferentially maintained relative

to genic CHG or CHH methylation following the loss of AtCMT3 expression. The results provide new insights to the mechanism of CMT3-initiation of gbM by demonstrating that CMT3 promotes the establishment of genic CG epimutations which can be maintained even in the absence of CMT3.

RESULTS

Expression of AtCMT3 in *E. salsugineum* results in increased CHG methylation

To gain insights into the mechanisms of CMT3-targeting of DNA methylation, full-length genomic CMT3 from *A. thaliana* (AtCMT3) was expressed in *E. salsugineum* under the native *A. thaliana* promoter. In total, we assessed plants derived from six transformation events. Two of these lineages, referred to as AtCMT3-L1 and AtCMT3-L2, were propagated by single seed descent for six generations following transformation and serve as the main focus of this study (Figure 2.1A; see Figure S1 and Table S1 for a complete description of all plant lines and associated data described in this study). Whole genome bisulfite sequencing was completed on individual plants for each generation (numbered T1-T6) of each line to assess the impact of AtCMT3 expression on DNA methylation. Plants expressing AtCMT3 showed increased levels of CHG methylation in intergenic regions, associated with preexisting methylation, as well as over some gene bodies with no prior methylation (Figure 2.1B and S2.2).

Differentially methylated regions (DMRs) were identified in each sequence context (CG, CHG, and CHH), comparing each generation of AtCMT3-L1 and AtCMT3-L2, and wild type (Table S2 and S3). Consistent with AtCMT3 activity, the majority of DMRs identified were CHG DMRs, with 25,096 identified that were characterized by a median size of 407 base pairs (bp) (Table S2). A smaller number of CG (3,502 DMRs, median size: 189 bp) and CHH (1,763

DMRs, median size: 243 bp) were also identified (Table S2). The majority of CHG DMRs (>99%, in each lineage) were hypermethylated CHG DMRs in AtCMT3 expressing lineages relative to wild type (Table S3). Many of the CG and CHH DMRs (between 45-72% CG and 86-98% CHH, depending on the line) were also hypermethylated DMRs in AtCMT3 expressing lines relative to wild type, and overlapped with CHG DMRs, suggesting they result from cross-talk between CMT3-mediated CHG methylation and other pathways (Figure S2.3, Table S3). There was also a significant overlap of hypermethylated CHG DMRs between all individuals, suggesting that AtCMT3 was not methylating DNA randomly (Figure S2.4).

The levels of CHG methylation varied between individuals expressing AtCMT3 (Figure 2.1B, Table S2 and S3). Plants were transformed using the floral dip method [154], which results in random and potentially multiple transgene insertions that can segregate out over generations. Therefore, we predicted that AtCMT3 transgene expression likely varied between individuals, which could explain differences in CHG methylation levels. To test this, we performed RNA-sequencing in the T3 - T5 generations of AtCMT3-L1, the T3 - T6 generations of AtCMT3-L2, as well as two, T2 generation plants from an additional, independently transformed lineage (AtCMT3-L3) (Figure S2.1). The results demonstrated a significant correlation between the levels of AtCMT3 expression and genome-wide CHG methylation levels ($R^2 = 0.8828$, $p = 1.669 \times 10^{-4}$, Figure S2.5, see Table S4 for FPKM values). This was especially notable for two T2 generation plants of AtCMT3-L3, which had the highest AtCMT3 expression levels (233.7 and 372.5 FPKM for AtCMT3-L3T2 and T2b, respectively) and the highest genome-wide percent CHG methylation (27% and 32%, for T2a and T2b, respectively). These results also revealed that expression of the AtCMT3 transgene was progressively lost in the AtCMT3-L2 lineage following the T4 generation (Table S4). We also examined the relationship between genome-

wide CHG methylation and AtCMT3 expression using qRT-PCR, including plants from additional lineages, and found a weaker although significant relationship ($R^2 = 0.5932$, $p = 0.025$, Figure S2.1 and S2.5). Taken together, AtCMT3 expression is likely one factor contributing to genome-wide CHG levels in these lines.

Despite lacking CMT3, wild-type *E. salsugineum* does exhibit residual levels of CHG methylation, likely deposited by other DNA methylation pathways such as RdDM [116] (Figure 2.1B and 2.2A). As the maize ortholog of CMT3 has been shown to preferentially act on hemimethylated CHG cytosines [147], we predicted that AtCMT3 would be preferentially targeted to regions with pre-existing CHG methylation. To test this, CHG DMRs were ranked based on wild-type CHG methylation levels. Results showed that the majority of CHG DMRs (18,812 / 25,096 or 75%) were characterized by the presence of low levels of CHG methylation in wild type that increased upon the expression of AtCMT3 (Figure 2.2A). Characterizing the DMRs based on overlap with genomic features revealed that 19,576 out of 25,096 (78%) total CHG DMRs overlapped with repetitive elements or intergenic regions, which are expected targets of all DNA methylation pathways (Figure 2.2A, Table S2). Summarizing CHG methylation over all repetitive elements confirmed that plants expressing AtCMT3 exhibited increased CHG methylation over these regions from ~15% methylation in wild type to ~30-70% methylation in AtCMT3 lines (Figure 2.2B).

Expression of AtCMT3 in *E. salsugineum* results in increased CHG methylation in a subset of gene bodies

In addition to methylated regions, a subset of regions (6,284 out of 25,096 (25%) CHG DMRs) that gained CHG methylation in AtCMT3-expressing lines had no pre-existing CHG

methylation, and the majority (4,725 / 6,284 or 75%) of these regions overlapped annotated genes (Figures 2.1B, 2.2A, and S2.2, Table S2). Examination of percent CHG methylation over all annotated genes revealed increases in CHG methylation over gene bodies, ranging from ~2-10% CHG methylation on average, compared to no CHG methylation in wild-type plants (Figure 2.2C). The patterns of CHG methylation in several of the generations of the AtCMT3-L2 lineage, which exhibited higher levels of AtCMT3 transgene expression (Table S4), were reminiscent of that seen for gbM [121], in that the percent methylation levels were highest towards the center of the gene bodies, decreasing towards the transcription start sites (TSS) and transcription termination sites (TTS) (Figure 2.2C).

To further characterize the genes that gained CHG methylation, genes that had less than 1% total DNA methylation in any context in wild type and that gained a minimum of 5% CHG methylation were identified in each line (CHG-gain genes) (Figure 2.2D, Table S5). The number of CHG-gain genes varied between individuals and lineages. In AtCMT3-L1, the number of CHG-gain genes generally increased over generational time from 99 genes in T1 to 583 genes in T6 (Figure 2.2D, Table S5). AtCMT3-L2 individuals were characterized by higher numbers of CHG-gain genes relative to AtCMT3-L1, with the T1 generation showing gains in 718 genes, 2,996 genes in T2, 3,553 genes in T3, and 4,055 genes in T4 (Figure 2.2D, Table S5). Following the T4 generation, the T5 and T6 generations of AtCMT3-L2 showed a decline in the number of CHG-gain genes, with 1,937 in T5 and only 88 genes in T6 (Figure 2.2D, Table S5). Despite the variation in the number of CHG-gain genes, the overlap of CHG-gain genes between all individuals was significantly higher than expected by chance (Figure S2.6). As CHG-gain genes had no prior CHG methylation, these results demonstrate a *de novo* methyltransferase activity of CMT3 *in vivo*.

We hypothesized that the variation in the number of CHG-gain genes between individuals could be related to variation in the expression levels of the AtCMT3 transgene, similar to genome-wide CHG methylation levels. Indeed, based on RNA-seq assessment of AtCMT3 expression, the levels of AtCMT3 expression and number CHG-gain genes were correlated ($R^2 = 0.7951$, $p = 0.001$) (Figure S2.7, Table S4). Again, two T2 generation individuals of the AtCMT3-L3 lineage, which had the highest AtCMT3 expression (Table S4), also had the highest number of CHG-gain genes (5,566 and 6,346 CHG-gain genes for AtCMT3-L3T2 and T2b, respectively) (Table S5). This result was also confirmed utilizing qRT-PCR ($R^2 = 0.9503$, $p = 3.91 \times 10^{-5}$) (Figure S2.7). Therefore, the decline in the number of CHG-gain genes in the AtCMT3-L2 T5 and T6 generations is correlated with the progressive loss of AtCMT3 transgene expression.

CHG methylation in gene bodies is not associated with stable H3K9 methylation

CMT3 can directly bind to H3K9me2 [18]. To determine if increases of CHG methylation detected in plants expressing AtCMT3 were correlated with H3K9me2, we conducted H3K9me2 chromatin immunoprecipitation and sequencing (ChIP-seq) in wild-type *E. salsugineum*. Consistent with CMT3 binding of H3K9me2, regions enriched for H3K9me2 in wild type showed increased CHG methylation in all lines expressing AtCMT3 (Figure 2.3A-B and S8A).

We next sought to determine how gains in CHG methylation affected the distribution of H3K9me2 by conducting H3K9me2 ChIP-seq in the T3 and T5 generation plants for AtCMT3-L1 and AtCMT3-L2. We found plants expressing AtCMT3 also showed enrichment for H3K9me2 across H3K9me2 ChIP peaks identified in wild type, but there were no gains in

H3K9me2 over these regions, despite the increase in CHG methylation (Figures 2.3A, 2.3C-F, and S2.8A). In contrast to repeat regions, gains in CHG methylation over gene bodies in AtCMT3-expressing lines were not associated with pre-existing H3K9me2 in wild-type plants (Figures 2.3A, 2.3C-F, and S8A). Also unexpected, the establishment of CHG methylation following AtCMT3 expression did not result in detectable H3K9me2 across CHG-gain genes (Figures 2.3A, 2.3C-F, and S2.8A).

CMT3 can also directly bind to H3K9me1, which distinguishes it from CMT2 [14, 18]. Therefore, we also conducted H3K9me1 ChIP-seq in transgenic and non-transgenic lines. Similar to findings in *A. thaliana* [20], H3K9me1 was enriched over regions characterized by H3K9me2 (Figure S2.9A-B). H3K9me1 was preferentially enriched in heterochromatin relative to CHG-gain genes before or after introduction of the transgene and the H3K9me1 signal over CHG-gain genes was indistinguishable from genes that remained unmethylated (Figure S2.9B-C). Overall, we conclude that gains of CHG methylation at genic loci resulting from expression of AtCMT3 are not associated with stable H3K9 methylation.

CHG methylation in gene bodies is not associated with transcriptional silencing

CMT3-mediated CHG methylation in heterochromatin is associated with transcriptional silencing [13, 14, 144]. To determine if gains in CHG methylation over gene bodies resulted in transcriptional changes, we compared expression of CHG-gain genes between wild-type and transgenic plants. In each line there was only a small proportion of CHG-gain genes (less than or equal to 10%) that showed a greater than 2 log₂-fold change and a similar number were both down and up-regulated (Figures 2.4A and S2.10, Table S5). There was also no relationship between the levels of CHG methylation gain and expression changes (Figures 2.4A and S2.10).

Additionally, we defined up- and down-regulated genes genome wide in AtCMT3-expressing lines (defined as greater than a 2 log₂-fold change relative to wild type), and CHG gain genes were not significantly enriched in either up- or down-regulated genes ($p > 0.05$, Fishers exact test) (Tables S6 and S7). We also considered that the changes in gene expression may be indirect effects of AtCMT3 expression and assessed up- and down-regulated genes for significant enrichments in Gene Ontology (GO) biologic processes. We found that many of the enriched terms for up- and down-regulated genes were involved in abiotic stress responses (Table S8). However, there were no consistently identified GO-term enrichments across lineages suggesting that AtCMT3 expression was unlikely the direct or indirect cause of these changes (Table S8). Lack of transcriptional changes directly related to CHG methylation is consistent with the lack of stable H3K9me₂ at these loci (Figure 2.3C-F). We conclude that genic CHG methylation in AtCMT3-expressing lines is uncoupled from heterochromatin formation and transcriptional silencing, similar to gbM.

Genes that gain CHG methylation in AtCMT3-expressing lines are orthologs of *A. thaliana* gbM genes and possess similar characteristics

GbM is present on conserved orthologous genes across diverse plant species [41, 46, 123-125]. To determine if genes that gain CHG methylation in AtCMT3-expressing lineages are genes that would be predicted to have gbM based on orthology, orthologs of CHG-gain genes in *E. salsugineum* were identified in *A. thaliana*. Among the generations of AtCMT3-L1 and L2, there was a total of 4,769 CHG-gain genes identified in at least one individual (Table S5). Among these genes, 4,104 have an orthologous gene encoded in the *A. thaliana* genome, and, out

of those, 1,526 were classified as gbM in the *A. thaliana* Col-0 accession, which is significantly more than expected by chance ($p = 2.83 \times 10^{-222}$, hypergeometric test) (Table S9).

Relative to unmethylated genes, gbM genes are generally characterized as being constitutively expressed at moderate levels, they tend to be longer and have more exons, and have a higher frequency of CMT3-preferred CHG sites [9, 10, 41, 46, 116, 123, 125, 155]. To determine whether the CHG-gain genes had similar characteristics, we compared the 4,769 CHG-gain genes identified to the remaining unmethylated *E. salsugineum* genes (UM genes). On average, CHG-gain genes were longer, contained more exons, exhibited a more moderate, but on average higher, range of expression, and had a higher frequency of CHG cytosines relative to gene length compared to unmethylated genes (Figure 2.4B-E). Therefore, AtCMT3 in *E. salsugineum* methylates genes that are orthologs and/or characterized by similar properties of gbM loci in *A. thaliana*.

Gains in CHG methylation over gene bodies are associated with gains in non-CHG methylation

GbM is defined as strictly CG context methylation [121], which contrasts with the predominantly CHG methylation observed over gene bodies in the AtCMT3-Lines (Figures 2.1B, 2.3A, S2.2, and S2.8A). To determine if AtCMT3 expression resulted in non-CHG methylation over gene bodies as well, we focused on the AtCMT3-L2 lineage, which showed higher AtCMT3 expression levels and CHG-gain genes relative to AtCMT3-L1. We analyzed all CHG DMRs that overlapped CHG-gain genes identified in the T4 generation (4,312 CHG DMRs overlapping 4,055 CHG-gain genes) (Table S2 and S5), and determined the levels of CG, CHG, and CHH methylation over these regions (Figure 2.5A). Re-focusing the analysis to the level of

DMRs over genes, rather than whole genes, revealed some residual methylation present on genes below our original cutoff of 1% gene-wide methylation. This residual methylation was almost exclusively in the CG context and is consistent with CMT3 preferentially localizing to methylated DNA (Figure 2.5A). Importantly, 2,094 of the 4,312 regions had no detectable CHG methylation in wild type and 597 of the 4,312 had no methylation in any sequence context (see also Figures 2.5A, S2.2, and S2.8).

As observed previously, the major gains in genic methylation in AtCMT3-Lines occurred in the CHG contexts (Figure 2.5A). Further dividing CHG methylation into CWG and CCG contexts revealed that the CHG methylation was highly enriched for CWG methylation, consistent with the preferred substrates of CMT3 [146, 147] (Figure 2.5A). Widespread, but lesser gains were also identified for CHH cytosines (Figure 2.5A). The CHH methylation was almost exclusively in the CWA context, which is indicative of CMT2 activity and not RdDM [146], suggesting that CMT3 activity likely leads to the recruitment of CMT2. Gains in genic CG methylation were comparatively lower than the other sequence contexts, but did occur appreciably over most loci, including regions with no prior CG methylation (Figure 2.5A). Further parsing of the data into all 16 trinucleotide contexts did not reveal any additional trends in the patterns of methylation (Figure S2.11A).

Ectopic genic CG methylation is preferentially maintained following loss of AtCMT3 expression

Establishment of methylation in non-CHG sequence contexts over CHG-gain genes is consistent with CMT3 activity recruiting additional methyltransferase pathways. The natural transgene silencing that occurred in the AtCMT3-L2 lineage following the T4 generation

provided an opportunity to test this hypothesis and determine whether non-CHG methylation remained following loss of AtCMT3 expression. To do so, we analyzed AtCMT3-L2 individuals for which we had both RNA-seq and whole genome bisulfite sequencing data (wild type, AtCMT3-L2 T3-T6) relative to the T4 generation, which showed the highest number of CHG-gain genes. The number of methylated cytosines in each sequence context in the CHG DMRs overlapping the 4,055 CHG-gain genes identified in AtCMT3-L2T4 were identified for each individual (same regions as Figure 2.5A; cytosines analyzed were corrected for coverage in all lines). Then the ratio of methylated cytosines was calculated relative to the T4 generation.

Results showed that the proportional CHG methylation levels on genes mirrored AtCMT3 expression levels, with a slight increase in proportional methylation from 92% in AtCMT3-L2T3 to 100% (by definition) in AtCMT3-L2T4 (Figure 2.5B). Following the T4 generation, and consistent with progressive silencing of the AtCMT3 transgene, the number of CHG cytosines methylated showed a marked decrease, with only 62% of T4 levels remaining methylated in the T5 generation and 4% remaining methylated in the T6 generation where AtCMT3 expression was greatly reduced (Figure 2.5B). CHH context cytosines showed a similar trend with 97% cytosines methylated in the T3 relative to the T4 generation, followed by a steep decline, with only 24% and 5% methylated remaining in the T5 and T6 generations, respectively, relative to the T4 (Figure 2.5B). Further dividing CHG and CHH cytosines into CWG vs. CCG or CWA vs. other CHH contexts did not reveal any differences in these trends (Figure S2.11B-C).

CG context cytosines showed a more substantial increase from the T3 to T4 generation, with CG cytosines in the T3 generation at 54% of the levels measured in T4 (Figure 2.5B). Also contrasting with CHG and CHH methylation, following the T4 generation, the proportion of

methyated CG sites remaining following the progressive AtCMT3 silencing was relatively high and much more stable, with 68% in T5 and 51% in T6 of the CG cytosines remaining methylated, relative to T4 (Figure 2.5B).

The results of the analysis of genic methylation for AtCMT3-L2 contrast with the identical analysis conducted on AtCMT3-L1, which did not show evidence of silencing of the AtCMT3 transgene. In this case, the relative methylation levels were maintained in all sequence contexts across generations (Figure S2.11D). Furthermore, to verify that the remaining genic CG methylation in AtCMT3-L2T6 did not result from residual AtCMT3 activity, we also completed an alternative approach where we crossed the AtCMT3-L1T5 transgene expressing line to wild type (non-transgenic) to segregate out the transgene. We analyzed two F2 progeny in which the transgene was segregated out and one F2 progeny that still contained the transgene (Figure S2.1). In the F2 progeny that did not encode the transgene, the ectopic genic CHG and CHH methylation that was present in the in the transgenic parent was lost but the genic CG methylation was maintained, similar to the result from AtCMT3-L2 when the transgene was silenced (Figure S2.11E). In contrast, the F2 progeny that still encoded the transgene maintained similar levels of genic methylation in all contexts as the transgenic parent (Figure S2.11E).

Importantly, the levels of newly methylated CG cytosines on CHG-gain genes in both the lines that lost AtCMT3 through silencing or crossing out were higher than background due to bisulfite non-conversion, estimated by comparing them to randomly sampled, un-methylated genes that did not gain CHG methylation (Figure S2.12A-B). The genic CG methylation is also unlikely to have occurred independently of AtCMT3, as the gains in genic CG methylation at these loci are also higher than those found by comparing an additional, non-transgenic *E. salusugineum* accession (Yukon) (Figure S2.12A-B). Thus, genic CG methylation showed a lesser

dependency on AtCMT3 expression and was preferentially maintained following loss of transgene expression, either through natural silencing or crossing out the transgene, consistent with maintenance of CG methylation by other methyltransferases recruited by initial AtCMT3 activity.

AtCMT3 preferentially methylates heterochromatin relative to genes

The majority CHG DMRs resulting from AtCMT3 expression were annotated as repeats or intergenic regions with prior methylation in wild type (~75% compared to ~25% genic loci with no prior methylation) (Figure 2.2A). Furthermore, the levels of CHG methylation in AtCMT3 expressing lines were much higher on repeats relative to genes (30-70% CHG methylation on repeats compared to 2-10% on genes) (Figure 2.2B-C). Therefore, we next sought to determine whether this difference was a result of AtCMT3 preferentially methylating heterochromatic loci relative to genic loci. We reasoned that favorable activity of AtCMT3 on heterochromatin could be revealed in Lines where AtCMT3 transgene expression was reduced, as they would be expected to show a greater proportional loss of CHG methylation over genic regions relative to heterochromatic loci when compared to Lines with high AtCMT3 expression.

To test this, we performed the same analysis reported in Figure 2.5B, except instead of genic loci, we focused on regions defined as hyper-CHG DMRs in AtCMT3-L2T4 that did not overlap genes (i.e. repeats and intergenic loci). Results showed that the relative number of methylated cytosines was much more robustly maintained in the AtCMT3-L2 T5 and T6 generations at these loci relative to genic loci following loss of AtCMT3 transgene expression (compare Figure 2.5B and 2.5C). This was especially evident in the T6 generation that showed a -4.6 log₂ fold change in AtCMT3 expression relative to AtCMT3-L2T4 (Figure 2.5C). Despite

substantial loss of transgene expression, 55% and 57% of CHG and CHH cytosines, respectively, remained methylated in the T6 generation relative to T4 across repeats and intergenic loci (Figure 2.5C). In contrast, only 4% of CHG and 5% of CHH cytosines remaining methylated across genes (Figure 5B). Methylated CG cytosines were also slightly more robust to loss of AtCMT3 transgene expression, dropping to 66% of T4 levels over repeats compared to 51% over genes in the T6 generation (Figure 2.5B-C).

An alternative explanation for these results is that methylation induced by AtCMT3 expression is more readily maintained by other methylation pathways in heterochromatin relative to genes, as other methyltransferases preferentially target these regions. To consider this possibility, we also analyzed the relative maintenance of DNA methylation in heterochromatin in lines in which the transgene was removed via crossing to wild type. In the F2 progeny of the AtCMT3-L1T5 X wild type cross that no longer encoded the transgene, both CG and CHH methylation were maintained at relatively similar levels as the transgenic parent, in contrast to CHG methylation that was only maintained at ~25% of the levels of the transgenic parent (Figure S2.11F). This is consistent with the possibility that AtCMT3-induced CG and CHH methylation, and to a lesser extent CHG methylation, can be perpetuated following loss of AtCMT3 in heterochromatin in preference to genes by other methylation pathways. However, the relative levels of CHG methylation in particular are lower than those detected over heterochromatin for lines in which the transgene was silenced, consistent with residual AtCMT3 preferentially targeting heterochromatin in these lines (Figure 2.5C). We conclude that AtCMT3 preferentially targets heterochromatin and does not readily methylate genic loci until expressed at high levels.

DISCUSSION

We have provided experimental evidence that CMT3 can initiate epimutations in the form of gene body CG methylation, which are maintained over generational time, even after loss of AtCMT3 expression. This finding has provided new insights into CMT3 localization and function by showing CMT3 is associated with de novo DNA methylation activity in vivo at genic loci lacking prior DNA methylation. The results also revealed a mechanism for the establishment of gbM that is consistent with the hypothesis that gbM is a passive effect of self-reinforcing positive feedback loops inherent to the heterochromatin machinery.

The natural loss of CMT3 in *E. salsugineum* and other species associated with the loss of gbM [41, 116], combined with prior work demonstrating that CMT3 targeting and activity requires binding to H3K9me [18, 147], suggests that both CMT3 and histone methyltransferases are important in the establishment of gbM. Furthermore, a plausible means for CMT3 activity to recruit additional methyltransferase pathways that deposit methylation in additional sequence contexts, such as CMT2, is indirectly through the promotion of H3K9me2 [17]. We therefore propose that both enzymes work in concert to provide de novo methylation of transcribed loci, to initially establish gbM through the model described in Figure 2.6, which expands on prior models [137].

In this model, nucleosomes possessing H3K9me are on rare occasions incorporated into transcribed genes. Initially, genic histone methylation is restricted to H3K9me1, which is bound by CMT3 and not CMT2 [14], and could thus explain the correlation between CMT3 and gbM. Most H3K9me1 associated nucleosomes are efficiently removed via the histone de-methylase, IBM1, in a transcription coupled mechanism. It is notable that encoded in the *E. salsugineum* genome are several expressed orthologs of the *A. thaliana* IBM1 (Table S10), which could

destabilize H3K9me and plausibly explain the lack of detection of H3K9me over genes that gain CHG methylation (Figures 2.3, S2.8 and S2.9). However, CMT3 binding to H3K9me1 may transiently stabilize H3K9me through the establishment of de novo CHG methylation, which activates the feedback loop between DNA and histone methyltransferases, promoting the establishment of H3K9me2. This contrasts with CMT3 activity at heterochromatin, where CMT3 preferentially methylates hemi-methylated CWG sites and is re-enforced by additional methylation pathways. As de novo methylation is a less favored activity of CMT3, this process is predicted to occur rarely, but can be promoted with high levels of CMT3 expression. Transient stabilization of H3K9me2 recruits additional methyltransferases which establish DNA methylation in additional sequences contexts. Finally, removal of H3K9me2 by IBM1 disrupts the feedback loop between H3K9me2 and CMTs resulting in the loss of non-CG methylation following DNA replication. CG methylation is maintained, however, due to the preferential recruitment of CG maintenance methyltransferases to hemi-methylated sites following DNA replication (Figure 2.6).

From a mechanistic standpoint, it is most parsimonious to conclude that gbM is a passive byproduct inherent to the function of CMT3 in the maintenance of heterochromatin. The presence of pathways that work to uncouple gbM from transcriptional silencing, such as the IBM1 pathway, further support this line of reasoning, as they may have evolved to counteract negative consequences of “spillage” of the heterochromatin machinery into genic space. Why, then, are some genes consistently characterized by body methylation across species and others not? It is telling that the genes that gain methylation in *E. salisugineum* are homologs of gbM genes in *A. thaliana* and/or retain similar characteristics including gene length, expression profile, and relative frequency of CHG sites. Rather than an exact determinant of gbM status, it

is likely that gene length and constitutive expression contribute to the exposure of a locus to incorporation of H3K9me1/2 nucleosomes, which, combined with the frequency of CMT3-preferred CWG sites and CMT3 levels, make a gene susceptible to methylation by CMT3 in a probabilistic manner. Under this model, gbM can be thought of as an epigenetic scar resulting from transient localization of the heterochromatin machinery that is likely to be present on a given gene as a function of these factors. This model does not exclude functional consequences of gbM, but it also does not require them.

MATERIALS AND METHODS

Experimental Models and Subject Details

Eutrema salsugineum Shandong ecotype was grown on soil at 21 °C in long day conditions (16 hours light, 8 hours dark). Plant transgenesis was conducted using the floral dip method [154]. The pEarleyGate 302 vector containing genomic sequence of *Arabidopsis thaliana* CMT3, including the native promoter, published by [18] was transformed into *Agrobacterium tumefaciens* strain C58C1. Bacteria were grown for 2 days at 30 °C in 200 ml cultures containing gentamicin (25 µg/mL), kanamycin (50 µg/mL), and rifampicin (50 µg/mL) and pelleted by centrifugation at 4 °C. Bacterial pellets were resuspended in 5% sucrose and 0.03% Silwet L-77 (Phyto Technology Laboratories) and used to dip open *E. salsugineum* inflorescences. Transgenic plants were selected for using Finale (BASTA, Bayer).

Whole genome bisulfite sequencing

Whole genome bisulfite sequencing libraries were generated based on methods described in [156]. DNA was extracted from cauline leaves of individual plants flash-frozen in liquid

nitrogen using the Qiagen DNeasy Kit according to the manufacturer's instructions. DNA was fragmented via sonication to a peak size of ~200bp and further size selected using AMPure beads (Beckman Coulter) to between 150 bp and 500 bp. Fragment end repair was performed using End-It from Lucigen incubated at room temperature for 45 minutes, followed by purification using AMPure beads. Next, A-tailing was conducted using Klenow Fragment from NEB in NEBNext dA-tailing reaction buffer at 37 °C for 30 minutes, followed by purification with AMPure beads and Illumina indexed adapter (NEXTFLEX Bisulfite-Seq Barcodes) ligation using T4 DNA ligase from NEB. Ligation was conducted at 16 °C for 16 hours. Ligation products were purified twice using AMPure beads and bisulfite converted using the EZ DNA Methylation-Gold Kit from Zymogen. Bisulfite converted DNA was then amplified using KAPA HiFi Uracil+ and universal primers with the following parameters: 95 °C for 2 minutes, 98 °C for 30 seconds, 8 cycles of 98 °C for 15 seconds, 60 °C for 30 seconds, 72 °C for 4 minutes, and a final extension time of 72 °C for 10 minutes. PCR products were purified using AMPure beads and sequenced with an Illumina NextSeq500 instrument by the Georgia Genomics & Bioinformatics Core . Adapters and primers used were those provided in the NEXTFLEX Bisulfite Sequencing Library Prep Kit (Bioo Scientific).

RNA-sequencing

Total RNA was extracted from cauline leaves of individual plants using the Direct-Zol RNA Mini-prep plus kit from Zymogen according to the manufacturer's instructions. Sequencing libraries were prepared from 1.3 µg input RNA with the Illumina TruSeq mRNA Stranded Library Kit according to the manufacturer's instructions, except all volumes were reduced to 1/3

of the recommended quantity. Sequencing was completed using an Illumina NextSeq500 instrument by the Georgia Genomics & Bioinformatics Core.

Chromatin immunoprecipitation and sequencing (ChIP-seq)

ChIP was conducted based on the protocol described in [157]. Cauline leaves were harvested from individual plants and submerged in cross-linking buffer (10 mM Tris-HCl pH 8, 1 mM EDTA, 0.4 M sucrose, 100 mM PMSF, 1 % formaldehyde). Tissues were vacuum infiltrated for 5 minutes at 85 kPa, followed by release of vacuum and 10 additional minutes at 85 kPa. Crosslinking was quenched with the addition of glycine to a concentration of 100 mM followed by 5 minutes under vacuum at 85 kPa. Tissues were then washed 5 times in water, flash frozen in liquid nitrogen, and ground to a fine powder with mortar and pestle. Powder was suspended in 10 mls extraction buffer 1 (0.4 M sucrose, 10 mM Tris-HCl pH 8, 10 mM MgCl₂, 5 mM BME, 0.1 mM PMSF, 1 mM EDTA, 1 tab / 10 ml Pierce Protease Inhibitors (ThermoFisher)). The suspension was filtered through 2 layers of Miracloth to enrich nuclei and pelleted by centrifugation for 20 minutes at 4000 rpm at 4 °C. Pellets were resuspended in 1 ml extraction buffer 2 (0.25 M sucrose, 10 mM Tris-HCl pH 8, 10 mM MgCl₂, 1% Triton X-100, 1 mM EDTA, 5 mM BME, 0.1 mM PMSF, 1 tab / 10 ml protease inhibitors) and centrifuged for 10 minutes at 12,000 g at 4 °C. Pellets were then suspended in 300 µl extraction buffer 3 (1.7 M sucrose, 10 mM Tris-HCl pH 8, 0.15% Triton X-100, 2 mM MgCl₂, 1 mM EDTA, 5 mM BME, 0.1 mM PMSF, 1 tab / 10 ml protease inhibitors) and layered on top of 300 µl extraction buffer 3. Samples were then centrifuged for 1 hour at 16,000 g at 4 °C, supernatant was removed, and chromatin pellets were resuspended in 100 µl nuclei lysis buffer (50 mM Tris-HCl pH 8, 10 mM EDTA, 1% SDS, 0.1 mM PMSF, 1 tab / 10ml protease inhibitors). Chromatin was fragmented

via sonication to a fragment size of ~200 base pairs and debris were removed by centrifugation at 16,000 g for 5 minutes at 4 °C. Ten µl of the supernatant was removed for input controls and the remaining supernatant was diluted 1:10 in ChIP dilution buffer (1.1 % Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-HCl pH 8, 167 mM NaCl, 0.1 M PMSF, 1 tab / 10 ml protease inhibitors).

Protein A Dynabeads (Invitrogen) were prepared by washing 25 µl beads three times with 1 ml ChIP dilution buffer (1.1 % Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-HCl pH 8, 167 mM NaCl). Beads were then resuspended in 100 µl ChIP dilution buffer with 5 µg anti-H3K9me2 (Cell Signaling Technology, Cat. # 9753s) or anti-H3K9me1 (Abcam # 8896) added. Antibodies were bound to beads at 4 °C with end over end rotation for 3 hours. Beads were then washed three times with ChIP dilution buffer (1.1 % Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-HCl pH 8, 167 mM NaCl, 0.1 M PMSF, 1 tab / 10 ml protease inhibitors) and resuspended in the diluted chromatin samples from above. Samples were incubated with rotation over night at 4 °C.

Following incubation, beads were washed twice with 1 ml each of low salt wash buffer (150 mM NaCl, 0.1% SDS, 1% TritonX-100, 2 mM EDTA, 20 mM Tris-HCl pH 8), high salt wash buffer (500 mM NaCl, 0.1% SDS, 1% TritonX-100, 2 mM EDTA, 20 mM Tris-HCl pH 8), and LiCl wash buffer (0.25 M LiCl, 1% NP40, 1% sodium deoxycholate, 1 mM EDTA, 10 mM Tris-HCl pH 8). Beads were then washed one time with 1 ml TE buffer (10 mM Tris-HCl pH 8, 1 mM EDTA) and resuspended in 250 µl elution buffer (1% SDS, 0.1 M NaHCO₃). Samples were eluted with incubation at 65 °C for 15 minutes with gentle agitation. The supernatant was removed and saved, and the elution was repeated with an additional 250 µl elution buffer. Supernatants were then combined and 20 µl 5 M NaCl was added. 500 µl elution buffer and 20

μ l 5 M NaCl were also added to the input controls. Crosslinks were reversed over night at 65 °C. Following crosslink reversal, 10 μ l of 0.5 M EDTA, 20 μ l 1 M Tris-HCl (pH 6.5), and 2 μ l of 10 mg/ml proteinase K (ChIP grade, Thermo Fisher Scientific) was added to each sample and incubated at 45 °C for 1 hour. DNA was extracted with phenol/chloroform/isoamyl alcohol (25:24:1) and resuspended in water.

ChIP sequencing libraries were prepared by conducting end repair, A-tailing, and adaptor ligation steps identical to those described for bisulfite sequencing library preparation, except that with the substitution of Illumina TruSeq adaptors and indexed primers. Libraries were amplified with Phusion DNA Polymerase (NEB) with the following parameters: 95 °C for 2 minutes, 98 °C for 30 seconds, 15 cycles of 98 °C for 15 seconds, 60 °C for 30 seconds, 72 °C for 4 minutes, and a final extension step of 72 °C for 10 minutes. Libraries were sequenced on an Illumina NextSeq500 instrument by the Georgia Genomics and Bioinformatics Core.

qRT-PCR

RNA was extracted from cauline leaves of individual plants using the Direct-Zol RNA Mini-prep plus kit from Zymogen according to the manufacturer's instructions. Synthesis of cDNA was completed using SuperScript III with random hexamers (Invitrogen) according to the manufacturer's instructions. Real time qRT-PCR was conducted using LightCycler 480 SYBR green master mix in a Light Cyler 480 instrument (Roche). Primers used include: AtCMT3 qRT-PCR FP: TGGTTTGAACCTCGTCACTAAA; AtCMT3 qRT-PCR RP: CGTTTGTCTCTGGGTGGTTAT; EsTUB4 qRT-PCR FP: CCTCCATATCCAAGGCGGTC; EsTUB4 qRT-PCR RP: GTACTGGCCGGTGTGATCAA.

Whole genome bisulfite sequencing mapping and analyses

WGBS data were processed using “single-end-pipeline” function from Methylypy as described in [158]. Briefly, quality-filtering and adapter-trimming were performed using cutadapt v1.9.dev1 [159]. Qualified Reads were aligned to the *E. salisugineum* v1.0 reference genome [160] (downloaded from: <https://phytozome.jgi.doe.gov>) using bowtie 2.2.4 [161]. Only uniquely aligned and non-clonal reads were retained. Chloroplast DNA (which is fully unmethylated) was used as a control to calculate the sodium bisulfite reaction non-conversion rate of unmodified cytosines. A binomial test was used to determine the methylation status of cytosines with a minimum coverage of three reads.

Identification of DMRs (Differential Methylated Regions) was performed using “DMRfind” function from Methylypy pipeline as described in [158]. Default parameters were adopted and only DMRs with at least 5 DMSs (Differential Methylated Sites) were reported and used for subsequent analysis.

To produce metaplots, 1 kb regions upstream and downstream features of interest were divided into 20 bins each. Features of interest were also divided into total of 20 bins. Weighted methylation levels were computed as the number of methylated reads divided by the total reads for each bin.

To generate the heatmap shown in Figure 2.2A, weighted percent CHG methylation was calculated for all significant CHG DMRs with a minimum of 5 cytosines with 3 read coverage each in all lines (Listed in Table S2). DMRs were then ranked by weighted %CHG methylation levels in wild type (vertical axis) and arranged by lineage (horizontal axis). To identify called DMRs reported in Table S3, all significant DMRs with minimum coverage requirements were filtered by cutoffs of a minimum of 10% change, relative to wild type, to be considered a hypo-

or hyper- CHH or CHG DMR, and a minimum of 20% change, relative to WT, to be considered a hypo- or hyper- CG DMR. Overlaps of hyper-CG, CHG, and CHH DMRs and generation of upset plots, shown in Figure S2.3, were calculated using Intervene v.0.6.1 [162]. To calculate the overlap of hyper-CHG DMRs between all individuals of the AtCMT3-L1 and AtCMT3-L2 lineages reported in Figure S2.4, coordinates of the called hyper-CHG DMRs reported in Table S3 were input to the bedtools v2.27.1 fisher command [163]. P values reported were calculated from Fisher's Exact Test with significance set at $p < 0.0004$ based on the Bonferroni correction for multiple testing.

To identify CHG-gain genes, genes were first filtered for those that had less than 1% cytosine methylation in any sequence context in wild-type. Genes were then filtered for coverage and only those with at least 10 informative CHG cytosines (min. 5 read coverage) in each line were assessed. Coverage corrections were separately done for: 1. wild type and all individuals of the AtCMT3-L1 and AtCMT3-L2 lineages described in Figure 2.1A; 2. wild type and all other individuals besides AtCMT3-L3T2c and the individuals of the AtCMT3-L1 and AtCMT3-L2 described in Figure 2.1A; and 3. wild type and AtCMT3-L3T2c. Genes were then called as CHG-gain genes if they showed a minimum of 5% increase in CHG methylation in *AtCMT3* expressing lines relative to wild-type. Percent CHG methylation was calculated as the number of methylated reads mapping to CHG sites at the gene of interest divided by the total number of reads mapping to CHG sites. CHG-gain genes are listed in Table S5. The significance of the overlap of CHG-gain genes between individuals shown in Figure S2.6 was calculated with a hypergeometric test with significance set at $p < 0.0004$ based on the Bonferroni correction for multiple testing.

To generate the heatmap shown in Figure 2.5A and S2.11A, all CHG DMRs identified in Table S2 that overlapped the CHG-gain genes AtCMT3-L2^{T4} were analyzed. Weighted methylation levels of trinucleotide sub-contexts (CNN, N=A/T/C/G) were calculated for each qualified DMR in wild type and AtCMT3-L2, T1-T6 derived from WGBS. The methylation levels of sub-contexts in wild type were used to determine their hierarchical clustering relationships. Corresponding methylation levels in AtCMT3-L2, T1-T6 were plotted in the same order.

To determine the relative number of methylated cytosines over CHG-gain genes in individuals of the AtCMT3-L2 lineage relative to AtCMT3-L2^{T4} (shown in Figure 2.5B, S2.11B-C), the analysis was limited to cytosines located within the regions analyzed in Figure 2.5A that had a minimum of 3 read coverage in each line and were not methylated in wild type. Methylation status was determined with a binomial test. Number of methylated cytosines in each sequence context were reported as a ratio, relative to the number of methylated cytosines in AtCMT3-L2^{T4}. In Figure 2.5C, the same analysis was performed as described in Figure 2.5B, except cytosines analyzed were those found in hyper-CHG DMRs defined in AtCMT3-L2T4 that did not overlap genes and were thus annotated as repeats or intergenic regions. In Figure S2.11D, the same analysis was conducted except the regions analyzed were they hyper-CHG DMRs identified in AtCMT3-L1T4 that overlapped AtCMT3-L1T4 CHG-gain genes and the ratios of methylated cytosines are relative to AtCMT3-L1T4. The same analysis was also conducted to produce Figure S2.11E and F, except the regions analyzed were the hyper CHG DMRs identified in AtCMT3-L1T5 that overlapped AtCMT3-L1T5 CHG gain genes (S11E) or did not overlap genes (S11F) and the ratios of methylated cytosines are relative to AtCMT3-L1T5.

To identify the background levels of CG methylation shown in Figure S2.12, an equal number of genes as the number of CHG-gain genes identified in AtCMT3-L2T4 (for S12A) or AtCMT3-L1T5 (for S12B) was randomly selected from all genes that were classified as unmethylated in wild type and did not gain CHG methylation in AtCMT3-expressing lineages using the R command `sample`. Then an equal amount of sequence as analyzed in Figure 2.5B (for S12A) or Figure S2.11E (for S2.12B) was extracted from the randomly chosen genes, with the total number of nucleotides distributed evenly across each gene. All CG cytosines with less than three read coverage in each line assessed and cytosines found to be methylated in wild type (Shandong accession) were removed from the analysis. Coverage filtering was completed individually for the lineages assessed in each panel of Figure S2.12. Methylated CG cytosines were then identified in each lineage using a binomial test and percent CG methylation was calculated as the number of methylated CG cytosines divided by the total number of CG cytosines. This was completed for 5 randomly chosen sets of unmethylated genes and compared to the identical analysis completed on the CHG gain gene regions assessed in Figure 2.5B (for S2.12A) or Figure S2.11E (for S2.12B).

RNA sequencing mapping and analyses

Quality-filtering and adapter-trimming were performed using Trimmomatic v0.33 with default parameters [164]. Qualified reads were aligned to the *E. salicigineum* v1.0 reference genome using HISAT2 v2.0.5 [165]. Gene expression (calculated as fragments per kilobase million; FPKM) values were computed using StringTie v1.3.3b [166]. To compare expression between wild type and AtCMT3-expressing lines, Log2 fold change values were calculated as $\text{Log}_2(\text{FPKM AtCMT3 line} / \text{FPKM wild-type})$. Genes with zero FPKM values were removed

from expression analyses. A cutoff of $\pm 2 \text{ Log}_2$ fold change was used to identify genes undergoing substantial changes in expression.

To conduct Gene Ontology enrichment analyses of up- and down-regulated genes, *A. thaliana* orthologs of *E. salsugineum* genes identified in [41] were utilized to extract GO annotations from TAIR (www.arabidopsis.org). These annotations were then used to identify significantly enriched GO terms with an elim Fisher's exact test (p-value < 0.01) using the R package topGO (<https://bioconductor.org/packages/release/bioc/html/topGO.html>).

ChIP-sequencing mapping and analyses

Quality-filtering and adapter-trimming were performed using Trimmomatic v0.33 [164] with default parameters. The remaining reads were aligned to the *E. salsugineum* v1.0 reference genome using Bowtie v1.1.1 [167] with the following parameters: "bowtie -m 1 -v 2 --best --strata --chunkmbs 1024 -S". Aligned reads were sorted using SAMtools v1.2 and duplicated reads were removed using SAMtools v0.1.19 [168]. ChIP-peaks were defined in wild type relative to input using HOMER 4.10 [169] with the following parameters: "-region -tagThreshold 10 -size 1000 -minDist 2500 -tbp 0". Identified peaks that were directly connected together were merged into a single region. Peaks were then further filtered by read density. The density for each merged region was defined as follows: aligned reads divided by region length. Only merged regions with density greater than 0.05 were outputted as peaks and used for subsequent analysis.

We used these ChIP-peak regions and the coordinates of CHG-gain genes to create metaplots to compare enrichment between samples. In the metaplots, mapped reads were normalized to total mapped reads for each locus of interest, and were averaged over 4 bins

representing 2kb upstream, 2kb from the transcription start site into the gene, 2kb from the transcription stop site into the gene, and 2kb downstream. Finally, the average bin values were normalized to account for the number of loci.

qRT-PCR analyses

Relative expression of the *AtCMT3* transgene to *TUB4* (Thhalv10003210m) was calculated using the double delta threshold cycle (Ct) method as $2^{-((\text{Average } AtCMT3 \text{ Ct}) - (\text{Average } TUB4 \text{ Ct}))}$ [170]. Average Ct values were calculated from three technical replicates.

Characterization CHG-gain genes

Gene length and exon number for CHG-gain genes and UM genes were derived from *E. salsugineum* reference annotation files. CHG site number was calculated by scanning the whole gene sequence with a 3 base window and step size of 1 base. Both positive strand and negative strand were considered. The CHG frequency was calculated by normalizing CHG site number to gene length. Significance for differences in characteristics between CHG-gain and UM genes was calculated with a Wilcoxon rank-sum in the stats package of R v3.44.

Orthologs of CHG-gain genes in *A. thaliana* and their gbM status reported in Table S9 were those identified in [41]. To determine if gbM-gain genes had more *A. thaliana* orthologs that were classified as gbM than expected by chance, a hypergeometric test was conducted using the values of 20,211 total *A. thaliana* genes with an *E. salsugineum* ortholog, with 4,532 of those being classified as gbM in *A. thaliana* by [41].

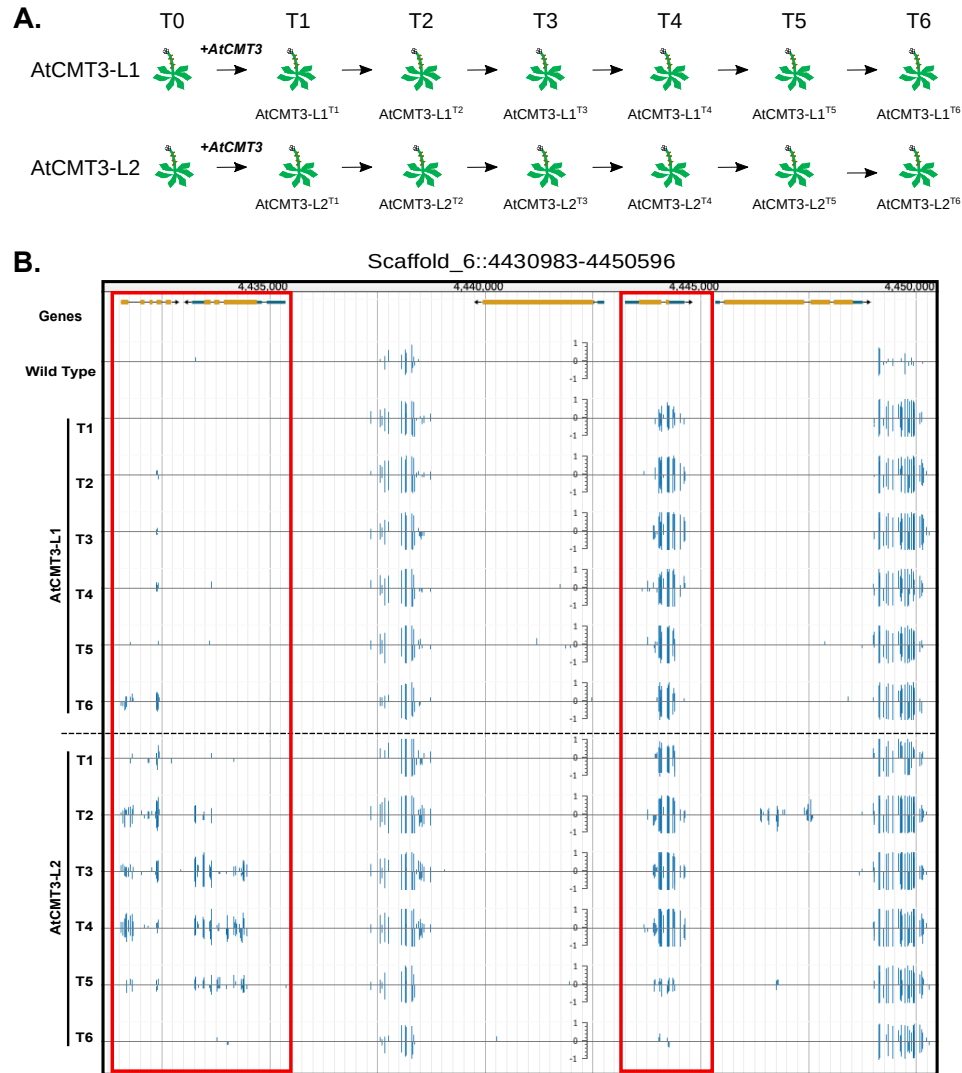


Figure 2.1. Expression of *AtCMT3* in *E. salsugineum* results in increased CHG methylation.

A. Schematic of the experiment. Two *E. salsugineum* lines derived from transformations with genomic *A. thaliana* *CMT3* were propagated by single seed decent for six generations (T1-T6). The two lines are referred to as AtCMT3-L1 and AtCMT3-L2, followed by the generation (T1-T6). For additional lines analyzed in this study see Figure S2.1 and Table S1. **B.** Genome browser view of CHG methylation levels derived from whole genome bisulfite sequencing. The image illustrates the gains in CHG methylation that occur in regions that are methylated in wild type, as well as over gene bodies with no pre-existing DNA methylation (boxed in red). Scales on tracks designate the weighted percent methylation, with 1 = 100% on the top strand and -1 = 100% on the bottom strand. Only CHG methylation is shown. For methylation in all contexts see Figure S2.2.

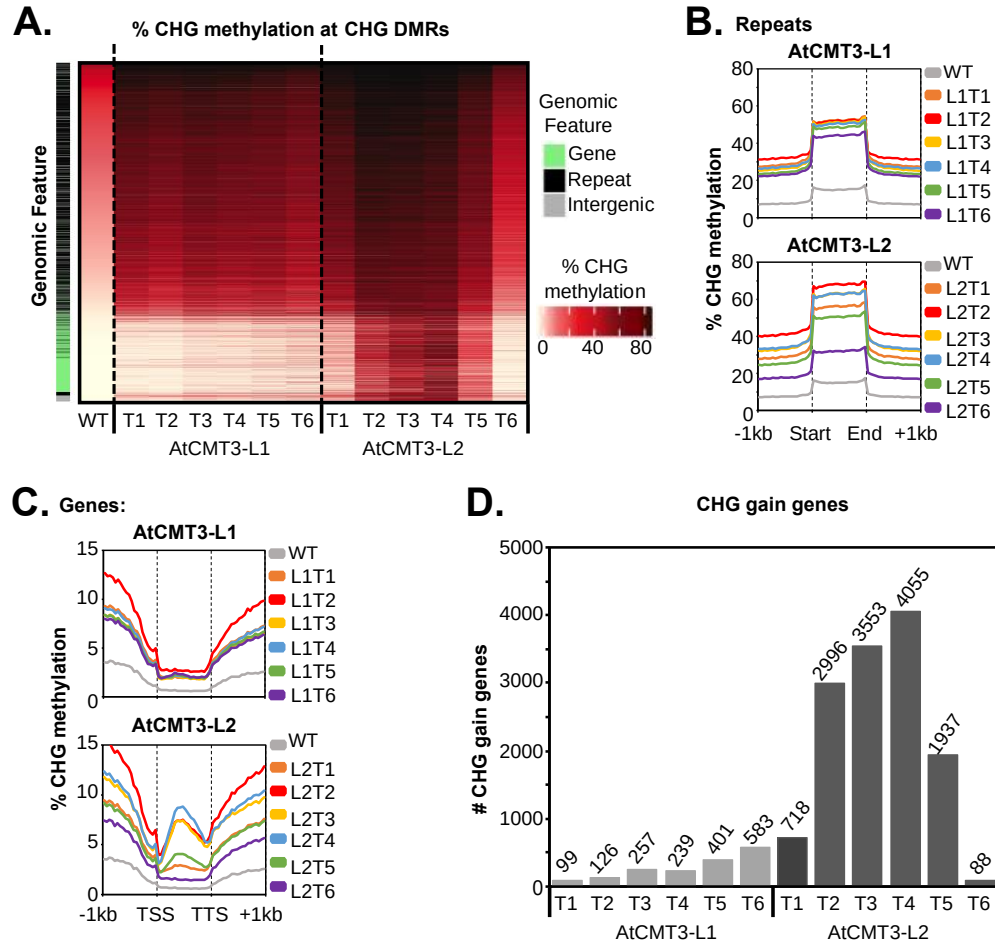


Figure 2.2. AtCMT3 expression results in CHG methylation over repeats and a subset of gene bodies.

A. Heatmap of % CHG methylation over hyper-CHG differentially methylated regions (DMRs) defined by comparing all AtCMT3-L1 and L2 generations (T1-T6) to wild type. DMRs are ranked by % CHG methylation levels in wild type plants showing that the majority of gains in CHG methylation occurred over regions with pre-existing CHG methylation classified as repeats or intergenic regions. A subset of regions with no pre-existing CHG methylation, mainly classified as genes, also showed gains in CHG methylation, especially in AtCMT3-L2 individuals. See also Tables S2 and S3. **B.** Metaplot summarizing % CHG methylation over repetitive sequences for each Line. **C.** Metaplot summarizing % CHG methylation over gene bodies for each Line. **D.** The number of genes gaining a minimum of 5% CHG methylation (CHG-gain genes) in each lineage. See Table S5 for lists of CHG-gain genes.

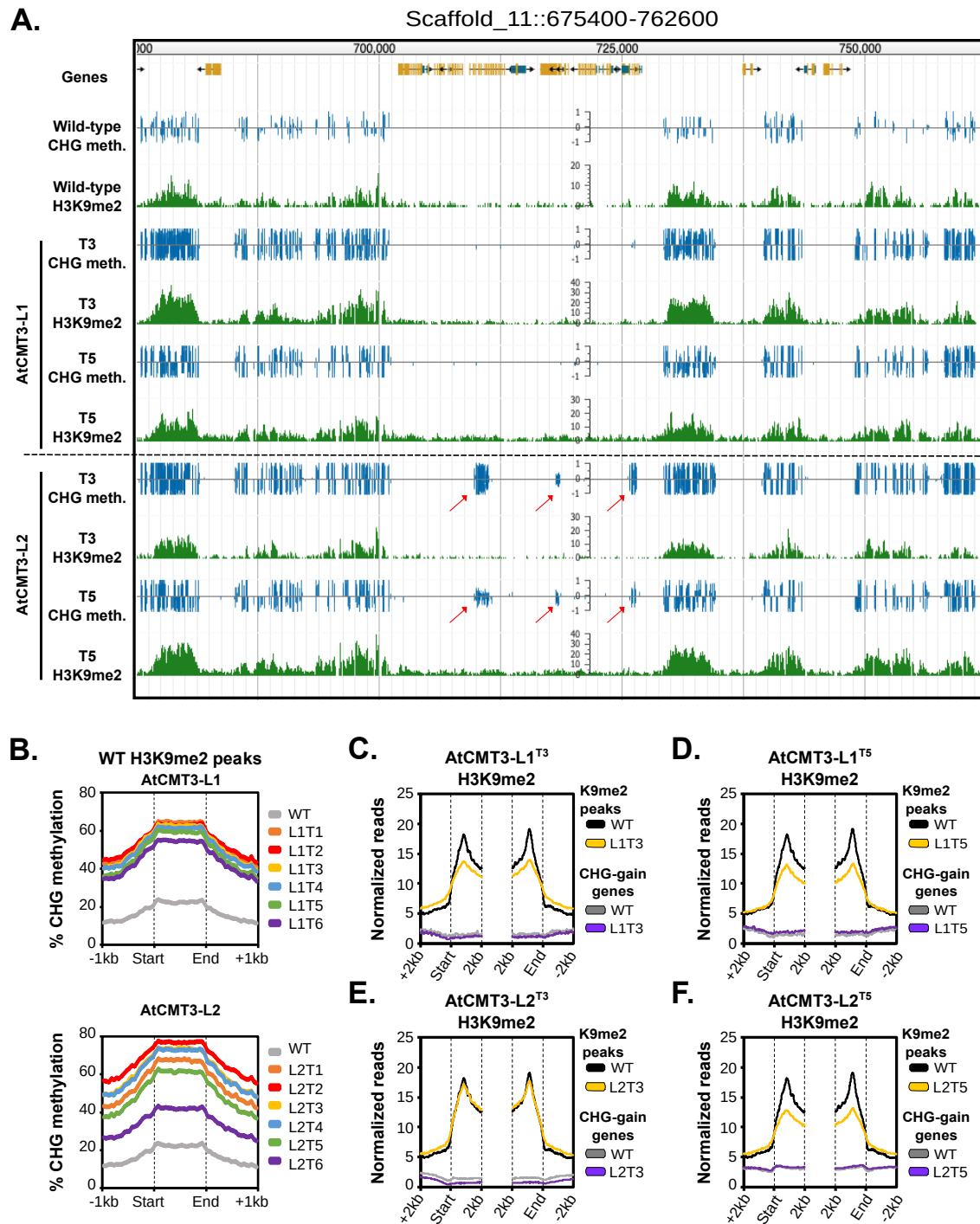


Figure 2.3. Gains in CHG methylation do not alter H3K9me2 levels or distribution.

A. Genome browser view of CHG methylation levels and H3K9me2 ChIP sequencing levels in the T3 and T5 generations of the AtCMT3-L1 and L2 lineages. Arrows indicate gains of CHG methylation over gene bodies in the AtCMT3-L2 generations that do not show H3K9me2

enrichment. Scales on methylation tracks designate the weighted percent methylation, with 1 = 100% on the top strand and -1 = 100% on the bottom strand. Scales on the H3K9me2 tracks indicate the number of mapped reads and are not adjusted for library size (See C-F for comparison of normalized reads). DNA methylation is only shown in the CHG context. For DNA methylation in all contexts see Figure S2.8. **B.** Metaplot of % CHG methylation over H3K9me2 ChIP peaks identified in wild type plants. **C-F.** Metaplot of H3K9me2 ChIP-sequencing enrichment over H3K9me2 ChIP peaks defined in wild type plants and over CHG-gain genes in AtCMT3-L1^{T3} (**C**), AtCMT3-L1^{T5} (**D**), AtCMT3-L2^{T3} (**E**), and AtCMT3-L2^{T5} (**F**). Reads were normalized to library size. See Table S5 for lists of CHG-gain genes in each lineage.

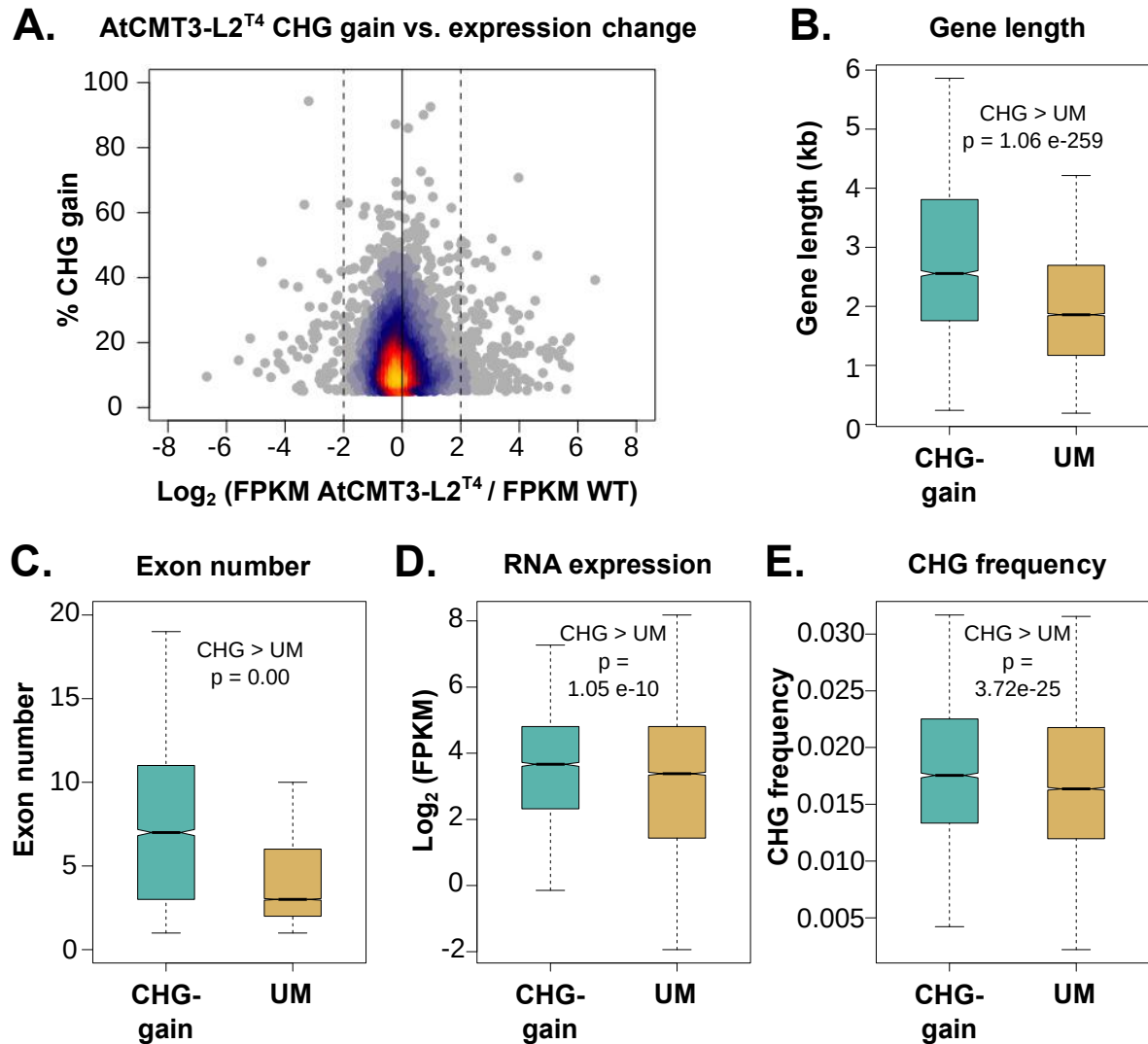


Figure 2.4. Genes that gain CHG methylation have *A. thaliana* gbM gene characteristics.

A. Relationship between levels of CHG methylation gain and log₂ fold change (FC) in expression for CHG-gain genes in AtCMT3-L2^{T4} relative to wild type. Genes with zero FPKM values were removed from the analysis. See Figure S2.10 for additional individuals analyzed. **B-E.** Comparison of the **B.** distribution of gene lengths, **C.** number of exons, **D.** expression levels, and **E.** frequency of CHG sites between CHG-gain genes and unmethylated genes (UM). P-values were calculated using a Wilcoxon rank-sum test. Boxes indicate the first and third quartiles, with the center line indicating the median and notches the 95% confidence interval of the median. Whiskers show 1.0 times the interquartile range and outliers beyond this range were excluded for visualization purposes, but included in all calculations.

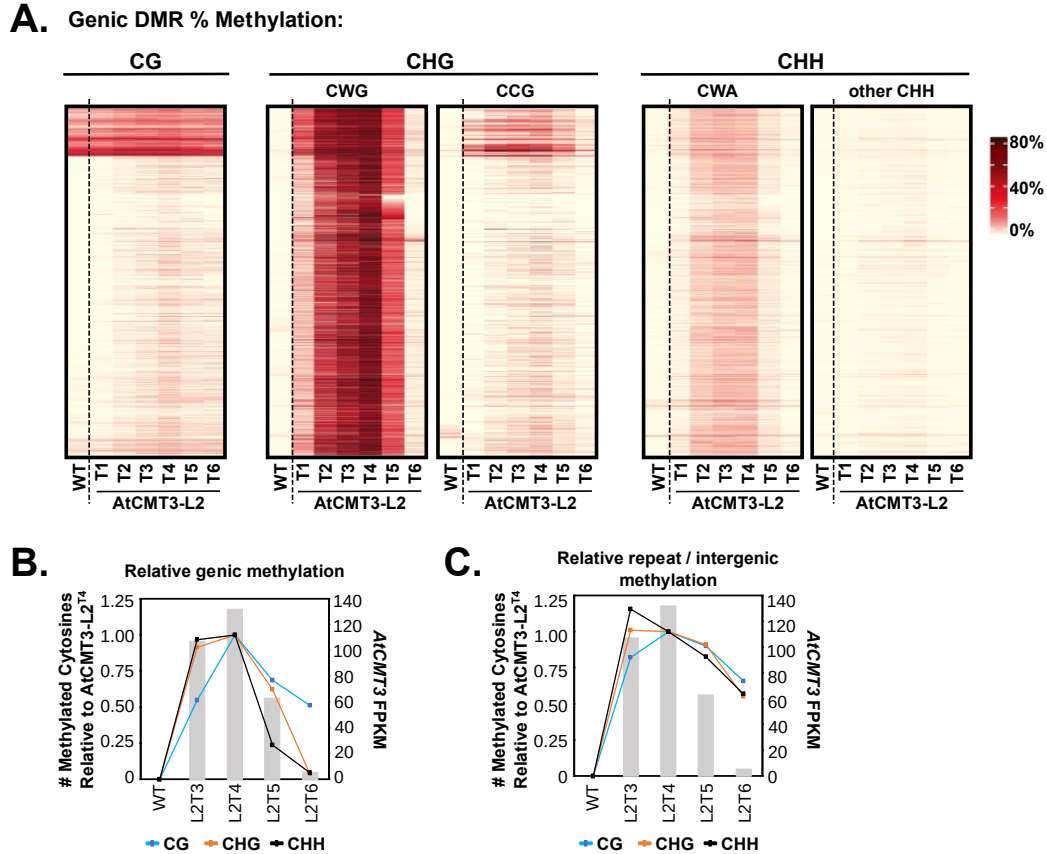


Figure 2.5. CHG methylation over gene bodies is associated with gains in non-CHG methylation.

A. Heatmap of % methylation levels across CHG DMRs overlapping the CHG-gain genes in *AtCMT3-L2*^{T4} divided into selected trinucleotide contexts. See Figure S2.11A for further parsing of the data into all 16 possible trinucleotide contexts. **B.** Assessment of the relationship of genic cytosine methylation to *AtCMT3* expression across *AtCMT3-L2* generations. Line plots show the number of methylated cytosines in each context relative to *AtCMT3-L2*^{T4} across CHG DMRs overlapping CHG-gain genes identified in *AtCMT3-L2*^{T4}. Bar plots show the expression of the *AtCMT3* transgene. For further parsing of the CHG and CHH contexts into CWG vs. CCG and CWA vs. other CHH, see Figure S2.11B-C. **C.** Assessment of the relationship of repeat and intergenic cytosine methylation to *AtCMT3* expression across *AtCMT3-L2* generations. Line plots show the number of methylated cytosines relative to *AtCMT3-L2*^{T4} across hyper CHG DMRs identified in *AtCMT3-L2*^{T4} that overlap repeats or intergenic regions. Bar plots show the expression of the *AtCMT3* transgene.

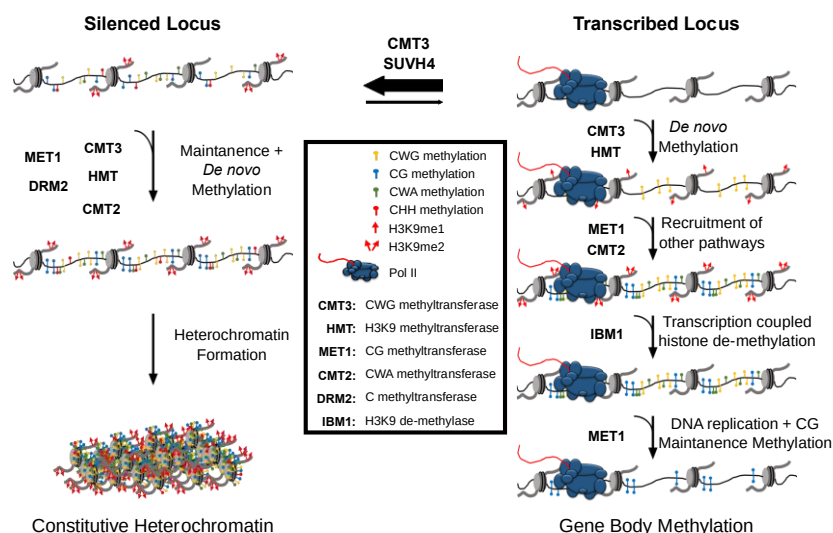


Figure 2.6. Hypothetical model for CMT3 establishment of gbM.

The activity of CMT3 and the histone methyltransferase, SUVH4, maintains CWG methylation and H3K9me1/2, respectively, and is most readily detected at silenced loci. At silenced loci, methylation by CMT3 and SUVH4 is reinforced by other methyltransferase pathways, including MET1 (CG methylation), CMT2 (CWA methylation), and DRM2 (methylation in all contexts), which maintain constitutive heterochromatin. In contrast to silenced loci, CMT3 and SUVH4 can only transiently establish *de novo* CWG methylation and H3K9me at transcribed loci characterized by gbM. This process may be initiated by incorporation of H3K9me1 nucleosomes, which are bound exclusively by CMT3 and not CMT2 and normally removed by the H3K9 de-methylase IBM1 in a transcription coupled mechanism. However, on rare occasions CMT3 may bind H3K9me1 located in genes and *de novo* methylate CWG cytosines. *De novo* methylation of CMT3 is not a favored activity of the enzyme, so this happens only very rarely or when CMT3 is expressed at high levels. This temporally stabilizes H3K9me due to the self-reinforcing feedback loop between histone and DNA methyltransferases. Transient stabilization of H3K9me promotes H3K9me2 that can then recruit additional methyltransferases including MET1 and CMT2 to methylate CG and CWA cytosines, respectively. Heterochromatin formation is inhibited, however, through eventual removal of H3K9me by IBM1. Loss of H3K9me and/or loss of available CMT3 results in the failure of maintenance of DNA methylation in all contexts except CG following DNA replication. CG methylation is maintained due to the preferential targeting of the CG maintenance methyltransferase, MET1, to hemi-methylated cytosines following replication.

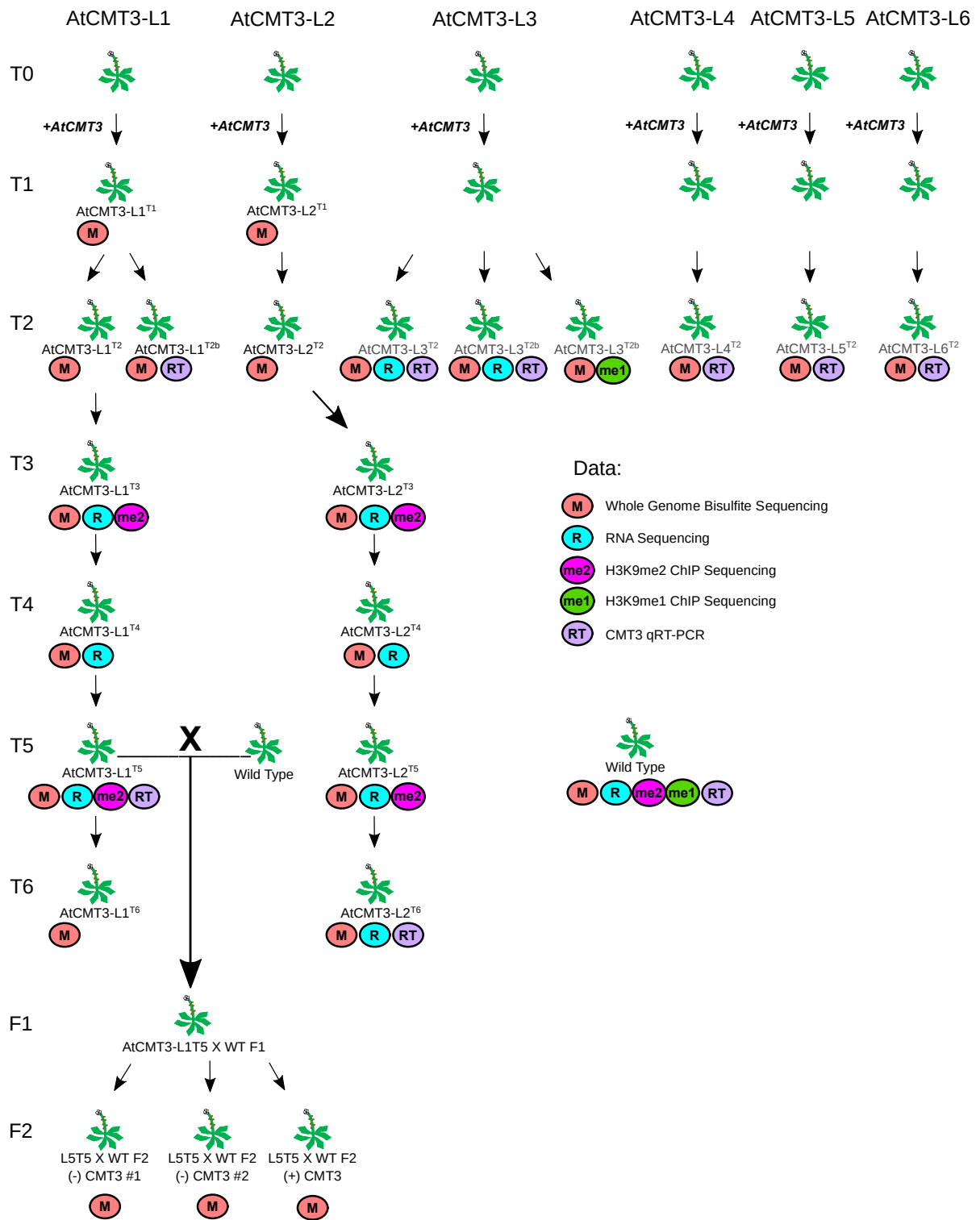


Figure S2.1. Diagram of all experimental *E. salsgineum* lines analyzed in this study with associated data collected.

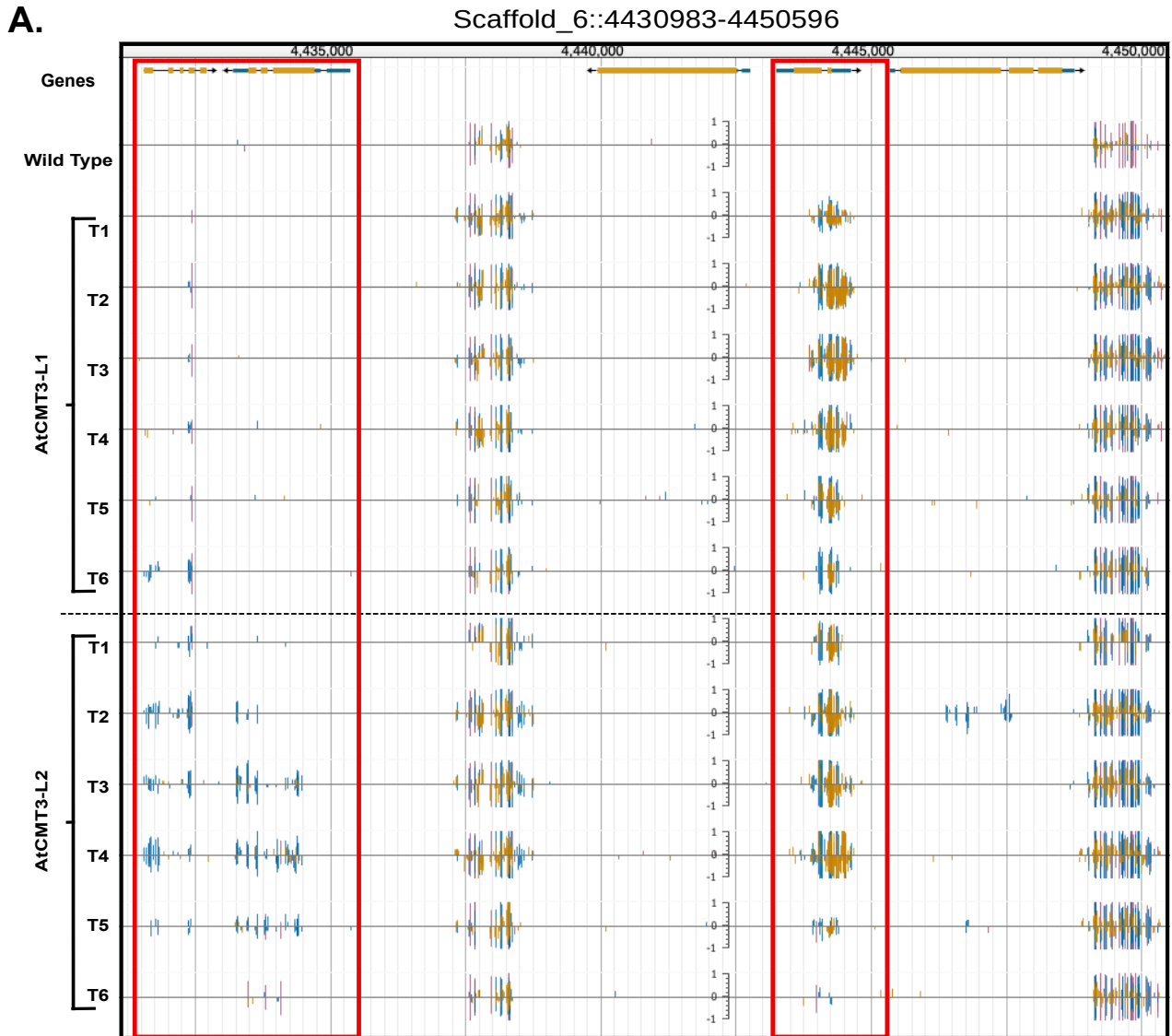


Figure S2.2. CHG methylation increases in *AtCMT3*-expressing *E. salsgineum* lines.

A. Genome browser view of CHG methylation levels derived from whole genome bisulfite sequencing. The image illustrates the gains in CHG methylation that occur in regions that are methylated in wild type, as well as over gene bodies with no pre-existing DNA methylation (boxed in red). Scales on tracks designate the weighted percent methylation, with 1 = 100% on the top strand and -1 = 100% on the bottom strand. Cytosine methylation is divided into CG (red), CHG (blue), and CHH (yellow) sequence contexts. Corresponds to the region shown in Figure 2.1B.

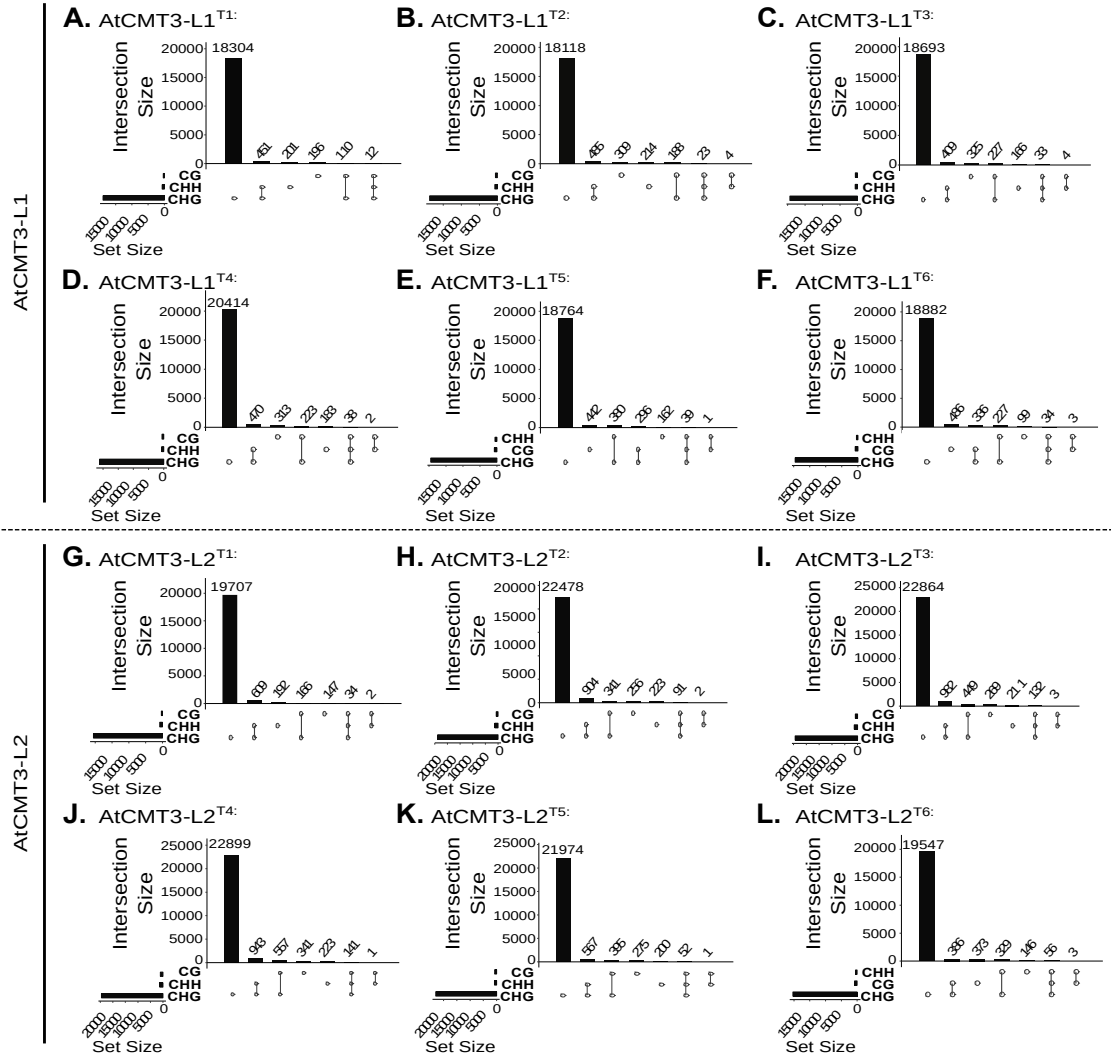


Figure S2.3. CG, CHG, and CHH DMRs co-localize.

A-L. Upset plots showing the number and overlap between hyper DMRs identified for each sequence context, CG, CHG, and CHH, within each generation of each lineage, AtCMT3-L1 and AtCMT3-L2: **A.** AtCMT3-L1^{T1}; **B.** AtCMT3-L1^{T2}; **C.** AtCMT3-L1^{T3}; **D.** AtCMT3-L1^{T4}; **E.** AtCMT3-L1^{T5}; **F.** AtCMT3-L1^{T6}; **G.** AtCMT3-L2^{T1}; **H.** AtCMT3-L2^{T2}; **I.** AtCMT3-L2^{T3}; **J.** AtCMT3-L2^{T4}; **K.** AtCMT3-L2^{T5}; **L.** AtCMT3-L2^{T6}.

	AtCMT3-L1T1 (18877)	AtCMT3-L1T2 (18811)	AtCMT3-L1T3 (19357)	AtCMT3-L1T4 (21143)	AtCMT3-L1T5 (19474)	AtCMT3-L1T6 (19521)	AtCMT3-L2T1 (20516)	AtCMT3-L2T2 (23812)	AtCMT3-L2T3 (24419)	AtCMT3-L2T4 (24545)	AtCMT3-L2T5 (22984)	AtCMT3-L2T6 (20308)	# CHG DMRs overlap
AtCMT3-L1T1 (18877)		18488	18659	18827	18590	18423	18786	18837	18841	18835	18775	18529	
AtCMT3-L1T2 (18811)	***		18657	18776	18533	18381	18599	18759	18740	18722	18672	18487	
AtCMT3-L1T3 (19357)	***	***		19315	18961	18816	19010	19268	19266	19254	19170	18819	
AtCMT3-L1T4 (21143)	***	***	***		19369	19286	19890	20864	20951	20935	20510	19550	
AtCMT3-L1T5 (19474)	***	***	***	***		19076	19060	19430	19358	19367	19260	18813	
AtCMT3-L1T6 (19521)	***	***	***	***	***		18991	19359	19395	19401	19267	18751	
AtCMT3-L2T1 (20516)	***	***	***	***	***	***		20434	20434	20445	20153	19308	
AtCMT3-L2T2 (23812)	***	***	***	***	***	***	***		23524	23416	22258	20107	
AtCMT3-L2T3 (24419)	***	***	***	***	***	***	***	***		24062	22663	20198	
AtCMT3-L2T4 (24545)	***	***	***	***	***	***	***	***	***		22903	20217	
AtCMT3-L2T5 (22984)	***	***	***	***	***	***	***	***	***	***		20047	
AtCMT3-L2T6 (20308)	***	***	***	***	***	***	***	***	***	***	***		

*** p < 0.0004

Figure S2.4. *AtCMT3* expression results in gains in CHG methylation over similar regions between lineages and across generations. Matrix showing the number and overlap between hyper-CHG DMRs identified in each individual in *AtCMT3*-L1 and *AtCMT3*-L2. The top half of the matrix shows the number of overlapping DMRs and the bottom half shows the p-value calculated based on fisher's exact test (**p significant < 0.0004, Bonferroni correction). See also Figure 2.2A and Table S3.

A. *AtCMT3* expression vs. genome % CHG methylation

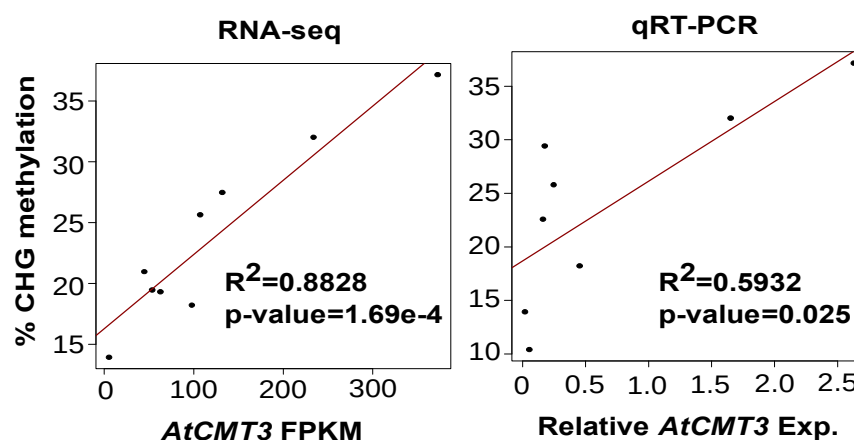


Figure S2.5. Genome wide percent CHG methylation is correlated with *AtCMT3* transgene expression levels. *AtCMT3* expression was determined by RNA-seq (first panel) or qRT-PCR (second panel). For qRT-PCR *AtCMT3* expression was normalized to *TUB4* expression (Thhalv10003210m).

	AtCMT3- L1T1 (99)	AtCMT3- L1T2 (126)	AtCMT3- L1T3 (257)	AtCMT3- L1T4 (239)	AtCMT3- L1T5 (401)	AtCMT3- L1T6 (583)	AtCMT3- L2T1 (718)	AtCMT3- L2T2 (2996)	AtCMT3- L2T3 (3553)	AtCMT3- L2T4 (4055)	AtCMT3- L2T5 (1937)	AtCMT3- L2T6 (88)
AtCMT3- L1T1 (99)		66	80	73	76	77	83	92	93	91	80	17
AtCMT3- L1T2 (126)	***		103	93	94	99	94	121	115	114	103	23
AtCMT3- L1T3 (257)	***	***		175	194	214	185	235	237	244	222	37
AtCMT3- L1T4 (239)	***	***	***		203	208	159	217	214	223	202	34
AtCMT3- L1T5 (401)	***	***	***	***		355	242	349	355	359	315	45
AtCMT3- L1T6 (583)	***	***	***	***	***		312	493	507	520	451	59
AtCMT3- L2T1 (718)	***	***	***	***	***	***		679	670	674	580	61
AtCMT3- L2T2 (2996)	***	***	***	***	***	***	***		2596	2582	1588	81
AtCMT3- L2T3 (3553)	***	***	***	***	***	***	***	***		3169	1721	82
AtCMT3- L2T4 (4055)	***	***	***	***	***	***	***	***	***		1883	88
AtCMT3- L2T5 (1937)	***	***	***	***	***	***	***	***	***	***		84
AtCMT3- L2T6 (88)	***	***	***	***	***	***	***	***	***	***	***	

CHG-gain genes overlap

p value of overlap: *** p < 0.0004

Figure S2.6. *AtCMT3* expression results in gains in CHG methylation over similar genes between lineages and across generations. Matrix showing the number and overlap between genes gaining a minimum of 5% CHG methylation identified in each individual in *AtCMT3*-L1 and *AtCMT3*-L2. The top half of the matrix shows the number of overlapping genes and the bottom half shows the p-value calculated based on a hypergeometric test (***p significant < 0.0004, Bonferroni correction).

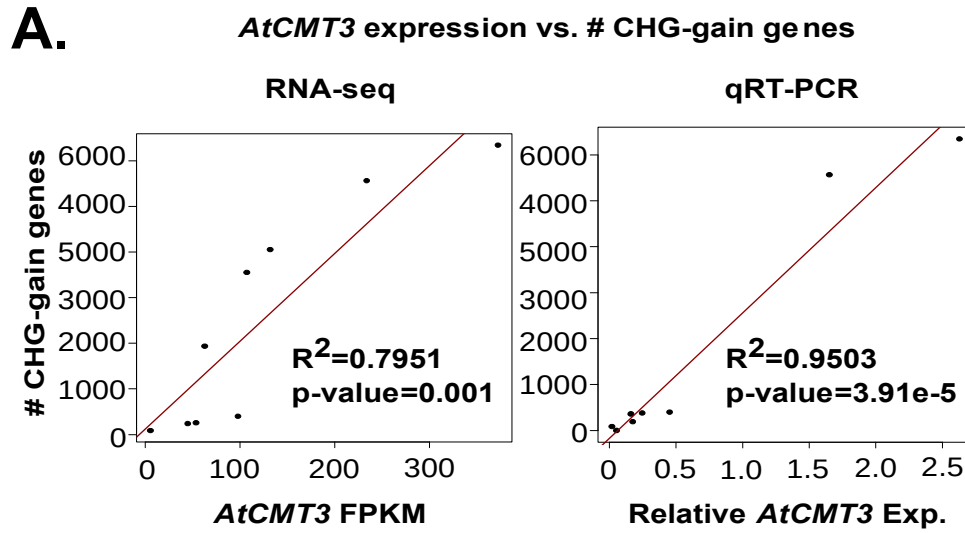


Figure S2.7. The number of genes gaining a minimum of 5% CHG methylation (CHG-gain genes) is correlated with *AtCMT3* transgene expression levels. *AtCMT3* expression was determined by RNA-seq (first panel) or qRT-PCR (second panel). For qRT-PCR *AtCMT3* expression was normalized to *TUB4* expression (Thhalv10003210m).

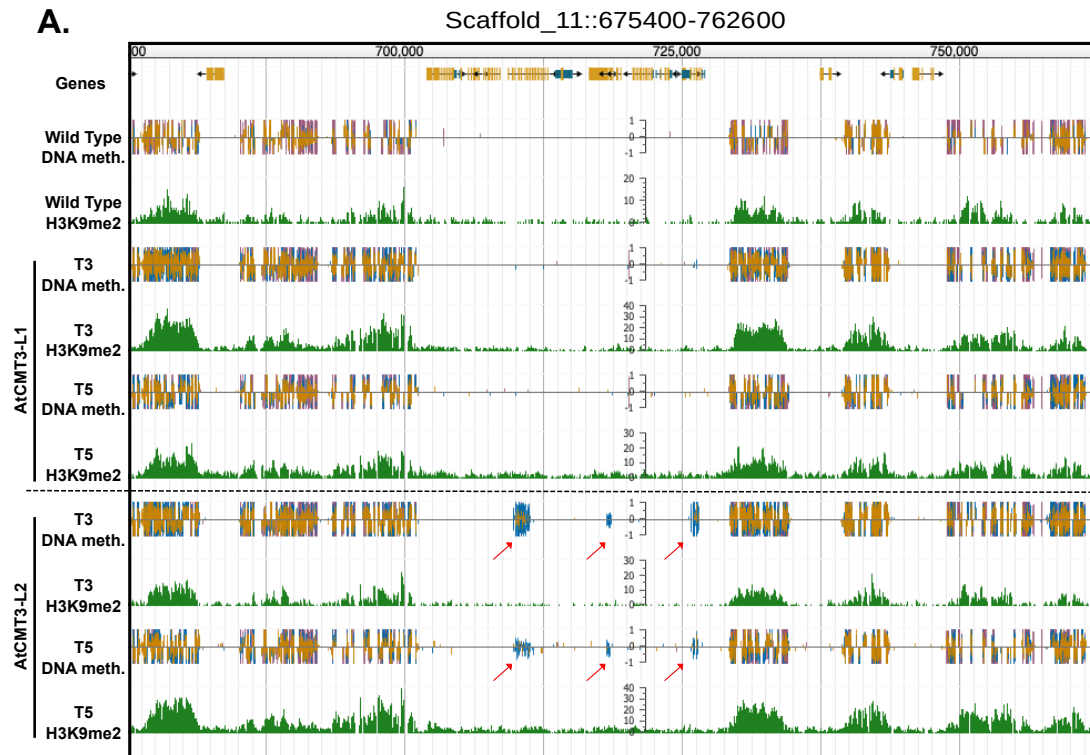


Figure S2.8. Gains in CHG methylation do not alter H3K9me2 levels or distribution.

A. Genome browser view of CHG methylation levels and H3K9me2 ChIP sequencing levels in the T3 and T5 generations of the AtCMT3-L1 and AtCMT3-L2 lineages. Arrows indicate gene bodies that gain CHG methylation in the AtCMT3-L2 generations but do not show H3K9me2 enrichment. Scales on methylation tracks designate the weighted percent methylation, with 1 = 100% on the top strand and -1 = 100% on the bottom strand. Scales on the H3K9me2 tracks indicate the number of mapped reads and are not normalized. Cytosine methylation is divided into CG (red), CHG (blue), and CHH (yellow) sequence contexts. Corresponds to the region shown in Figure 2.3A.

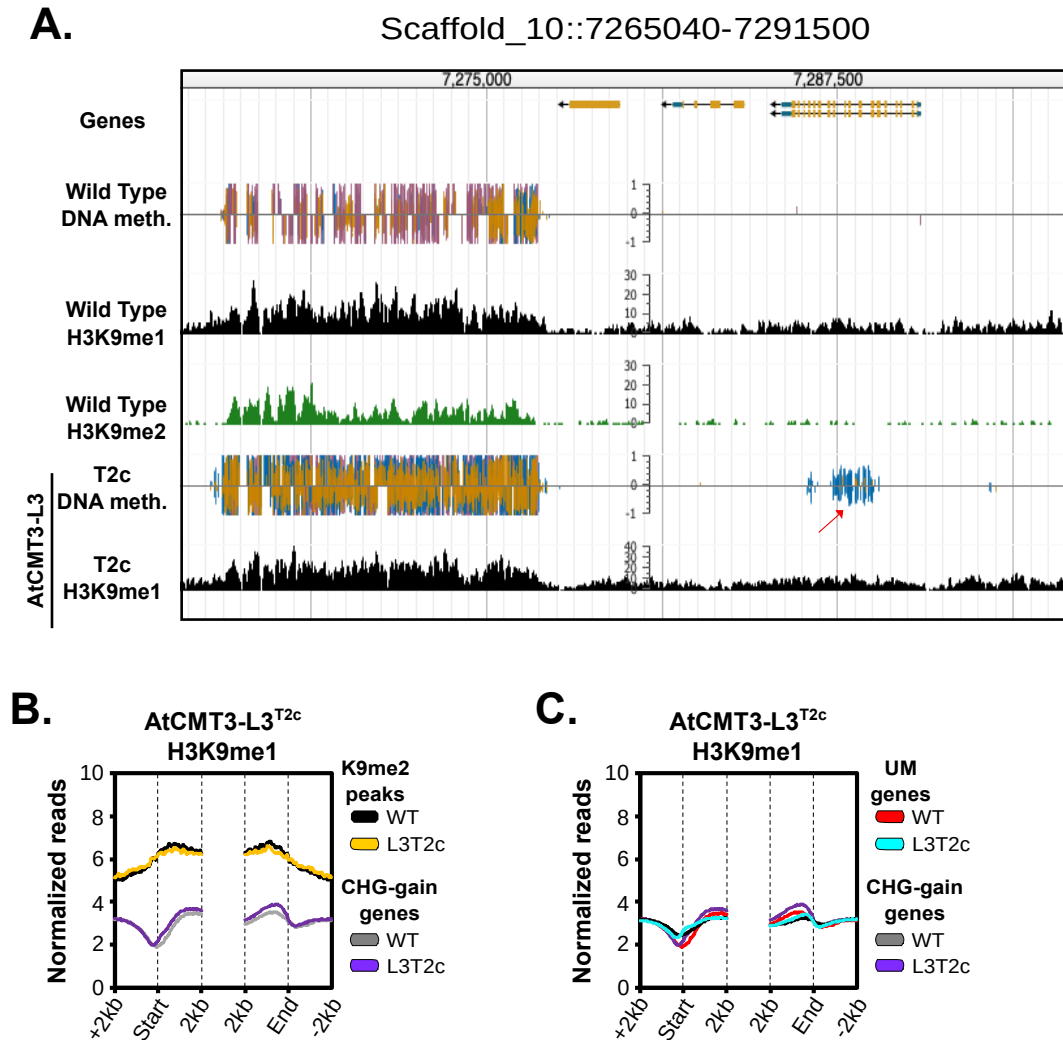


Figure S2.9. Gains in CHG methylation do not alter H3K9me1 levels or distribution.

A. Genome browser view of CHG methylation levels and H3K9me1 ChIP sequencing levels in the T2c generation of the AtCMT3-L3 lineage. In wild type and AtCMT3-L3^{T2c}, H3K9me1 can be found enriched in regions also enriched for H3K9me2. The arrow indicates a gene body that gains CHG methylation in AtCMT3-L3 but does not show H3K9me1 enrichment. Scales on methylation tracks designate the weighted percent methylation, with 1 = 100% on the top strand and -1 = 100% on the bottom strand. Scales on the H3K9me1 tracks indicate the number of mapped reads and are not normalized. Cytosine methylation is divided into CG (red), CHG (blue), and CHH (yellow) sequence contexts. **B.** Metaplot of H3K9me1 ChIP-sequencing enrichment over H3K9me2 ChIP peaks defined in wild type plants and over CHG-gain genes in AtCMT3-L3^{T2c}. **C.** Metaplot of H3K9me1 ChIP-sequencing enrichment comparing enrichment over CHG-gain genes in AtCMT3-L3^{T2c} to genes that do not gain CHG methylation (UM genes).

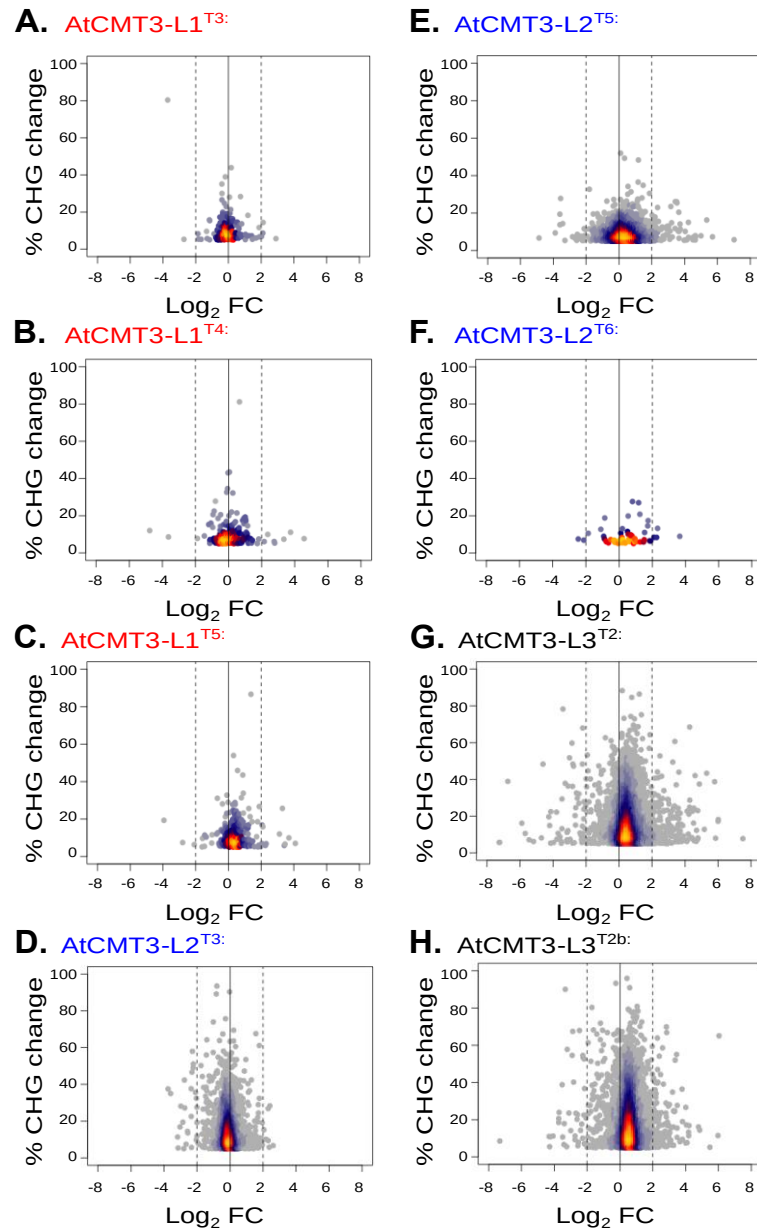


Figure S2.10. Changes in expression are not related to CHG methylation levels.

Relationship between fold change in expression relative to wild type and CHG methylation levels for genes identified as gaining a minimum of 5% CHG methylation in each respective line. Lines analyzed include those with both RNA sequencing and whole genome bisulfite sequencing data: **A.** AtCMT3-L1^{T3}; **B.** AtCMT3-L1^{T4}; **C.** AtCMT3-L1^{T5}; **D.** AtCMT3-L2^{T3}; **E.** AtCMT3-L2^{T5}; **F.** AtCMT3-L2^{T6}; **G.** AtCMT3-L3^{T2}; **H.** AtCMT3-L3^{T2b}. Genes with zero FPKM values were removed from the analysis.

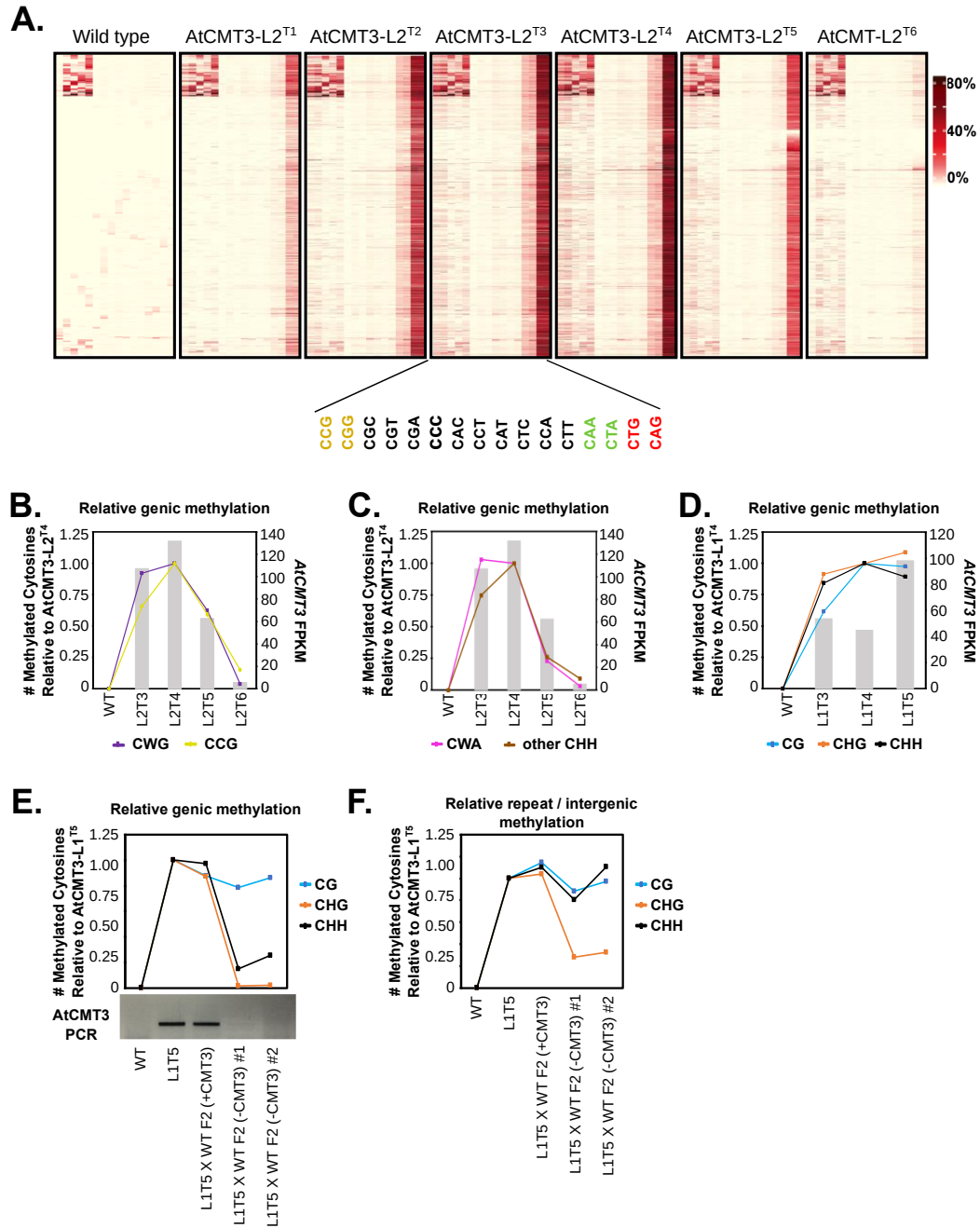


Figure S2.11. CHG methylation over gene bodies is associated with gains in non-CHG methylation.

A. Heatmap of % methylation levels across CHG hyper-DMRs overlapping the CHG-gain genes in AtCMT3-L2^{T4} divided into all 16 possible trinucleotide contexts. **B-C.** Assessment of the relationship of genic cytosine methylation to *AtCMT3* expression across AtCMT3-L2 generations. Line plots show the number of methylated cytosines in **B.** CWG vs. CCG contexts and **C.** CWA vs.

other CHH contexts relative to *AtCMT3-L2*^{T4} across CHG DMRs overlapping CHG-gain genes defined in *AtCMT3-L2*^{T4}. Bar plots show the expression of the *AtCMT3* transgene. See also Figure 2.5A-B. **D.** Assessment of the relationship of genic cytosine methylation to *AtCMT3* expression across *AtCMT3-L1* generations. Line plots show the number of methylated cytosines in each context relative to *AtCMT3-L1*^{T4} across CHG DMRs overlapping CHG-gain genes identified in *AtCMT3-L1*^{T4}. Bar plots show the expression of the *AtCMT3* transgene. **E-F.** Assessment of the number of methylated cytosines relative to *AtCMT3-L1*^{T5} across hyper CHG DMRs identified in *AtCMT3-L1*^{T5} that overlap CHG-gain genes identified in *AtCMT3-L1*^{T5} (**E**) or repeats or intergenic regions (**F**). *AtCMT3-L1*^{T5} was crossed to wild type (non-transgenic) and 3 F2 progeny were assessed: one progeny that retained the transgene (*L1T5XWT* F2 (+CMT3)) and two where the transgene segregated out (*L1T5XWT* F2 (-CMT3) #1 and #2). The gel image below the line plot is the result of PCR conducted on genomic DNA from each line with primers to detect the *AtCMT3* transgene.

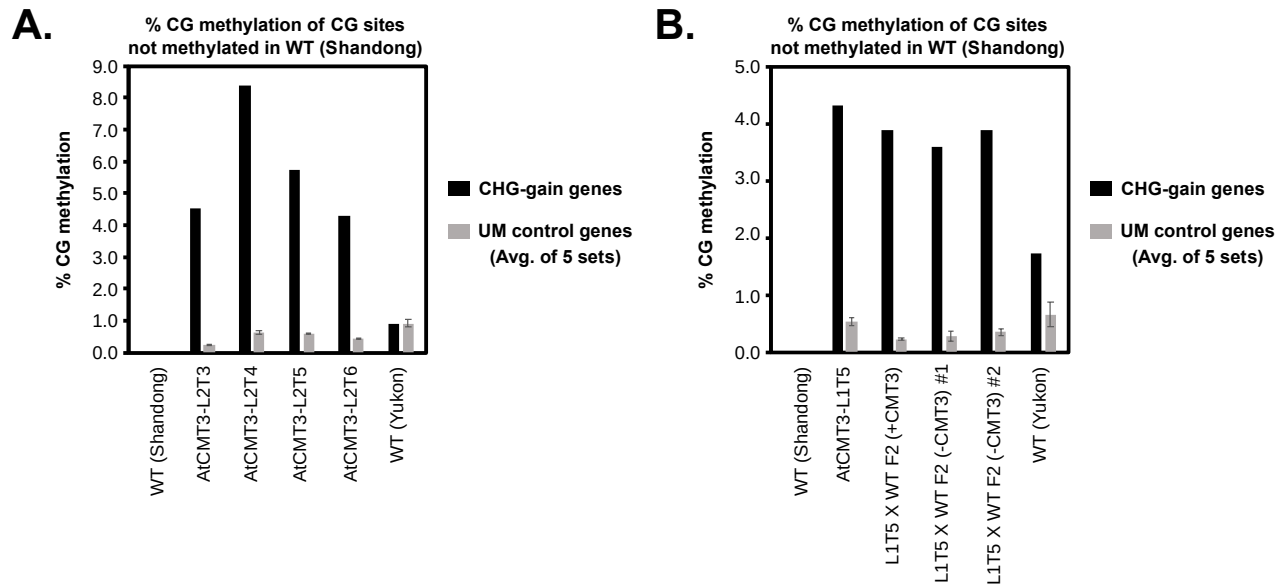


Figure S2.12. AtCMT3-induced genic CG methylation is maintained at higher levels than background following loss of AtCMT3 expression.

A. Comparison of the % methylation of CG sites that were not found to be methylated in the non-transgenic wild type (Shandong ecotype) parent of AtCMT3 transgenic lines. Shown is the % CG methylation calculated in each lineage across hyper CHG DMRs identified in AtCMT3-L2^{T4} that overlap CHG-gain genes identified in AtCMT3-L2^{T4} (black bars; same regions assessed in Figure 5B) compared to the average % CG methylation of an equal amount of sequence space extracted from 5 randomly chosen sets of unmethylated genes that did not gain CHG methylation in AtCMT3-L2^{T4}. The number of genes chosen in each set was equal to the number of AtCMT3-L2^{T4} CHG gain genes. Note that AtCMT3-L2T5 and T6 exhibited silencing of the *AtCMT3* transgene, yet still maintained CG methylation levels higher than that detected on unmethylated genes. Methylation over the same regions was also assessed in additional, non-transgenic *E. salsugineum* accession (Yukon) to demonstrate that the levels of CG methylation over CHG gain genes in transgenic lines were unlikely to have occurred independently of AtCMT3. **B.** Comparison of the % methylation of CG sites that were not found to be methylated in the non-transgenic wild type (Shandong ecotype) parent of AtCMT3 transgenic lines. Shown is the % CG methylation calculated in each lineage across hyper CHG DMRs identified in AtCMT3-L1^{T5} that overlap CHG-gain genes identified in AtCMT3-L1^{T5} (black bars; same regions assessed in Figure S11E) compared to the average % CG methylation of an equal amount of sequence space extracted from 5 randomly chosen sets of unmethylated genes that did not gain CHG methylation in AtCMT3-L1^{T5}. The number of genes chosen in each set was equal to the number of AtCMT3-L1^{T5} CHG gain genes. AtCMT3-L1^{T5} was crossed to wild type (non-transgenic) and 3 F2 progeny were assessed: one progeny that retained the transgene (L1T5XWT F2 (+CMT3)) and two where the transgene segregated out (L1T5XWT F2 (-CMT3) #1 and #2). Note that the F2 progeny where the transgene segregated out still maintained CG methylation levels higher than that detected on unmethylated genes. As in (A), methylation over the same regions was also assessed in additional, non-transgenic *E. salsugineum* accession (Yukon) to demonstrate that the levels of CG methylation over CHG gain genes in transgenic lines were unlikely to have occurred independently of AtCMT3.

CHAPTER 3

NATURAL VARIATION IN DNA METHYLATION HOMEOSTASIS AND THE
EMERGENCE OF EPIALLELES ¹

¹ Yinwen Zhang, Jered M. Wendte, Lexiang Ji, Robert J. Schmitz. Proceedings of the National Academy of Sciences, 2020, 117(9): 4874-4884.

Reprinted here with permission of the publisher.

ABSTRACT

In plants and mammals, DNA methylation plays a critical role in transcriptional silencing by delineating heterochromatin from transcriptionally active euchromatin. A homeostatic balance between heterochromatin and euchromatin is essential to genomic stability. This is evident in many diseases and mutants for heterochromatin maintenance, which are characterized by global losses of DNA methylation coupled with localized ectopic gains of DNA methylation that alter transcription. Furthermore, we have shown that genome-wide methylation patterns in *Arabidopsis thaliana* are highly stable over generations, with the exception of rare epialleles. However, the extent to which natural variation in the robustness of targeting DNA methylation to heterochromatin exists, and the phenotypic consequences of such variation, remain to be fully explored. Here we describe the finding that both heterochromatin and genic DNA methylation are highly variable among 725 *A. thaliana* accessions. We found genic DNA methylation is inversely correlated with that in heterochromatin, suggesting that certain methylation pathway(s) may be redirected to genes upon the loss of heterochromatin. This redistribution likely involves a feedback loop involving the DNA methyltransferase, CHROMOMETHYLASE 3 (CMT3), H3K9me2 and histone turnover, as highly expressed, long genes with a high density of CMT3-preferred CWG sites are more likely to be methylated. Importantly, although the presence of CG methylation in genes alone may not affect transcription, genes containing CG methylation are more likely to become methylated at non-CG sites and silenced. These findings are consistent with the hypothesis that natural variation in DNA methylation homeostasis may underlie the evolution of epialleles that alter phenotypes.

INTRODUCTION

Heterochromatin is abundant in eukaryotic genomes and it is important for the transcriptional silencing of repeats and transposons (TEs), as well as for centromere function [171]. Heterochromatin is composed of more tightly condensed chromatin compared to euchromatin. At the molecular level, heterochromatin is typically demarcated by Histone H3 lysine 9 methylation (H3K9me) and there are a variety of enzymes that “read” this covalent histone modification to help establish and maintain heterochromatin [172-175]. In plant and mammalian genomes, heterochromatin is associated with an additional chromatin modification, DNA cytosine methylation, which in plants is established in a feedback loop with H3K9 methylation [7, 13, 18, 19, 23, 24, 52, 175, 176]. Loss of maintenance of heterochromatin leads to genome instability, resulting in numerous phenotypic consequences. In humans, disease progression in many cancer types is characterized by global losses of DNA methylation, accompanied by ectopic gains of DNA methylation in CpG islands that may silence transcription of tumor suppressors [120]. Similarly, in the model plant, *Arabidopsis thaliana*, mutations in the nucleosome remodeler, *DECREASE IN DNA METHYLATION 1 (DDM1)*, result in a global decrease in DNA methylation in heterochromatin and a redistribution of DNA methylation to genes that results in pleiotropic developmental defects [106, 177]. Therefore, it is critical to maintain chromatin homeostasis, or proper targeting of chromatin modifications that demarcate heterochromatin from euchromatin, through both mitotic and meiotic cell divisions. Consistent with this, methylated regions were found to be faithfully propagated genome-wide over multiple generations in *A. thaliana* with 99.998% accuracy [96]. Yet, despite this demonstrated stability in an individual accession of *A. thaliana*, DNA methylation has been widely characterized both within and between plant species and found to be highly variable in genome-wide levels and distribution [75, 178, 179]. This

variation is due, at least in part, to genetic variation in genes encoding the machinery responsible for the maintenance and targeting of DNA methylation [46, 118, 178-181]. As a whole, however, epigenetic diversity may be more generally conceptualized as consequence of natural, population-level variation in chromatin homeostasis, the extent, causes, and phenotypic consequences of which remain to be fully explored.

Most Angiosperm (flowering) plants studied to date encode multiple, functionally distinct DNA methyltransferase enzymes that combined methylate cytosines in all sequence contexts, including CG, CHG, and CHH (H = A, C, or T) [7, 46, 52, 179]. Methylation in all contexts is characteristic of silenced repeats in constitutive heterochromatin [9, 69]. Protein-coding genes located in euchromatin can be characterized by DNA methylation as well, though the patterns of methylation can vary. Many constitutively expressed genes in flowering plant genomes are characterized by strictly CG context methylation, which is referred to as gene body methylation (gbM) [115]. Genes can also be characterized by TE-like methylation (teM) that occurs in multiple sequence contexts, and, when teM is also associated with H3K9me₂, teM genes are generally transcriptionally silent. Finally, many genes in flowering plant genomes are essentially devoid of DNA methylation and thus referred to as unmethylated (UM) [Reviewed in: 182].

From an evolutionary standpoint, gbM has garnered much interest, as it is a feature of not just flowering plants, but many animal species as well, suggesting it may represent an ancestral state conserved over long periods of evolutionary time [44, 128, 129, 183]. Curiously, it is absent from fungal genomes that contain cytosine DNA methylation [45]. In general, although there is accumulation of methylation variation at the level of individual cytosines within genes, gbM is faithfully inherited over generational timescales [96, 184, 185]. Furthermore, among flowering plant species, gbM is often found on conserved orthologous genes between species [125, 179, 181,

186-190]. GbM is also consistently associated with genes with characteristic features including constitutive, but moderate expression levels and reduced expression variance between species relative to UM genes [10, 69, 77, 118, 191]. The correlations of gbM with higher but less variable levels of expression have been interpreted as evidence gbM is functionally implicated in gene expression and maintained in populations by natural selection [183, 191, 192]. However, numerous studies of experimentally induced and natural losses of gbM have found no evidence for expression changes directly associated with the loss of gbM, suggesting the correlation between gbM and expression is not a functional relationship [10, 77, 178-180, 184, 187, 189]. Also, although the presence of gbM on some conserved orthologous genes between species is well established, there is wide variation in the percentage of orthologous genes characterized by gbM, both within and between species, with species differences ranging from ~0 to 60% of all genes characterized by gbM in a given species [179].

Mechanistic studies that have identified the molecular processes that underlie the establishment of gbM have provided evidence for an alternative hypothesis for the conservation of gbM, which does not invoke a conserved, universal function [46, 107, 111, 122, 138, 180, 193, 194]. The establishment of gbM requires the activity of the DNA methyltransferase, CHROMOMETHYLASE 3 (CMT3) [122, 180-182, 193, 194]. CMT3 preferentially methylates cytosines in the CHG sequence context (with a preference for CWG relative to CCG, W = A or T) in a self-reinforcing feedback loop with H3K9me2 to maintain constitutive heterochromatin [13, 144, 145, 176, 195, 196]. CMT3 physically binds H3K9me2 to activate methyltransferase activity on nearby cytosines, whereas histone methyltransferases (HMTs) can bind DNA methylation to methylate nearby histones [23, 24, 148, 176, 195]. In addition to heterochromatin, evidence suggests that CMT3 and HMTs are localized to transcribed genes with gbM in euchromatin where

they may transiently establish DNA methylation and H3K9me2 that is characteristic of heterochromatin [107, 193]. This activity can recruit additional methyltransferases that methylate cytosines in CG and CHH contexts as well [107, 193]. However, heterochromatin and transcriptional silencing are not ultimately established at these genic loci due to the activity of the H3K9 de-methylase, INCREASED BONSAI METHYLATION 1 (IBM1), which removes H3K9me2 in a co-transcriptional process [107-109]. IBM1 activity disrupts the feedback loop between CMT3 and H3K9me2, and DNA methylation in all sequence contexts is lost passively following DNA replication, except at CG sites, which is characteristic of gbM [193]. CG methylation is maintained due to the preferential activity of the CG maintenance methyltransferase, METHYLTRANSFERASE 1 (MET1), for hemimethylated CG sites following DNA replication, to which it is recruited to methylate the complementary strand [1, 12, 101, 197].

The mechanism of gbM establishment is consistent with the possibility that gbM may be a passive byproduct related to perturbations to or constraints on chromatin homeostasis that promote the transient off-targeting of the heterochromatin machinery to genes. Indeed, variation of levels of gbM within and between species has been correlated with genome-wide levels of CMT3-preferred CHG methylation, and, in *A. thaliana*, gbM genes have been found to on average be localized closer to dense, pericentromeric heterochromatin relative to UM genes [178, 179, 191]. GbM genes also tend to be longer and have a higher frequency of CMT3-preferred CWG context cytosines relative to UM genes, which may increase the probability of CMT3 activity [180, 193]. Thus, a model has emerged that posits gbM is a byproduct resulting from machinery that facilitates heterochromatin formation. The presence of molecular pathways that uncouple DNA methylation from H3K9me2 at PolII transcribed loci, such as IBM1, reduce deleterious consequences of gbM by preventing transcriptional silencing, such that gbM is not eliminated from populations via

selection, but rather maintained passively by maintenance methyltransferases. Under this model, the various patterns of DNA methylation that characterize genes (UM, gbM, or teM) are on a continuous spectrum, such that factors that influence the balance between pathways that promote or remove methylation dictate the methylation state of a given gene (Figure 3.1) [178, 194, 198].

We conducted a comprehensive, within species analysis of gbM in 725 natural *A. thaliana* accessions derived from across the globe to explore the relationship between the maintenance of heterochromatin DNA methylation and gbM [178]. We found that the number of genes characterized by gbM varied widely in these accessions, ranging from ~9 to 20% of all genes, and that some of the variation in the number of gbM genes likely has a genetic basis. We categorized genes based on the conservation of gbM status within *A. thaliana* and found that genic features associated with gbM genes relative to UM genes, including expression characteristics, gene length, and CWG frequency, among others, were generally correlated with gbM status conservation. Using machine learning, we created a model that could utilize genic features alone, independent of methylation data, to accurately predict gbM status, with gene length and CWG frequency being the most important predictors. Importantly, we also experimentally demonstrated genes with conserved gbM status in *A. thaliana* are preferentially methylated by CMT3 relative to other genes and CMT3 is further biased towards genes with pre-existing gbM. Finally, to explore possible phenotypic implications of gbM, we examined the epiallelic states of gbM genes within these accessions. Although the majority alternate epiallelic state of gbM genes was UM, the most conserved gbM genes were more likely than other gbM genes to exist as polyepialleles in the form of teM in these accessions, and epiallelic shifts to teM were at times associated with transcriptional silencing. All together, we interpret these findings to be consistent with a passive model for the conservation of gbM as a trait across Angiosperm species, arising as a byproduct of population-

level variation in DNA methylation homeostasis (Figure 3.1). We further posit that the evolutionary relevance of gbM derives from the predisposition of gbM loci to epiallelic shifts from gbM to teM, which affects transcription and could lead to phenotypic consequences.

RESULTS

Distribution of genic DNA methylation patterns within *A. thaliana*

To gain insights into the within species level variation in gbM, we first classified all genes according to three genic DNA methylation patterns across 725 natural accessions of *A. thaliana*: 1. gbM (strictly CG context methylation); 2. teM (TE-like methylation, or methylation in CHH, CHG, or multiple sequence contexts); and 3. UM (unmethylated) (Dataset S1). Methylation patterns were defined based on cytosine methylation in exons only—see Methods. We found that, similar to prior results [178], the methylation pattern of many genes was highly conserved across accessions, whereas a subset of genes showed variability in their methylation pattern within the population (termed polyepialleles) (Dataset S1, see also representative genome browser views of genes with various epiallelic classes in Figure 3.2A). On average in a given accession, ~79.2% of genes are classified as UM, ~3.6% genes are classified as teM and ~17.2% are classified as gbM (Figure 3.2B, Dataset S2). However, the number of genes in each category in a given accession showed a wide range of variability across accessions. The number of UM genes ranged from 18,488 to 21,605 (68.9% to 80.5% of all genes), the number of teM genes ranged from 593 to 1,668 (2.2% to 6.2% of all genes), and the number of gbM genes ranged from 2,428 to 5,221 (9.1% to 19.5% of all genes) (Dataset S2). Thus, although the presence of gbM as a genic methylation pattern is conserved across species [179, 190], there is

considerable variation in the proportion of genes classified as gbM within a single species, similar to results found between species [179, 181].

We next classified all gbM genes according to their epiallele frequency, or the proportion of the population in which a given gene was classified as gbM. Within the accessions analyzed, more than one third of all coding genes (10,940 of 26,834 total genes) were classified as gbM in at least one accession (Dataset S3). For the remaining analyses we refer to these genes as gbM genes. Of these gbM genes, approximately one fifth (2,078) were found to be highly conserved and classified as gbM in more than 90% of the accessions analyzed (hereafter referred to as “core” gbM genes) (Figure 3.2C, Dataset S3). In contrast, approximately half of the gbM genes (4,607) were rare events as they had a gbM frequency of less than 10% (Figure 3.2C, Dataset S3). The remaining gbM genes were fairly evenly distributed across the frequencies between 10%-90% (Figure 3.2C, Dataset S3). In comparison, 10,852 of 26,834 total genes were categorized as UM across accessions and 4,712 of 26,834 total genes were classified as teM genes (Dataset S3). For the remainder of the study, this highly conserved set of unmethylated genes is referred to as the UM genes and utilized for comparison to the gbM genes described above. The teM genes were excluded from further analyses that focused on gbM genes, since the major alternative epiallelic status of gbM genes is UM and not teM (Dataset S4). The remaining genes were not classified due to missing data or low coverage.

Genetic and epigenomic features associated with gbM gene number

There is a wide variation in the number of gbM genes across accessions (Figure 3.2B). Thus, we considered number of gbM genes as a trait and sought to determine whether there was a genetic basis for this variation. First, we performed a SNP based heritability analysis, which

determined that genetic variance explained 10.5% of the variation in gbM gene number. Similarly, mapping the number of gbM genes on to a phylogenetic tree was consistent with gbM gene number being affected by genetic background (Figure 3.3A). We next conducted a genome-wide association analysis utilizing gbM gene number as a trait, but found no clear association signal (*SI Appendix*, Figure S3.1). This is likely indicative that gbM gene number is a multi-factorial trait and the influence of any one genetic factor could vary widely depending on the genetic background. To address this possibility, we refined the analysis to a subset of 198 accessions with a similar mCG level (0.22 ± 0.01) since the number of gbM genes is positively correlated with a gene's mCG level [178]. Within this subpopulation, we identified a QTL [75] on chromosome 5, nucleotides 2,811,722 to 2,890,303 (Figure 3.3B). The association signal reveals a clear peak that passes the threshold of $1E-06$. Candidate genes in the peak region include RUG1 (AT5G08710), a putative regulator of chromosome condensation and an RCC1 family protein (the top SNP is found in the intron: chr5: 2,835,403, $-\log_{10} p=6.40$) [199, 200], and a linker histone H1 and H5 family protein (AT5G08780, the top SNP is a missense variant, Lys to Asn: chr5: 2857876, $-\log_{10} p=5.23$) [201]. Intriguingly, similar, but not identical, genes encoding an RCC1 family protein and a histone H1 family protein were identified in GWAS that considered levels of gbM, rather than number of gbM genes [118].

Because gbM status and the number of CWG sites within a gene are correlated [180], we also considered that genetic variation resulting in changes in CWG frequency could influence gbM gene number across accessions. However, we found no relationship between the total number of CWG sites within genes and the gbM gene number across accessions (*SI Appendix*, Figure S3.2A). Another possible cause of variation in gbM gene number is the expression level of *CMT3*, which is thought to facilitate the origins of gbM [180, 182, 193]. However, we found little correlation

between *CMT3* expression and gbM gene number across natural accessions of *A. thaliana* (Figure 3.3C). Correspondingly, variation of *CMT3* expression also showed little correlation with whole genome CHG methylation levels (Dataset S5 and *SI Appendix*, Figure S3.2B). These results may be consistent with the possibility that, given a uniform genetic background, altering expression of *CMT3* could influence gbM gene number and CHG methylation levels, but there are likely many other factors influencing this trait at the population level.

Finally, we considered the possibility that there may be a relationship between whole genome CHG methylation and gbM gene number. This was tested using phylogenetically independent contrasts (PIC), a method which removes the phylogenetic relationships of accessions to account for genetic relatedness. Comparing genome-wide CHG levels and gbM gene numbers across accessions revealed that these two traits are negatively correlated (Pearson correlation = 0.453, $p = 4.034\text{E-}32$) (Figure 3.3D). Importantly, the significance of this correlation remained whether whole genome CHG methylation levels were estimated from reads mapped to the reference genome or independently of mapping, suggesting this variation was not significantly influenced by variation in repetitive regions between accessions (*SI Appendix*, Figure S3.2C-D). Since the majority of CHG methylation co-occurs with H3K9me2 in dense pericentromeric heterochromatin [18], and if we assume that gbM is an indirect readout of CMT3 activity on genes, as is suggested by other studies [180, 182, 193], this result is consistent with the conclusion that there is a tradeoff between CMT3 activity in heterochromatin and genic loci. This trade-off is not necessarily related to the intrinsic enzymatic activity of CMT3, but rather the robustness of properly targeting CMT3 to heterochromatin.

Genic features are predictive of gbM status and correlate with epiallele frequency

GbM genes are generally distinguished from UM genes by being characterized as having longer gene lengths, more moderate, but on average higher expression levels, a greater number of associated transcripts, lower substitution rates (dN/dS), a higher frequency of CMT3-preferred CWG (W = A or T) context cytosines, and a lower frequency of MET1-preferred CG context cytosines [109, 180, 190, 202]. Thus, we next sought to determine the relationship between these features and the epiallele frequency of gbM genes in *A. thaliana* and to identify the extent to which these genomic features were predictive of gbM status. The gbM genes were divided into ten groups based on gbM frequency within *A. thaliana*. Next, pairwise comparisons between adjacent groups based on their gbM frequencies were completed for each genic feature, with gbM loci with the lowest gbM frequency (0-0.1 in the population) compared to UM genes (Figure 3.4A-F and Dataset S6). The results revealed that all genic features show general trends relative to gbM frequency. This was most pronounced with gene length and CWG frequency, which demonstrated the most significant differences in pairwise comparisons, with higher gbM frequencies associated with higher values for each feature (Figure 3.4A-B). For CG frequency, dN/dS and transcript number, significant differences between pairwise comparisons were limited to genes with lower epiallele frequencies and UM genes (Figure 3.4C-E). Finally, for gene expression levels, no significant differences were detected between pairwise comparisons, but a general trend towards more moderate and slightly higher expression was evident as gbM frequency increased (Figure 3.4F).

Recent work has demonstrated that the establishment of gbM is initiated by CMT3 [180, 193], which mainly methylates cytosines in dense, pericentromeric heterochromatin [18, 24]. We therefore hypothesized that location on the chromosome relative to pericentromeric

heterochromatin may be related to a gene's gbM status, as a correlation was noted previously using methylation data from microarrays [191]. We first compared the distance to the centromere of gbM genes relative to UM genes using seven *A. thaliana* accessions, including the reference accession, Col-0 (note that gene positions were based on the reference genome, but methylation status was identified individually in each accession). Results demonstrated that on all or nearly all five chromosomes in each accession, gbM genes were significantly closer, on average, than UM genes to the centromere (*SI Appendix*, Figure S3.3 and Dataset S7). A notable exception was Chromosome 3, in which gbM genes were found to be significantly closer to the centromere than UM genes in only two out of the seven accessions (*SI Appendix*, Figure S3.3 and Dataset S7). Next, we examined the relationship between the distance to the centromere and the epiallele frequency of gbM genes. Similar to other genic features, a significant difference in distance to the centromere was found in the pairwise comparison between gbM genes with the lowest gbM frequency and UM genes, yet no additional significant differences were detected between the additional gbM frequency groups (Figure 3.4G). These results further demonstrate that gbM is correlated with multiple genic features, which are also related to the gbM frequency of a locus in *A. thaliana*.

We next sought to determine whether genic properties could be utilized to predict the gbM status of a gene. To do so, we combined the 10,940 gbM genes identified in *A. thaliana* with the 10,852 UM genes and randomly divided the genes into two equal groups that maintain the same proportion of gbM genes and UM genes. We utilized one group as a training data set to build a machine learning binary prediction model based on the seven genic features described above, which we then used to predict the gbM vs UM status of the remaining test group (*SI Appendix*, Figure S3.4A). We evaluated machine learning prediction models trained using 12 different

machine learning algorithms. The random forest algorithm was found to have the highest prediction accuracy by cross-validation of the training data set (*SI Appendix*, Figure S3.4B) and was chosen to predict gene's methylation status in the test group.

For all genes in the test group, the methylation status of ~79% of the genes could be accurately predicted, which is greater than 50% accuracy that would be expected by chance (Figure 3.4H). The sensitivity of the model was also high, with ~77% of all gbM genes correctly identified (Figure 3.4H). The sensitivity of the model increased with increasing gbM frequency, and genes with a gbM frequency of more than 90% could be predicted with ~97% accuracy (Figure 3.4I and Dataset S8). In contrast, for genes with less than 10% gbM frequency, the prediction accuracy was ~59%, and there was much more uncertainty associated with the prediction (the probability of random prediction for these genes was 50.2%) (Figure 3.4I and Dataset S8). Within the model, the various genic features differed in their importance for predicting gbM status (see methods for how importance was determined). Most important was gene length, followed by CWG and CG frequency, which showed similar levels of importance (Figure 3.4J). Next, in order of importance, was dN/dS, relative distance to the centromere, expression level, and, finally, transcript number was the least important factor (Figure 3.4J). Collectively these results show that several genomic features are correlated with gbM and frequency of gbM in *A. thaliana* and can be used to accurately predict the gbM status of a given gene.

Genes with high gbM frequencies are preferentially methylated by CMT3

Given the role of CMT3 in the establishment and maintenance of gbM [180, 182, 193], we hypothesized that the genic features associated with gbM status and gbM frequency would also be predictive of CMT3 targeting. To test this hypothesis, we utilized data from previously described

experimental systems that promote persistent CMT3 targeting to genes [177, 193] to determine whether CMT3 is preferentially targeted to genes with high gbM frequencies and associated genic features.

The first experimental system we examined was *A. thaliana ddm1* mutants [177]. DDM1 is a nucleosome remodeler and it is required for DNA methylation on heterochromatin [103, 203, 204]. Mutation of *ddm1* results in a hypomethylation of heterochromatin and a previous study discovered that propagation of *ddm1* mutants for nine generations resulted in a genome-wide redistribution of DNA methylation to euchromatin [177]. These results support a model whereby a reduction in the ability to maintain heterochromatin can lead to widespread ectopic off-targeting to genes of the heterochromatin maintenance machinery. This mutant phenotype is reminiscent of our finding of a negative correlation between genome-wide CHG methylation levels (a proxy for heterochromatin) and gbM gene numbers, which could be a symptom of variation in the ability to maintain homeostatic targeting of DNA methylation across accessions. Therefore, we used this *ddm1* mutant data to test the hypothesis that disruption to maintenance of heterochromatin could lead to preferential CMT3-mediated ectopic CHG methylation of high frequency gbM genes. We observed the gradual accumulation of CHG methylation on 3,372 genes over nine generations (see Methods). Enrichment tests of these 3,372 genes showed that 2,435 genes are classified as gbM genes ($p = 0$, Fisher's exact test) (Figure 3.5A Dataset S9). Importantly, these genes are specifically enriched for genes with high gbM frequencies in the population (>0.9) ($p = 0$, Fisher's exact test) (Figure 3.5A Dataset S9), suggesting that genic features predictive of gbM status are also predictive of CMT3 targeting.

An important caveat of the *ddm1* system is that gbM genes that gained CHG methylation possessed pre-existing gbM which could influence CMT3-targeting via a self-reinforcing feedback

loop between cytosine methylation and H3K9me2 [176], as H3K9 methyltransferases are recruited to targets via their SRA-domain which binds cytosine methylation [23, 24]. Therefore, we also examined a second experimental system, *Eutrema salsugineum*, which is an Angiosperm species that has naturally lost the gene encoding CMT3 and gbM, and thus provides a model to examine the targeting to CMT3 to genes independent of pre-existing gbM [180, 193]. Heterologous expression of *CMT3* in *E. salsugineum* results in CHG methylation on a subset of genes, and it was previously shown that this subset of genes was significantly enriched in orthologs of gbM genes in *A. thaliana*, Col-0 accession [193]. We further expanded this analysis and identified 4,076 *A. thaliana* orthologs of the genes that gain CHG methylation in *CMT3*-expressing *E. salsugineum* and examined their gbM status across all 725 *A. thaliana* accessions (Dataset S10). Of these, 2,706 were classified as gbM in at least one *A. thaliana* accession, which is significantly more than expected by chance ($p = 5.50 \times 10^{-162}$, Fisher's exact test) (Figure 3.5B, Dataset S9). These gbM loci were also specifically enriched in the core gbM genes ($p = 5.57 \times 10^{-99}$, Fisher's exact test) (Figure 3.5B, Dataset S9). Thus, even without prior gbM, CMT3 preferentially methylates orthologs of gbM loci in *A. thaliana*, with a significant enrichment found at the core gbM genes, again suggesting there are intrinsic features of these genes that promote CMT3 targeting.

GbM genes are more susceptible to epiallelic shifts that lead to transcriptional silencing

Similar to proposals by others, we hypothesized that gbM loci and teM loci might represent a continuous spectrum of chromatin states (Figure 3.1) [71, 137], resulting from the balance between factors that promote or disrupt DNA methylation homeostasis [198]. Factors that push this balance to one end of the spectrum or to the other may therefore explain the formation of some naturally occurring epialleles, such as those previously described to be dependent on *CMT3* [205-

207]. We reasoned that the core gbM genes, which are preferentially methylated by CMT3, may be more prone to transitions to teM status characteristic of heterochromatin, which is promoted by CMT3. Indeed, genes with the highest gbM frequency showed a significantly higher teM frequency than genes with lower gbM frequencies (Figure 3.6A). In addition, the number of gbM genes and teM genes are positively correlated in the population (*SI Appendix*, Figure S3.5A), further supporting the conclusion that gbM and teM-status are on a continuous spectrum.

We next sought to more carefully examine examples of genes that exist as a teM gene in certain accessions but are otherwise categorized as a gbM gene in the majority of accessions (>0.9, or the core gbM genes), as these could provide clues to mechanisms underlying polyepiallelic states. We identified several accessions that possessed greater than one hundred teM genes that were otherwise classified as core gbM genes across all accessions (Dataset S11). We focused on the three accessions with the highest number of transitions from core gbM to teM methylation states, Cnt-1, Monte-1 and UKSE06-533 (Figure 3.6B, *SI Appendix*, Figure S3.5B and Dataset S11), and sought to identify possible causative factors for these epimutations. Previous studies have demonstrated that mutations in the H3K9me2 demethylase, *IBM1*, results in gbM to teM transitions that affect transcription [107]. Functional expression of *IBM1* requires the presence of intronic methylation, the removal of which promotes expression of a truncated, non-functional transcript [208]. Interestingly, *IBM1* intronic methylation in Cnt-1, Monte-1 and UKSE06-533 is reduced compared to Col-0, which we predicted may preclude proper accumulation of functional *IBM1* transcripts in these lines (Figure 3.6C). By conducting qRT-PCR to detect truncated (non-functional) vs full-length (functional) *IBM1* transcripts in these accessions, we found that ratio of non-functional vs. functional transcripts was higher in Cnt-1 and Monte-1 compared to Col-0 (Figure 3.6D). The ratio was also slightly higher in UKSE06-533, although not significantly

higher. Also consistent with loss of IBM1 function in Cnt-1 and Monte-1, genes gaining CHG methylation also gained methylation in additional contexts, including CG and demonstrated reduced expression (Figure 3.6E-H). Importantly, Cnt-1 and Monte-1 are located on different branches of the phylogenetic tree (*SI Appendix*, Figure S3.5C), suggesting this reduction of methylation in the intron of *IBM1* happened independently in multiple lineages. Although alternative mechanisms are possible, especially in the case of UKSE06-533, these results suggest that the ectopic, non-CG DNA hypermethylation on genes in these accessions may be caused by abnormal processing of *IBM1* transcripts, which promotes epiallelic transitions from gbM to teM preferentially on core gbM genes.

DISCUSSION

The comprehensive analysis of gbM in *A. thaliana* revealed gbM to be a highly variable trait in this species. We found that, within the accessions tested, 10,940 of 26,834 total genes (~41% of coding genes) were classified as gbM in at least one accession, yet only ~2,078 genes (~8% of coding genes) were found to be characterized by gbM in the majority (>90%) of the accessions. These data are consistent with other within and between species comparisons of the conservation of gbM on orthologous genes, which also found a wide range of variation [77, 118, 178, 179, 186, 189]. We interpret these results to be consistent with the hypothesis that gbM is the result of a passive process related to the variation in the robustness of the maintenance of heterochromatin. Overall, our analyses suggest the establishment and maintenance of gbM is likely a multi-factorial trait and is likely the sum result of genetic and / or environmental factors that influence DNA methylation homeostasis, including but not limited to:

1. the activity and sub-genomic targeting accuracy of DNA methyltransferases and other enzymes involved in establishing heterochromatin-associated modifications exclusively to sequences that require heterochromatinization.
2. the efficiency with which cells can sequester genomic regions targeted for silencing into higher order structures associated with dense heterochromatin that distinguish them from transcriptionally active regions.
3. genic features that expose genes to targeting by the heterochromatin machinery, including especially gene length and frequency of CMT3-preferred CWG context cytosines, but also expression levels and proximity to TEs and pericentromeric heterochromatin.
4. the efficiency of the co-transcriptional machinery, including IBM1, that removes H3K9me2 from genes.

The interplay between these factors dictates the spectrum of loci that are gbM and may have increased susceptibility to epiallelic shifts to teM, which can alter expression and thus have phenotypic consequences (Figure 3.1). Under this model, what are the evolutionary implications of gbM as a trait? As gbM loci are generally housekeeping genes without a consensus molecular function, most alterations to transcription would be expected to have negative consequences. In support of this, *A. thaliana* mutant backgrounds (e.g. *ibm1* mutants) that result in gbM to teM shifts have altered transcriptional states associated with pleiotropic developmental defects [107, 108]. Yet, it is also clear that polyepialleles are present within this species (Figure 3.5B) [82, 89, 178, 205, 206, 209-211]. In well-studied examples of genes that exist as polyepialleles, polyepiallelic states have been found to be associated with genomic rearrangements that position genes closer to TEs or with genes that have undergone duplications [82, 178, 205, 206]. In the case of *HISN6* (a histidine biosynthesis gene) and *TAD3* (a transfer RNA deaminase), which are essential genes that

have undergone duplications in some accessions, the duplication event is associated with CMT3-dependent methylation and silencing of one of the paralogs [205, 206]. In this case silencing of one paralog may be beneficial in correcting for gene dosage. However, these epialleles may also contribute to speciation, as they have been demonstrated to contribute to hybrid incompatibility in crosses to accessions where the gene is not duplicated and inheritance of the silent copy is lethal [205, 206]. Thus, there are clear precedents for polyepialleles potentially affecting fitness. However, it is also important to consider that it is currently unclear whether an epiallele can act as a substrate for natural selection and change frequencies in populations over time, as epialleles are also prone to reversions from silent to expressed states [205, 206, 212, 213]. Given the current data, it is parsimonious to conclude that, as a whole, gbM, and the associated susceptibility to epiallelic switching associated with silencing, is a mildly deleterious condition present more or less in populations due to trade-offs with selective pressures related to the maintenance of DNA methylation homeostasis, and associated constraints to this process imposed by environmental factors and genome architecture. Thus, variation in the robustness of targeting heterochromatin DNA methylation is likely a symptom of the broader phenomena of population level variation of chromatin homeostasis, which may contribute to a diverse array of epigenetic phenomena, including epigenetic drivers of disease.

MATERIALS AND METHODS

Data acquisition

Methylomes, transcriptomes and genomic variants (SNPs) of 725 accessions used in this analysis was obtained from published datasets of 1001 *A. thaliana* database (<http://signal.salk.edu/1001.php>) [178, 214] and reanalyzed.

The methylomes of ninth generation of *ddm1* mutants including four lines of *ddm1* mutants that were independently self-pollinated eight times (9G *ddm1*) [177] were obtained and reanalyzed. The list of CHG-gain genes that showed a minimum of 5% increase in CHG methylation in *E. salsugineum* AtCMT3-L2 expressing lines relative to wild type, was obtained from [193], and orthologs gene pairs between *E. salsugineum* and *A. thaliana* were those identified in [179].

Methylome mapping

WGBS data were processed using “single-end-pipeline” function from Methylypy as described in [158]. Quality-filtering and adapter-trimming were performed using cutadapt v1.9.dev1 [215]. Qualified reads were aligned to the *A. thaliana* TAIR10 reference genome [203] (downloaded from: <https://phytozome.jgi.doe.gov>) using bowtie 2.2.4 [212]. Only uniquely aligned and non-clonal reads were retained. Chloroplast DNA (which is fully unmethylated) was used as a control to calculate the sodium bisulfite reaction non-conversion rate of unmodified cytosines. A binomial test was used to determine the methylation status of cytosines with a minimum coverage of three reads.

RNA-seq mapping

Quality-filtering and adapter-trimming were performed using Trimmomatic v0.33 with default parameters [164]. Qualified reads were aligned to the *A. thaliana* TAIR10 reference genome using HISAT2 v2.0.5 [165]. Gene expression (FPKM) values were computed using StringTie v1.3.3b [166]. Genes with zero FPKM values among all investigated accessions were removed from expression analyses.

Gene methylation status classification

The coding genes for each of 725 accessions were classified as gbM, teM or UM by applying a binomial test to the number of methylated sites in a gene, similar to the methods described in [202]. To classify genes as gbM, teM, or UM, the total number of cytosines and the methylated cytosines were counted for cytosines in each context (CG, CHG, and CHH) for the coding sequences (CDS) of the primary transcript for each gene. The percentage of methylated sites for each sequence context in all coding regions from each accession were used as the background probability of having methylation on a single site. Given a background probability and the total number of cytosines and methylated cytosines, a p-value was calculated using a binomial distribution to show the cumulative probability of having higher number of methylated cytosines on a given gene. Then a q-values were calculated by adjusting p values by Benjamin–Hochberg FDR to control the false discovery rate.

Genes were classified as gbM if they had reads mapping to at least 20 CG sites and had a q-value < 0.05 for mCG and a q-value > 0.05 for mCHG and mCHH. Genes were classified as mCHG if they had reads mapping to at least 20 CHGs, a mCHG q-value < 0.05 , and a mCHH q-value > 0.05 . As mCG is commonly associated with mCHG, the q-value for mCG could be significant or insignificant in mCHG genes. Genes were classified as mCHH if they had reads mapping to at least 20 mCHH sites and a mCHH q-value < 0.05 . Q-values for mCG and mCHG could be anything as both types of methylation are associated with mCHH. mCHG and mCHH genes were collectively referred to as teM genes. Genes were classified as unmethylated (UM) if they had reads mapping to at least 20 mCHH sites and had a q-value > 0.05 for all sequence contexts. For other cases, genes were classified into a low coverage category, since they did not

meet the minimum requirement of number of reads mapping to the sequence context illustrated above.

To characterize the difference between gbM genes and UM genes, a gene's methylation status was further summarized based on its status among the 725 accessions. For each gene, the percentage of four methylation classes (UM, teM, gbM and low coverage) were summarized over 725 accessions. Genes were defined as UM genes if they are UM in over 90% of the accessions, without having been identified as gbM or teM in the remaining accessions. A gene was defined as gbM gene if it is gbM in at least one accession, unless there are more accessions with teM than gbM or the proportion of low coverage accessions was more than 50%. All gbM genes and UM genes defined in this way were used for the following gene feature comparisons and training of prediction models.

Gene features preparation

Gene length, exon length, exon number and number of transcripts for all coding genes were obtained directly from the *A. thaliana* reference protein-coding gene annotation file in the GFF format (TAIR11) [216]. Number of CWG (CAG or CTG) and CG (CGG or CGT or CGC or CGA) sites were calculated by scanning a gene's primary transcript sequence with a three base window and step size of one base. Then, CWG and CG site frequencies were calculated by normalizing the number of each context to a gene's primary transcript length. Gene's relative location to the pericentromere was obtained from calculating the absolute distance between a gene's location on the chromosome and the location of pericentromere, which was obtained from *A. thaliana* genome assembly annotation file (<https://www.arabidopsis.org/download/>). Gene expression levels used in

the gene feature comparison between gbM genes and UM genes as shown in Figure 3.3 were the averaged FPKM values across all investigated 725 accessions for each gene.

Substitution rates were calculated between CDS pairs between *A. thaliana* and *A. lyrata*. Reciprocal best BLAST with an e-value cutoff of $\leq 1e-08$ was used to identify orthologs between *A. thaliana* and *A. lyrata*. Individual CDS pairs were aligned using PRANK, and Gblocks was applied to eliminate poorly aligned positions in an alignment with a cutoff of eight contiguous non-conserved positions and no gap positions allowed. The yn00 package in the program PAML for pairwise sequence comparison was used to estimate synonymous substitution rates, non-synonymous substitution rates, and adaptive evolution (dS, dN and w respectively). Differences in gene features between UM genes and genes with varied gbM epiallele frequencies (the proportion of gbM accessions) was assessed using the Welch two sample t-test statistic in R, except for transcripts number which was tested using a Chi square goodness of fit test as this was a discrete datatype.

Prediction model for gene's methylation status

A total of 10 features (Dataset S6) that had shown different distributions between gbM and UM genes from previous study [190, 202] and our preliminary analysis were selected to train a prediction model of a gene's methylation status. To build prediction models for a gene's methylation status, a correlation matrix was created from the 10 genic features listed in Dataset S6, and three of them (exon number, number of CWG/CG sites on the coding sequences) were removed before the model training process, since they showed an absolute correlation of 0.6 or higher with gene length. Before model training, the features were transformed to normalized data by subtracting the mean for each feature and dividing by the standard deviation. Among all gbM

and UM genes, training and testing datasets were created by random sampling of 50% of the genes into a training set balancing the proportion of gbM genes and UM genes based on the original gene set. A dozen machine learning algorithms (Figure S3.4B) were used to train the prediction models. Five default values were tried for the main parameters of each algorithm during the training process and values with the best performance were chosen in the final evaluation. However, it should be noted that all algorithms performed well (Figure S3.4B).

The accuracy of each algorithm is determined by a 10-fold cross-validation method. The 10-fold cross-validation method involves randomly splitting the training dataset into 10 equal sized subsets. Of the 10 subsets, a single subset was retained as the validation data for testing the model, and the remaining 9 subsets were used as training data. the cross-validation process was then repeated 10 times, with each of the 10 subsets used exactly once as the validation data. The process of randomly splitting the training dataset into 10 subsets was repeated 5 times. Together, the final model accuracy was taken as the mean of 50 validations. Since the model trained by the random forest algorithm showed the highest prediction accuracy, it was applied to predict a gene's methylation status in the testing set. The importance of features was estimated during model training. The prediction accuracy for all features was recorded, then testing was performed in the same way by permutating each feature. The difference between the prediction accuracy of the training dataset, including one permuted feature and that of the original dataset, were then averaged over all decision trees generated by the random forest algorithm and normalized by the standard deviation of the differences. The model training and performance evaluation process was done with the support of the caret package (<https://github.com/topepo/caret/>) in R

Additional genome-wide analyses

A neighbor-joining tree of 725 accessions was constructed by using MEGA7 [217] using SNPs [214] from coding sequences. SNPs were filtered using a minor allele frequency cutoff of 0.05 and a minimum data integrity cutoff of 0.5, then the VCF formatted SNPs were converted into sequence alignments readable by MEGA7.

The number of CWG sites in the CDS was calculated by replacing the reference sequence with each accession's genotype and then the number of CWG sites were tabulated based on the replaced CDS for each accession.

Reference mapping-based genome-wide percent methylation was calculated by dividing the total number of aligned methylated reads to the genome by the total number of methylated plus unmethylated reads. To evaluate the accuracy of this estimate, we compared percent methylation values of 17 accessions that possess publicly available reference genomes [218, 219]. We observed that the methylation levels estimated by mapping sequence reads to each accession's own assembly is generally higher than that estimated from mapping to the reference (*SI Appendix*, Figure S3.2C).

To address the possibility of sequence variation among accessions, reads-based genome methylation levels were also calculated by using a nonreference-based DNA methylation predictive model, FASTmC [220]. This method estimates methylation levels directly from WGBS reads. Methylation levels estimated in this way generally show a higher value than the value estimated by alignments to their own assembly, but this method also reveals more variation (*SI Appendix*, Figure S3.2C). Because of computational efficiency, this method was based on random sampling on a subset of 10,000 reads, which may include sampling bias. Regardless, this

method used to estimate genome CHG methylation levels also shows a significant negative correlation with the number of gbM genes within each accession (*SI Appendix*, Figure S3.2D).

Correlations between number of gbM genes, methylation levels, number of CWG sites and *CMT3* gene expression were performed using R and the data were corrected for phylogenetic signals among accessions using a phylogenetically independent contrasts (PIC) method in APE [221]. Because extremely low pair-wise distances (little genetic variance) between accessions will generate outliers after PIC correction, a cutoff of 0.01 pair-wise distance was applied to prune clades with almost no genetic differences, which resulted in a final set of 620 accessions that were used in the correlation tests. This same set of individuals was also used for GWAS where the number of gbM genes was used as the phenotype. GWAS analysis was performed with a compressed mixed linear model [222] implemented in the GAPIT package [223] of R.

SNP-based heritability (h^2) of the trait is the proportion of the total variance ($\sigma_e^2 + \sigma_a^2$) explained by the genetic variance (σ_a^2), σ_e^2 is the residual variance. Both σ_e^2 and σ_a^2 were estimated by maximum likelihood method in the compressed mixed linear model during GWAS analysis.

Genome-wide SNPs that satisfied a minor allele frequency of 0.05 were used for association studies, and the genome-wide threshold was modified using the Bonferroni method. If a significant SNP, which passed a threshold of $1e-5$, lied within 10kb of another significant SNP, they were combined into a block. Using this block as a starting point, all other significant SNPs within 10kb of either end of the block, were further combined into the block. The procedure was repeated until no significant SNPs could be found within 10kb of the block ends. These blocks were referred to as QTLs [75]. Genes that located within the QTL were identified as candidate genes of the trait.

Characterizing genes that are predisposed to mCHG

The methylomes of ninth generation *ddm1* mutants [177] were mapped to reference genome using Methylypy [158] and the methylation status of each gene was determined using a binomial test as described above. Genes that had reads mapping to at least 20 CHGs, a mCHG q -value < 0.05 in ninth generation *ddm1* mutants and mCHG q -value > 0.05 in wild type were identified as mCHG-gain genes. The remaining genes were classified as genes that do not gain mCHG. All coding genes of *A. thaliana* were also categorized into gbM, teM and UM based on their methylation status in *A. thaliana* population as defined above. Then a Fisher's exact test was used to examine the significance of the association between gene's mCHG-gain status and their methylation categories. Similarly, all gbM genes defined from *A. thaliana* population were categorized into different groups with varied gbM epiallele frequency. Those gbM genes were also classified into two groups based on whether they gain mCHG in ninth generation *ddm1* mutants. A Fisher's exact test was used to examine the significance of the association between gene's mCHG-gain status and groups with varied gbM epiallele frequency.

The genes that gain CHG methylation in *E. salsugineum* AtCMT3-L2 expressing lines were obtained from published data [193]. *A. thaliana* orthologs of the genes in *E. salsugineum* were classified into two classes based on whether they gain mCHG in CMT3-expressing *E. salsugineum*. Those *A. thaliana* orthologs were also categorized into gbM, teM and UM based on their methylation status in *A. thaliana* population (Dataset S10). Then a Fisher's exact test was used to examine the significance of the association between gene's mCHG-gain status and their methylation categories. The significance of the association between gene's mCHG-gain status and groups with varied gbM epiallele frequency were also examined as that in *ddm1* mutants.

q-RT PCR

RNA was extracted from cauline leaves of individual plants using Trizol according to the manufacturer's instructions. Synthesis of cDNA was completed using SuperScript III with random hexamers (Invitrogen) according to the manufacturer's instructions. Real time qRT-PCR was conducted using LightCycler 480 SYBR green master mix in a Light Cycler 480 instrument (Roche). Primers used to detect short and long *IBM1* transcripts are those published in [208]. Relative expression of the short (non-functional) to full-length (functional) *IBM1* transcripts was calculated using the double delta threshold cycle (Ct) method [224]. Average Ct values were calculated from three technical replicates.

Analysis of gene expression and CG methylation levels in Cnt-1, Monte-1 and UKSE06-533

For gene expression analysis in Cnt-1, the Cnt-1 library and FPKM values were averaged across all accessions to create a “control library” and used for further comparisons. Only genes with an FPKM >0 in both libraries were retained. Then, Cnt-1 and control libraries were collectively normalized using quantile normalization method to remove global variation across samples [225]. Then the expression level of genes that exist as teM in Cnt-1 but are otherwise categorized as gbM gene in majority of accessions (>0.9 or the core gbM genes) were compared with their gene pairs in the control library. A Wilcoxon signed-rank test was used for comparisons between pairs of genes. In addition to statistical significance, Cohen's *d* was used to assess effect size of differences in gene expression. The same procedure was applied to Monte-1 and UKSE06-533.

For mCG level analysis in Cnt-1, mCG level of core gbM genes in Cnt-1 and the averaged mCG levels across all accessions were normalized using quantile normalization method [225].

Then, among those core gbM genes, genes that exist as teM gene in Cnt-1 were compared with their gene pairs in the control group. Statistical analysis was the same as that in expression comparison. The same procedure was applied to Monte-1 and UKSE06-533.

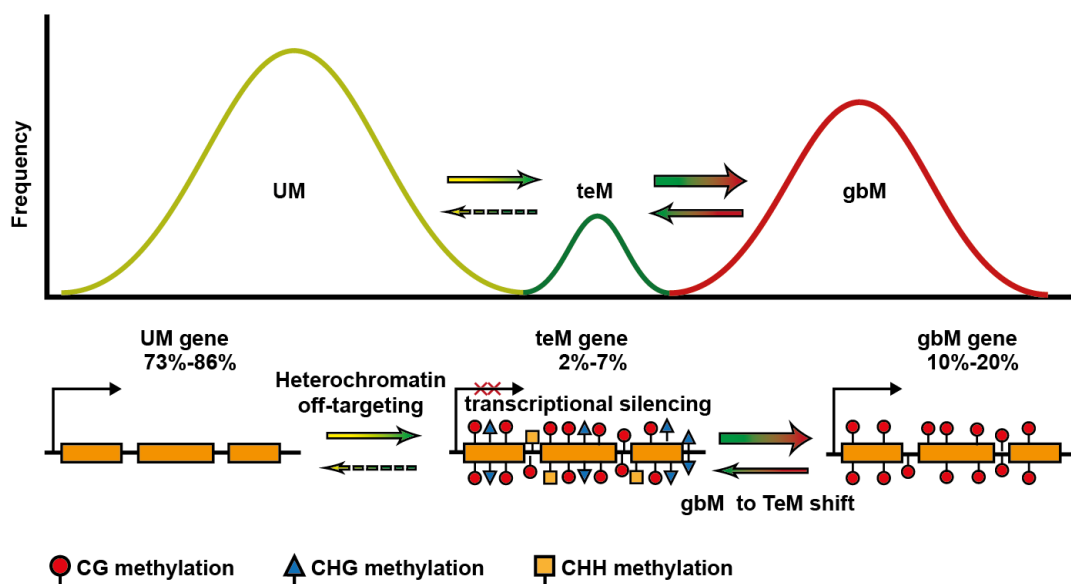


Figure 3.1. Genic DNA methylation patterns represent a continuum of epigenetic states.

In *A. thaliana*, most genes located within euchromatin lack DNA methylation (UM). Some genes are prone to off-targeting by the heterochromatin machinery, which can result in transposable element-like DNA methylation (teM), characterized by cytosine methylation in all sequence contexts, CG, CHG, and CHH (H = A, C, or T), and transcriptional silencing. To combat the negative consequences of off-targeting of DNA methylation to genes, cells encode pathways that disrupt heterochromatin formation at genes by promoting the loss of non-CG methylation. This allows CG methylation to be maintained passively by maintenance methyltransferases, as CG methylation alone is inconsequential for transcription, which results in the strictly CG methylation pattern characteristic of gene body methylation (gbM). GbM then increases the probability of a gbM to teM shift due to feedback loops associated with DNA and histone methylation. The proportion of genes characterized by each methylation state within an individual vary due to factors that influence homeostatic targeting of DNA methylation. Percentages indicate the variation in the proportion of genes classified with each methylation pattern across 725 *A. thaliana* accessions.

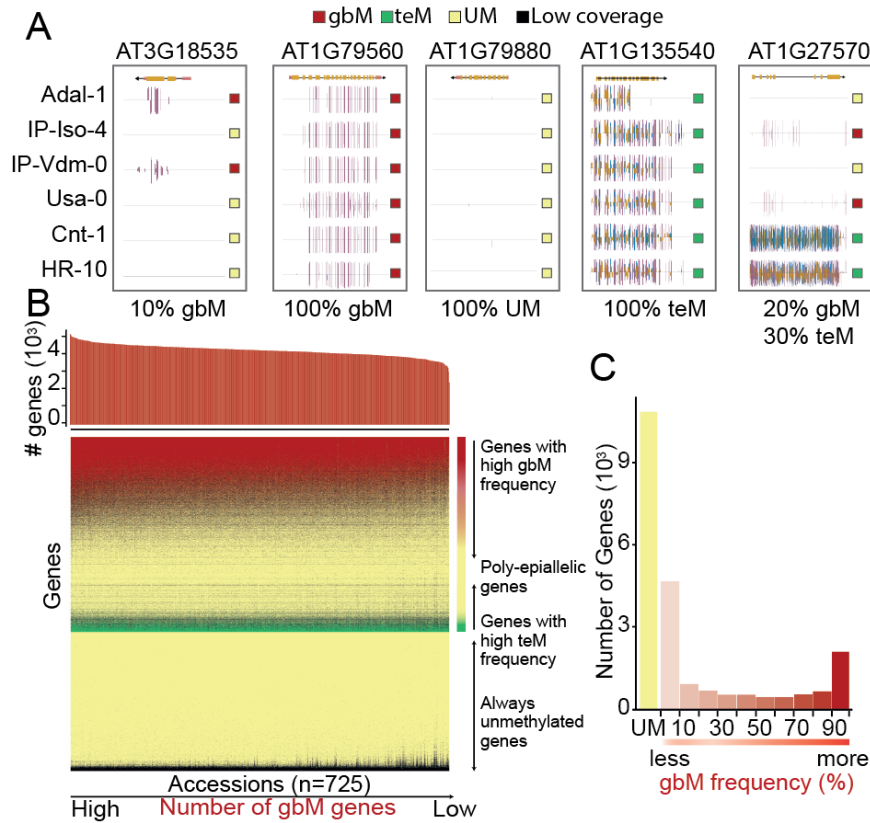


Figure 3.2. Distribution of genic DNA methylation patterns within *A. thaliana*. (A) Genome browser views demonstrating the three genic DNA methylation patterns and distributions within *A. thaliana*, represented by six accessions. Genic methylation patterns detected include: gene body methylation (gbM), which is restricted to mCG only (denoted by a red square); transposable element-like methylation (teM), where methylation is detected in sequence contexts other than or in addition to mCG (mCG and/or mCHG and/or mCHH, denoted by a green square); and unmethylated genes (UM, denoted by a yellow square). The methylation pattern of a given gene can vary within the 725 *A. thaliana* accessions, with some genes predominantly UM, whereas others are rarely gbM in some individuals (first panel). Some genes are predominantly gbM (second panel). Some genes are always UM (third panel). Some genes are always teM (fourth panel). Finally, some genes are variable between all methylation patterns (fifth panel). Within the browser views, mCG is denoted by red lines, mCHG by blue lines, and mCHH by yellow lines. (B) Classification of genes based on DNA methylation patterns within *A. thaliana*. All annotated genes are listed on the y-axis of the heatmap, ordered by the percentage of accessions the genes were classified as a given DNA methylation pattern. Genes were classified as UM (yellow), gbM (red) and teM (green) based on a binominal test. Black indicates genes that were not classified due to low coverage. Accessions are arranged on the x-axis by the number of gbM genes identified, denoted in the histogram in the top panel. (C) Number of genes classified as gbM within *A. thaliana*. Genes identified as gbM in at least one accession were categorized into ten groups based on gbM frequency, or the percentage of accessions a given gene was classified as gbM. The histogram shows the number of genes classified in each category and UM genes are also shown for comparison.

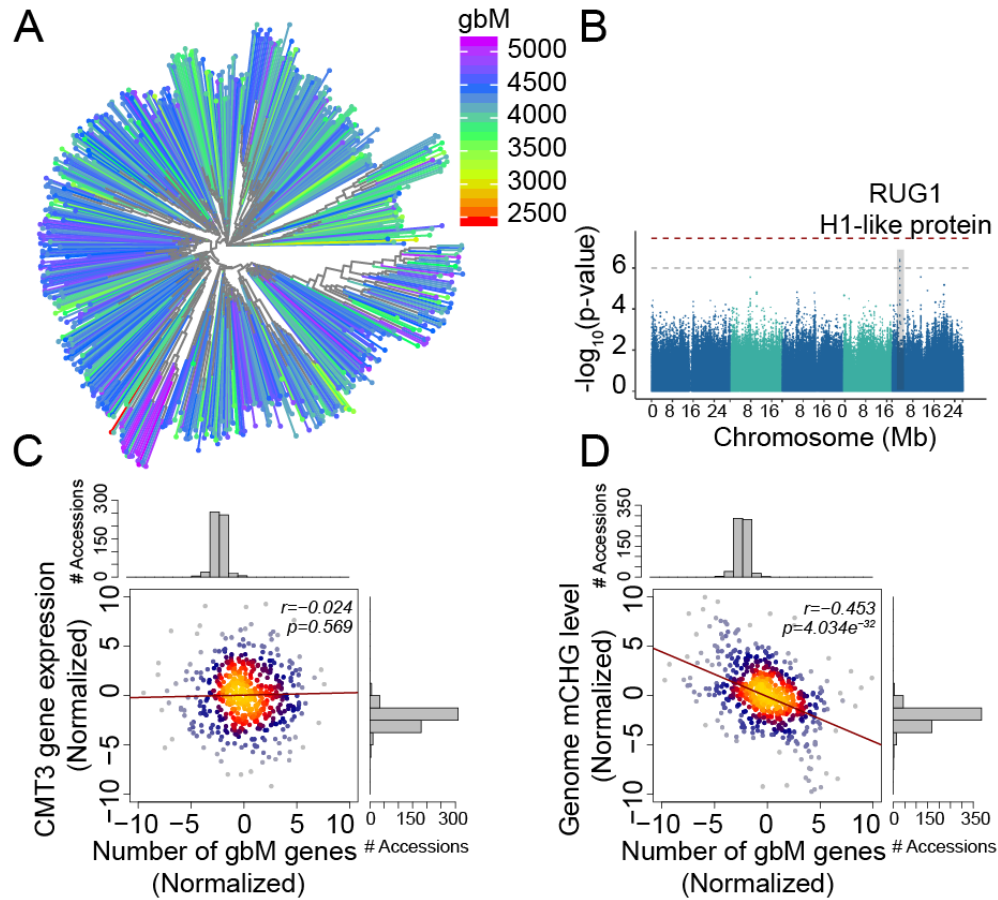


Figure 3.3. Genetic and epigenetic factors associated with gbM gene number. (A) Close genetic relatives have similar gbM gene numbers within *A. thaliana*. The number of gbM genes, indicated by a color gradient, is projected on a Neighbor-Joining tree of 725 *A. thaliana* accessions. (B) Manhattan plots of GWAS of gbM gene number using the CMLM model in the GAPIT package in R. The horizontal gray line indicates a threshold of $1e^{-6}$, and the red line indicates genome-wide threshold $p=0.05$ with a Bonferroni correction. The candidates in the peak region include *RUG1* (AT5G08710), regulator of chromosome condensation *RCC1* family protein (the top SNP is found in the intron: chr5: 2,835,403, $-\log_{10} p=6.40$), and a linker histone H1 and H5 family protein (AT5G08780, the top SNP is a missense variant from Lys to Asn: chr5: 2857876, $-\log_{10} p= 5.23$). (C) The correlation between the number of gbM genes and *CMT3* gene expression levels. The density of samples is represented by brighter colors on the scatter plot. (D) The correlation between the number of gbM genes and the genome-wide mCHG levels estimated by mapping bisulfite sequenced reads to the Col-0 reference genome.

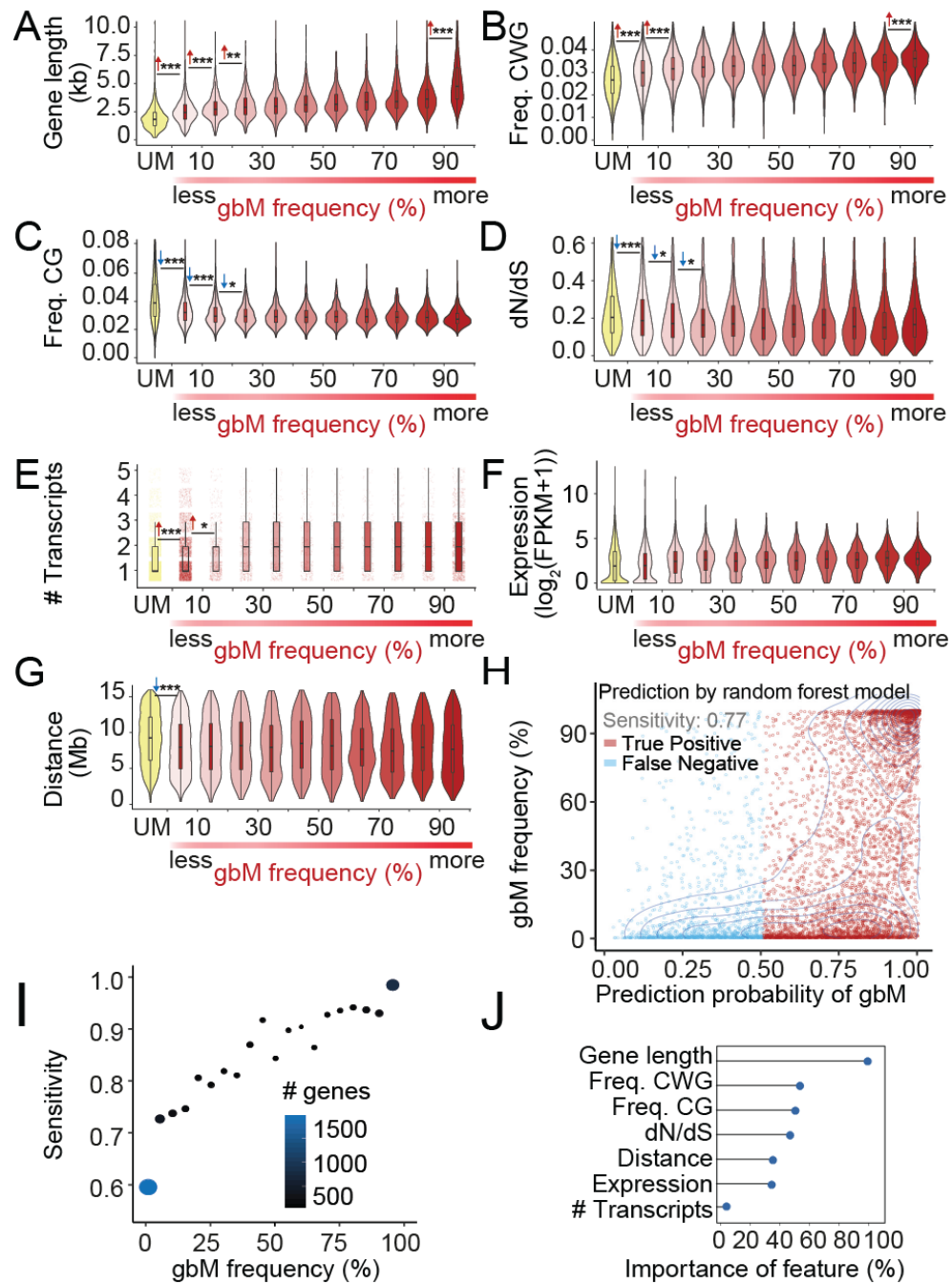


Figure 3.4. Genic features are predictive of gbM status and correlate with gbM frequency.

Figure 3.4 (A-G) Association of seven gene features across the gbM frequency categories. **(A)** gene length; **(B)** frequency of CWGs relative to gene length; **(C)** frequency of CGs relative to gene length; **(D)** dN/dS rate; **(E)** transcript number; **(F)** gene expression level; **(G)** chromosomal location relative to the centromere. The significance of differences between neighboring groups was calculated using a Welch's Two Sample t-test for most of features with normal distributions except for number of transcripts category, which was tested using a Chi Square goodness of fit test. Significance and direction are indicated. **(H)** Predictability of gbM based on seven gene features. The scatter plot shows the prediction probability of a gene being classified as gbM on the x-axis based on a random forest prediction model and each gene's empirically determined gbM frequency in *A. thaliana* on the y-axis. Red dots show gbM genes that were successfully predicted as gbM, and blue dots show gbM genes that failed to be predicted as gbM. **(I)** Plot of gbM genes divided into 20 groups in 5% intervals based on their gbM frequency (x-axis) relative to the sensitivity of the random forest prediction model (the proportion of successfully predicted gbM genes, y-axis), shows that genes that are classified as gbM in a higher percentage of individuals in the population are more likely to be successfully predicted as gbM based on gene features alone. Dot size and color are proportional to the number of genes in each group. **(J)** Importance ranking of gene features in predicting gbM status shows that gene length and frequency of CWG context cytosines are the most predictive features.

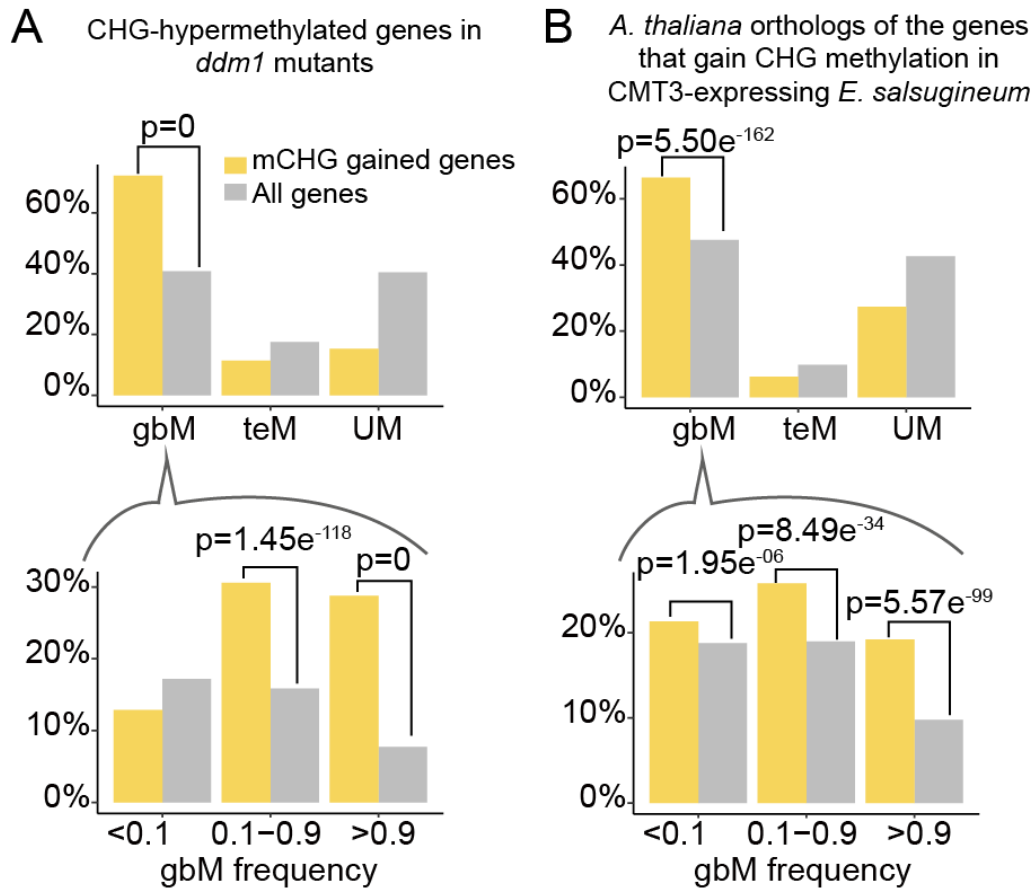


Figure 3.5. Genes with high gbM frequencies are preferentially methylated by CMT3.

(A) Distribution of CHG-hypermethylated genes in ninth generation *ddm1* mutants. The top panel shows that most of the CHG-gain genes in *ddm1* mutants are classified as gbM genes in *A. thaliana*. The bottom panel shows that among gbM genes, CHG-gain genes are specifically enriched for genes with high gbM frequencies in the population. **(B)** Distribution of *A. thaliana* orthologs of the genes that gain CHG methylation following heterologous expression of *CMT3* in *Eutrema salsugineum*, which is a species that has naturally lost *CMT3* and gbM. The top and bottom panels show that CHG-gain genes in the *E. salsugineum* system are enriched for gbM genes, specifically high-frequency gbM genes, as well.

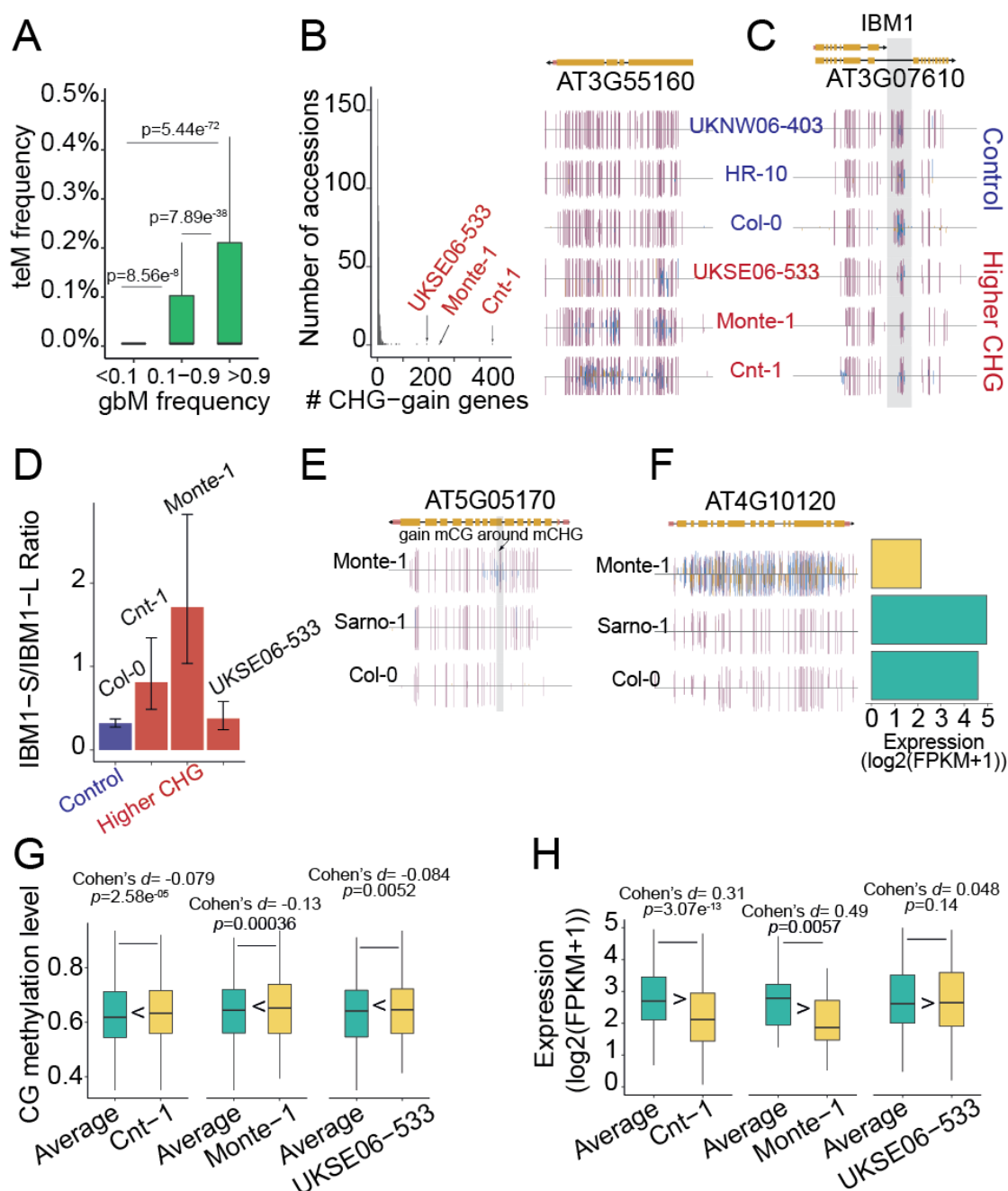


Figure 3.6. GbM genes are more susceptible to epiallelic shifts that lead to transcriptional silencing.

Figure 3.6. (A) Boxplots show the distribution of teM epialleles for gbM genes in different gbM frequency categories. Genes with the highest gbM frequency showed a significantly higher teM frequency compared to genes with lower gbM frequencies. Wilcoxon rank sum test was used for comparisons between groups. (B) The left panel shows a frequency distribution of the number of CHG-gain genes within the core gbM genes (genes with $\geq 90\%$ gbM frequency within the population) in accessions. The top three accessions (Cnt-1, Monte-1 and UKSE06-533) with highest CHG-gain genes are indicated with their name and an arrow. The right panel shows a genome browser view of a representative core gbM gene possessing high CHG methylation in Cnt-1, Monte-1 and UKSE06-533 relative to other accessions shown that do not have these patterns of methylation (UKNW06-403 and HR-10, most closely related to Cnt-1 on the phylogenetic tree) in addition to the reference accession, Col-0, as controls. (C) A genome browser view in the left panel shows that accessions with high CHG methylation within core gbM genes show reduced methylation on the longest intron of *IBM1*, which is known to be important for proper splicing of *IBM1* transcripts. Functional *IBM1* is required to prevent CHG accumulation at gbM loci. (D) A qRT-PCR result which shows that the three highest CHG-gain accessions tend to have a higher proportion of short *IBM1* transcripts compared to functional, full-length transcripts that encode a functional protein. (E) A genome browser view shows a representative core gbM gene with additional CG methylation (highlighted with grey background) around CHG methylation in Monte-1 relative to a control (Sarno-1 is most closely related to Monte-1 on the phylogenetic tree). (F) A genome browser view shows a representative core gbM gene possessing teM status in Monte-1 and lower expression level relative to other accessions. (G) In the first panel, the distribution of CG methylation level of genes that exist as teM genes in Cnt-1 but are otherwise categorized as gbM genes in majority of accessions (shown in yellow bar) were compared with their pairs in a control (green bar), which were the averaged values across all accessions. Significance of the comparison was given by Wilcoxon signed-rank test and effect size of the difference was shown as Cohen's *d*. The same analysis for Monte-1 and UKSE06-533 were shown in the second and third panel respectively. (H) In the first panel, the distribution of expression level of genes that exist as teM genes in Cnt-1 were compared with their pairs in a control group, which were the averaged values across all accessions. The same analysis for Monte-1 and UKSE06-533 are shown in the second and third panel respectively.

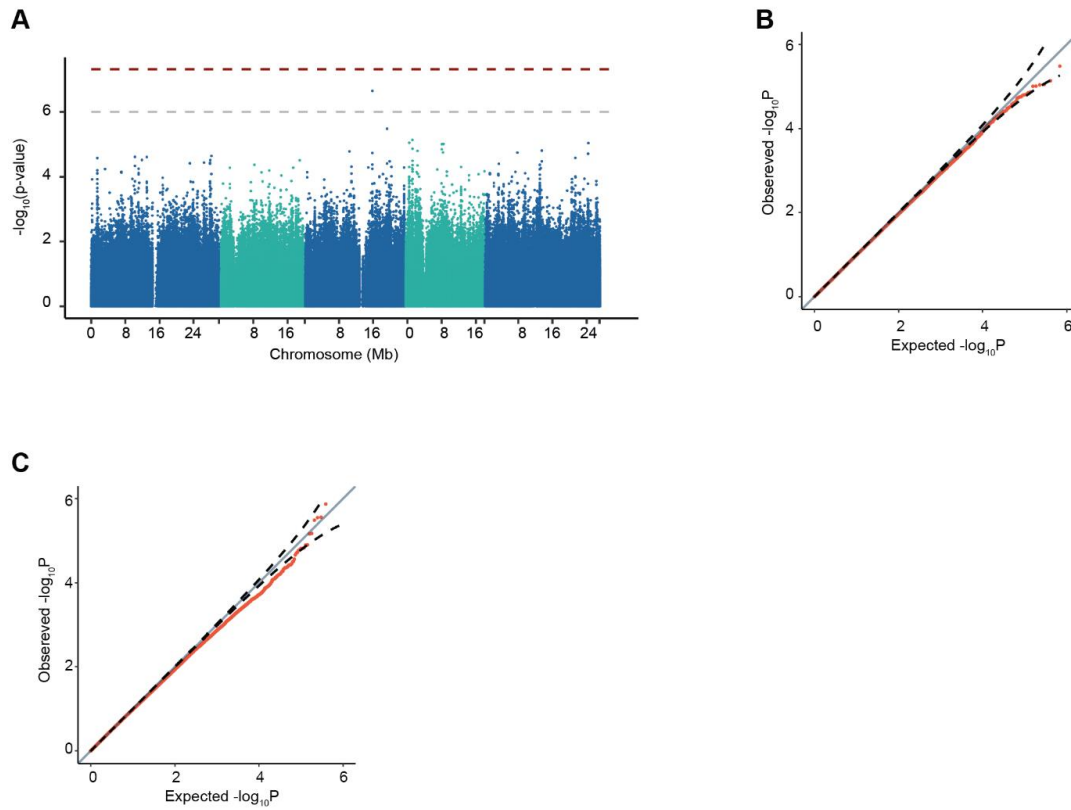


Figure S3.1. GWAS analysis for 620 accessions and corresponding QQ plots. (A) Manhattan plots of GWAS analysis for 620 accessions using gbM gene number as the phenotype. **(B)** QQ plot of GWAS analysis for 620 accessions. **(C)** QQ plot of GWAS analysis for 198 accessions in Figure 3.2B.

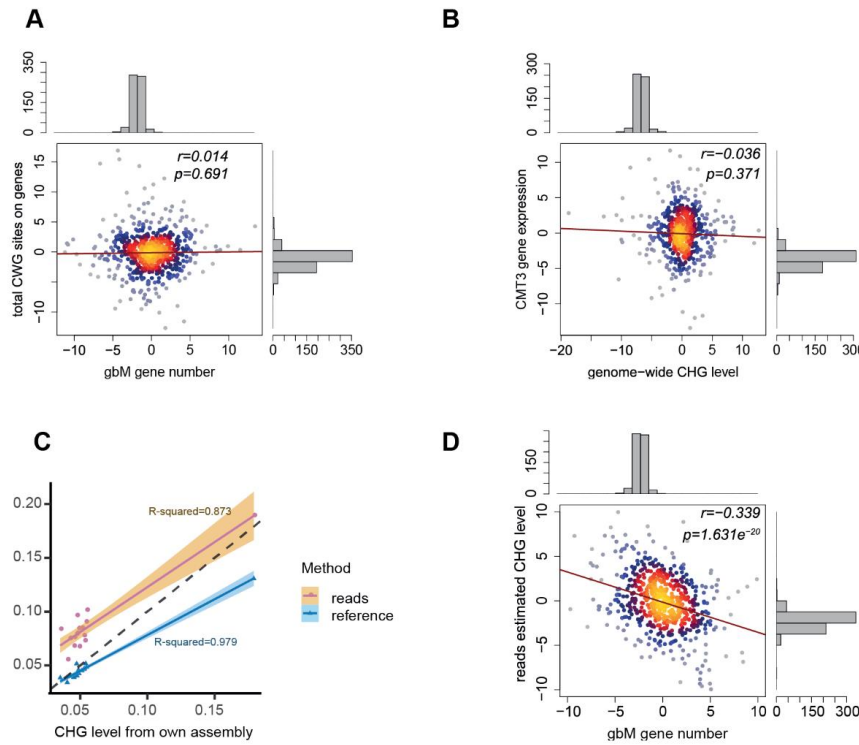


Figure S3.2. Correlations of gbM gene number with various genomic features. (A) The correlation between the number of gbM genes and total CMT3 sites number on coding genes. (B) The correlation between CMT3 gene expression and genome wide percent CHG methylation calculated by a nonreference-based DNA methylation predictive model. (C) The scatter plot shows whole genome CHG methylation levels estimated by the different methods. The x-axis is the whole genome CHG methylation level of 17 accessions estimated from mapping the methylomes to their own genome assemblies. The blue line shows the whole genome CHG methylation level estimated from mapping to the Col-0 reference. The yellow line shows the whole genome CHG methylation levels calculated by a nonreference-based DNA methylation predictive model, FASTmC. (D) The correlation between the number of gbM genes and genome-wide CHG methylation levels calculated by the nonreference-based method.

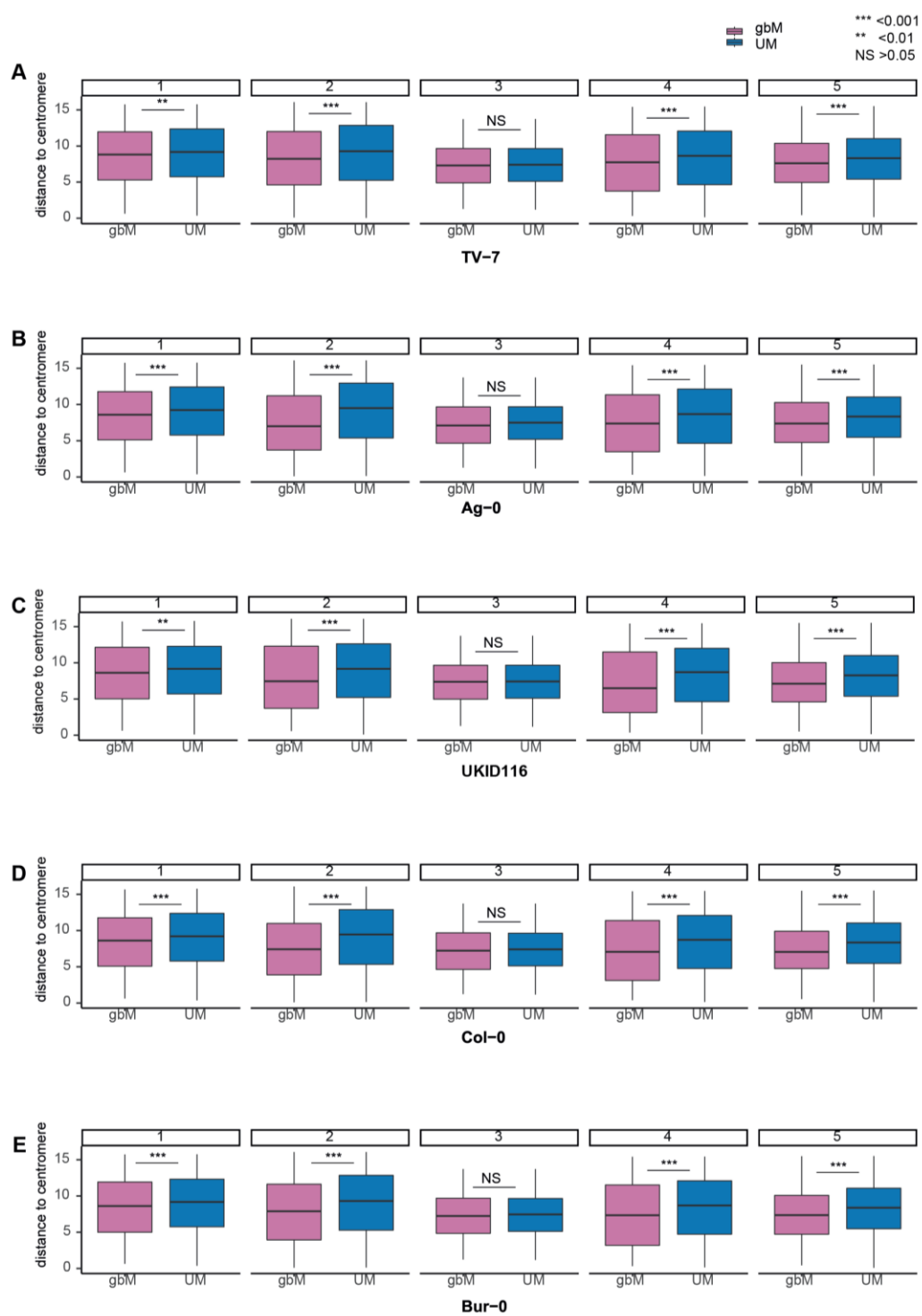


Figure S3.3. Gene location comparison between gbM and UM genes for nature accessions.

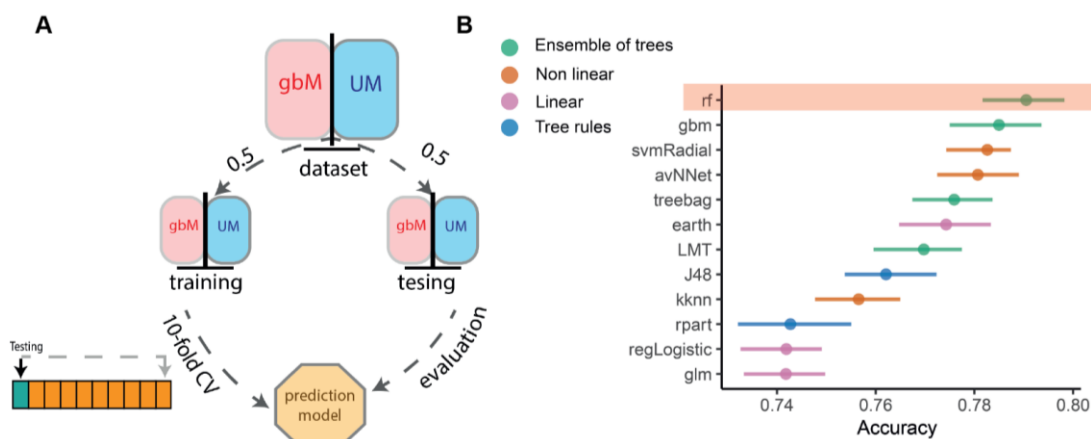


Figure S3.4. Gene methylation status prediction model construction and algorithm selection. (A) The pipeline for the machine learning model data sampling and model training process. **(B)** Rank of 12 machine learning algorithms based on the average prediction accuracy taken from 50 cross-validations during the model training process.

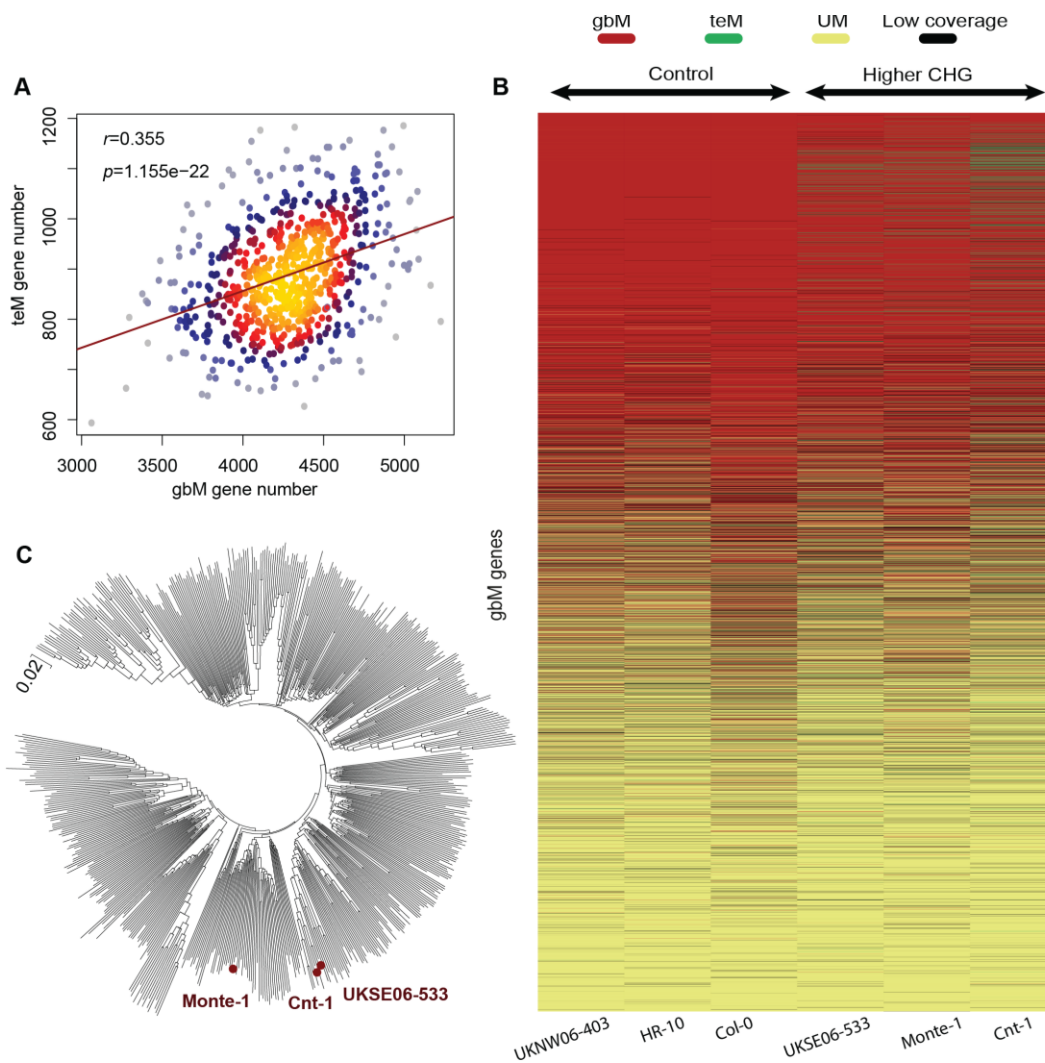


Figure S3.5. gbM and teM-status are on a continuous spectrum. (A) The number of gbM genes and TE-like genes are positively correlated in the population. **(B)** Heatmap shows the methylation status of gbM genes for six accessions including Cnt-1, Monte-1 and UKSE06-533 (top 3 accessions with highest number of genes gain CHG methylation within core gbM genes (with $\geq 90\%$ gbM frequency), and 3 more typical accessions. **(C)** Neighbor-joining tree of all 725 accessions, with the location of Cnt-1 and Monte-1 and UKSE06-533 noted.

CHAPTER 4

HETEROCHROMATIN IS A QUANTITATIVE TRAIT ASSOCIATED WITH
SPONTANEOUS EPIALLELE FORMATION ¹

¹ Yinwen Zhang, Hosung Jang, Rui Xiao, Ioanna Kakoulidou, Robert S. Piecyk, Frank Johannes, Robert J. Schmitz. Nature Communications, 2021, 12(1): 1-13.

Reprinted here with permission of the publisher.

ABSTRACT

Epialleles are meiotically inherited variations in expression states that are not linked to changes in DNA sequence. Although they are well documented to persist in plant genomes, their molecular origins are unknown. Here, we show using a variety of mutant and experimental populations that epialleles in *Arabidopsis thaliana* that result from ectopic hypermethylation are due to feedback regulation of pathways that primarily function to maintain DNA methylation at heterochromatin. Perturbations to maintenance of heterochromatin methylation leads to feedback regulation of DNA methylation in genes, with a preference for genes with pre-existing DNA methylation. Using single base resolution methylomes from epigenetic recombinant inbred lines (epiRIL), we show that epiallelic variation is enriched in euchromatin, yet, associated with QTL primarily located in heterochromatin. Mapping three-dimensional chromatin contacts reveals that genes that are hotspots for ectopic hypermethylation have increased contact frequencies with regions possessing H3K9me2. Altogether, these data show that feedback regulation of pathways that evolved to maintain heterochromatin silencing leads to the origins of spontaneous hypermethylated epialleles.

INTRODUCTION

Genetic variation is the primary driver of phenotypic variation, yet, there are well-documented examples of phenotypes arising independent of changes to the DNA sequence. These variants are referred to as epigenetic alleles (epialleles) and they can be transmitted across generations, especially in plants [226, 227]. However, the extent and the contribution of natural epialleles underlying certain traits is less clear [228]. This lack of knowledge is mostly a

consequence of the challenges to studying epialleles, both in their identification and in demonstrating causality.

The best characterized epialleles causally linked to phenotypes were discovered in plants and most were associated with changes in DNA methylation [84, 90, 229-232]. For example, *colorless non-ripening* (*cnr*) epiallele results from spontaneous formation of DNA methylation in its upstream regulatory region and a decrease in gene expression, resulting in tomato fruits with a mealy discolored skin compared to wild type [90]. Once formed, these particular epialleles are stably inherited in subsequent generations, with rare reversion events observed along the way. The use of whole genome bisulfite sequencing (WGBS) [9, 69] to identify differentially methylated regions has rapidly advanced our understanding of epiallele frequency, stability and molecular nature. For example, WGBS of a population of mutation accumulation lines of *Arabidopsis thaliana* shows that epialleles are rare, enriched in transcribed regions (genes, repeats and transposons) and are relatively stable once formed [79, 88, 96]. Importantly, the single base resolution nature of WGBS also reveals which DNA methylation pathways are most often associated with epiallele formation [9, 69].

In plants, cytosine DNA methylation is present at CG, CHG (H = A, C or T) and CHH sites [52]. Multiple independent pathways/enzymes coordinately reinforce DNA methylation, which ensures its stability through mitotic and meiotic cell divisions. For example, METHYLTRANSFERASE 1 (MET1) is recruited to hemimethylated CGs during S-phase of DNA replication to maintain their methylation [8, 233]. Upon exiting S-phase, CHROMOMETHYLASE2 and 3 (CMT2/3) are recruited to DNA associated with histone 3 lysine 9 dimethylation (H3K9me2) to methylate CWA (W = A or T) and CHG sites, respectively [13, 14, 19, 146]. After the cell exits the cell cycle, the RNA-directed DNA methylation (RdDM)

pathway, which is guided by small RNAs, directs DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) to methylate cytosines in any sequence context [234]. Often, these pathways are found acting upon the same sequences and their coordinated efforts result in the relatively stable DNA methylation patterns that are observed between distinct cell types and over generational timescales in plants [235].

Although epialleles are readily identified and characterized using WGBS to detect losses and gains of DNA methylation, their spontaneous origins are unknown. This has led to use experimental approaches to follow that fate of epialleles and to increase the frequency of epiallele formation to study their molecular origins and their association with traits. For example, silenced epialleles can be induced at a family of *MuDR* transposons in maize by exposing them to a silencing trigger *Muk* [236]. Although *MuDR* elements are silenced after exposure to a silencing trigger, one specific *MuDR* element reverts over time in the absence of the trigger. Other examples of experimental induced epialleles includes the creation of epigenetic recombinant inbred lines (epiRILs) in *Arabidopsis thaliana* [110, 111], whereby RILs are created between wild type and a mutant defective in maintenance of DNA methylation (*met1* or *DECREASE IN DNA METHYLATION 1 – ddm1*) [1, 103]. The epialleles in the epiRIL populations are mostly due to losses or gains of parental methylation states, however, epialleles also form at novel regions that were not present in either parent. The molecular basis for the origins of epialleles is an active area of investigation and there are some clues from previous published studies. For example, the creation of hybrids between *met1* and wild type reveals widespread redistribution of DNA methylation [237, 238]. Wild type chromosomes experienced hypomethylation, which was complemented by greater methylation on the *met1* chromosomes [237]. Moreover, a comparison of the DNA methylomes between 1st generation *ddm1* mutants

and a 9th generation *ddm1* mutants shows increasing reductions of DNA methylation over generations [119]. Unexpectedly, novel epialleles, in the form of ectopic DNA methylation, are abundant in the 9th generation mutants even though maintenance of DNA methylation was unaffected [119]. These data point to a model whereby epigenomes are maintained by some unknown mechanisms that involves some level of feedback regulation in coordination of epigenomic states between chromosomes.

There are multiple examples of feedback regulation of epigenomic states in plants. For example, expression of the DNA demethylase *REPRESSOR OF SILENCING 1 (ROS1)* is sensitive to genome-wide levels of DNA methylation [239]. Similarly, the expression of the full-length transcript encoding an H3K9 demethylase, named *INCREASE IN BONSAI METHYLATION 1 (IBM1)* [107], is sensitive to H3K9me2/DNA methylation levels within an intronic repeat [208]. Additional clues of epigenomic feedback regulation are provided by increasing phenotypic variation with each additional generation of selfing in the *Arabidopsis thaliana* mutants *met1*, *ddm1* and *ibm1* [99, 103, 107, 119].

Evidence for DNA methylation feedback regulation is even observed in natural populations of *Arabidopsis thaliana* where a negative correlation was found between the abundance of CHG methylation and the frequency of gene body DNA methylated (gbM) genes [240]. This leads to a model whereby the abundance of heterochromatin, demarcated by H3K9me2 and CHG methylation, is linked to the origins of epialleles in the form by transitioning unmethylated genes to gbM genes [240]. However, the natural variation in gbM is dependent on a functioning *CMT3* pathway, as without *CMT3*, gbM cannot be established and maintained [116, 193, 241].

Although evidence for DNA methylation feedback regulation between chromosomes and its role in epiallele formation is mounting [242], there is still a lack of mechanistic understanding of this process. How is it that unmethylated sequences become newly methylated spontaneously? What provides the trigger and why are some regions more susceptible for epiallele formation than others? In this study, we follow the fate of epialleles in genes that are either converted from unmethylated to methylated or from gbM to a transposon like methylation (teM) pattern in a variety of experimental mutant lines and populations. We show that loss of *ibm1* leads to ectopic CMT2 and 3 activity in the form of CWA and CHG methylation, respectively, and almost exclusively at gbM genes compared to unmethylated genes. This ectopic activity also occurred, albeit at a lower frequency, when methylation of these gbM genes was erased using a *met1* epiRIL line. This shows the importance of the CMT3-H3K9me2 feedback loop in the establishment of epialleles. This was further evaluated using the *ddm1* epiRIL population where the abundance of heterochromatin was a variable trait in these lines. In total, single base resolution methylomes for 169 *ddm1* epiRILs was produced. Using these new data, we discovered that variation in the genome-wide levels of CHG methylation, which serves as a proxy for heterochromatin, was negatively correlated with the abundance of epiallele formation, most of which was targeted to gbM genes. Methylation QTL analysis further confirmed the importance of heterochromatin for epiallelic variation, as pericentromeres were hotspots for QTL. Collectively, this study demonstrates that a positive feedback loop between H3K9me2 and CMT2/3 is a major contributing factor to the origins of spontaneous epialleles and that heterochromatin is a quantitative trait locus that influences epiallele formation.

RESULTS

Ectopic hypermethylation accumulates in gbM genes in methylation mutants

Phenotypic variation of certain mutants, namely *met1*, *ddm1* and *ibm1*, becomes stronger over repeated generations of self-crossing [99, 103, 107, 119]. To understand the possible reasons for this observation, previous studies examined the DNA methylomes of *ibm1* and *ddm1* separated by up to eight generations of single seed descent growth, respectively [109, 119, 243, 244]. These analyses uncovered non-parental epialleles in the form of ectopic hypermethylation of CHG specifically in genes, which was expected for *ibm1* [108], but was rather unexpected for *ddm1*. To explore the fate of these non-parental epialleles further, we measured CG, CHG and CHH methylation in *ibm1* mutants selfed for one and three generations using previously published data (**Figure 4.1A**). We categorized genes into gene body DNA methylated genes (gbM, N = 5,314) and unmethylated genes (UM, N = 12,684) (**Table S3**). Although it was known that ectopic CHG methylation occurs in genes, our analysis reveals that this ectopic methylation is essentially exclusive to gbM genes (**Figures 4.1A-D**). Over 40% (2,265/5,314) of the gbM genes acquired ectopic CHG methylation in the first-generation mutant, which is highly statistically significant (Fisher's exact test, p-value = 0), compared to less than 1% (31/12,684) of the UM genes (Fisher's exact test, p-value = 1, **Figure 4.1D; Table S4**). Furthermore, the levels of ectopic methylation increased over generations and was also found at CWA sites, which is catalyzed by CMT2 (**Figures 4.1B-D, S4.1A, S4.2A and S4.2B**). Together, these results are consistent with a proposed model whereby loss of IBM1 activity leads to ectopic H3K9me2 specifically in gene bodies of gbM genes, which recruits CMT2 and CMT3 to methylate CWA and CHG sites, respectively [108, 137]. The specificity to gbM versus UM genes is most likely due to ectopic H3K9me2 that results from the activity of the histone lysine nine

methyltransferases, SUV4/5/6, which are targeted to methylated DNA via their SRA domains [245]. The activity of this positive feedback loop specifically at gbM genes is consistent with the increase in DNA methylation and with the increased phenotypic variation of *ibm1* mutants over time.

The increase in phenotypic variation over generational time is much stronger in *ddm1* compared to *ibm1*, although the *ibm1* mutant used has not been propagated for as many generations as *ddm1*. Furthermore, much of the phenotypic variation in *ddm1* is mostly due to the loss of maintenance of DNA methylation over generational time. A similar analysis was performed for data from 9th generation selfed *ddm1*, which revealed a reduction of CG methylation in gbM genes (**Figure 4.1D**) and showed ectopic CHG and CWA in genes (**Figures 4.1B, 4.1E, S4.1B, S4.2C and S4.2D**). However, although there was a greater probability of gbM genes acquiring ectopic methylation (2.5%, 130/5,314, Fisher's exact test, p-value = 8.85×10^{-10}), it was much less than what is observed in *ibm1* and the ectopic methylation was also observed in UM genes albeit at a much lower frequency (0.17%, 22/12,684, Fisher's exact test, p-value = 1) (**Figure 4.1F; Table S4**). Combined, these results show that loss of IBM1 activity immediately and directly affects ectopic methylation of gbM genes almost exclusively, whereas loss of DDM1 activity leads to preferential accumulation of non-parental epialleles via ectopic methylation with a preference for gbM genes compared to UM genes.

Abundance of heterochromatin is associated with the frequency of spontaneous epialleles

DNA methylation feedback regulation between chromosomes is an emerging theme and has major implications for the formation of spontaneous epialleles [242]. One of the best examples of this process is nicely demonstrated using the *ddm1* epiRIL population that had variation in the number of chromosomal regions inherited from the *ddm1* parent versus the wild-

type parent. It was observed that hypomethylated chromosomes from *ddm1* led to ectopic *de novo* methylation at regions inherited from wild type and that the amount of hypermethylation was negatively correlated with the abundance of heterochromatin [119].

The initial observations of the *ddm1* epiRIL population were examined using microarray technology [119], which does not provide the opportunity to evaluate methylation in specific contexts. As a result, WGBS of three epiRILs was used on a subset of three lines to further reveal the type of hypermethylation present [119]. To further expand on these intriguing initial observations, we performed WGBS on 169 individuals from the *ddm1* epiRIL population (**Table S1**) that are unique from the previously studied set of 123 lines using tiling arrays [114]. Using differentially methylated regions (**Table S5**) we were able to create a new haplotype map of each line that differentiated whether a chromosomal segment was inherited from wild type versus *ddm1* (**Figure 4.2A**). Each *ddm1* epiRIL was assigned a hypomethylation index based on the amount of DNA inherited from the *ddm1* parent and the haplotype map (**Figure 4.2B**). The higher the hypomethylation index indicates increased amounts of DNA from the *ddm1* parent is present in the line. Next, we examined ectopic DNA methylation in gbM genes (**Figure S4.3**) and observed an enrichment of ectopic CHG and to a certain extent CWA methylation, which was most apparent in lines with a very high hypomethylation index (**Figure 4.2C**). Collectively, these data support that hypomethylated chromosomes from the *ddm1* parent led to an imbalance in DNA methylation patterns, resulting in ectopic activity of CMT2 and CMT3 at gbM genes.

Heterochromatin impacts spontaneous epiallele formation in trans

To further evaluate the impact to the disruption to normal DNA methylation states, we analyzed genes based on whether they were inherited from wild-type versus *ddm1* derived

haplotypes. Ectopic activity of CMT3 and CHG methylation was observed at gbM genes (**Figure 4.3A**) regardless of which haplotype they were present within and this ectopic activity correlated with the hypomethylation index (**Figures 4.3B, 4.3C and S4.4; Table S6**). Importantly, this ectopic methylation was not due to disruption to maintenance of methylation within the 7th intron of IBM1, which is known to lead to ectopic mCHG in genes (**Figure S4.4**). In total, ectopic CHG methylation was observed at 1,384 and 1,059 genes in wild-type and *ddm1* haplotypes, respectively (**Figure 4.3D; Table S4**). A deeper inspection of these genes revealed a significant enrichment for ectopic activity specifically at gbM genes compared to teM genes and UM genes (which had a strong depletion compared to background expectations – **Figure 4.3E**). Next, we evaluated genes on a case-by-case basis to determine how many of them were susceptible to feedback regulation of DNA methylation. For example, the gene presented in **Figure 4.3A** shows a positive correlation between the change of CHG methylation and the hypomethylation index regardless of which haplotype the gene resided within (**Figure 4.3F**). Therefore, we performed a correlation analysis between the difference in CHG methylation between epiRILs and wild type for each gene that has evidence for ectopic CHG methylation in any line. These results revealed, yet again, a strong preference for gbM genes compared to UM and teM genes (**Figure 4.3G**). Collectively, these results demonstrate that a greater amount of DNA hypomethylation, which is associated with a greater reduction to heterochromatin methylation, leads to ectopic activity of CMT3 in genes, with a clear preference for gbM loci.

Heterochromatin is a hotspot for methylation QTL^{epi}

The fact that genes in both wild-type and *ddm1* haplotypes are almost equally affected indicates that they are regulated in trans (by distant loci). To test this hypothesis, we used a

methylation QTL^{epi} (meQTL^{epi}) approach to explore the potential causal basis for the observed spontaneous hypermethylation in the *ddm1* epiRILs. Traditional genetic markers could not be used given the reduced genetic variations in these lines. Instead, we used differentially methylated regions (DMRs) that segregate in expected patterns across the population (**Table S5**), as markers for QTL analysis [232]. Using the single base resolution methylomes from 169 *ddm1* epiRILs we verified, improved and increased coverage of the existing map compared to previous attempts. We scanned each chromosome for associations with global mCHG, average mCHG of genes and the number of mCHG genes within each epiRIL line (**Figure 4.4A; Table S7**). There were no obvious associations with global levels of mCHG, however, the pericentromeric regions of multiple chromosomes were associated with the average mCHG levels of genes as well as the number of ectopic mCHG regions in genes (**Figure 4.4A**). To study this further, we performed a comprehensive QTL analysis for each of the 1,595 genes that had evidence of ectopic mCHG. In total, 701 of these genes were associated with meQTL^{epi} and the majority (718/975) of the QTL resided in heterochromatin regions (**Figure 4.4B**). A genome-wide view of these events shows that distant (trans) meQTL^{epi} are especially enriched in the pericentromeric regions (**Figures 4.4B and 4.4C; Table S7**). Curiously, chromosome 3 was somewhat devoid of meQTL^{epi} in contrast to other chromosomes, which was unexpected given the presence of well-known repeats such as a 5 kb chloroplast insertion, an rDNA and a telomeric repeat. Of the associations detected, most were linked with a single meQTL^{epi}, although numerous loci that gained mCHG in the gene body were found to be associated with 2-3 meQTL^{epi} (**Figure 4.4D**). Collectively, these data show that disruption to DNA methylation at heterochromatin that was triggered by the initial loss of *ddm1* leads to spontaneous epiallele formation at hundreds of loci across the chromosomes.

H3K9me2 is important to the formation of spontaneous epialleles at unmethylated loci

The ectopic CHG and CWA methylation at gbM genes observed in *ibm1*, *ddm1* and *ddm1* epiRILs demonstrates independent paths for how positive feedback regulation between H3K9me2 and CMT2/CMT3 is established *de novo*. However, understanding the origins of spontaneous ectopic hypermethylation from these experiments is complicated by preexisting DNA methylation at gbM genes. Previous research has shown that ectopic CHG accumulates in gene bodies in *met1*, which is devoid of CG methylation [9]. We re-evaluated *met1* WGBS data and found 1,161 genes that accumulate CHG methylation compared to wild type, with a strong statistical enrichment at gbM genes (**Figure S4.5**). However, loss of *MET1* leads to wide range of disruption to normal nuclear processes that could lead to indirect effects that explain these observations. Therefore, to understand how genes can spontaneously transition from unmethylated to methylated in a wild-type genotype, we took advantage of the *met1* epiRIL population [111]. In the *met1* epiRILs, large segments of chromosomes have lost gbM because they were derived from the *met1* parent where CG methylation is lost. Therefore, using these lines we can compare genes that maintain gbM versus those that were once gbM, but are now UM due to inheritance via *met1*.

This design enables us to test the hypothesis that H3K9 methylation can establish itself *de novo* to recruit CMT2/3 activity to unmethylated genes. To increase the prevalence of H3K9me2, we crossed *ibm1-6* into the *met1* epiRIL-12 and isolated a homozygous *ibm1* line. We produced or used publicly available WGBS data from Col-0, *met1-3*, *ibm1-6*, *met1* epiRIL-12 and *ibm1;met1* epiRIL-12 (**Table S2** - [25, 116]). Using CG methylation patterns, we were able to identify regions within each genotype that were derived from the *met1* versus the wild-type and/or *ibm1-6* parents (**Figures 4.5A and S4.6**). We identified 9.1 Mbs of sequence in

ibm1;met1 epiRIL-12 where 256 gbM genes had lost all CG methylation on chromosome 2 and were homozygous for *ibm1-6* (**Figure 4.5B**). This *met1*-derived region also possessed 625 genes that were UM. We compared *ibm1*-induced ectopic DNA methylation of this collection of 881 genes, all of which have no CG methylation, yet have different histories of CG methylation (either gbM or UM - **Figure 4.5B**). No obvious increase in ectopic CHG or CWA methylation was observed in *ibm1;met1* epiRIL-12 using a meta-analysis of these genes (**Figure 4.5B**). Therefore, we performed a gene specific analysis, which identified a significant enrichment (46/256, Fisher's exact test, $p\text{-value} = 1.81\text{e}^{-16}$) of ectopic CHG and/or CWA methylation in genes that were at one time gbM compared to UM genes (12/625, Fisher's exact test, $p\text{-value} = 1$) (**Figures 4.5C-E; Table S4**). The genes that acquired CHG methylation were on average much longer than genes that did not regardless of their historical gbM versus UM status (**Figure 4.5F**). In summary, these results show that spontaneous epialleles can form in unmethylated genes and that there is a preference for genes that were gbM in previous generations. The reason for this preference is unknown, but it is likely in part due to the length and transcriptional activity known to be associated with gbM genes.

Hotspots for epiallele formation have increased chromatin contacts with H3K9me2 regions

Why certain genes are hotspots for epiallele formation versus others is unknown. One possible explanation for the preferential accumulation of mCHG in certain genes is that they are in nuclear neighborhoods/compartments that have higher concentrations of nucleosomes that possess H3K9me2. This increased localized pool of H3K9me2 could lead to spontaneous and rare incorrect incorporation of these nucleosomes into gene bodies when chromatin is reassembled after DNA replication. Once present in a gene body, CMT3 would bind H3K9me2

to methylate associated DNA [14, 18]. Presumably, H3K9me2 would be removed by IBM1 once the gene is transcribed, but mCHG would still be present. To test the hypothesis, we used Hi-C data [246], which incorporates proximity ligation, to reveal chromatin interactions. We identified all regions that have at least one edge of the chromatin interaction that overlaps with H3K9me2 (**Figure 4.6A**). We discovered that gbM and teM genes are enriched for contacts with H3K9me2, whereas UM genes were not enriched (**Figure 4.6B**). The distance between H3K9me2 regions that contacted gbM and UM genes were not significantly different from one another, yet the teM genes were significantly further away (**Figure 4.6C**). What did distinguish gbM genes that were hotspots for spontaneous epiallele formation (ectopic mCHG) were that they had a greater contact frequency with H3K9me2 regions compared to UM genes (**Figure 4.6D**). These results show that one possible mechanism by which certain loci are susceptible for spontaneous epiallele formation could be through association with H3K9me2 regions of the genome in three-dimensional space.

DISCUSSION

One possible reason for the prevalence of epialleles in plants is that there is no comprehensive erasure of DNA methylation from one generation to the next like there is in mammalian genomes [103, 247-249]. Instead, extensive reinforcement of DNA methylation occurs during gamete production [250], with the exception of rare events important for genomic imprinting. This feature of flowering plants enables the occurrence of rare spontaneous epialleles in a single cell to propagate in subsequent cell divisions and in some cases to become a major allele in the next generation if it is present in a cell that leads to the production of gametes. Therefore, once plant epialleles form, they can be studied to understand their stability over

generations, interaction with alleles with different epigenomic states and association with phenotypes.

Another likely explanation that is consistent with the epigenetic nature of epialleles is the involvement of positive feedback loops. These positive feedback loops are fundamental to maintenance of DNA methylation in plants. Hemimethylated CGs are recognized by VARIANT IN METHYLATION 1 [251], which recruits MET1 to maintain CG methylation. CMT2 and CMT3 are directed to methylate DNA through their binding to regions with H3K9me₂ and H3K9 methyltransferases bind methylated DNA and methylate H3K9 [13, 14, 18, 19, 148, 252, 253]. Similarly, the H3K9 methyl binding protein SAWADEE HOMEDOMAIN HOMOLOG1 [254] and the DNA methylation readers SUVH2/9 recruit the RdDM pathway to target sequences for DNA methylation reinforcement [245]. There are multiple examples in the literature that show the effectiveness of these feedback loops at re-establishing patterns of DNA methylation [235]. For example, multiple proteins have recently been used as triggers in epigenome editing whereby the protein is targeted to the FWA promoter to establish a DNA methylation [245, 255]. Once a positive feedback loop is established the transgene triggers can be segregated away as they are no longer needed for maintenance [245]. Another recent study illustrated how RNAi-independent pathway(s) effectively re-establishes non-CG methylation at transposons that had lost all non-CG methylation due to loss of H3K9me₂ and CMT2/3 activity [256]. Because CG methylation was still present at these transposons, it functioned to recruit SUVH4/5/6 to re-establish H3K9me₂ and non-CG methylation. It is well known how these pathways maintain DNA methylation, but how feedback regulation is established *de novo* at new regions in the genome is challenging to study given their spontaneous nature.

One possibility that is consistent with the spontaneous nature of epialleles is that the feedback regulation important to maintenance of DNA methylation have a low rate of off-targeting activity [253]. Although their main function is to target repeats and transposons for silencing, biochemical features of many of these components, such as domains that recognize methylated DNA (SRA, MBD) and histones (CHROMO, BAH, SAWADEE) leads to improper establishment of these pathways at unintended regions in the genome [253]. Given the function of IBM1, a histone demethylase that removes H3K9me2 from PolII-dependent transcribed regions of the genome, it seems likely that plant epigenomes have evolved mechanisms to reduce off-targeting activity [107, 108, 137, 138]. In this particular case, pre-existing CG methylation at gbM genes serves as a substrate for the SRA domains in SUVH4/5/6 to catalyze H3K9me2 [148], which is counteracted by the activities of IBM1.

Our analysis of the *ibm1;met1 epiRIL-12*, shows that the absence of IBM1 does result in ectopic CMT3 activity at completely unmethylated genes, which suggests that H3K9me2 is likely a trigger for inducing spontaneous epialleles. Evidence that H3K9me2 can seed *de novo* methylation independent from RdDM and pre-existing DNA methylation has also been observed at the *BONSAI* locus in *A. thaliana* [244]. Therefore, understanding H3K9me2 dynamics and why certain genes are hotspots for H3K9me2 activity compared to others will be essential to understanding the origin of epialleles. The ectopic hypermethylated epialleles identified in the genotypes used in this study show that there is a non-random process that makes certain loci more susceptible to spontaneous epiallele formation (**Figure S4.7**). This has led us to speculate that gbM genes in angiosperms arise from off-targeting activity of positive feedback loops established between CMT3-H3K9me2 and that gbM is evolutionarily neutral. These genes are susceptible to this pathway because of their inherent functions. These genes are generally

‘housekeeping’ genes that are long, moderately expressed in all cells and typically not regulated by development/environment. Transcription at these loci is likely a prerequisite to becoming a gbM gene, transcription could lead to incorporation of H3K9me2 nucleosomes at a low rate. Typically, this H3K9me2 is removed by IBM1, but over evolutionary time mistakes occur which lead to ectopic activity of CMT3 leading to CHG methylation and eventually CG methylation. How CHG methylation transitions to CG methylation to provide transgenerational stability of epialleles is currently unknown and was not addressed in this study. There are at least two possible mechanisms by which this occurs [122]. One includes the activity of the RdDM pathway, which is recruited to regions possessing H3K9me2 via SHH enabling DRM2 to methylate CGs. The second possibility includes recruitment of VIM1 to methylated DNA via its SRA domain (in this case CHG methylation) [143], which could lead to rare de novo methylation of CGs by MET1. Regardless of the exact mechanism of transitioning CHG to CG methylation, the incorporation of these feedback loop would be prevented at many developmental/environmental regulated genes as they are targeted by the Polycomb Repression Complex and H3K27me3 to precisely limit their expression [257]. H3K27me3 is incompatible with maintenance of DNA methylation in *A. thaliana*, as it colocalizes with H2A.Z [258]. Given H3K9me2 associates with H2A.W [259], the presence of H2A.Z would prevent the CMT3-H3K9me2 feedback loop from establishing gbM at many genes in the genome.

We hypothesize that spontaneous epialleles form as a byproduct of enzymatic activities that are dedicated to the maintenance of heterochromatin and that variation in heterochromatin abundance and methylation influences the rate of spontaneous epiallele formation [240]. If genes that maintain heterochromatin sometimes act on inappropriate targets, causing epialleles, natural selection cannot reduce epialleles by reducing DNA methylation because that would lead to loss

of maintenance of heterochromatin and result in genome instability. In this way, the evolution of epialleles is similar to the evolution of chromosome rearrangements. In that case, molecular recombination, is favored to generate gamete diversity (or for some other reason) that has an unintended, deleterious consequences, namely the production of genome rearrangements by ectopic recombination.

Numerous major questions remain, but the most important one is how is H3K9me2 initially incorporated into an unmethylated region *de novo*. Transcription coupled incorporation of H3K9me2 nucleosomes is one possibility given the role of nucleosome eviction and reincorporation during transcription [260]. It's also possible that H3K9me2 is mis-incorporated to certain regions of the genome during the establishment of chromatin upon DNA replication. In this study, we used Hi-C data to show the spontaneous epialleles we identified interact with H3K9me2 regions at greater frequencies than unmethylated genes in the genome. This result supports the hypothesis that there are sub-nuclear compartments that could have different concentrations of H3K9me2 nucleosome pools and that spontaneous epialleles are more likely to occur in pools with high concentrations of H3K9me2 nucleosomes. However, the current evidence to support these conclusions are premature to make stronger conclusions. Regardless, the involvement of CMT3-H3K9me2 feedback regulation is a major factor in the origins of spontaneous epialleles and future studies will be required to test these proposed models.

MATERIALS AND METHODS

Plant material

The only new plant material generated for this study was the create of *ibm1;met1 epiRIL-12*. These individuals were produced by crossing *ibm1-6* (SALK_006042) [261] to a *met1 epiRIL-*

12 [111] and isolating homozygous *ibm1-6* lines in the F2. The complete set of *ddm1*-epiRILs from Johannes et al. (2009) was obtained from the Versailles Arabidopsis Stock center of INRA (<http://publiclines.versailles.inra.fr/>). All epiRIL lines were propagated in a greenhouse at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK). The plants were grown in single seed pots and at a later developmental stage six siliques per plant were left to dry. Seeds for a selection of 169 epiRILs were sown in the IPK fully automatic phenotyping facility for small plants (Junker et al., 2015). Three independent experiments in three consecutively months were performed. In each of the 3 cultivation experiments, 6 individual plants from each line were grown in 2 separate trays completely randomized in the chamber. The sown pots were firstly placed for 3 days at 4°C in darkness and then the plants were acclimated for 2 days under 16/14°C with reduced light conditions. Following this, they were grown under long day conditions (16h light, 8h dark) at 20/18°C, 60-75% humidity and 180-240 µE light intensity. The pots were watered with 55, 30 and 20 ml water the 2, 7 and 8 day after sowing (DAS) and then every other day with 55 ml.

Library construction

All epiRIL lines were harvested at 27 DAS in a time frame of 3 hours. Flowering stems and roots were removed and all 18 individual plants from each epiRIL line were pooled in 50 ml tubes. They were immediately frozen in liquid nitrogen and stored to -80°C until processing. Genomic DNA was extracted from each pooled sample using the DNAeasy plant mini kit from Qiagen. 169 epiRIL lines with at least 1 µg of DNA were submitted to the Beijing Genome Institute where they were prepared for WGBS libraries. Sequencing was performed on an Illumina HiSeq

X ten instrument. Clean raw paired-end files were obtained from BGI and used for downstream analysis. The *ibm1;met1 epiRIL-12* library was prepared following the MethylC-seq protocol without use of the bisulfite conversion step. MethylC-seq libraries were prepared as previously described [156]. Briefly, genomic DNA was sonicated to 200 bp using a Covaris S-series focused ultrasonicator, and end-repaired using End-It DNA end-repair kit (Epicentre). End-repaired DNA was subjected to A-tailing using Klenow 3'–5' exo– (NEB) and ligated to methylated adapters using T4 DNA ligase (NEB). Ligated DNA was subsequently bisulfite converted using the EZ DNA methylation-Gold kit as per the manufacturer's instructions and amplified using KAPA HiFi uracil + Readymix Polymerase.

Methylome mapping

The WGBS data from the epiRILs was analyzed using the Methylstar pipeline [262]. Methylation state calls at the level of single cytosines as well as at the level of regions (200 bp bins, step size = 50 bps, at least 10 cytosines per bin) were obtained with Methimpute [263]. We only used Methimpute to obtain region-level methylation calls (~200 bp windows), which we then used to re-construct the recombination map. We used Methimpute's 2-state Hidden Markov Model to classify a given region as either homozygous methylated or homozygous unmethylated. These state calls were used downstream for the construction of an augmented linkage map in the epiRIL panel. For all other methylome analyses, such as the analysis of CHG methylation in the epiRILs, we did not use Methimpute. The *ibm1;met1 epiRIL-12* data used in this study were processed by Methylypy v1.3 as described in [158]. Quality filtering and adapter trimming were performed using cutadapt v1.9.dev1 [264]. Qualified reads were aligned to the *A. thaliana* TAIR10 reference genome [203] (downloaded from <https://phytozome.jgi.doe.gov>) using bowtie

2.2.4 [161]. Only uniquely aligned and nonclonal reads were retained. Chloroplast DNA (which is fully unmethylated) was used as a control to calculate the sodium bisulfite reaction non-conversion rate of unmodified cytosines. A binomial test was used to determine the methylation status of cytosines with a minimum coverage of three reads.

Data acquisition

WGBS data of *ddm1* and *ibm1* mutants and their self-crossing offspring used in this analysis were obtained from published datasets [119]. WGBS data of *met1-3*, *ibm1-6* and *met1* epiRIL-12 were obtained from previously published datasets [25, 116]. Hi-C data and H3K9me2 ChIP-seq reads used in this study were obtained from a previously published dataset [246, 265].

Gene body methylation status classification

To explore the fate of ectopic methylation of genes that have different DNA methylation states, we categorized genes into one of three classes including gbM, teM and UM based on CG, CHG and CHH methylation in wild-type Col-0. The total number of cytosines and the methylated cytosines were counted for cytosines in each context (CG, CHG, and CHH) for the coding sequences (CDS) of the primary transcript for each gene. The percentage of methylated sites for each sequence context in all coding regions were used as the background probability of having methylation on a single site. Given a background probability and the total number of cytosines and methylated cytosines, a *p* value was calculated using a binomial distribution to show the cumulative probability of having a higher number of methylated cytosines on a given gene [155]. Then *q*-values were calculated by adjusting *p* values by Benjamin–Hochberg FDR to control the false discovery rate.

Genes were classified as gbM if they had reads mapping to at least 20 CG sites and had a q-value < 0.05 for mCG and a q-value > 0.05 for mCHG and mCHH. Genes were classified as mCHG if they had reads mapping to at least 20 CHGs, a mCHG q-value < 0.05 , and a mCHH q-value > 0.05 . As mCG is commonly associated with mCHG, the q-value for mCG could be significant or insignificant in mCHG genes. Genes were classified as mCHH if they had reads mapping to at least 20 mCHH sites and a mCHH q-value < 0.05 . Q-values for mCG and mCHG could be anything as both types of methylation are associated with mCHH. mCHG and mCHH genes were collectively referred to as teM genes. Genes were classified as unmethylated (UM) if they had reads mapping to at least 20 mCHH sites and had a q-value > 0.05 for all sequence contexts. To make sure the selected UM genes are truly unmethylated, we further remove genes with more than 2 symmetric mCG from UM genes list. Only genes that fit the definition of gbM, teM and UM above were used for the downstream analysis in this study.

DNA methylation patterns on genes

For *ibm1*, *ddm1* and *ddm1* epiRILs lines, their DNA methylation patterns on gbM and UM genes were explored. Each gene was divided into 20 windows. Additionally, regions 1,000 bp upstream and downstream were each divided into 20 50-bp windows. Methylation levels were calculated for each window according to previous recommendations [158], and a mean value for each window was averaged over the methylation level of the same window from genes in the same group (gbM/UM). The mean methylation levels and their corresponding window number was applied for generating metaplot by R package ggplot2.

Locating the *met1* derived region in *ibm1;met1 epiRIL-12*

The *met1 epiRIL-12* line used in this study was carefully selected to ensure the *IBM1* locus was derived from the Col-0 parent instead of the *met1* parent. The *met1 epiRIL-12* derived regions of *ibm1;met1 epiRIL-12* was identified by comparing CG methylation levels on the exons of gbM genes between the epiRIL line and its parents (*ibm1-6* and *met1 epiRIL-12*). Each chromosome sequence was divided into continuous bins and each bin included at least 10 exons. A bin with an average mCG level that was greater than 25% reduced in *ibm1;met1 epiRIL-12*, as compared to that in *ibm1-6* was used to differentiate the *ibm1-6* and *met1 epiRIL-12* derived regions. All downstream analyses of this line used previously described criteria for defining gbM and UM genes. Regions that possessed *met1 epiRIL-12* derived regions in the *ibm1-6* mutant background were used to evaluate the consequence of mutant *ibm1-6* on genes that had lost gbM due to previous loss of *met1*. Ectopic CHG or CWA methylation on gbM and UM genes required a minimum of three sites with an average 10% higher methylation level when compared to the wild-type Col-0 parent.

Construction of an augmented epiRIL linkage map

Previously, we used tiling array data from 123 epiRILs and their *ddm1-2* and Col-wt founder lines and identified differentially methylated regions (DMRs) in the founders that were stably inherited in the epiRILs through at least 10 rounds of meiosis [114]. Using these ultra-stable DMRs as physical markers we were able to construct a linkage map involving 126 DMRs in this isogenic experimental system [114], and later used this map for QTL^{epi} analysis [232]. The 169 epiRIL methylomes from the present study were employed to augment the existing linkage map, with the goal to improve marker spacing and mapping resolution. Among the 169

epiRILs measured in the present work, only 37 overlapped with the 123 epiRILs used in the previous our studies [114, 232]. We used the tiling-array-derived linkage map of 126 DMRs as a starting point. For the 37 overlapping epiRILs, the WGBS derived methylation state calls at these 126 DMR position were consistent with the previous tiling array calls, indicating that the these DMRs are robust. For a given chromosome, we stepped in new DMRs into the existing linkage map. DMRs that correlated significantly across chromosomes with any of the core DMRs were rejected. For map cleaning we employed Rqtl's tutorial on map construction [266]. For DMRs with identical cM positions we selected those that best optimized overall marker spacing in terms of base pair coverage. DMRs were kept or rejected by manual inspection in conjunction with the likelihood score from the drop.one function in Rqtl. The final linkage map contains 144 well-spaced markers, which provide improved coverage in chromosome arms.

Methylation analysis for *ddm1* epiRILs

For methylation analysis in *ddm1* epiRILs, we need to figure out whether a chromosomal segment was inherited from wild type or *ddm1* for each line. Therefore, the methylation levels on 144 selected DMR markers was calculated for each line. DMRs was classified as unmethylated *ddm1* marker if their methylation level less than 0.5. Otherwise, DMRs was classified as methylated WT maker if their methylation level greater or equal to 0.5. Then, haplotype map in figure 4.2A was generated based on the location of 144 DMRs and their methylation status (methylated, unmethylated) for each line by R package ggplot2. We also use a measurement of hypomethylation index to indicates amounts of DNA inherited from the *ddm1* parent. Hypomethylation index is calculated as an average of the values on 144 DMR markers (methylated WT marker =0, unmethylated *ddm1* marker =1). So higher the value represents a more

hypomethylated genome. To further evaluate the impact to the disruption to normal DNA methylation states, we classified genes into wild-type versus *ddm1* derived regions based on haplotype map for each line. Then, we explored how ectopic CHG methylation are distributed on genes from WT versus *ddm1* parent, respectively. Fisher's Exact test was used to evaluate the enrichment of mCHG-gain genes in three types of genes (gbM, UM, teM). Pearson correlation coefficient was used to evaluate how CHG methylation changes against hypomethylation index for each gene.

QTL mapping analysis

Using the new map together with the different mCHG molecular phenotypes described above, we performed interval mapping with Rqtl as previously described [232].

Characterizing chromatin interactions between genes and H3K9me2

We used previously published Hi-C data [246] to identify chromatin contacts between two genomic regions in nonadjacent locations along the genome. The HiC-Pro v2.11.4 pipeline was used to process Hi-C data, from raw reads to normalized contact matrices for selected windows using the parameter '--binsize 2000', since 2 kb is an approximate value of the average gene length of *A. thaliana* [267]. The contact matrices were then transformed to Fit-Hi-C readable input files with hicpro2fithic.py. Then, significant contacts were identified using Fit-Hi-C by assessing the enrichment of observed from the expected contact counts [268]. Raw reads of H3 and H3K9me2 ChIP-seq were trimmed with Trim Galore v0.6.5 (<https://github.com/FelixKrueger/TrimGalore>) with default parameters. The remaining reads were aligned to the *A. thaliana* TAIR10 reference genome using Bowtie2 v2.3.5.1 with default

setting. Aligned reads were sorted using SAMtools v1.10 [269]. Then, unmapped reads, duplicated reads and multiple mapped reads were filtered out using Sambamb v0.7.1 [270]. Lastly, the remaining uniquely mapped reads were used in MACS2 for peak calling with parameters ‘-g 1.35e+8 --broad’ [271]. H3 ChIP-seq was used as a control for identification of H3K9me2 enriched regions. Significant contacts that had at least one edge that overlaps with a H3K9me2 enriched region and another edge that overlapped with a gbM (n=875), teM (n=332) or UM (n=2424) gene was selected for downstream analysis. Fisher’s Exact test was used to evaluate the enrichment of genes that have contacts with H3K9me2 peaks in three types of genes (gbM, UM, teM). Distribution of distance and the number of contacts between three types of genes and H3K9me2 peaks were compared by Two-sample Wilcoxon tests. The selected gbM genes (n=875) were further classified into a ectopic mCHG-gain group (gbM; mCHG+) and a group without ectopic mCHG (gbM; mCHG-) based on methylome of *ddm1* mutant. The distribution of the number of contacts between these two groups and H3K9me2 peaks were also compared by Two-sample Wilcoxon tests.

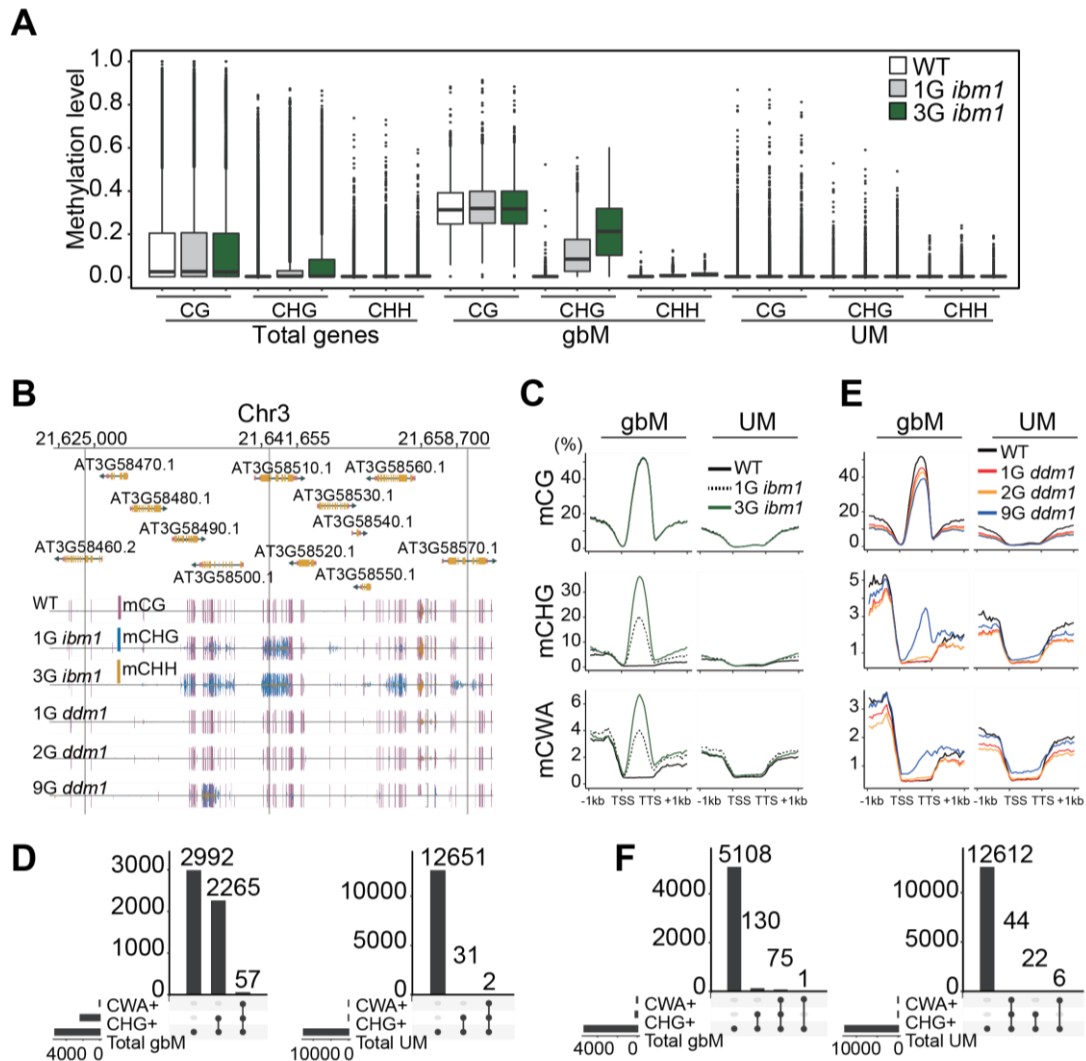


Figure 4.1. Ectopic CHG methylation of gbM genes increases over generations in *ibm1* and *ddm1* mutants. (A) The DNA methylation patterns of WT (Col-0), 1G *ibm1*, and 3G *ibm1* mutants for all genes, gbM and UM genes. (B) A genome browser view of genes with ectopic non-CG methylation in *ibm1* and *ddm1* mutants (C) DNA methylation patterns of WT (Col-0), 1G *ibm1* and 3G *ibm1* mutants for gbM and UM genes. (D) mCHG-gain (>0.1) and mCWA-gain (>0.1) in 1G *ibm1* and enriched in gbM genes. (E) DNA methylation patterns of WT (Col-0), 1G *ddm1*, 2G *ddm1*, and 9G *ddm1* mutants for gbM and UM genes. (F) Gain of non-CG methylation in 9G *ddm1* mutants is enriched in gbM genes.

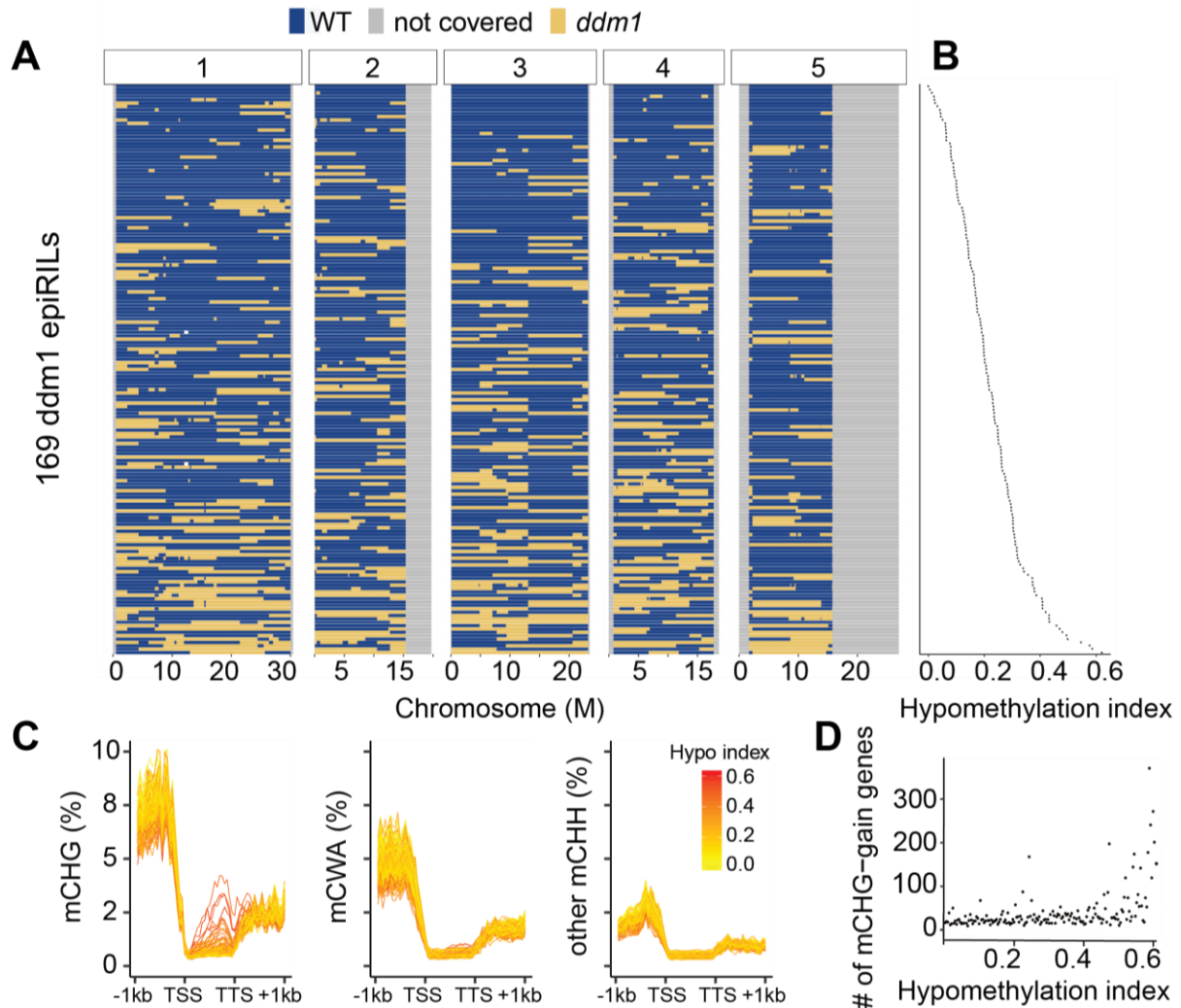


Figure 4.2. Hypomethylated chromosomes from the *ddm1* parent led to ectopic DNA methylation in gbM genes. (A) A haplotype map for 169 *ddm1* epiRIL lines. Each row represents the genome composition of one epiRIL line which was ordered based on hypomethylation index which indicates amounts of DNA inherited from the *ddm1* parent. (B) Hypomethylation index is calculated as an average of the values on 144 DMR markers (methylated WT marker =0, unmethylated *ddm1* marker =1) derived from the haplotype map. (C) Metaplot shows non-CG methylation patterns of gbM genes for 169 *ddm1* epiRIL lines. Methylation patterns for each sample is shown by one line that was colored based on its hypomethylation index. (D) The scatter plot indicates the number of gbM genes that gain CHG methylation for each sample.

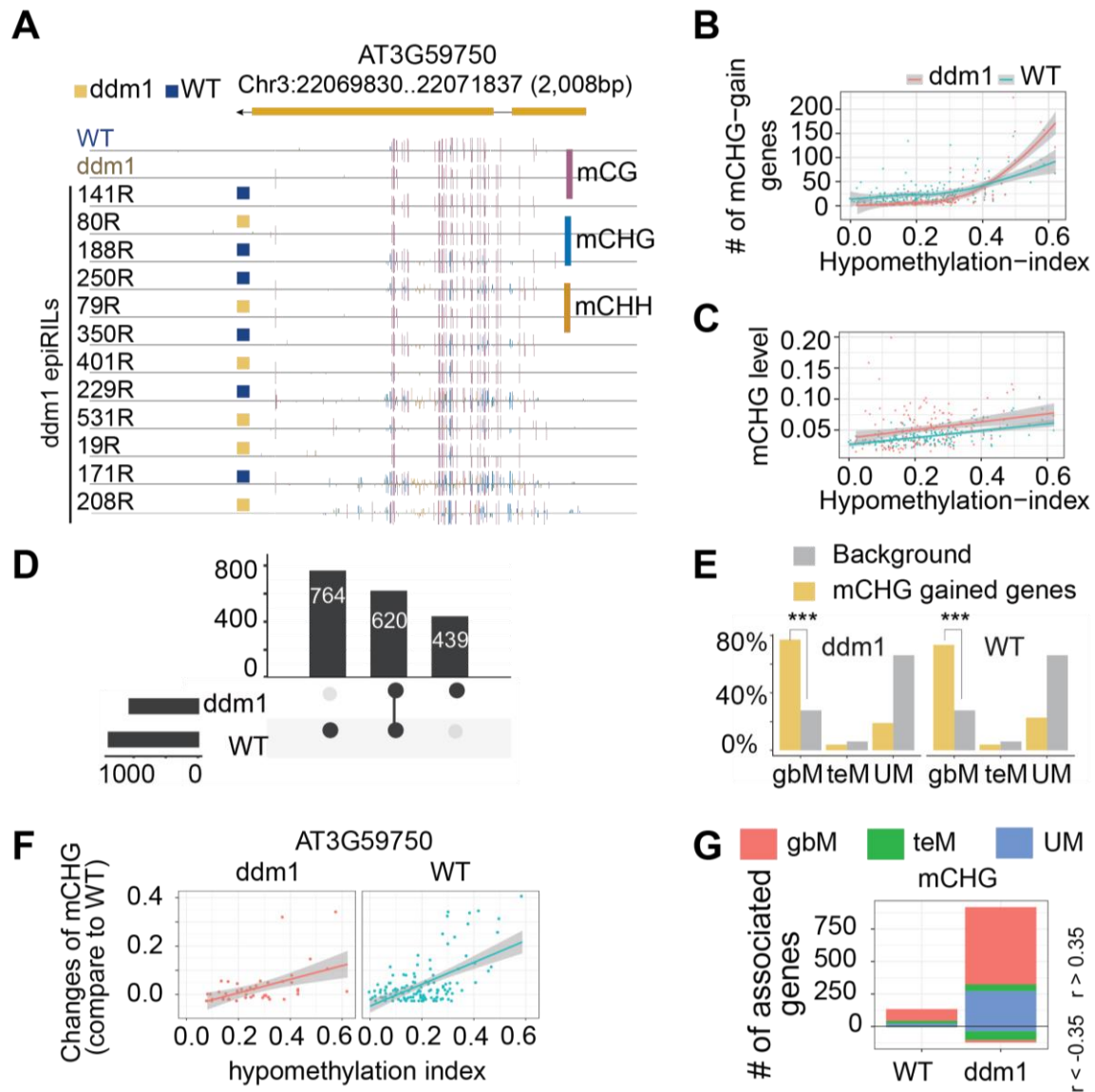


Figure 4.3. Global hypomethylation leads to ectopic activity of CMT3 at genes present within wild-type and *ddm1* haplotypes. (A) Genome browser view of a representative gene AT3G59750, in which non-CG methylation was observed in epiRILs with either wild-type or *ddm1* derived haplotypes. (B) Correlation between hypomethylated genomes and the number of gbM genes that gain CHG methylation. Genes inherited from WT and *ddm1* parents were calculated separately for each epiRIL. (C) Correlation between hypomethylated genomes and the average accumulated CHG methylation level on mCHG-gain genes. (D) An union set of total mCHG-gain genes from all epiRIL lines. It shows the number of mCHG-gain genes from each haplotype type as well as those shared by both haplotype. (E) An enrichment analysis using Fisher's Exact test shows that mCHG-gain genes that either from *ddm1* (left) or WT haplotypes (right) are enriched in gbM genes. (F) Genome-wide association of methylation level changes (difference of value compared to WT) at an example gene compared to its hypomethylation index. An example of positive correlation is shown for AT3G59750. For each gene, the epiRILs were grouped based on haplotypes (either WT or *ddm1* derived haplotypes). (G) Correlation test in (F) was applied for all genes. The barplot shows the number of genes with CHG methylation level changes are strongly correlated with the hypomethylation index (with Pearson correlation coefficient either >0.35 (above the line) or <-0.35 (below the line)). The number of associated genes is calculated separately based on haplotypes (either WT or *ddm1* derived haplotypes) and gene type (gbM, teM and UM).

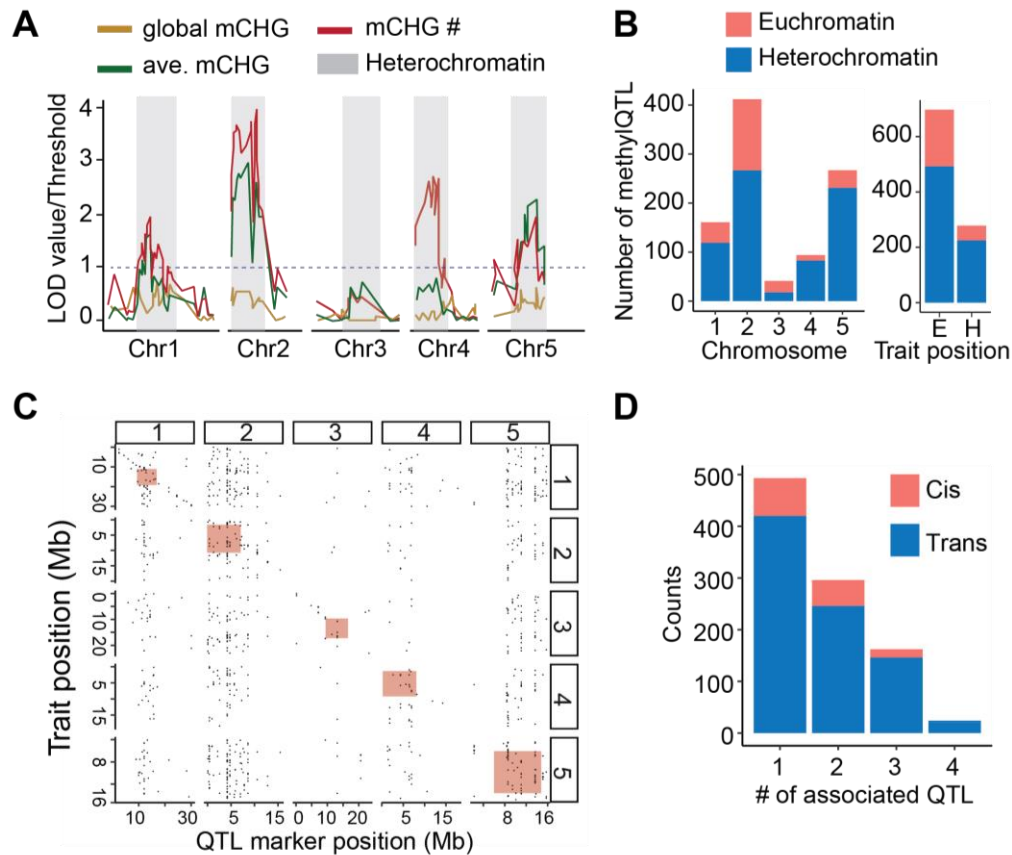


Figure 4.4. Methylation QTL for genes with ectopic mCHG are especially enriched in the pericentromeric regions. (A) Standard interval QTL mapping for genic mCHG related phenotypes. Methylation status (methylated, unmethylated) on 144 stably inherited DMR makers was used for QTL mapping. The results for three phenotypes, global mCHG level (global mCHG), the number of mCHG-gain genes (mCHG #), averaged mCHG level on genes (ave. mCHG), were plotted as likelihood-ratio test statistic (LOD score divided by threshold value of 3) against physical location of markers (Mb). (B) Barplots shows that methylQTL are especially enriched in the pericentromeric regions in comparison to euchromatin, whereas affected genes are enriched in euchromatin (right barplot). E=Euchromatin and H=Heterochromatin. (C) The scatter plot shows the location of the 701 genes against the location of their corresponding methylQTL, respectively. (D) A comprehensive QTL analysis for each of the 1595 genes that had evidence of ectopic mCHG. Significantly associated QTL loci were identified for 701 genes. For association detected, half of them were linked with a single methylQTL, and others were found to be associated with 2-4 methylQTL. The majority of methylQTL are distant (trans) to associated genes.

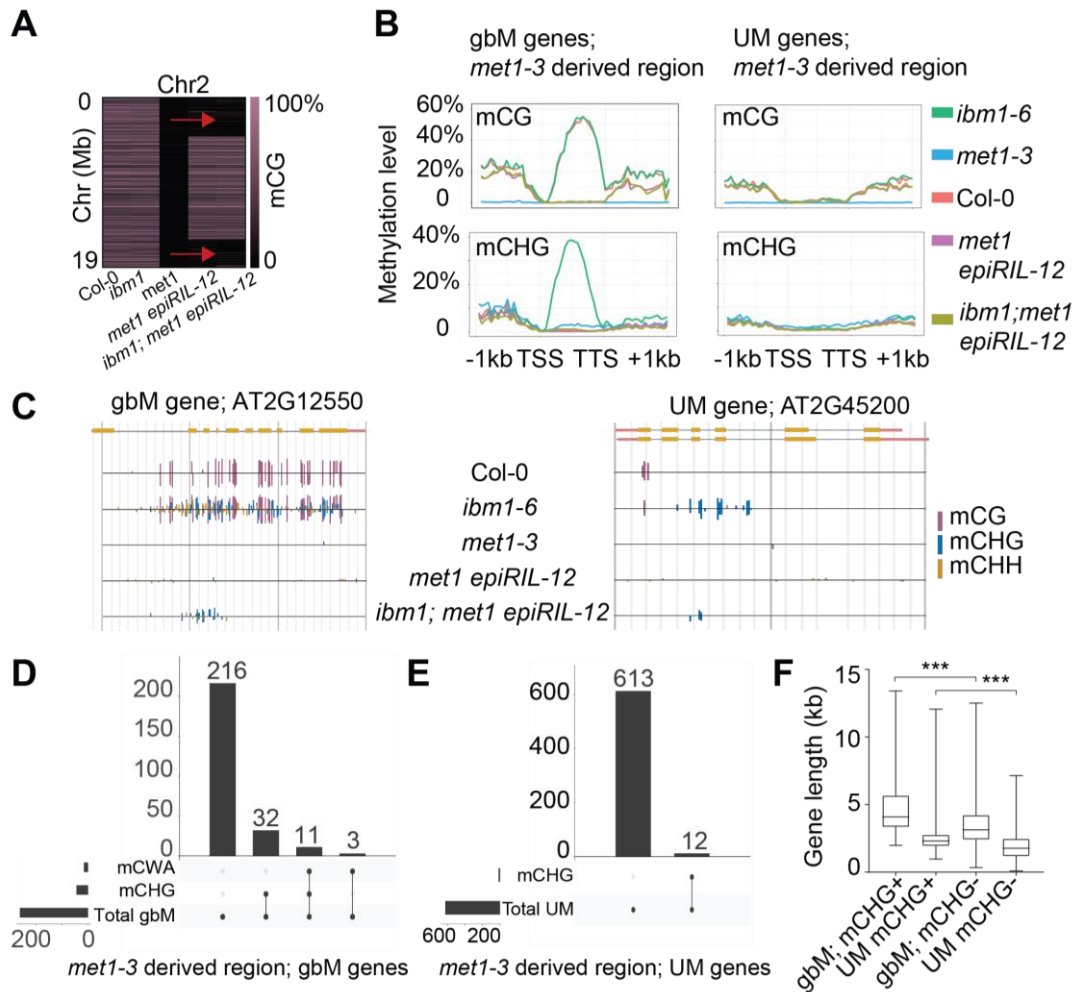


Figure 4.5. Mutations in *ibm1* lead to ectopic methylation preferentially in genes that previously possessed gbM. (A) A schematic showing regions that inherited mCG from the *met1 epiRIL-12* parent or the *ibm1-6* parent on Chromosome 2 (Col-0, *ibm1-6* and *met1-3* are shown as controls). Purple to black colors show the mCG percentage of genes that possess gbM in Col-0 from 100% to 0%, respectively. Red arrows indicate a region in *ibm1; met1 epiRIL-12* derived from the original *met1-3* parent. (B) DNA methylation plot of gbM and UM genes including 1kb up and downstream for indicated genotypes and methylation contexts. Depicted regions shown are derived from *met1-3*. TSS: Transcription Start Site. TTS: Transcription Termination Site. (C) Genome browser examples of gbM and UM genes from *met1-3* derived region of the *met1 epiRIL-12* genotypes. (D and E) The number of gbM and UM genes that gained mCHG, mCWA or non-mCWA for in the *met1-3* derived region of *ibm1; met1 epiRIL-12*. Vertical bar plot shows count of overlapping genes between different methylation contexts. Horizontal bar plot shows count of specific type of genes. (F) Gene length distribution of genes that gained mCHG in gbM and UM genes derived from the *met1-3* region of *ibm1; met1 epiRIL-12*. Gain is depicted by + and no gain is depicted by - *** $p < 0.002$; pairwise student T-test.

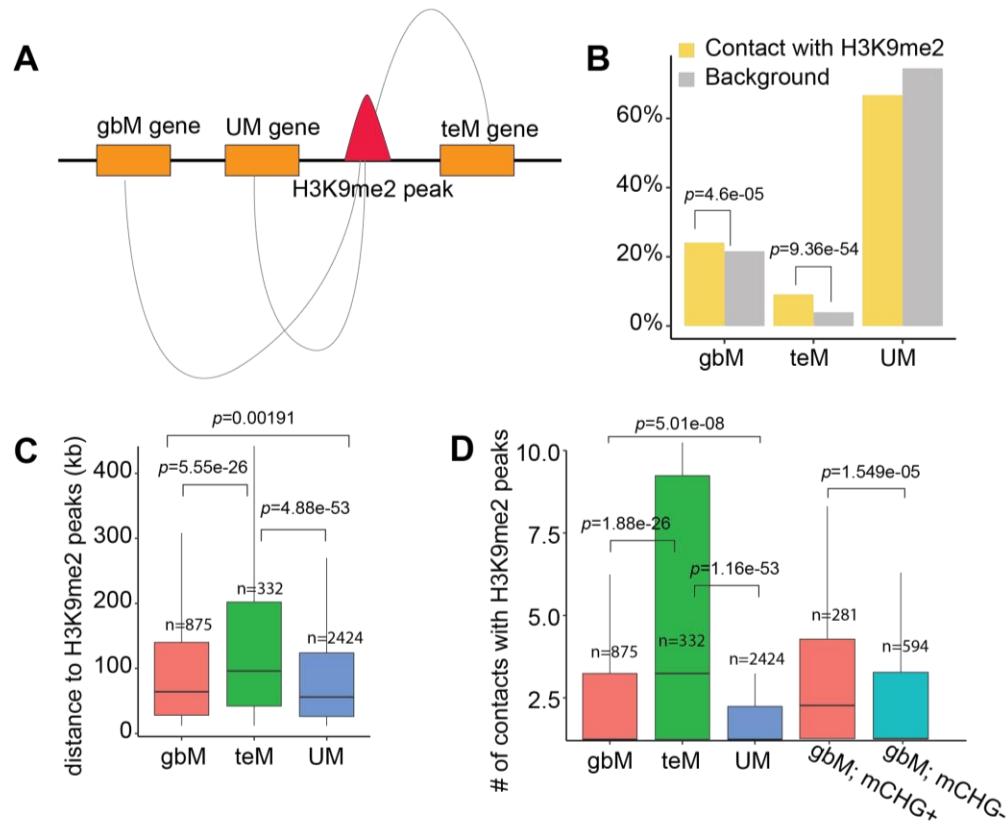


Figure 4.6. Loci that are susceptible to spontaneous epiallele formation have a greater contact frequency with H3K9me2 region. (A) Hi-C data was used for identifying significant three-dimensional contacts between genomic regions. The schematic plot depicts the interaction from regions possessing H3K9me2 and genes. **(B)** Enrichment test for genes that have contacts with H3K9me2. Yellow bar show the proportion of each type of genes that contacts with H3K9me2. Grey bar show the proportion of each type genes in comparison with all coding genes. Fisher's exact test was used for enrichment test. **(C)** Distribution of distance between three types of genes (gbM, UM, teM) and H3K9me2 regions. Two-sample Wilcoxon tests was used for pairwise comparison of each two groups. **(D)** Distribution of the number of contacts between three types of genes (gbM, UM, teM) and H3K9me2 regions. gbM genes were further classified into a group with ectopic mCHG (gbM; mCHG+) and a group without ectopic mCHG (gbM; mCHG-) based on ddm1 mutant line data. Two-sample Wilcoxon tests was used for pairwise comparison of each two groups.

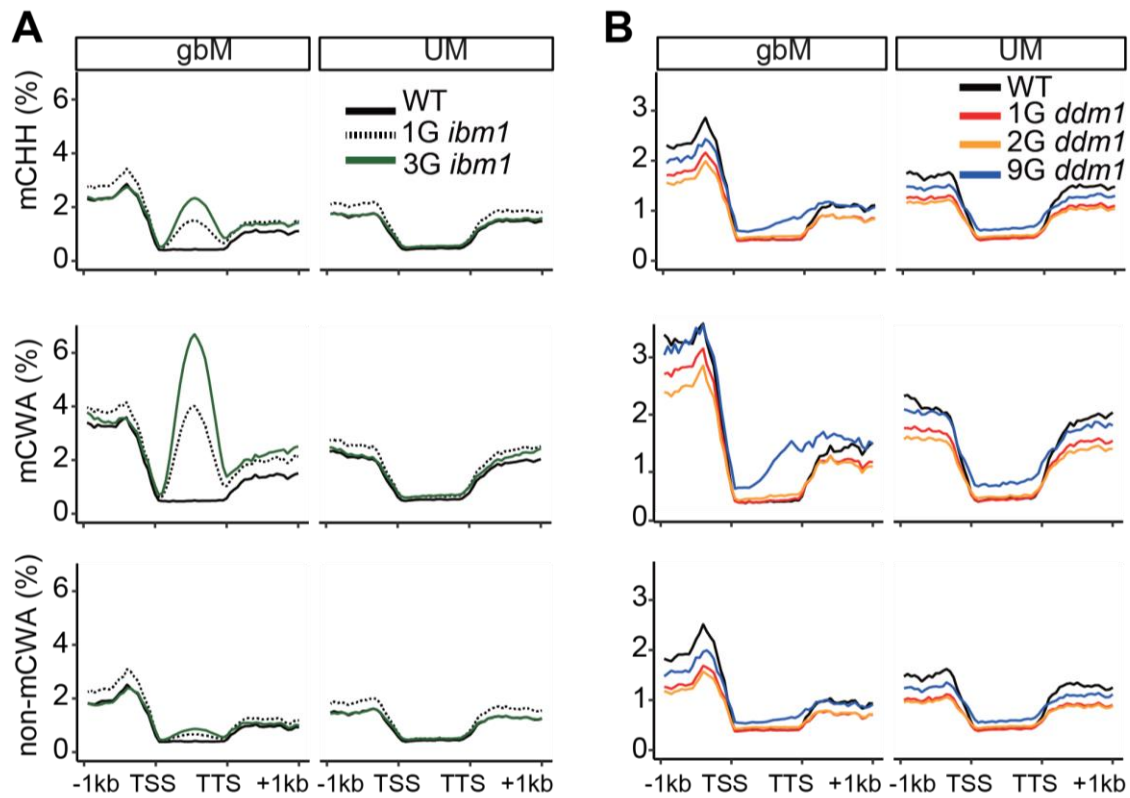


Figure S4.1. Ectopic CHH methylation on genes are mainly at CWA sites in *ibm1* and *ddm1* mutants. (A) CHH methylation of WT (Col-0), 1G *ibm1*, and 3G *ibm1* mutants for gbM and UM genes. **(B)** CHH methylation of WT (Col-0), 1G *ddm1*, 2G *ddm1*, and 9G *ddm1* mutants for gbM and UM genes.

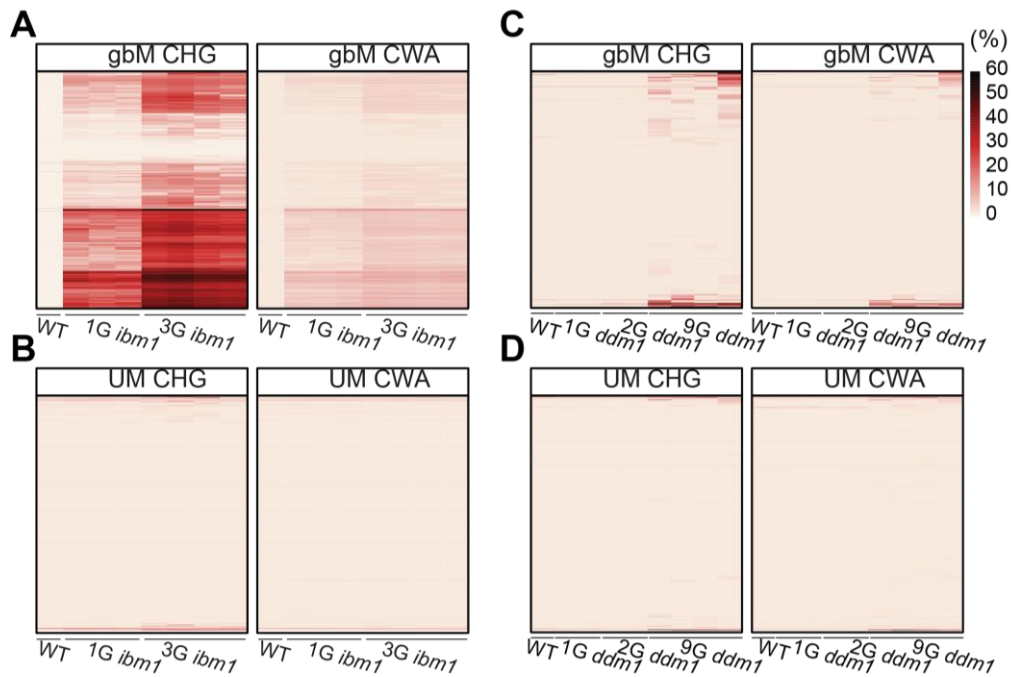


Figure S4.2. Gradual increase of non-CG methylation at gbM genes in *ibm1* and *ddm1* mutants over generations. (A) Heatmaps show mCHG and mCWA increasing over generations on gbM genes in *ibm1* mutants. (B) mCHG and mCWA patterns on UM genes in *ibm1* mutants. (C) mCHG and mCWA increasing over generations on gbM genes in *ddm1* mutants. (D) mCHG and mCWA on UM genes in *ddm1* mutants.

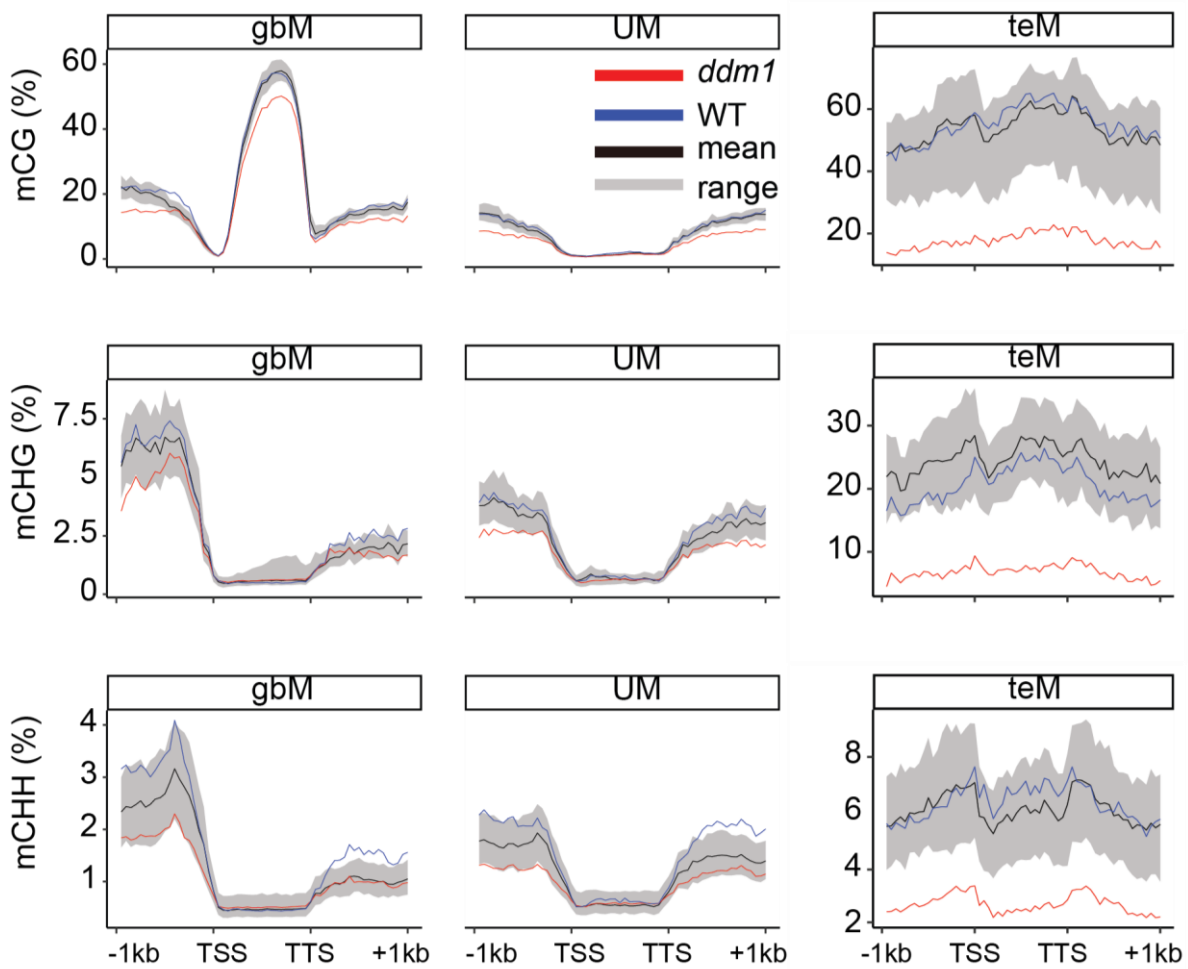


Figure S4.3. DNA methylation patterns at gbM, UM and teM genes for *ddm1* epiRILs. Red and blue lines show the methylation pattern for *ddm1* and WT parents. Black line shows the averaged methylation level of 169 epiRILs. Grey bands show the maximum and minimum value of those epiRILs.

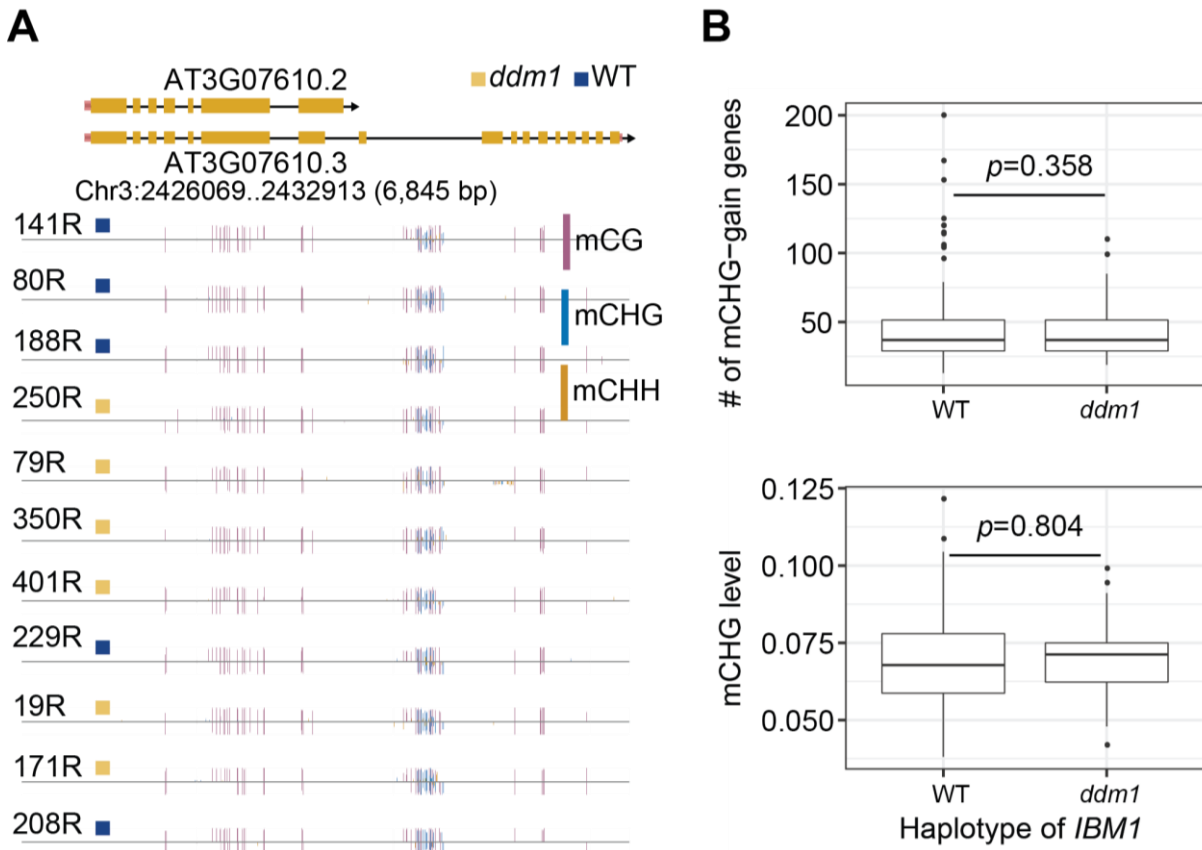


Figure S4.4. Ectopic non-CG methylation was not influenced by intronic methylation of *IBM1* in *ddm1* epiRILs. (A) Mutations in the H3K9me2 demethylase, *IBM1*, results in ectopic non-CG methylation mainly on gbM genes. Functional expression of *IBM1* requires the presence of intronic methylation, the removal of which promotes expression of a truncated, non-functional transcript. However, no significant reduction of intronic methylation has been shown in *ddm1* epiRILs. (B) To further exploring the potential influence of *IBM1* on ectopic genic non-CG methylation, epiRILs were classified into two groups based on *IBM1* belonging to WT or *ddm1* haplotype, since *IBM1* inherited from a *ddm1* parent may suffer a reduction of intronic methylation. However, there are no significant difference between two groups in terms of the number of ectopic mCHG genes and genic mCHG level.

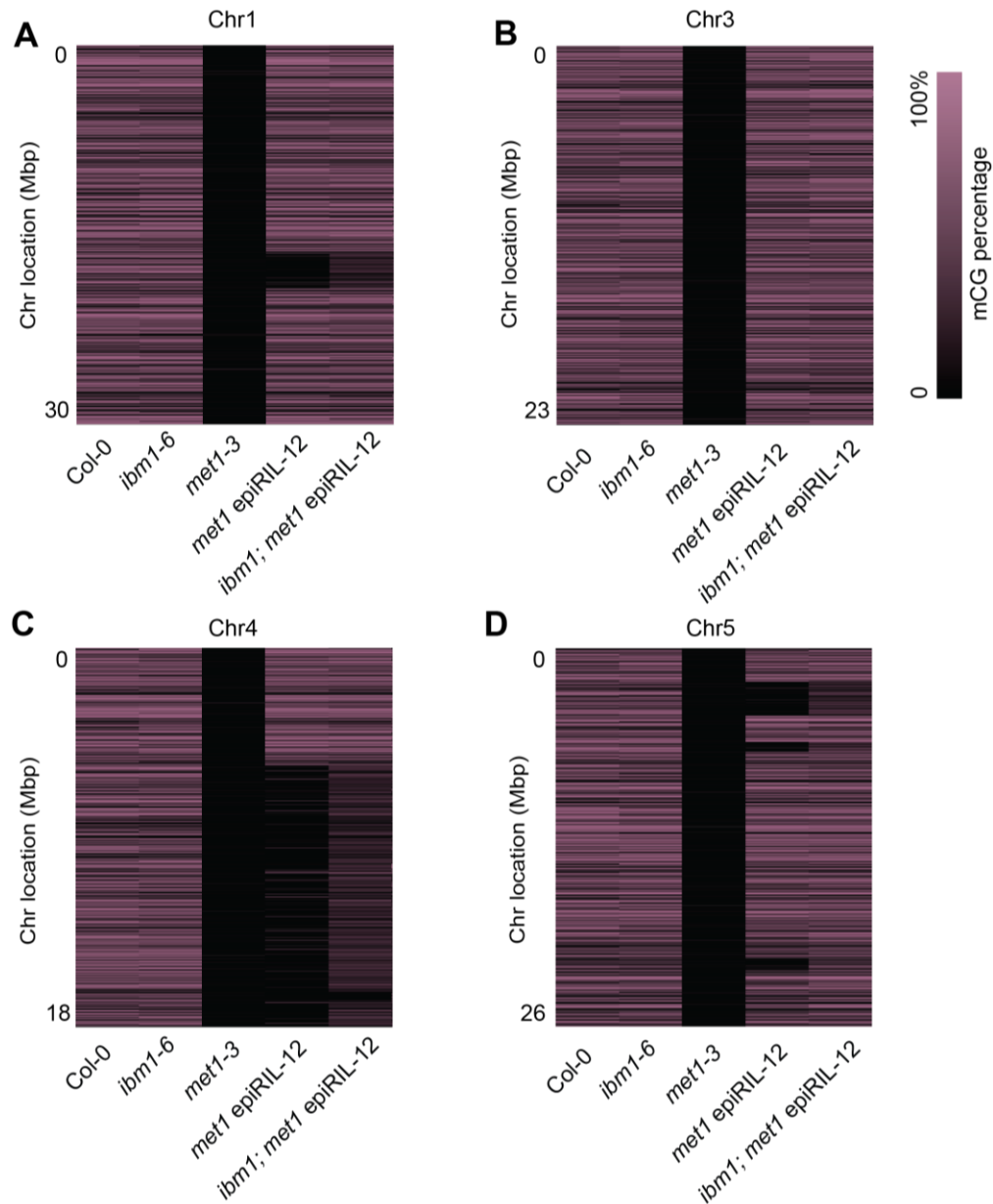


Figure S4.5. Except chromosome 2, no obvious large region in *ibm1; met1 epiRIL-12* is derived from the original *met1-3* parent. (A) Schematic shows regions that inherited mCG from the *met1 epiRIL-12* parent or the *ibm1-6* parent on Chromosome 1 (*Col-0*, *ibm1-6* and *met1-3* are shown as controls). Purple to black colors show the mCG percentage of genes that possess gbM in *Col-0* from 100% to 0%, respectively. **(B)** on Chromosome 3 **(C)** on Chromosome 4 **(D)** on Chromosome 5

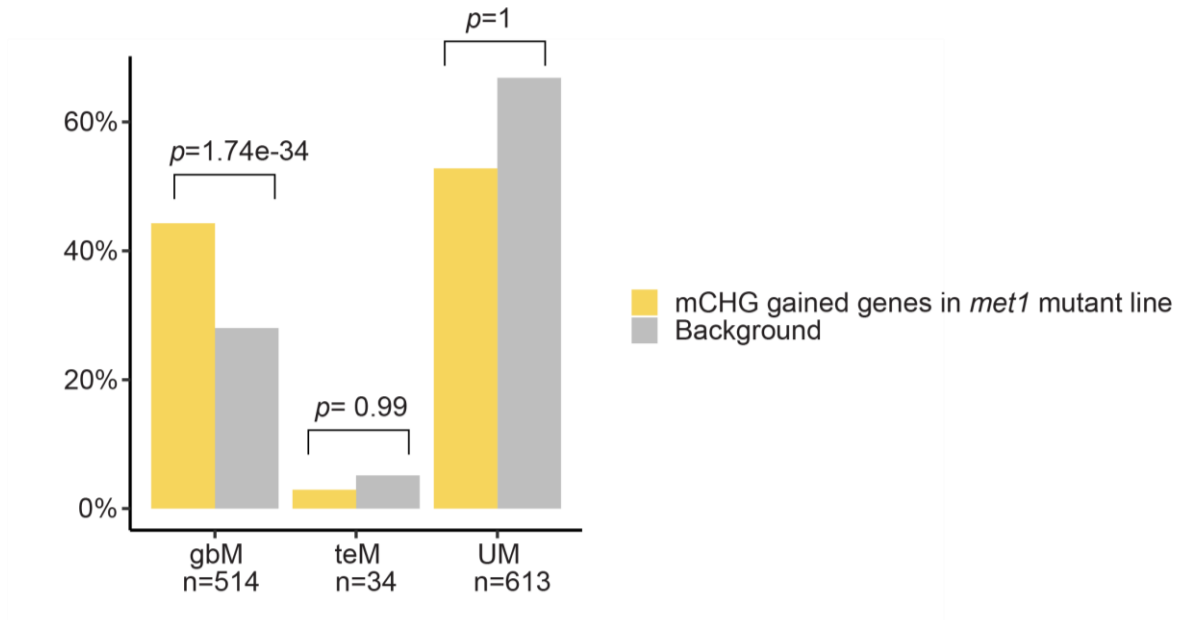


Figure S4.6. Ectopic mCHG-gain genes in *met1* mutant are enriched in gbM genes. In total, there are 1161 mCHG-gain genes in *met1* mutant line. Based on an Fisher's Exact enrichment test, those genes are enriched in gbM genes. Yellow bar show the ratio of each type of genes with ectopic mCHG. Grey bar show the ratio of each type of genes in comparison with all coding genes.

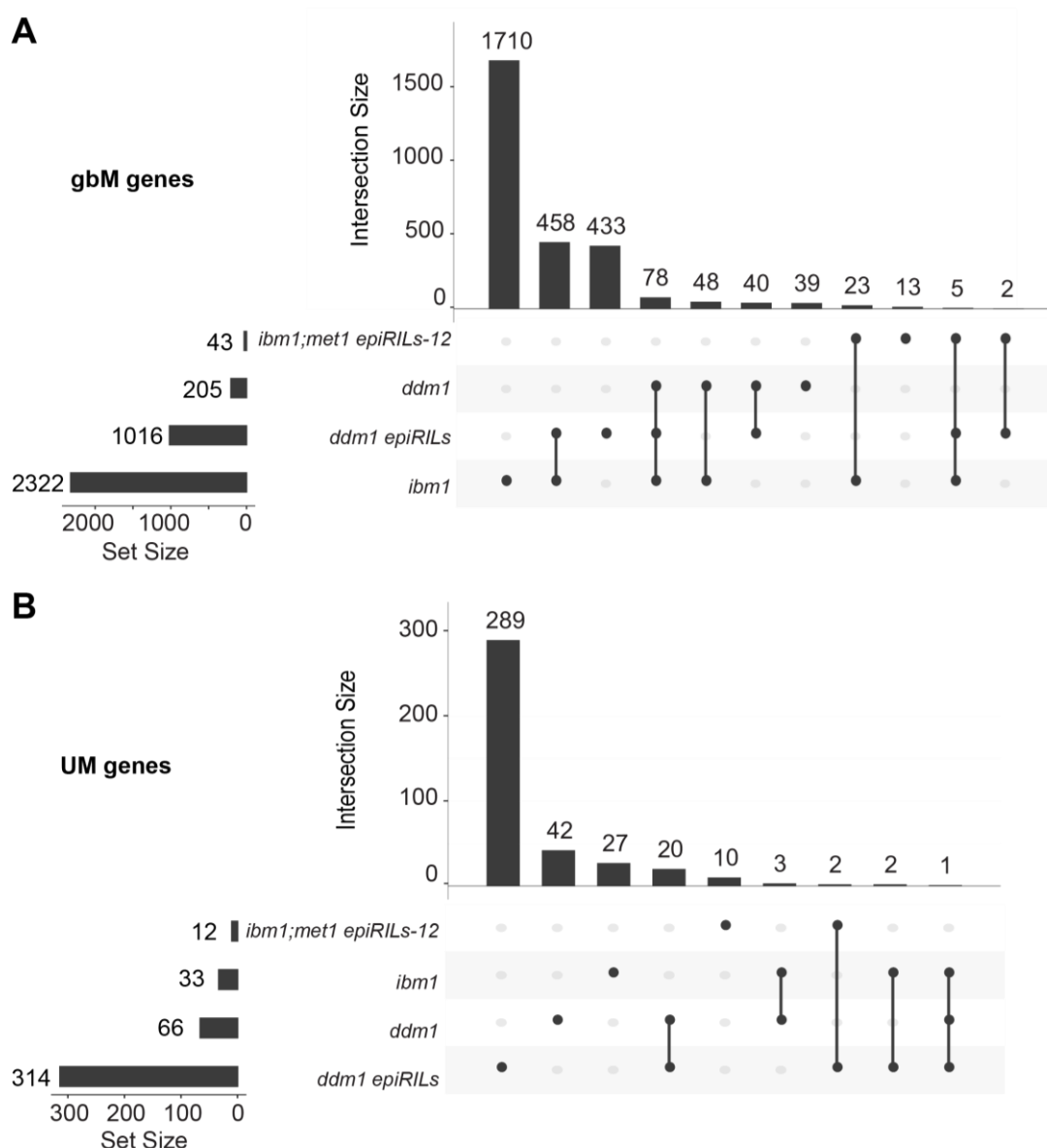


Figure S4.7. Total size and intersections of ectopic mCHG-gain genes among four datasets in this study. (A) An Upset plot shows the total number of mCHG-gain gbM genes for each of four datasets (barplot at the left side panel). Bars in the middle show the number of overlapped genes for each combination of datasets that indicated by linked black dots in the bottom panel. For bars with a single dot at bottom, it gives the number of genes that uniquely exist in a given dataset. **(B)** Same as above, but for ectopic mCHG happened on UM genes.

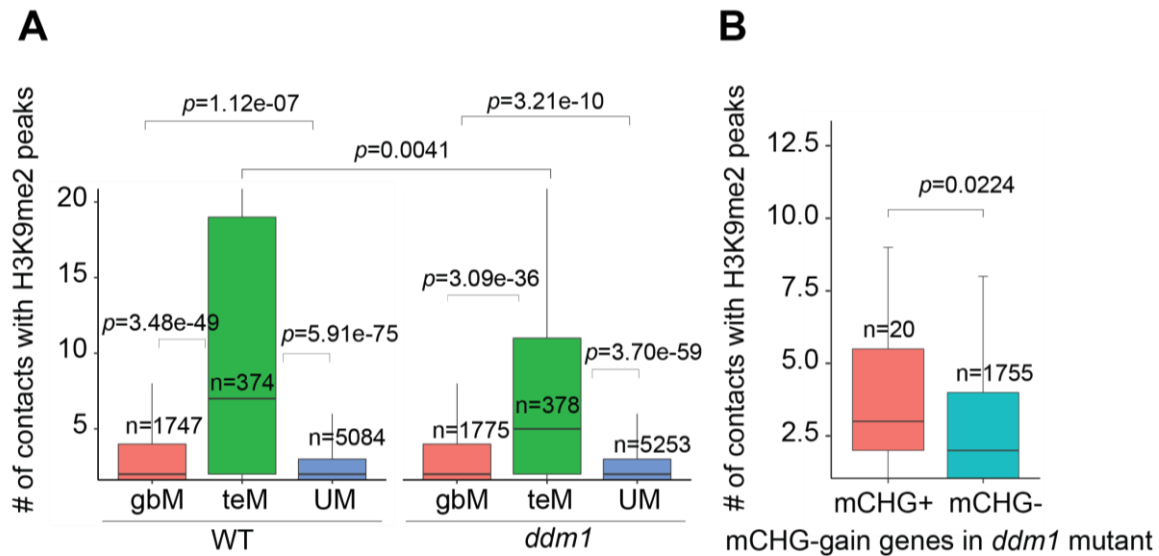


Figure S4.8. Loci that are susceptible to spontaneous epiallele formation in *ddm1* mutant have a greater contact frequency with H3K9me2 region. (A) Distribution of the number of contacts between three types of genes (gbM, UM, teM) and H3K9me2 regions. Hi-C data for both wild-type (left panel) and *ddm1* mutant (right panel) were used for identifying significant interactions between genomic regions. (B) gbM genes were further classified into a group with ectopic mCHG (mCHG+) and a group without ectopic mCHG (mCHG-) based on *ddm1* 1G mutant line data. Distribution of the number of contacts between the two group of genes and H3K9me2 regions was generated based on Hi-C data of *ddm1* mutant. Two-sample Wilcoxon tests was used for pairwise comparison of each two groups in this analysis.

REFERENCES

1. Finnegan, E.J., W.J. Peacock, and E.S. Dennis, Reduced DNA methylation in *Arabidopsis thaliana* results in abnormal plant development. *Proc Natl Acad Sci U S A*, 1996. 93(16): p. 8449-54.
2. Robertson, K.D., DNA methylation and human disease. *Nat Rev Genet*, 2005. 6(8): p. 597-610.
3. Slotkin, R.K. and R. Martienssen, Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet*, 2007. 8(4): p. 272-85.
4. Coulondre, C., et al., Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature*, 1978. 274(5673): p. 775-80.
5. Lister, R., et al., Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 2009. 462(7271): p. 315-22.
6. He, X.J., T. Chen, and J.K. Zhu, Regulation and function of DNA methylation in plants and animals. *Cell Res*, 2011. 21(3): p. 442-65.
7. Zhang, H., Z. Lang, and J.K. Zhu, Dynamics and function of DNA methylation in plants. *Nat Rev Mol Cell Biol*, 2018. 19(8): p. 489-506.
8. Bostick, M., et al., UHRF1 plays a role in maintaining DNA methylation in mammalian cells. *Science*, 2007. 317(5845): p. 1760-4.
9. Lister, R., et al., Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*, 2008. 133(3): p. 523-36.
10. Zhang, X., et al., Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell*, 2006. 126(6): p. 1189-201.
11. Kankel, M.W., et al., *Arabidopsis* MET1 cytosine methyltransferase mutants. *Genetics*, 2003. 163(3): p. 1109-22.
12. Woo, H.R., et al., VIM1, a methylcytosine-binding protein required for centromeric heterochromatinization. *Genes Dev*, 2007. 21(3): p. 267-77.
13. Lindroth, A.M., et al., Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. *Science*, 2001. 292(5524): p. 2077-80.
14. Stroud, H., et al., Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*. *Nat Struct Mol Biol*, 2014. 21(1): p. 64-72.

15. Pontes, O., et al., The Arabidopsis chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell*, 2006. 126(1): p. 79-92.
16. Zhong, X., et al., Molecular mechanism of action of plant DRM de novo DNA methyltransferases. *Cell*, 2014. 157(5): p. 1050-60.
17. Du, J., et al., DNA methylation pathways and their crosstalk with histone methylation. *Nat Rev Mol Cell Biol*, 2015. 16(9): p. 519-32.
18. Du, J., et al., Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. *Cell*, 2012. 151(1): p. 167-80.
19. Jackson, J.P., et al., Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature*, 2002. 416(6880): p. 556-60.
20. Jackson, J.P., et al., Dimethylation of histone H3 lysine 9 is a critical mark for DNA methylation and gene silencing in *Arabidopsis thaliana*. *Chromosoma*, 2004. 112(6): p. 308-15.
21. Ebbs, M.L., L. Bartee, and J. Bender, H3 lysine 9 methylation is maintained on a transcribed inverted repeat by combined action of SUVH6 and SUVH4 methyltransferases. *Mol Cell Biol*, 2005. 25(23): p. 10507-15.
22. Ebbs, M.L. and J. Bender, Locus-specific control of DNA methylation by the Arabidopsis SUVH5 histone methyltransferase. *Plant Cell*, 2006. 18(5): p. 1166-76.
23. Du, J., et al., Mechanism of DNA methylation-directed histone methylation by KRYPTONITE. *Mol Cell*, 2014. 55(3): p. 495-504.
24. Johnson, L.M., et al., The SRA methyl-cytosine-binding domain links DNA and histone methylation. *Curr Biol*, 2007. 17(4): p. 379-84.
25. Stroud, H., et al., Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. *Cell*, 2013. 152(1-2): p. 352-64.
26. Zhang, X., et al., Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis thaliana*. *Genome Biol*, 2009. 10(6): p. R62.
27. Searle, I.R., et al., JMJ14, a JmjC domain protein, is required for RNA silencing and cell-to-cell movement of an RNA silencing signal in *Arabidopsis*. *Genes Dev*, 2010. 24(10): p. 986-91.
28. Bannister, A.J. and T. Kouzarides, Regulation of chromatin by histone modifications. *Cell Res*, 2011. 21(3): p. 381-95.

29. Allis, C.D. and T. Jenuwein, The molecular hallmarks of epigenetic control. *Nat Rev Genet*, 2016. 17(8): p. 487-500.
30. Ernst, J. and M. Kellis, Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol*, 2010. 28(8): p. 817-25.
31. Liu, Y., et al., PCSD: a plant chromatin state database. *Nucleic Acids Res*, 2018. 46(D1): p. D1157-D1167.
32. Jones, P.A., Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat Rev Genet*, 2012. 13(7): p. 484-92.
33. Heyn, H. and M. Esteller, An Adenine Code for DNA: A Second Life for N6-Methyladenine. *Cell*, 2015. 161(4): p. 710-3.
34. Zhao, Z., et al., Accurate prediction of DNA N(4)-methylcytosine sites via boost-learning various types of sequence features. *BMC Genomics*, 2020. 21(1): p. 627.
35. Colot, V. and J.L. Rossignol, Eukaryotic DNA methylation as an evolutionary device. *Bioessays*, 1999. 21(5): p. 402-11.
36. Jeltsch, A., Molecular biology. Phylogeny of methylomes. *Science*, 2010. 328(5980): p. 837-8.
37. Tang, Y., et al., Widespread existence of cytosine methylation in yeast DNA measured by gas chromatography/mass spectrometry. *Anal Chem*, 2012. 84(16): p. 7249-55.
38. Greer, E.L., et al., DNA Methylation on N6-Adenine in *C. elegans*. *Cell*, 2015. 161(4): p. 868-78.
39. Deshmukh, S., et al., Levels of DNA cytosine methylation in the *Drosophila* genome. *PeerJ*, 2018. 6: p. e5119.
40. Smith, Z.D. and A. Meissner, DNA methylation: roles in mammalian development. *Nat Rev Genet*, 2013. 14(3): p. 204-20.
41. Niederhuth, C.E., et al., Widespread natural variation of DNA methylation within angiosperms. *Genome Biol*, 2016. 17(1): p. 194.
42. Nai, Y.S., et al., Diversity of Fungal DNA Methyltransferases and Their Association With DNA Methylation Patterns. *Front Microbiol*, 2020. 11: p. 616922.
43. Lyko, F., et al., The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biol*, 2010. 8(11): p. e1000506.

44. Bewick, A.J., et al., Evolution of DNA Methylation across Insects. *Mol Biol Evol*, 2017. 34(3): p. 654-665.
45. Bewick, A.J., et al., Diversity of cytosine methylation across the fungal tree of life. *Nat Ecol Evol*, 2019. 3(3): p. 479-490.
46. Bewick, A.J., et al., The evolution of CHROMOMETHYLASES and gene body DNA methylation in plants. *Genome Biol*, 2017. 18(1): p. 65.
47. Feng, S., S.E. Jacobsen, and W. Reik, Epigenetic reprogramming in plant and animal development. *Science*, 2010. 330(6004): p. 622-7.
48. Surani, M.A., K. Hayashi, and P. Hajkova, Genetic and epigenetic regulators of pluripotency. *Cell*, 2007. 128(4): p. 747-62.
49. Sasaki, H. and Y. Matsui, Epigenetic events in mammalian germ-cell development: reprogramming and beyond. *Nat Rev Genet*, 2008. 9(2): p. 129-40.
50. Hemberger, M., W. Dean, and W. Reik, Epigenetic dynamics of stem cells and cell lineage commitment: digging Waddington's canal. *Nat Rev Mol Cell Biol*, 2009. 10(8): p. 526-37.
51. Messerschmidt, D.M., B.B. Knowles, and D. Solter, DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes Dev*, 2014. 28(8): p. 812-28.
52. Law, J.A. and S.E. Jacobsen, Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet*, 2010. 11(3): p. 204-20.
53. Niederhuth, C.E. and R.J. Schmitz, Covering your bases: inheritance of DNA methylation in plant genomes. *Mol Plant*, 2014. 7(3): p. 472-80.
54. Duncan, B.K. and J.H. Miller, Mutagenic deamination of cytosine residues in DNA. *Nature*, 1980. 287(5782): p. 560-1.
55. Shen, J.C., W.M. Rideout, 3rd, and P.A. Jones, The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res*, 1994. 22(6): p. 972-6.
56. Schmutte, C., et al., Base excision repair of U:G mismatches at a mutational hotspot in the p53 gene is more efficient than base excision repair of T:G mismatches in extracts of human colon tumors. *Cancer Res*, 1995. 55(17): p. 3742-6.
57. Krokan, H.E. and M. Bjoras, Base excision repair. *Cold Spring Harb Perspect Biol*, 2013. 5(4): p. a012583.

58. Bird, A.P., DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res*, 1980. 8(7): p. 1499-504.
59. Simmen, M.W., Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals. *Genomics*, 2008. 92(1): p. 33-40.
60. Lander, E.S., et al., Initial sequencing and analysis of the human genome. *Nature*, 2001. 409(6822): p. 860-921.
61. Lee, H., et al., Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc Natl Acad Sci U S A*, 2012. 109(41): p. E2774-83.
62. Xia, J., L. Han, and Z. Zhao, Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome. *BMC Genomics*, 2012. 13 Suppl 8: p. S7.
63. Weng, M.L., et al., Fine-Grained Analysis of Spontaneous Mutation Spectrum and Frequency in *Arabidopsis thaliana*. *Genetics*, 2019. 211(2): p. 703-714.
64. Kusmartsev, V., et al., Cytosine Methylation Affects the Mutability of Neighboring Nucleotides in Germline and Soma. *Genetics*, 2020. 214(4): p. 809-823.
65. Robertson, K.D. and P.A. Jones, DNA methylation: past, present and future directions. *Carcinogenesis*, 2000. 21(3): p. 461-7.
66. Greenblatt, M.S., et al., Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. *Cancer Res*, 1994. 54(18): p. 4855-78.
67. Laird, P.W., Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet*, 2010. 11(3): p. 191-203.
68. Schmitz, R.J. and X. Zhang, High-throughput approaches for plant epigenomic studies. *Curr Opin Plant Biol*, 2011. 14(2): p. 130-6.
69. Cokus, S.J., et al., Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. *Nature*, 2008. 452(7184): p. 215-9.
70. Ziller, M.J., et al., Charting a dynamic DNA methylation landscape of the human genome. *Nature*, 2013. 500(7463): p. 477-81.
71. Kawakatsu, T., et al., Epigenomic Diversity in a Global Collection of *Arabidopsis thaliana* Accessions. *Cell*, 2016. 166(2): p. 492-505.

72. Smallwood, S.A., et al., Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods*, 2014. 11(8): p. 817-820.
73. Clark, S.J., et al., Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. *Genome Biol*, 2016. 17: p. 72.
74. Clark, S.J., et al., Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat Protoc*, 2017. 12(3): p. 534-547.
75. Schmitz, R.J., et al., Patterns of population epigenomic diversity. *Nature*, 2013. 495(7440): p. 193-8.
76. Schmitz, R.J. and J.R. Ecker, Epigenetic and epigenomic variation in *Arabidopsis thaliana*. *Trends Plant Sci*, 2012. 17(3): p. 149-54.
77. Vaughn, M.W., et al., Epigenetic natural variation in *Arabidopsis thaliana*. *PLoS Biol*, 2007. 5(7): p. e174.
78. Richards, E.J., Inherited epigenetic variation--revisiting soft inheritance. *Nat Rev Genet*, 2006. 7(5): p. 395-401.
79. Schmitz, R.J., et al., Transgenerational epigenetic instability is a source of novel methylation variants. *Science*, 2011. 334(6054): p. 369-73.
80. Saze, H., O. Mittelsten Scheid, and J. Paszkowski, Maintenance of CpG methylation is essential for epigenetic inheritance during plant gametogenesis. *Nat Genet*, 2003. 34(1): p. 65-9.
81. Penterman, J., R. Uzawa, and R.L. Fischer, Genetic interactions between DNA demethylation and methylation in *Arabidopsis*. *Plant Physiol*, 2007. 145(4): p. 1549-57.
82. Bender, J. and G.R. Fink, Epigenetic control of an endogenous gene family is revealed by a novel blue fluorescent mutant of *Arabidopsis*. *Cell*, 1995. 83(5): p. 725-34.
83. Jacobsen, S.E. and E.M. Meyerowitz, Hypermethylated SUPERMAN epigenetic alleles in *arabidopsis*. *Science*, 1997. 277(5329): p. 1100-3.
84. Soppe, W.J., et al., The late flowering phenotype of *fwa* mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. *Mol Cell*, 2000. 6(4): p. 791-802.
85. Liu, J., et al., siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in *Arabidopsis*. *Genes Dev*, 2004. 18(23): p. 2873-8.

86. Saze, H. and T. Kakutani, Heritable epigenetic mutation of a transposon-flanked *Arabidopsis* gene due to lack of the chromatin-remodeling factor DDM1. *EMBO J*, 2007. 26(15): p. 3641-52.
87. Tsukahara, S., et al., Bursts of retrotransposition reproduced in *Arabidopsis*. *Nature*, 2009. 461(7262): p. 423-6.
88. Becker, C., et al., Spontaneous epigenetic variation in the *Arabidopsis thaliana* methylome. *Nature*, 2011. 480(7376): p. 245-9.
89. Cubas, P., C. Vincent, and E. Coen, An epigenetic mutation responsible for natural variation in floral symmetry. *Nature*, 1999. 401(6749): p. 157-61.
90. Manning, K., et al., A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nat Genet*, 2006. 38(8): p. 948-52.
91. Stam, M., et al., Differential chromatin structure within a tandem array 100 kb upstream of the maize b1 locus is associated with paramutation. *Genes Dev*, 2002. 16(15): p. 1906-18.
92. Johannes, F. and R.J. Schmitz, Spontaneous epimutations in plants. *New Phytol*, 2019. 221(3): p. 1253-1259.
93. Zhang, X., et al., Global analysis of genetic, epigenetic and transcriptional polymorphisms in *Arabidopsis thaliana* using whole genome tiling arrays. *PLoS Genet*, 2008. 4(3): p. e1000032.
94. Zheng, X., et al., Transgenerational epimutations induced by multi-generation drought imposition mediate rice plant's adaptation to drought condition. *Sci Rep*, 2017. 7: p. 39843.
95. Shaw, R.G., D.L. Byers, and E. Darmono, Spontaneous mutational effects on reproductive traits of *Arabidopsis thaliana*. *Genetics*, 2000. 155(1): p. 369-78.
96. Hofmeister, B.T., et al., Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. *Genome Biol*, 2017. 18(1): p. 155.
97. Yao, N., R.J. Schmitz, and F. Johannes, Epimutations Define a Fast-Ticking Molecular Clock in Plants. *Trends Genet*, 2021. 37(8): p. 699-710.
98. Jiang, C., et al., Environmentally responsive genome-wide accumulation of de novo *Arabidopsis thaliana* mutations and epimutations. *Genome Res*, 2014. 24(11): p. 1821-9.
99. Mathieu, O., et al., Transgenerational stability of the *Arabidopsis* epigenome is coordinated by CG methylation. *Cell*, 2007. 130(5): p. 851-62.

100. Deleris, A., et al., Loss of the DNA methyltransferase MET1 Induces H3K9 hypermethylation at PcG target genes and redistribution of H3K27 trimethylation to transposons in *Arabidopsis thaliana*. *PLoS Genet*, 2012. 8(11): p. e1003062.
101. Ronemus, M.J., et al., Demethylation-induced developmental pleiotropy in *Arabidopsis*. *Science*, 1996. 273(5275): p. 654-7.
102. Jeddeloh, J.A., T.L. Stokes, and E.J. Richards, Maintenance of genomic methylation requires a SWI2/SNF2-like protein. *Nat Genet*, 1999. 22(1): p. 94-7.
103. Vongs, A., et al., *Arabidopsis thaliana* DNA methylation mutants. *Science*, 1993. 260(5116): p. 1926-8.
104. Zemach, A., et al., The *Arabidopsis* nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. *Cell*, 2013. 153(1): p. 193-205.
105. Lippman, Z., et al., Role of transposable elements in heterochromatin and epigenetic control. *Nature*, 2004. 430(6998): p. 471-6.
106. Kakutani, T., et al., Developmental abnormalities and epimutations associated with DNA hypomethylation mutations. *Proc Natl Acad Sci U S A*, 1996. 93(22): p. 12406-11.
107. Saze, H., et al., Control of genic DNA methylation by a *jmjC* domain-containing protein in *Arabidopsis thaliana*. *Science*, 2008. 319(5862): p. 462-5.
108. Miura, A., et al., An *Arabidopsis* *jmjC* domain protein protects transcribed genes from DNA methylation at CHG sites. *EMBO J*, 2009. 28(8): p. 1078-86.
109. Inagaki, S., et al., Autocatalytic differentiation of epigenetic modifications within the *Arabidopsis* genome. *EMBO J*, 2010. 29(20): p. 3496-506.
110. Johannes, F., et al., Assessing the impact of transgenerational epigenetic variation on complex traits. *PLoS Genet*, 2009. 5(6): p. e1000530.
111. Reinders, J., et al., Compromised stability of DNA methylation and transposon immobilization in mosaic *Arabidopsis* epigenomes. *Genes Dev*, 2009. 23(8): p. 939-50.
112. Teixeira, F.K., et al., A role for RNAi in the selective correction of DNA methylation defects. *Science*, 2009. 323(5921): p. 1600-4.
113. Kooke, R., et al., Epigenetic basis of morphological variation and phenotypic plasticity in *Arabidopsis thaliana*. *Plant Cell*, 2015. 27(2): p. 337-48.

114. Colome-Tatche, M., et al., Features of the Arabidopsis recombination landscape resulting from the combined loss of sequence variation and DNA methylation. *Proc Natl Acad Sci U S A*, 2012. 109(40): p. 16240-5.
115. Tran, R.K., et al., DNA methylation profiling identifies CG methylation clusters in Arabidopsis genes. *Curr Biol*, 2005. 15(2): p. 154-9.
116. Bewick, A.J., et al., On the origin and evolutionary consequences of gene body DNA methylation. *Proc Natl Acad Sci U S A*, 2016. 113(32): p. 9111-6.
117. Zilberman, D., et al., Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet*, 2007. 39(1): p. 61-9.
118. Dubin, M.J., et al., DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. *Elife*, 2015. 4: p. e05255.
119. Ito, T., et al., Genome-wide negative feedback drives transgenerational DNA methylation dynamics in Arabidopsis. *PLoS Genet*, 2015. 11(4): p. e1005154.
120. Flavahan, W.A., E. Gaskell, and B.E. Bernstein, Epigenetic plasticity and the hallmarks of cancer. *Science*, 2017. 357(6348).
121. Bewick, A.J. and R.J. Schmitz, Gene body DNA methylation in plants. *Curr Opin Plant Biol*, 2017. 36: p. 103-110.
122. Wendte, J.M. and R.J. Schmitz, Specifications of Targeting Heterochromatin Modifications in Plants. *Mol Plant*, 2018. 11(3): p. 381-387.
123. Takuno, S. and B.S. Gaut, Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proc Natl Acad Sci U S A*, 2013. 110(5): p. 1797-802.
124. Seymour, D.K., et al., Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. *PLoS Genet*, 2014. 10(11): p. e1004785.
125. Takuno, S., J.H. Ran, and B.S. Gaut, Evolutionary patterns of genic DNA methylation vary across land plants. *Nat Plants*, 2016. 2: p. 15222.
126. Regulski, M., et al., The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. *Genome Res*, 2013. 23(10): p. 1651-62.
127. Wang, H., et al., CG gene body DNA methylation changes and evolution of duplicated genes in cassava. *Proc Natl Acad Sci U S A*, 2015. 112(44): p. 13729-34.

128. Zemach, A., et al., Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science*, 2010. 328(5980): p. 916-9.
129. Feng, S., et al., Conservation and divergence of methylation patterning in plants and animals. *Proc Natl Acad Sci U S A*, 2010. 107(19): p. 8689-94.
130. Huff, J.T. and D. Zilberman, Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. *Cell*, 2014. 156(6): p. 1286-1297.
131. Vial-Pradel, S., et al., Arabidopsis Zinc-Finger-Like Protein ASYMMETRIC LEAVES2 (AS2) and Two Nucleolar Proteins Maintain Gene Body DNA Methylation in the Leaf Polarity Gene ETTIN (ARF3). *Plant Cell Physiol*, 2018. 59(7): p. 1385-1397.
132. Zilberman, D., et al., Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. *Nature*, 2008. 456(7218): p. 125-9.
133. Lorincz, M.C., et al., Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat Struct Mol Biol*, 2004. 11(11): p. 1068-75.
134. Maunakea, A.K., et al., Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, 2010. 466(7303): p. 253-7.
135. Choi, J., et al., DNA methylation and histone H1 cooperatively repress transposable elements and aberrant intragenic transcripts. *bioRxiv*, 2019.
136. Bewick, A.J., et al., Evolutionary and Experimental Loss of Gene Body Methylation and Its Consequence to Gene Expression. *G3 (Bethesda)*, 2019.
137. Inagaki, S. and T. Kakutani, What triggers differential DNA methylation of genes and TEs: contribution of body methylation? *Cold Spring Harb Symp Quant Biol*, 2012. 77: p. 155-60.
138. Teixeira, F.K. and V. Colot, Gene body DNA methylation in plants: a means to an end or an end to a means? *EMBO J*, 2009. 28(8): p. 997-8.
139. Matzke, M.A. and R.A. Mosher, RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat Rev Genet*, 2014. 15(6): p. 394-408.
140. Zhou, M. and J.A. Law, RNA Pol IV and V in gene silencing: Rebel polymerases evolving away from Pol II's rules. *Curr Opin Plant Biol*, 2015. 27: p. 154-64.
141. Wendte, J.M. and C.S. Pikaard, The RNAs of RNA-directed DNA methylation. *Biochim Biophys Acta Gene Regul Mech*, 2017. 1860(1): p. 140-148.

142. Haag, J.R. and C.S. Pikaard, Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nat Rev Mol Cell Biol*, 2011. 12(8): p. 483-92.
143. Woo, H.R., T.A. Dittmer, and E.J. Richards, Three SRA-domain methylcytosine-binding proteins cooperate to maintain global CpG methylation and epigenetic silencing in Arabidopsis. *PLoS Genet*, 2008. 4(8): p. e1000156.
144. Bartee, L., F. Malagnac, and J. Bender, Arabidopsis cmt3 chromomethylase mutations block non-CG methylation and silencing of an endogenous gene. *Genes Dev*, 2001. 15(14): p. 1753-8.
145. Papa, C.M., et al., Maize chromomethylase Zea methyltransferase2 is required for CpNpG methylation. *Plant Cell*, 2001. 13(8): p. 1919-28.
146. Gouil, Q. and D.C. Baulcombe, DNA Methylation Signatures of the Plant Chromomethyltransferases. *PLoS Genet*, 2016. 12(12): p. e1006526.
147. Stoddard, C.I., et al., A Nucleosome Bridging Mechanism for Activation of a Maintenance DNA Methyltransferase. *Mol Cell*, 2019. 73(1): p. 73-83 e6.
148. Li, X., et al., Mechanistic insights into plant SUVH family H3K9 methyltransferases and their binding to context-biased non-CG DNA methylation. *Proc Natl Acad Sci U S A*, 2018. 115(37): p. E8793-E8802.
149. Bernatavichute, Y.V., et al., Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis thaliana. *PLoS One*, 2008. 3(9): p. e3156.
150. Malagnac, F., L. Bartee, and J. Bender, An Arabidopsis SET domain protein required for maintenance but not establishment of DNA methylation. *EMBO J*, 2002. 21(24): p. 6842-52.
151. Soppe, W.J., et al., DNA methylation controls histone H3 lysine 9 methylation and heterochromatin assembly in Arabidopsis. *EMBO J*, 2002. 21(23): p. 6549-59.
152. Tariq, M., et al., Erasure of CpG methylation in Arabidopsis alters patterns of histone H3 methylation in heterochromatin. *Proc Natl Acad Sci U S A*, 2003. 100(15): p. 8823-7.
153. Mathieu, O., A.V. Probst, and J. Paszkowski, Distinct regulation of histone H3 methylation at lysines 27 and 9 by CpG methylation in Arabidopsis. *EMBO J*, 2005. 24(15): p. 2783-91.
154. Clough, S.J. and A.F. Bent, Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant J*, 1998. 16(6): p. 735-43.
155. Takuno, S. and B.S. Gaut, Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly. *Mol Biol Evol*, 2012. 29(1): p. 219-27.

156. Urich, M.A., et al., MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. *Nat Protoc*, 2015. 10(3): p. 475-83.
157. Schubert, D., et al., Silencing by plant Polycomb-group genes requires dispersed trimethylation of histone H3 at lysine 27. *EMBO J*, 2006. 25(19): p. 4638-49.
158. Schultz, M.D., et al., Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature*, 2015. 523(7559): p. 212-6.
159. Martin, M., Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011, 2011. 17(1): p. 3.
160. Yang, R., et al., The Reference Genome of the Halophytic Plant *Eutrema salsugineum*. *Front Plant Sci*, 2013. 4: p. 46.
161. Langmead, B. and S.L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 2012. 9(4): p. 357-9.
162. Khan, A. and A. Mathelier, Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinformatics*, 2017. 18(1): p. 287.
163. Quinlan, A.R. and I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 2010. 26(6): p. 841-2.
164. Bolger, A.M., M. Lohse, and B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 2014. 30(15): p. 2114-20.
165. Kim, D., B. Langmead, and S.L. Salzberg, HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*, 2015. 12(4): p. 357-60.
166. Pertea, M., et al., Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc*, 2016. 11(9): p. 1650-67.
167. Langmead, B., et al., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 2009. 10(3): p. R25.
168. Li, H., et al., The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 2009. 25(16): p. 2078-9.
169. Heinz, S., et al., Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, 2010. 38(4): p. 576-89.

170. Livak, K.J. and T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the 2⁻(Delta Delta C(T)) Method. *Methods*, 2001. 25(4): p. 402-8.
171. Elgin, S.C., Heterochromatin and gene regulation in *Drosophila*. *Curr Opin Genet Dev*, 1996. 6(2): p. 193-202.
172. Lachner, M., et al., Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature*, 2001. 410(6824): p. 116-20.
173. Bannister, A.J., et al., Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature*, 2001. 410(6824): p. 120-4.
174. Nakayama, J., et al., Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science*, 2001. 292(5514): p. 110-3.
175. Lindroth, A.M., et al., Dual histone H3 methylation marks at lysines 9 and 27 required for interaction with CHROMOMETHYLASE3. *EMBO J*, 2004. 23(21): p. 4286-96.
176. Du, J.M., et al., DNA methylation pathways and their crosstalk with histone methylation. *Nature Reviews Molecular Cell Biology*, 2015. 16(9): p. 519-532.
177. Ito, T., et al., Genome-Wide Negative Feedback Drives Transgenerational DNA Methylation Dynamics in *Arabidopsis*. *Plos Genetics*, 2015. 11(4).
178. Kawakatsu, T., et al., Epigenomic Diversity in a Global Collection of *Arabidopsis thaliana* Accessions. *Cell*, 2016. 166(2): p. 492-505.
179. Niederhuth, C.E., et al., Widespread natural variation of DNA methylation within angiosperms. *Genome Biology*, 2016. 17(1): p. 194.
180. Bewick, A.J., et al., On the origin and evolutionary consequences of gene body DNA methylation. *Proceedings of the National Academy of Sciences of the United States of America*, 2016. 113(32): p. 9111-9116.
181. Kiefer, C., et al., Interspecies association mapping links reduced CG to TG substitution rates to the loss of gene-body methylation. *Nature Plants*, 2019. 5(8): p. 846-855.
182. Bewick, A.J. and R.J. Schmitz, Gene body DNA methylation in plants. *Current Opinion in Plant Biology*, 2017. 36: p. 103-110.
183. Zilberman, D., An evolutionary case for functional gene body methylation in plants and animals. *Genome Biology*, 2017. 18(1): p. 87.

184. Picard, C.L. and M. Gehring, Proximal methylation features associated with nonrandom changes in gene body methylation. *Genome Biol*, 2017. 18(1): p. 73.
185. van der Graaf, A., et al., Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proc Natl Acad Sci U S A*, 2015. 112(21): p. 6676-81.
186. Seymour, D.K. and B.S. Gaut, Phylogenetic shifts in gene body methylation correlate with gene expression and reflect trait conservation. *Molecular Biology and Evolution*, 2019.
187. Bewick, A.J., et al., Evolutionary and Experimental Loss of Gene Body Methylation and Its Consequence to Gene Expression. *G3-Genes Genomes Genetics*, 2019. 9(8): p. 2441-2445.
188. Takuno, S., D.K. Seymour, and B.S. Gaut, The Evolutionary Dynamics of Orthologs That Shift in Gene Body Methylation between *Arabidopsis* Species. *Molecular Biology and Evolution*, 2017. 34(6): p. 1479-1491.
189. Seymour, D.K., et al., Evolution of DNA Methylation Patterns in the Brassicaceae is Driven by Differences in Genome Organization. *Plos Genetics*, 2014. 10(11).
190. Takuno, S. and B.S. Gaut, Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. *Proceedings of the National Academy of Sciences of the United States of America*, 2013. 110(5): p. 1797-1802.
191. Zilberman, D., et al., Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature Genetics*, 2007. 39(1): p. 61-9.
192. Muyle, A. and B.S. Gaut, Loss of Gene Body Methylation in *Eutrema salsugineum* Is Associated with Reduced Gene Expression. *Molecular Biology and Evolution*, 2019. 36(1): p. 155-158.
193. Wendte, J.M., et al., Epimutations are associated with CHROMOMETHYLASE 3-induced de novo DNA methylation. *Elife*, 2019. 8.
194. Inagaki, S. and T. Kakutani, What Triggers Differential DNA Methylation of Genes and TEs: Contribution of Body Methylation? *Cold Spring Harb Symp Quant Biol*, 2012. 77(155-160).
195. Stoddard, C.I., et al., A Nucleosome Bridging Mechanism for Activation of a Maintenance DNA Methyltransferase. *Molecular Cell*, 2019. 73(1): p. 73-83 e6.
196. Gouil, Q. and D.C. Baulcombe, DNA Methylation Signatures of the Plant Chromomethyltransferases. *Plos Genetics*, 2016. 12(12): p. e1006526.
197. Woo, H.R. and E.J. Richards, Natural variation in DNA methylation in ribosomal RNA genes of *Arabidopsis thaliana*. *BMC Plant Biol*, 2008. 8: p. 92.

198. Williams, B.P. and M. Gehring, Stable transgenerational epigenetic inheritance requires a DNA methylation-sensing circuit. *Nat Commun*, 2017. 8(1): p. 2124.
199. Tabata, S., et al., Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature*, 2000. 408(6814): p. 823-6.
200. Carazo-Salas, R.E., et al., Generation of GTP-bound Ran by RCC1 is required for chromatin-induced mitotic spindle formation. *Nature*, 1999. 400(6740): p. 178-81.
201. Hergeth, S.P. and R. Schneider, The H1 linker histones: multifunctional proteins beyond the nucleosomal core particle. *EMBO Rep*, 2015. 16(11): p. 1439-53.
202. Takuno, S. and B.S. Gaut, Body-Methylated Genes in *Arabidopsis thaliana* Are Functionally Important and Evolve Slowly. *Molecular Biology and Evolution*, 2012. 29(1): p. 219-227.
203. Berardini, T.Z., et al., The *Arabidopsis* information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis*, 2015. 53(8): p. 474-85.
204. Jeddeloh, J.A., T.L. Stokes, and E.J. Richards, Maintenance of genomic methylation requires a SWI2/SNF2-like protein. *Nature Genetics*, 1999. 22(1): p. 94-7.
205. Blevins, T., et al., Hybrid incompatibility caused by an epiallele. *Proc Natl Acad Sci U S A*, 2017. 114(14): p. 3702-3707.
206. Agorio, A., et al., An *Arabidopsis* Natural Epiallele Maintained by a Feed-Forward Silencing Loop between Histone and DNA. *Plos Genetics*, 2017. 13(1): p. e1006551.
207. Chen, W., et al., Requirement of CHROMOMETHYLASE3 for somatic inheritance of the spontaneous tomato epimutation Colourless non-ripening. *Sci Rep*, 2015. 5: p. 9192.
208. Rigal, M., et al., DNA methylation in an intron of the IBM1 histone demethylase gene stabilizes chromatin modification patterns. *EMBO J*, 2012. 31(13): p. 2981-93.
209. Silveira, A.B., et al., Extensive natural epigenetic variation at a de novo originated gene. *Plos Genetics*, 2013. 9(4): p. e1003437.
210. Durand, S., et al., Rapid establishment of genetic incompatibility through natural epigenetic variation. *Curr Biol*, 2012. 22(4): p. 326-31.
211. Luff, B., L. Pawlowski, and J. Bender, An inverted repeat triggers cytosine methylation of identical sequences in *Arabidopsis*. *Molecular Cell*, 1999. 3(4): p. 505-11.

212. Quadrana, L., et al., Natural occurring epialleles determine vitamin E accumulation in tomato fruits. *Nat Commun*, 2014. 5: p. 3027.
213. Manning, K., et al., A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nature Genetics*, 2006. 38(8): p. 948-52.
214. Alonso-Blanco, C., et al., 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell*, 2016. 166(2): p. 481-491.
215. Martin, M., Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 2011. 17(1): p. 10-12.
216. Cheng, C.Y., et al., Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J*, 2017. 89(4): p. 789-804.
217. Kumar, S., G. Stecher, and K. Tamura, MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, 2016. 33(7): p. 1870-1874.
218. Zapata, L., et al., Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. *Proc Natl Acad Sci U S A*, 2016. 113(28): p. E4052-60.
219. Gan, X., et al., Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*, 2011. 477(7365): p. 419-23.
220. Bewick, A.J., et al., FASTmC: A Suite of Predictive Models for Nonreference-Based Estimations of DNA Methylation. *G3 (Bethesda)*, 2015. 6(2): p. 447-52.
221. Paradis, E., J. Claude, and K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, 2004. 20(2): p. 289-90.
222. Zhang, Z., et al., Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, 2010. 42(4): p. 355-60.
223. Lipka, A.E., et al., GAPIT: genome association and prediction integrated tool. *Bioinformatics*, 2012. 28(18): p. 2397-9.
224. Livak, K.J. and T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the 2(T)(-Delta Delta C) method. *Methods*, 2001. 25(4): p. 402-408.
225. Bolstad, B.M., et al., A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 2003. 19(2): p. 185-93.

226. Heard, E. and R.A. Martienssen, Transgenerational epigenetic inheritance: myths and mechanisms. *Cell*, 2014. 157(1): p. 95-109.
227. Richards, E.J., Inherited epigenetic variation--revisiting soft inheritance. *Nat. Rev. Genet.*, 2006. 7(5): p. 395-401.
228. Springer, N.M. and R.J. Schmitz, Exploiting induced and natural epigenetic variation for crop improvement. *Nat Rev Genet*, 2017. 18(9): p. 563-575.
229. Cubas, P., C. Vincent, and E. Coen, An epigenetic mutation responsible for natural variation in floral symmetry. *Nature*, 1999. 401(6749): p. 157-61.
230. Ong-Abdullah, M., et al., Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature*, 2015. 525(7570): p. 533-7.
231. Henderson, I. and S. Jacobsen, Tandem repeats upstream of the Arabidopsis endogene SDC recruit non-CG DNA methylation and initiate siRNA spreading. *Genes & Development*, 2008. 22(12): p. 1597-1606.
232. Cortijo, S., et al., Mapping the epigenetic basis of complex traits. *Science*, 2014. 343(6175): p. 1145-8.
233. Borges, F., et al., Loss of Small-RNA-Directed DNA Methylation in the Plant Cell Cycle Promotes Germline Reprogramming and Somaclonal Variation. *Curr Biol*, 2021. 31(3): p. 591-600 e4.
234. Cao, X. and S.E. Jacobsen, Role of the arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing. *Curr Biol*, 2002. 12(13): p. 1138-44.
235. Papareddy, R.K., et al., Chromatin regulates expression of small RNAs to help maintain transposon methylome homeostasis in Arabidopsis. *Genome Biol*, 2020. 21(1): p. 251.
236. Singh, J., M. Freeling, and D. Lisch, A position effect on the heritability of epigenetic silencing. *PLoS Genet*, 2008. 4(10): p. e1000216.
237. Rigal, M., et al., Epigenome confrontation triggers immediate reprogramming of DNA methylation and transposon silencing in Arabidopsis thaliana F1 epihybrids. *Proc Natl Acad Sci U S A*, 2016. 113(14): p. E2083-92.
238. Lauss, K., et al., Parental DNA Methylation States Are Associated with Heterosis in Epigenetic Hybrids. *Plant Physiol*, 2018. 176(2): p. 1627-1645.
239. Williams, B.P., et al., Methylation-sensitive expression of a DNA demethylase gene serves as an epigenetic rheostat. *PLoS Genet*, 2015. 11(3): p. e1005142.

240. Zhang, Y., et al., Natural variation in DNA methylation homeostasis and the emergence of epialleles. *Proc Natl Acad Sci U S A*, 2020. 117(9): p. 4874-4884.
241. Kiefer, C., et al., Interspecies association mapping links reduced CG to TG substitution rates to the loss of gene-body methylation. *Nat Plants*, 2019. 5(8): p. 846-855.
242. Williams, B.P. and M. Gehring, Principles of Epigenetic Homeostasis Shared Between Flowering Plants and Mammals. *Trends Genet*, 2020. 36(10): p. 751-763.
243. Zabet, N.R., et al., Cytosine methylation at CpCpG sites triggers accumulation of non-CpG methylation in gene bodies. *Nucleic Acids Res*, 2017. 45(7): p. 3777-3784.
244. Sasaki, T., et al., RNAi-independent de novo DNA methylation revealed in *Arabidopsis* mutants of chromatin remodeling gene DDM1. *Plant J*, 2012. 70(5): p. 750-8.
245. Johnson, L.M., et al., SRA- and SET-domain-containing proteins link RNA polymerase V occupancy to DNA methylation. *Nature*, 2014. 507(7490): p. 124-8.
246. Liu, C., et al., Genome-wide analysis of chromatin packing in *Arabidopsis thaliana* at single-gene resolution. *Genome Res*, 2016. 26(8): p. 1057-68.
247. Ibarra, C.A., et al., Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. *Science*, 2012. 337(6100): p. 1360-1364.
248. Calarco, J.P., et al., Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. *Cell*, 2012. 151(1): p. 194-205.
249. Walker, J., et al., Sexual-lineage-specific DNA methylation regulates meiosis in *Arabidopsis*. *Nat Genet*, 2018. 50(1): p. 130-137.
250. Slotkin, R.K., et al., Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. *Cell*, 2009. 136(3): p. 461-72.
251. Woo, H.R., et al., VIM1, a methylcytosine-binding protein required for centromeric heterochromatinization. *Genes Dev.*, 2007. 21(3): p. 267-77.
252. Du, J., et al., Mechanism of DNA methylation-directed histone methylation by KRYPTONITE. *Mol Cell*, 2014. 55(3): p. 495-504.
253. Wendte, J.M. and R.J. Schmitz, Specifications of targeting heterochromatin modifications in plants. *Mol Plant*, 2017.

254. Law, J.A., et al., Polymerase IV occupancy at RNA-directed DNA methylation sites requires SHH1. *Nature*, 2013. 498(7454): p. 385-9.
255. Gallego-Bartolome, J., et al., Co-targeting RNA Polymerases IV and V Promotes Efficient De Novo DNA Methylation in Arabidopsis. *Cell*, 2019. 176(5): p. 1068-1082 e19.
256. To, T.K., et al., RNA interference-independent reprogramming of DNA methylation in Arabidopsis. *Nat Plants*, 2020. 6(12): p. 1455-1467.
257. Zhang, X., et al., Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS Biol*, 2007. 5(5): p. e129.
258. Luo, C., et al., Integrative analysis of chromatin states in Arabidopsis identified potential regulatory mechanisms for natural antisense transcript production. *Plant J*, 2012.
259. Yelagandula, R., et al., The histone variant H2A.W defines heterochromatin and promotes chromatin condensation in Arabidopsis. *Cell*, 2014. 158(1): p. 98-109.
260. Kujirai, T. and H. Kurumizaka, Transcription through the nucleosome. *Curr Opin Struct Biol*, 2020. 61: p. 42-49.
261. Alonso, J.M., et al., Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science*, 2003. 301(5633): p. 653-7.
262. Shahryary, Y., et al., AlphaBeta: computational inference of epimutation rates and spectra from high-throughput DNA methylation data in plants. *Genome Biol*, 2020. 21(1): p. 260.
263. Taudt, A., et al., METHimpute: imputation-guided construction of complete methylomes from WGBS data. *BMC Genomics*, 2018. 19(1): p. 444.
264. Martin, M., Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* 2011. 17(1): p. 10-12.
265. Inagaki, S., et al., Gene-body chromatin modification dynamics mediate epigenome differentiation in Arabidopsis. *EMBO J*, 2017. 36(8): p. 970-980.
266. Broman, K.W., et al., R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, 2003. 19(7): p. 889-90.
267. Servant, N., et al., HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol*, 2015. 16: p. 259.

- 268. Ay, F., T.L. Bailey, and W.S. Noble, Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res*, 2014. 24(6): p. 999-1011.
- 269. Li, H., et al., The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 2009. 25(16): p. 2078-9.
- 270. Tarasov, A., et al., Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, 2015. 31(12): p. 2032-4.
- 271. Zhang, Y., et al., Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 2008. 9(9): p. R137.