Widespread natural variation of DNA methylation within angiosperms

Chad E. Niederhuth[1*], Adam J. Bewick[1*], Lexiang Ji[2], Magdy S. Alabady[3], Kyung Do Kim[4], Justin T. Page[5], Qing Li[6], Nicholas A. Rohr[1], Aditi Rambani[6,] John M. Burke[3], Joshua A. Udall[5], Chiedozie Egesi[7], Jeremy Schmutz[8,9], Jane Grimwood[8], Scott A. Jackson[4], Nathan M. Springer[6] and Robert J. Schmitz[1]

[1]Department of Genetics, University of Georgia, Athens, GA 30602, USA

[2]Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

[3]Department of Plant Biology, University of Georgia, Athens, GA 30602, USA

[4]Center for Applied Genetic Technologies, University of Georgia, Athens, GA 30602, USA

[5]Plant and Wildlife Science Department, Brigham Young University, Provo, Utah 84602, USA

[6]Department of Plant Biology, Microbial and Plant Genomics Institute, University of Minnesota, Saint Paul, MN 55108, USA

[7]National Root Crops Research Institute (NRCRI), Umudike, Km 8 Ikot Ekpene Road, PMB 7006, Umuahia 440001, Nigeria

[8]HudsonAlpha Institute for Biotechnology, Huntsville, AL, 35806, USA

[9]Department of Energy Joint Genome Institute, Walnut Creek, California, USA

*Indicates these authors contributed equally to this work

Corresponding author:

Robert J. Schmitz
University of Georgia
120 East Green Street, Athens, GA 30602
Telephone: (706) 542-1887
Fax: (706) 542-3910
E-mail: schmitz@uga.edu

Abstract

DNA methylation can be faithfully inherited across generations in flowering plant genomes. Failure to properly maintain DNA methylation can lead to epigenetic variation and transposon reactivation. Plant genomes are dynamic, spanning large ranges in size and there is interplay between the genome and epigenome in shaping one another. To understand the variation in genomic patterning of DNA methylation between species, we compared methylomes of 34 diverse angiosperm species. By examining these variations in a phylogenetic context it becomes clear that there is extensive variation in mechanisms that govern gene body DNA methylation, euchromatic silencing of transposons and repeats, as well as silencing of heterochromatic transposons. Extensive variation is observed at all cytosine sequence contexts (CG, CHG and CHH, where H = A, C, T), with the Brassicaceae showing reduced CHG methylation levels and also reduced or loss of CG gene body methylation. The Poaceae are characterized by a lack or reduction of heterochromatic CHH methylation and enrichment of CHH methylation in genic regions. Reduced CHH methylation levels are found in clonally propagated species, suggesting that these methods of propagation may alter the epigenomic landscape over time, in the absence of sexual reproduction. These results show that DNA methylation targeting pathways have diverged functionally and that extant DNA methylation patterns are likely a reflection of the evolutionary and life histories of plant species.

Introduction

Biological diversity is established at multiple levels. Historically this has focused on studying the contribution of genetic variation. However, epigenetic variations manifested in the form of DNA methylation [1-3], histones and histone modifications [4], which together make up the epigenome, may also contribute to biological diversity. These components are integral to proper regulation of many aspects of the genome; including chromatin structure, transposon silencing, regulation of gene expression, and recombination [5-7]. Significant amounts of epigenomic diversity are explained by genetic variation [2, 3, 8-12], however, a large portion remains unexplained and in some cases these variants arise independently of genetic variation and are thus defined as "epigenetic" [2, 9-11, 13, 14]. Moreover, epigenetic variants are heritable and also lead to phenotypic variation [15-18]. To date, most studies of epigenomic variation in plants are based on a handful of model systems. Current knowledge is in particular based upon studies in *Arabidopsis thaliana*, which is tolerant to significant reductions in DNA methylation, a feature that enabled the discovery of many of the underlying mechanisms. However, *A. thaliana*, has a particularly compact genome that is not fully reflective of angiosperm diversity [19, 20]. The extent of natural variation of mechanisms that lead to epigenomic variation in plants, such as cytosine DNA methylation, is unknown and understanding this diversity is important to understanding the potential of epigenetic variation to contribute to phenotypic variation [21].

In plants, cytosine methylation occurs in three sequence contexts; CG, CHG, and CHH (H = A, T, or C), are under control by distinct mechanisms [22]. Methylation at CG (mCG) and CHG (mCHG) sites is typically symmetrical across the Watson and Crick strands [23]. mCG is maintained by the methyltransferase MET1, which is recruited to hemi-methylated CG sites and methylates the opposing strand [24, 25], whereas mCHG is maintained by the plant specific CHROMOMETHYLTRANSFERASE 3 (CMT3) [26], and is strongly associated with dimethylation of lysine 9 on histone 3 (H3K9me2) [27]. The BAH and CHROMO domains of CMT3 bind to H3K9me2, leading to methylation of CHG sites [27]. In turn, the histone methyltransferases KRYPTONITE (KYP), and Su(var)3-9 homologue 5 (SUVH5) and SUVH6 recognize methylated DNA and methylate H3K9 [28], leading to a self-reinforcing loop [29]. Asymmetrical methylation of CHH sites (mCHH) is established and maintained by another member of the CMT family, CMT2 [30, 31]. CMT2, like CMT3, also contains BAH and CHROMO domains and methylates CHH in H3K9me2 regions [30, 31]. Additionally, all three sequence contexts are methylated *de novo* via RNA-directed DNA methylation (RdDM) [32]. Short-interfering 24 nucleotide (nt) RNAs (siRNAs) guide the *de novo* methyltransferase DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) to target sites [33, 34]. The targets of CMT2 and RdDM are often complementary, as CMT2 in *A. thaliana* primarily methylate regions of deep heterochromatin, such as transposons bodies [30]. RdDM regions, on the other hand, often have the highest levels of mCHH methylation and primarily target the edges of transposons and the more recently identified mCHH islands [30, 31, 35]. The mCHH islands in *Zea mays* are associated with upstream and downstream of more highly expressed genes where they may function to prevent transcription of neighboring transposons [35, 36]. The establishment, maintenance, and

consequences of DNA methylation are therefore highly dependent upon the species and upon the particular context in which it is found.

Sequencing and array-based methods allow for studying methylation across entire genomes and within species [1, 3, 12, 14, 37]. Whole genome bisulfite sequencing (WGBS) is particularly powerful, as it reveals genome-wide single nucleotide resolution of DNA methylation [38-40]. WGBS has been used to sequence an increasing number of plant methylomes, ranging from model plants like *A. thaliana* [38, 39] to economically important crops like *Z. mays* [2, 10, 35, 41]. This has enabled a new field of comparative epigenomics, which places methylation within an evolutionary context [42-45]. The use of WGBS together with *de novo* transcript assemblies has provided an opportunity to monitor the changes in methylation of gene bodies among species [46] but does not provide a full view of changes in the patterns of context-specific methylation at different types of genomic regions [47].

Here, we report a comparative epigenomics study of 34 angiosperms (flowering plants). Differences in mCG and mCHG are in part driven by repetitive DNA and genome size, whereas in the Brassicaceae there are reduced mCHG levels and reduction/losses of CG gene body methylation (gbM). The Poaceae are distinct from other lineages, having low mCHH levels and a lineage-specific distribution of mCHH in the genome. Additionally, species that have been clonally propagated often have low levels of mCHH. Although some features, such as mCHH islands, are found in all species, their association with effects on gene expression is not universal. The extensive variation found reveals distinct mechanisms between species for how DNA methylation is established and maintained in flowering plant genomes.

Results

Genome-wide DNA methylation variation across angiosperms

We compared single-base resolution methylomes of 34 angiosperm species that have genome assemblies (**Table S1**). MethylC-seq [39, 48] was used to sequence 26 species and an additional eight species with previously published methylomes were downloaded and reanalyzed [11, 14, 35, 47, 49-51]. Different metrics were used to make comparisons at a whole-genome level. The genome-wide weighted methylation level [52] gives a single value for each context and each species (**Figure 1A-C**). The proportion that each methylation context makes up of all methylation indicates the predominance of specific methylation pathways **(Figure 1D**). The per-site methylation level is a distribution of methylation levels at each individual methylated site (**Figure 1E-G**), whereas symmetry is a comparison of per-site methylation levels at cytosines on the Watson versus the Crick strand for the symmetrical CG and CHG contexts (**Figure S1, and S2**). These two measures often provide insight into how well methylation is maintained [53] and how ubiquitously the sites are methylated across cell types (**Figure S3**).

These data revealed that DNA methylation is a major base modification in flowering plant genomes. However, there are numerous distinctions in how cytosine methylation is used by each species, which reflects the activities of distinct methylation

targeting pathways. Within each species, mCG had the highest levels of methylation genome-wide (**Figure 1A, Table S2**). Between species, levels ranged from a low of ~30.5% in *A. thaliana* to a high of ~92.5% in *Beta vulgaris.* NonCG methylation (mCHG and mCHH) was far more variable. Levels of mCHG varied as much as ~8 fold between species, from only ~9.3% in *Eutrema salsugineum* to ~81.2% in *B. vulgaris* (**Figure 1B, Table S2**). mCHH levels were universally the lowest, but also the most variable with as much as an ~16 fold difference. The highest being ~18.8% is in *B. vulgaris.* This was unusually high, as 85% of species had less than 10% mCHH and half had less than 5% mCHH (**Figure 1C, Table S2**). The lowest mCHH level was found in *Vitis vinifera* with only ~1.1% mCHH. mCG is the most predominant type of methylation making up the largest proportion of the total methylation in all examined species (**Figure 1D**). *B. vulgaris* was a notable outlier, having the highest levels of methylation in all contexts, and having particularly high mCHH levels. Multiple factors may be contributing to the differences between species observed, ranging from genome size and architecture, to differences in the activity of DNA methylation targeting pathways.

We examined these methylomes in a phylogenetic framework, which led to several novel findings regarding the evolution of DNA methylation pathways across flowering plants. The Brassicaceae, which includes *A. thaliana*, have lower levels of mCHG by all three measures: total methylation (~9.3-22% mCHG), proportion of methylation, and per-site methylation (**Figure 1B, 1D, 1F**). Symmetrical mCHG sites have a wider range of methylation levels and increased asymmetry, whereas, non-Brassicaceae species have very highly methylated symmetrical sites (**Figure S2, S3**), indicating that the CMT3 pathway is more effective in these genomes. *E. salsugineum*, with the lowest mCHG levels, is a natural *cmt3* mutant and *CMT3* is under relaxed selection in other Brassicaceae [54](Bewick, et al *in preparation*). Methylation of CG sites is less well maintained in the Brassicaceae, with *Capsella rubella* showing the lowest levels of per-site mCG methylation.

Within the Fabaceae, *Glycine max* and *Phaseolus vulgaris*, which are in the same lineage, show considerably lower per-site mCHH levels as compared to *Medicago truncatula* and *Lotus japonicus*, even though they have equivalent levels of genome-wide mCHH (**Figure 1C, 1G**). The Poaceae, in general, have much lower levels of mCHH (~1.4-5.8%), both in terms of total methylation level and as a proportion of total methylated sites across the genome. Per-site mCHH level distributions varied, with species like *Brachypodium distachyon* having amongst the lowest of all species, whereas others like *O. sativa* and *Z. mays* have levels comparable to *A. thaliana*. In *Z. mays*, CMT2 has been lost [30], and it may be that in other Poaceae, mCHH pathways are less efficient even though CMT2 is present. Collectively, these results indicate that different DNA methylation pathways may predominate in different lineages, with ensuing genome-wide consequences.

Several dicot species showed very low levels of mCHH (< 2%): *V. vinifera*, *Theobroma cacao*, *Manihot esculenta* (cassava), *Eucalyptus grandis.* No causal factor based on examined genomic features or examined methylation pathways was identified, however, these plants are commonly propagated via clonal methods [55]. Amongst non-Poaceae species, the six lowest mCHH levels were found in species with histories of

clonal propagation (**Figure S4**). Effects of micropropagation on methylation in *M. esculenta* using methylation-sensitive amplified polymorphisms have been observed before [56], so has altered expression of methyltransferases due to micropropagation in *Fragaria* x *ananassa* (strawberry) [57]. To test if clonal propagation was responsible for low mCHH, we examined a DNA methylome of a parental *M. esculenta* plant that previously undergone clonal propagation and a DNA methylome of its offspring that was germinated from seed. Additionally, the original *F. vesca* plant used for this study had been micropropagated for four generations. We germinated seeds from these plants, as they would have undergone sexual reproduction and examined these as well. Differences were slight, showing little substantial evidence of genome-wide changes in a single generation (**Figure S5**). Both of these results are based on one generation of sexual reproduction. This may be insufficient to fully observe any changes and will require further studies of samples collected over multiple generations.

Genome architecture of DNA methylation

DNA methylation is often associated with heterochromatin. Two factors can drive increases in genome size, whole genome duplication (WGD) events and increases in the copy number for repetitive elements. The majority of changes in genome size among the species we examined is due to changes in repeat content as the total gene number in these species only varies two-fold, whereas the genome size exhibits ~8.5 fold change. As genomes increase in size due to increased repeat content it is expected that DNA methylation levels will increase as well. This was tested using phylogenetic generalized least squares [58] (**Table S3**). Phylogenetic relationships were inferred from a species tree constructed using 50 single copy loci (**Figure S6**) [59]. Indeed, positive correlations were found between mCG (p-value $< 1.3 \times 10^{-4}$) and the strongest correlation for mCHG and genome size (p-value $< 1 \times 10^{-7}$) (**Figure 2A**). No correlation was observed between mCHH and genome size after multiple testing correction (p-value > 0.04) (**Figure 2A**). A relationship between genic methylation level and genome size has been previously reported [46]. We found that within coding sequences (CDS) methylation levels were correlated with genome size for both mCHG (p-value $< 2.3 \times 10^{-7}$) and mCHH (p-value $< 6.6 \times 10^{-7}$), but not for mCG (p-value > 0.18) (**Figure 2B**).

The highest levels of DNA methylation are typically found in centromeres and pericentromeric regions [38, 39, 47]. The distributions of methylation at chromosomal levels were examined in 100kb sliding windows (**Figure 2C, Figure S7**). The number of genes per window was used as a proxy to differentiate euchromatin and heterochromatin. Both mCG and mCHG have negative correlations between methylation level and gene number, indicating that these two methylation types are mostly found in gene-poor heterochromatic regions (**Figure 2D**). Most species also show a negative correlation between mCHH and gene number, even in species with very low mCHH levels like *V. vinifera.* However, several Poaceae species show no correlation or even positive correlations between gene number and mCHH levels. Only two grass species showed negative correlations, *Setaria viridis* and *Panicum hallii,*

which fall in the same lineage (**Figure 2D**). This suggests that heterochromatic mCHH is significantly reduced in many lineages of the Poaceae.

The methylome will be a composite of methylated and unmethylated regions. We implemented an approach (see Supplementary Methods) to identify methylated regions within a single sample to discern the average size of methylated regions and their level of DNA methylation for each species in each sequence context (**Figure S8**). For most species, regions of higher methylation are often smaller in size, with regions of low or intermediate methylation being larger (**Figure S9**). More small RNAs, in particular 24 nucleotide (nt) siRNAs map to regions of higher mCHH methylation (**Figure S10**). This may be because RdDM is primarily found on the edges of transposons whereas other mechanisms predominate in regions of deep heterochromatin [30]. Using these results, we can make inferences into the architecture of the methylome.

mCHG and mCHH regions are more variable in both size and methylation levels than mCG regions, as little variability in mCG regions was found between species (**Figure S8**). For mCHG regions, the Brassicaceae differed the most having lower methylation levels and *E. salsugineum* the lowest. This fits with *E. salsugineum* being a *cmt3* mutant and RdDM being responsible for residual mCHG [54]. However, the sizes of these regions are similar to other species, indicating that this has not resulted in fragmentation of these regions (**Figure S8**). The most variability was found in mCHH regions. Within the Fabaceae, the bulk of mCHH regions in *G. max* and *P. vulgaris* are of lower methylation in contrast to *M. truncatula* and *L. japonicus* (**Figure S8**). As these lower methylated mCHH regions are larger in size (**Figure S9**) and less targeted by 24 nt siRNAs (**Figure S10**), it would appear that deep heterochromatin mechanisms, like those mediated by CMT2, are more predominant than RdDM in these species as compared to *M. truncatula* and *L. japonicus.* Indeed the genomes of *G. max* and *P. vulgaris* are also larger than *M. truncatula* and *L. japonicus* (**Table S2**). In the Poaceae, we also find that mCHH regions are more highly methylated, even though genome-wide, mCHH levels are lower (**Figure S8**). This indicates that much of the mCHH in these genomes comes from smaller regions targeted by RdDM (**Figure S9, S10**), which is supported by RdDM mutants in *Z. mays* [41]. In contrast, previously discussed species like *M. esculenta*, *T. cacao*, and *V. vinifera* had mCHH regions of both low methylation and small size which could indicate that effect of all mCHH pathways have been limited in these species (**Figure S9, S10**).

Methylation of repeats

Genome-wide mCG and mCHG levels are related to the proliferation of repetitive elements. Although the quality of repeat annotations does vary between the species studied, correlations were found between repeat number and mCG (p-value = 5×10^{-5}) and mCHG levels (p-value = 7.5×10^{-3}) (**Figure 3A, Table S3**). This likely also explains the correlation of methylation with genome size, as large genomes often have more repetitive elements [60, 61]. No such correlation between mCHH levels and repeat numbers was found after multiple testing correction (p-value > 0.05) (**Figure 3A**). This was unexpected given that mCHH is generally associated with repetitive sequences. Although coding sequence (CDS) mCHG and mCHH correlates with genome size, only

CDS mCHG (p-value = 3.8 x 10^{-2}) correlated with the total number of repeats (**Figure S11A**). All methylation contexts, however, were found to be correlated to the percentage of genes containing repeats within the gene (exons, introns, and untranslated regions - mCG p-value = 2.2 x 10^{-3}, mCHG p-value = 3.6 x 10^{-4}, mCHH p-value = 2 x 10^{-4}) (**Figure S11B, Table S3**), including introns or untranslated regions. In fact, plotting the percentage of genes containing repeats against the total number of repeats showed a cluster of species possessing fewer genic repeats given the total number of repeats in their genomes (**Figure S11C**). This implies that the transposon load of a genome alone does not affect methylation levels in genes, rather, it is more likely a result of the distribution of transposons within a genome.

Considerable variation exists in methylation patterns within repeats. Across all species, repeats were heavily methylated at CG sequences, but were more variable in CHG and CHH methylation (**Figure 3B**). mCHG was typically high at repeats in most species, with the exception of the Brassicaceae, in particular *E. salsugineum.* Similarly, low levels of mCHH were found in most Poaceae. Across the body of the repeat, most species show elevated levels in all three methylation contexts as compared to outside the repeat (**Figure 3C, S12**). Again, several Poaceae species stood out, as *B. distachyon* and *Z. mays* showed little change in mCHH within repeats, fitting with the observation that mCHH is depleted in deep heterochromatic regions of the Poaceae.

CG gene body methylation

Methylation within genes in all three contexts is associated with suppressed gene expression [32], whereas genes that are only mCG methylated within the gene body are often constitutively expressed genes [62-64]. We classified genes using a modified version of the binomial test described by Takuno and Gaut [44] into one of four categories: CG gene body methylated (hereafter gbM), mCHG, mCHH, and unmethylated (UM) (**Table S4**). gbM genes are methylated at CG sites, but not at CHG or CHH. NonCG contexts are often coincident with mCG, for example RdDM regions are methylated in all three contexts. We further classified nonCG methylated genes as mCHG genes (mCHG and mCG, no mCHH) or mCHH genes (mCHH, mCHG, and mCG). Genes with insignificant amounts of methylation were classified as unmethylated.

Between species, the methylation status of gbM can be conserved across orthologs [45]. The methylation state of orthologous genes across all species was compared using *A. thaliana* as an anchor (**Figure 4A)**. *A. lyrata* and *C. rubella* are the most closely related to *A. thaliana* and also have the greatest conservation of methylation status, with many *A. thaliana* gbM gene orthologs also being gbM genes in these species (~86.3% and ~79.8% of *A. thaliana* gbM genes, respectively). However, they also had many gbM genes that had unmethylated *A. thaliana* orthologs (~ 18.6% and ~13.9% of *A. thaliana* genes, respectively). Although gbM is generally "conserved" between species, this conservation breaks down over evolutionary distance with gains and losses of gbM in different lineages. In terms of total number of gbM genes, *M. truncatula* and *Mimulus guttatus* had the greatest number (**Table S2**). However, when the percentage of gbM genes in the genome is taken into account (**Figure 4A**), *M.*

truncatula appeared similar to other species, whereas *M. guttatus* remained an outlier with ~60.7% of all genes classified as gbM genes. The reason why *M. guttatus* has unusually large numbers of gbM loci is unknown and will require further investigation. In contrast, there has been considerable loss of gbM genes in *Brassica rapa*, and *Brassica oleracea* and a complete loss in *E. salsugineum.* This suggests that over longer evolutionary distance, the methylation status of gbM varies considerably and is dispensable as it is lost entirely in *E. salsugineum*.

gbM is characterized by a sharp decrease of methylation around the transcriptional start site (TSS), increasing mCG throughout the gene body, and a sharp decrease at the transcriptional termination site (TTS) [63, 64]. gbM genes identified in most species show this same trend and even have comparable levels of methylation (**Figure 4B, S13**). Here, too, the decay and loss of gbM in the Brassicaceae is observed as *B. rapa* and *B. oleracea* have the second and third lowest methylation levels, respectfully in gbM genes and *E. salsugineum* shows no canonical gbM having only a few genes that passed statistical tests for having mCG in gene bodies. As has previously been found [63, 64], gbM genes are more highly expressed as compared to UM and nonCG (mCHG and mCHH) genes (**Figure 4C, S14**). The exception to this is *E. salsugineum* where gbM genes have almost no expression. A subset of unexpressed genes with mCG methylation was found, and in some cases, had higher mCG methylation around the TSS (mCG-TSS). Using previously identified mCG regions we identified genes with mCG overlapping the TSS, but lacking either mCHG or mCHH regions within or near genes. These genes had suppressed expression (**Figure 4C, S14**) showing that although mCG is not repressive in gene-bodies, it can be when found around the TSS.

gbM genes are known to have many distinct features in comparison to UM genes. They are typically longer, have more exons, the observed number of CG dinucleotides in a gene are lower than expected given the GC content of the gene ([O/E]), and they evolve more slowly [44, 45]. We compared gbM genes to UM genes for each of these characteristics, using *A. thaliana* as the base for pairwise comparison for all species except the Poaceae where *O. sativa* was used (**Table S5**). With the exception of *E. salsugineum*, which lacks canonical gbM, these genes were longer and had more exons than UM genes (**Table S5**). Most gbM genes also had a lower CG [O/E] than UM genes, except for six species, four of which had a greater CG [O/E]. These included both *M. guttatus* and *M. truncatula*, which had the greatest number of gbM genes of any species. Recent conversion of previously UM genes to a gbM status could in part explain this effect. Previous studies have shown that gbM orthologs between *A. thaliana* and *A. lyrata* [44] and between *B. distachyon* and *O. sativa* [45] are more slowly evolving than UM orthologs. Although this result holds up over short evolutionary distances, it breaks down over greater distances with gbM genes typically evolving at equivalent rates as UM and, in some cases, faster rates (**Table S5**).

NonCG methylated genes

NonCG methylation exists within genes and is known to suppress gene expression [15, 17, 65-67]. Although differences in annotation quality could lead to

some transposons being misannotated as genes and thus as targets of nonCG methylation, within-species epialleles demonstrates that significant numbers of genes are indeed targets [3, 11]. In many species there were genes with significant amounts of mCHG and little to no mCHH. High levels of mCHG within *Z. mays* genes is known to occur, especially in intronic sequences due in part to the presence of transposons [68]. Based on this difference in methylation, mCHG and mCHH genes were maintained as separate categories (**Table S4**). The methylation profiles of mCHG and mCHH genes often resembled that of repeats (**Figure 5A, S13**). Both mCHG and mCHH genes are associated with reduced expression levels (**Figure 5B, S13**). As mCHG methylation is present in mCHH genes, this may indicate that mCHG alone is sufficient for reduced gene expression. It was also observed that *Cucumis sativus* has an unusual pattern of mCHH in many highly expressed genes, although the result did not replicate after sequencing a second independent *C. sativus* sample (**Figure S15**). The number of genes possessing nonCG types of methylation ranged from as low as ~3% of genes (*M. esculenta*) to as high as ~32% of genes (*F. vesca*) (**Figure 5C**). In all the Poaceae, mCHG genes made up at least ~5% of genes and typically more. In contrast, mCHG genes were relatively rare in the Brassicaceae where mCHH genes were the predominant type of nonCG genes. In most of the clonally propagated species with low mCHH, there were typically few mCHH genes and more mCHG genes with the exception of *F. vesca.*

Unlike gbM genes, there was no conservation of methylation status across orthologs of mCHG and mCHH genes (**Figure S16**). For many nonCG methylated genes, orthologs were not identified based on our approach of reciprocal best BLAST hit. For example, orthologs were found for only 488 of 999 of *A. thaliana* mCHH genes across all species. Previous comparisons of *A. thaliana*, *A. lyrata*, and *C. rubella* have shown no conservation of nonCG methylation between orthologs within the Brassicaceae [47]. However, we did observe some conservation based on gene ontology (GO). The same GO terms were often enriched in multiple species (**Figure S17, Table S6**). The most commonly enriched terms were involved in processes such as proteolysis, cell death, and defense responses; processes that could have profound effects on normal growth and development and may be developmentally or environmentally regulated. There was also enrichment in many species for genes related to electron-transport chain processes, photosynthetic activity, and other metabolic processes. Further investigation of these genes revealed that many are orthologs to chloroplast or mitochondrial genes, suggesting that they may be recent transfers from the organellar genome. The transfer of organellar genes to the nucleus is a frequent and ongoing process [69, 70]. Although DNA methylation is not found in chloroplast genomes, transfer to the nucleus places them in a context where they can be methylated, contributing to the mutational decay of these genes via deamination of methylated cytosines [71].

mCHH islands

In *Z. mays,* high mCHH is enriched in the upstream and downstream regions of highly expressed genes and are termed mCHH islands [35, 36]. We identified mCHH islands 2kb upstream and downstream of annotated genes for each species, finding that

the percentage of genes with such regions varied considerably across species (**Figure 6A**) the fewest being in *V. vinifera, B. oleracea*, and *T. cacao*, each having mCHH islands associated with less than 2% of genes. As both *V. vinifera* and *T.* cacao have low genome-wide mCHH levels, this may explain the difference, however, this is not the case for *B. oleracea*. mCHH islands are thought to mark euchromatin-heterochromatin boundaries and are often associated with transposons, however, we found no correlation between the total number of repeats in the genome and the number of genes with mCHH islands (**Figure S18A, Table S3**). As in the case of CDS methylation levels, this lack of correlation was largely due to differences in the distribution of repeats (**Figure S18B, Table S3**). When correlated to the percentage of genes with repeats 2kb upstream or downstream, both upstream and downstream mCHH islands are correlated (upstream p-value = 7.9×10^{-6}, downstream p-value < 8.1×10^{-6}) (**Figure 6B**). *B. oleracea* in particular stood out as having few repeats in 2kb upstream or downstream of genes, explaining in part why it possess so few mCHH islands, despite its large genome and overall number of repeats.

In *Z. mays*, mCHH islands are more common in genes in the most highly expressed quartiles [35, 36]. This is true of several species, such as *P. persica* and all the Poaceae (**Figure 6C and S19**). However, many species showed no significant association between mCHH islands and gene expression (**Figure 6C and S19**). This was true for all the Brassicaceae. Other species such as *M. guttatus*, despite having high levels of mCHH islands associated with genes like the Poaceae (~46% upstream, ~40% downstream), also showed no association with gene expression (**Figure 6C**). As has been observed previously in *Z. mays*, mCG and mCHG levels are generally higher on the distal side of the mCHH island to the gene (**Figure 6D and S20**), marking a boundary of euchromatin and heterochromatin [36]. However, this difference in methylation level is much less pronounced in most other species as compared to *Z. mays* and much less evident for downstream mCHH islands than for upstream ones (**Figure S20**). These may indicate different preferences in transposon insertion sites or a need to maintain a boundary of heterochromatin near the start of transcription.

Discussion

We present the methylomes of 34 different angiosperm species in a phylogenetic framework using comparative epigenomics, which enables the study of DNA methylation in an evolutionary context. Extensive variation was found between species, both in levels of methylation and distribution of methylation, with the greatest variation being observed in nonCG contexts. The Brassicaceae show overall reduced mCHG levels and reduced numbers of gbM genes, leading to a complete loss in *E. salsugineum*, that is associated with loss of CMT3 [54]. Whereas in the Poaceae, mCHH levels are typically lower than that in other species. The Poaceae have a distinct epigenomic architecture compared to eudicots, with mCHH often depleted in deep heterochromatin and enriched in genic regions. We also observed that many species with a history of clonal propagation have lower mCHH levels. Epigenetic variation induced by propagation techniques can be of agricultural and economic importance [72], and understanding the effects of clonal propagation will likely future studies over

multiple generations. Evaluation of per-site methylation levels, methylated regions, their structure, and association with small RNAs indicates that there are differences in the predominance of various molecular pathways.

Variation exists within features of the genome. Repeats and transposons show variation in their methylation level and distribution with impacts on methylation within genes and regulatory regions. Although gbM genes do show many conserved features, this breaks down with increasing evolutionary distance and as gbM is gained or lost in some species. gbM is known to be absent in the basal plant species *Marchantia polymorpha* [46] and *Selaginella moellendorffii* [43]. That it has also been lost in the angiosperm *E. salsugineum*, indicates that it is dispensable over evolutionary time. That nonCG methylation shows no conservation at the level of individual genes, indicates that it is gained and lost in a lineage specific manner. It is an open question as to the evolutionary origins of nonCG methylation within genes. That these are correlated to repetitive elements in genes suggests transposons as one possible factor. That many nonCG genes lack orthologous genes could indicate a preferential targeting of *de novo* genes, as in the case of the QQS gene in *A. thaliana* [18]. At a higher order level, there appears to be a commonality in what categories of genes are targeted, as many of the similar functions are enriched across species. Other features, such as mCHH islands, also are not conserved and show extensive variation that is associated with the distribution of repeats upstream and downstream of genes.

This study demonstrates that widespread variation in methylation exists between flowering plant species. For many species, this is the first reported methylome and methylome browsers for each species have been made available to serve as a resource (http://schmitzlab.genetics.uga.edu/plantmethylomes). Historically, our understanding has come primarily from *A. thaliana*, which has served as a great model for studying the mechanistic nature of DNA methylation. However, the extent of variation observed previously [46, 47] and now shows that there is still much to be learned about underlying causes of variation in this molecular trait. Due to its role in gene expression and its potential to vary independently of genetic variation, understanding these causes will be necessary to a more complete understanding of the role of DNA methylation underlying biological diversity.

Methods

MethylC-seq and analysis

DNA was isolated from leaf tissue and MethylC-seq libraries for each species were prepared as previously described [48]. Previously published datasets were obtained from public databases and reanalyzed [11, 14, 35, 47, 49-51, 54]. Genome sequences and annotations for most species were downloaded from Phytozome 10.1 (http://phytozome.jgi.doe.gov/pz/portal.html) [73]. The *L. japonicus* genome was downloaded from the *Lotus japonicus* Sequencing Project (http://www.kazusa.or.jp/lotus/)[74], the *B. vulgaris* genome was downloaded from *Beta vulgaris* Resource (bvseq.molgen.mpg.de/)[75], and the *C. sativa* genome from *C. sativa* (*Cannabis*) Genome Browser Gateway (http://genome.ccbr.utoronto.ca/cgi-bin/hgGateway) [76]. As annotations for *S. viridis* were not available, gene models from

the closely related *S. italica* were mapped onto the *S. viridis* genome using Exonerate [77].

Sequencing data for each species was aligned to their respective genome (Table S1) and methylated sites called using previously described methods [78]. In brief, reads were trimmed for adapters and quality using Cutadapt [79] and then mapped to both a converted forward strand (all cytosines to thymines) and converted reverse strand (all guanines to adenines) using bowtie [80]. Reads that mapped to multiple locations and clonal reads were removed. The non-conversion rate (rate at which unmethylated cytosines failed to be converted to uracil) was calculated by using reads mapping to the lambda genome or the chloroplast genome if available (Table S1). Cytosines were called as methylated using a binomial test using the non-conversion rate as the expected probability followed by multiple testing correction using Benjamini-Hochberg False Discovery Rate (FDR). A minimum of three reads mapping to a site was required to call a site as methylated. Data are available at the Plant Methylome DB http://schmitzlab.genetics.uga.edu/plantmethylomes.

Phylogenetic Tree

A species tree was constructed using BEAST2 [81] on a set of 50 previously identified single copy loci [59]. Protein sequences were aligned using PASTA [82] and converted into codon alignments using custom Perl scripts. Gblocks [83] was used to identify conserved stretches of amino acids and then passed to JModelTest2 [84, 85] to assign the most likely nucleotide substitution model.

Genome-wide analyses

Genome-wide weighted methylation was calculated from all aligned data by dividing the total number of aligned methylated reads to the genome by the total number of methylated plus unmethylated reads [52]. Correlations between methylation levels, genome sizes, and gene numbers were done in R and corrected for phylogenetic signal using the APE [86], phytools [87], and NLME packages. In total, 22 comparisons were conducted (Table S3) and a p-value < 0.05 after Bonferroni Correction. Distribution of methylation levels and genes across chromosomes was conducted by dividing the genome into 100 kb windows, sliding every 50 kb using BedTools and custom scripts. Pearson's correlation between gene number and methylation level in each window was conducted in R. Weighted methylation levels for each repeat were calculated using custom python and R scripts.

Methylated-regions

Methylated regions were defined independent of genomic feature by methylation context (CG, CHG, or CHH) using BEDTools [88] and custom scripts. For each context, only methylated sites in that respective context were considered used to define the region. The genome was divided into 25bp windows and all windows that contained at least one methylated cytosine in the context of interest were retained. 25bp windows were then merged if they were within 100 bp of each other, otherwise they were kept separate. The merged windows were then refined so that the first methylated cytosine became the new start position and the last methylated cytosine new end position. Number of methylated sites and methylation levels for that region was then recalculated

for the refined regions. A region was retained if it contained at least five methylated cytosines and then split into one of four groups based on the methylation levels of that region: group 1, < 0.05%, group 2, 5-15%, group 3, 15-25%, group 4, > 25%. Size of methylated regions were determined using BedTools.

Small RNA (sRNA) cleaning and filtering

Libraries for *B. distachyon*, *C. sativus*, *E. grandis*, *E. salsugineum*, *M. truncatula*, *P. hallii,* and *R. communis* were constructed using the TruSeq Small RNA Library Preparation Kit (Illumina Inc). Small RNA-seq datasets for additional species were downloaded from GEO and the SRA and reanalyzed [14, 49, 50, 89-91]. The small RNA toolkit from the UEA computational Biology lab was used to trim and clean the reads [92]. For trimming, 8 bp of the 3' adapter was trimmed. Trimmed and cleaned reads were aligned using PatMan allowing for zero mismatches [93]. BedTools [88] and custom scripts were used to calculate overlap with mCHH regions.

Gene-level analyses

Genes were classified as mCG, mCHG, or mCHH by applying a binomial test to the number of methylated sites in a gene [44]. The total number of cytosines and the methylated cytosines were counted for each context for the coding sequences (CDS) of the primary transcript for each gene. A single expected methylation rate was estimated for all species by calculating the percentage of methylated sites for each context from all sites in all coding regions from all species. We restricted the expected methylation rate to only coding sequences as the species study differ greatly in genome size, repeat content, and other factors that impact genome-wide methylation. Furthermore, it is known that some species have an abundance of transposons in UTRs and intronic sequences, which could lead to misclassification of a gene. A single value was calculated for all species to facilitate comparisons between species and to prevent setting the expected methylation level to low, as in the case of *E. salsugineum* or to high, as in the case of *B. vulgaris*, which would further lead to misclassifications.

A binomial test was then applied to each gene for each sequence context and q-values calculated by adjusting p-values by Benjamini-Hochberg FDR. Genes were classified as mCG if they had reads mapping to at least 20 CG sites and has 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 was allowed to 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 were allowed to be anything as both types of methylation are associated with mCHH. mCG-TSS genes were identified by overlap of mCG regions with the TSS of each gene and the absence of any mCHG or mCHH regions within the gene or 1000 bp upstream or downstream.

GO terms for each gene were downloaded from phytozome 10. 1 (http://phytozome.jgi.doe.gov/pz/portal.html) [73]. GO term enrichment was performed using the parentCHILD algorithm [94] with the F-statistic as implemented in the topGO module in R. GO terms were considered significant with a q-value < 0.05.

Exon number, gene length and [O/E].

For each species the general feature format 3 (gff3) file from phytozome 10.1 (http://phytozome.jgi.doe.gov/pz/portal.html) [73] was used to determine exon number and coding sequence length (base pairs, bp) for each annotated gene (hereafter referred to as CDS). Additionally, for each full length CDS (starting with the start codon ATG, and ending with one of the three stop codons TAA/TGA/TAG), from the phytozome 10.1 primary CDS fasta file, the CG [O/E] ratio was calculated, which is the observed number of CG dinucleotides relative to that expected given the overall G+C content of a gene. Differences for these genic features between CG gbM and UM genes were assessed using permutation tests (100,000 replicates) in R, with the null hypothesis being no difference between the gbM and UM methylated genes.

Identifying orthologs and estimating evolutionary rates.

Substitution rates were calculated between CDS pairs of monocots to *Oryza sativa*, and dicots to *A. thaliana*. Reciprocal best BLAST with an e-value cutoff of ≤1E-08 was used to identify orthologs between dicot-*A. thaliana*, and monocot-*O. sativa* pairs. Individual CDS pairs were aligned using MUSCLE, insertion-deletion (indel) sites were removed from both sequences, and the remaining sequence fragments were shifted into frame and concatenated into a contiguous sequence. A ≥30 bp and ≥300 bp cutoff for retained fragment length after indel removal, and concatenated sequence length was implemented, respectively. Coding sequence pairs were separated into each combination of methylation (i.e., CG gbM-CG gbM, and UM-UM). The *yn00* (Yang-Neilson) [95] model in the program PAML (Phylogenetic Analysis by Maximum Likelihood) for pairwise sequence comparison was used to estimate synonymous and non-synonymous substitution rates, and adaptive evolution (*dS*, *dN*, and ω, respectively) [96]. Differences in rates of evolution between methylated and unmethylated pairs were assessed using permutation tests (100,000 replicates) in R, with the null hypothesis being no difference between the CG gbM and UM methylated genes.

RNA-seq mapping and analysis

RNA-seq datasets [11, 14, 47, 49, 54, 89, 91, 97-102] were downloaded from the Gene Expression Omnibus (GEO) and the NCBI Short Read Archive (SRA) for reanalysis. *B. distachyon* and *C. sativus* RNA-seq libraries were constructed using Illumina TruSeq Stranded mRNA Library Preparation Kit (Illumina Inc.) and sequenced on a NextSeq500 at the Georgia Genomics Facility. Reads were aligned using Tophat v2.0.13 [103] supplied with a reference genome feature file (GFF) with the following arguments -I 50000 --b2-very-sensitive --b2-D 50. Transcripts were then quantified using Cufflinks v2.2.1 [104] supplied with a reference GFF.

mCHH islands

mCHH islands were identified for both upstream and downstream regions as previously described [36]. Briefly, methylation levels were determined for 100bp windows across the genome. Windows of 25% or greater mCHH either 2 kb upstream or downstream of genes were identified and intersecting windows of that region were retained. Methylation levels were then plotted centered on the window of highest

mCHH. Genes associated with mCHH islands were categorized as non-expressed (NE) or divided into one of four quartiles based on their expression level.

Acknowledgements

We would like to thank Drs. J. Chris Pires, Scott T. Woody, Richard M. Amasino, Heinz Himmelbauer, Fred G. Gmitter, Timothy R. Hughes, Rebecca Grumet, CJ Tsai, Karen S. Schumaker, Kevin M. Folta, Marc Libault, Steve van Nocker, Steve D. Rounsely, Andrea L. Sweigart, Gerald A. Tuskan, Thomas E. Juenger, Douglas G. Bielenberg, Brian Dilkes, Thomas P. Brutnell, Todd C. Mockler, Mark J. Guiltinan, and Mallikarjuna K. Aradhya for providing tissue and DNA of various species used in this study. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 to J.S. We thank the Joint Genome Institute and collaborators for access to unpublished genomes of *B. rapa, S. viridis, P. virgatum*, and *P. hallii*. This work was supported by the National Science Foundation (NSF) (MCB – 1402183), by the Office of the Vice President of Research at UGA, and by The Pew Charitable Trusts to R.J.S. C.E.N was supported by a NSF postdoctoral fellowship (IOS – 1402183).

Author information: Genome browsers for all methylation data used in this paper is located at Plant Methylation DB (http://schmitzlab.genetics.uga.edu/plantmethylomes). Sequence data for MethylC-seq, RNA-seq, and small RNA-seq is located at the Gene Expression Omnibus, accession GSE79526.

Figure legends

Figure 1: Genome-wide methylation levels for **(A)** mCG, **(B)** mCHG, **(C)** and mCHH. **(D)** Using the genome-wide methylation levels, the proportion that each context contributes towards the total methylation (mC) was calculated. **(E)** The distribution of per-site methylation levels for mCG, **(F)** mCHG, **(G)** and mCHH. Species are organized according to their phylogenetic relationship.

Figure 2: (A) Genome-wide methylation levels are correlated to genome size for mCG (blue) and mCHG (green), but not for mCHH (maroon) Significant relationships indicated. **(B)** Coding region (CDS) methylation levels is not correlated to genome size for mCG (blue), but is for mCHG (green), and mCHH (maroon). Significant relationships indicated. **(C)** Chromosome plots show the distribution of mCG (blue), mCHG (green), and mCHH (maroon) across the chromosome (100kb windows) in relationship to genes. **(D)** For each species, the correlation (Pearson's correlation) in 100kb windows between gene number and mCG (blue), mCHG (green), and mCHH (maroon).

Figure 3: (A) Genome-wide methylation levels were correlated with repeat number for mCG (blue) and mCHG (green), but not for mCHH (maroon). Significant relationships indicated. **(B)** Distribution of methylation levels for repeats in each species. **(C)** Patterns of methylation upstream, across, and downstream of repeats for mCG (blue), mCHG (green), and mCHH (maroon).

Figure 4: (A) Heatmap showing methylation state of orthologous genes (horizontal axis) to *A. thaliana* for each species (vertical axis). Species are organized according to phylogenetic relationship. **(B)** Percentage of genes in each species that are gbM. The Brassicacea are highlighted in gold. **(C)** The levels of mCG in upstream, across, and downstream of gbM genes for all species. Species in gold belong to the Brassicaceae and illustrate the decreased levels and loss of mCG. **(D)** gbM genes are more highly expressed, while mCG over the TSS (mCG-TSS) has reduced gene expression.

Figure 5: (A) Methylation levels for mCG (blue), mCHG (green), and mCHH (maroon) were plotted upstream, across, and downstream of mCHG and mCHH genes. **(B)** Gene expression of mCHG and mCHH genes versus all genes. **(C)** The percentage of mCHG and mCHH genes per species. Species are arranged by phylogenetic relationship.

Figure 6: (A) Percentage of genes with mCHH islands 2kb upstream or downstream. **(B)** Upstream and downstream mCHH islands are correlated with upstream and downstream repeats (respectively). Significant relationships indicated. **(C)** Association of upstream mCHH islands with gene expression. Genes are divided into non-expressed (NE) and quartiles of increasing expression. **(D)** Patterns of upstream mCHH islands. Blue, green, and red lines represent mCG, mCHG and mCHH levels, respectively.

Bibliography

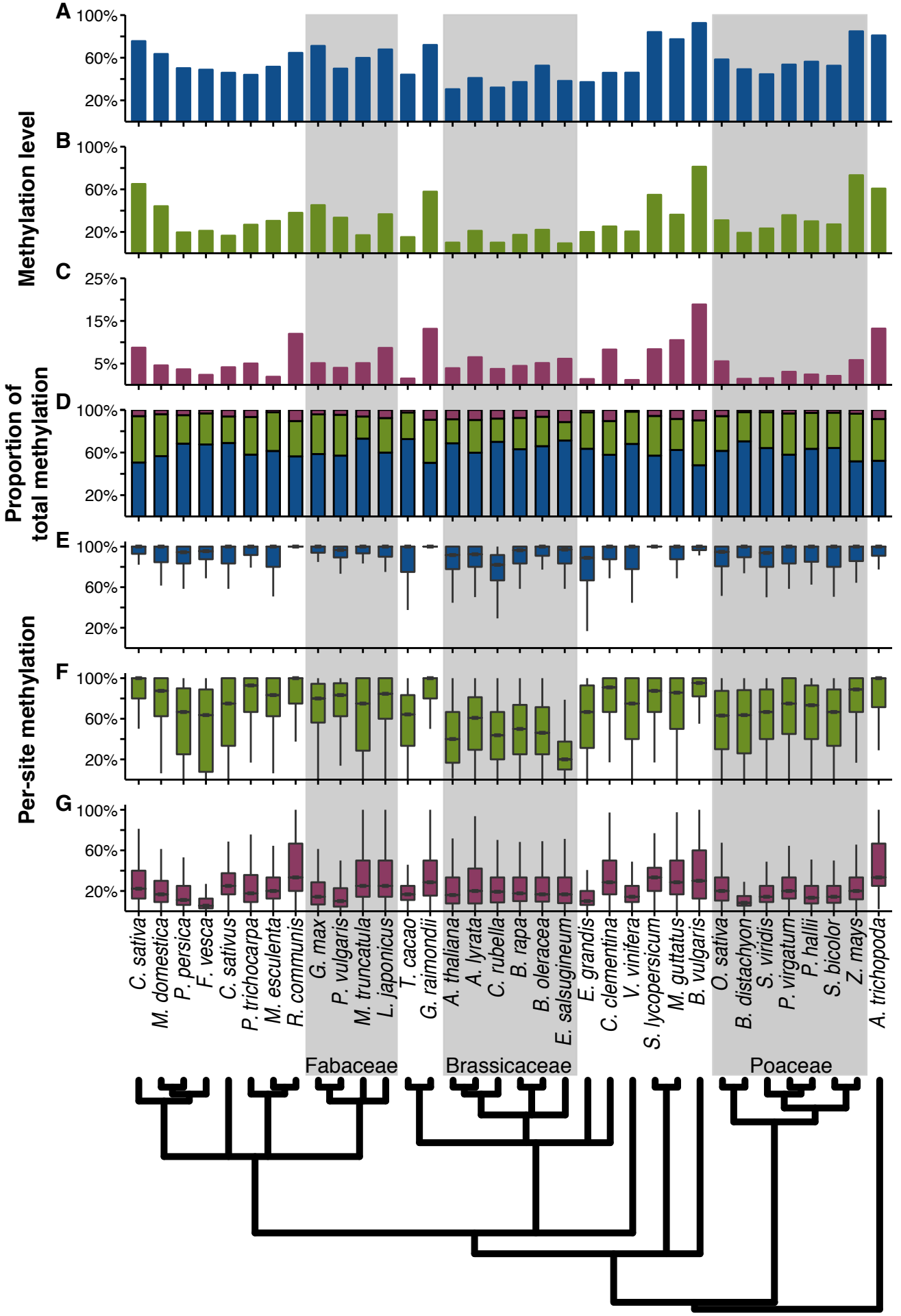
1. Vaughn, M.W., et al., Epigenetic natural variation in Arabidopsis thaliana. PLoS Biol, 2007. **5**(7): p. e174.
2. Eichten, S.R., et al., Heritable epigenetic variation among maize inbreds. PLoS Genet, 2011. **7**(11): p. e1002372.
3. Schmitz, R.J., et al., Patterns of population epigenomic diversity. Nature, 2013. **495**(7440): p. 193-8.
4. Filleton, F., et al., The complex pattern of epigenomic variation between natural yeast strains at single-nucleosome resolution. Epigenetics Chromatin, 2015. **8**: p. 26.
5. Colome-Tatche, M., et al., Features of the Arabidopsis recombination landscape resulting from the combined loss of sequence variation and DNA methylation. Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(40): p. 16240-16245.
6. Jones, P.A., Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet, 2012. **13**(7): p. 484-92.
7. Schmitz, R.J. and X. Zhang, Decoding the Epigenomes of Herbaceous Plants. 2014. **69**: p. 247-277.
8. Bell, J.T., et al., DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol, 2011. **12**(1): p. R10.
9. Eichten, S.R., et al., Epigenetic and genetic influences on DNA methylation variation in maize populations. Plant Cell, 2013. **25**(8): p. 2783-97.
10. 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.
11. Schmitz, R.J., et al., Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. Genome Res, 2013. **23**(10): p. 1663-74.
12. Dubin, M.J., et al., DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. Elife, 2015. **4**: p. e05255.
13. Becker, C., et al., Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature, 2011. **480**(7376): p. 245-9.
14. Schmitz, R.J., et al., Transgenerational epigenetic instability is a source of novel methylation variants. Science, 2011. **334**(6054): p. 369-73.
15. Cubas, P., C. Vincent, and E. Coen, An epigenetic mutation responsible for natural variation in floral symmetry. Nature, 1999. **401**(6749): p. 157-61.
16. Thompson, A.J., et al., Molecular and genetic characterization of a novel pleiotropic tomato-ripening mutant. Plant Physiol, 1999. **120**(2): p. 383-90.
17. 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.
18. Silveira, A.B., et al., Extensive natural epigenetic variation at a de novo originated gene. PLoS Genet, 2013. **9**(4): p. e1003437.
19. Arabidopsis Genome, I., Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature, 2000. **408**(6814): p. 796-815.
20. Flowers, J.M. and M.D. Purugganan, The evolution of plant genomes: scaling up from a population perspective. Curr Opin Genet Dev, 2008. **18**(6): p. 565-70.
21. Lane, A.K., et al., pENCODE: a plant encyclopedia of DNA elements. Annu Rev Genet, 2014. **48**: p. 49-70.

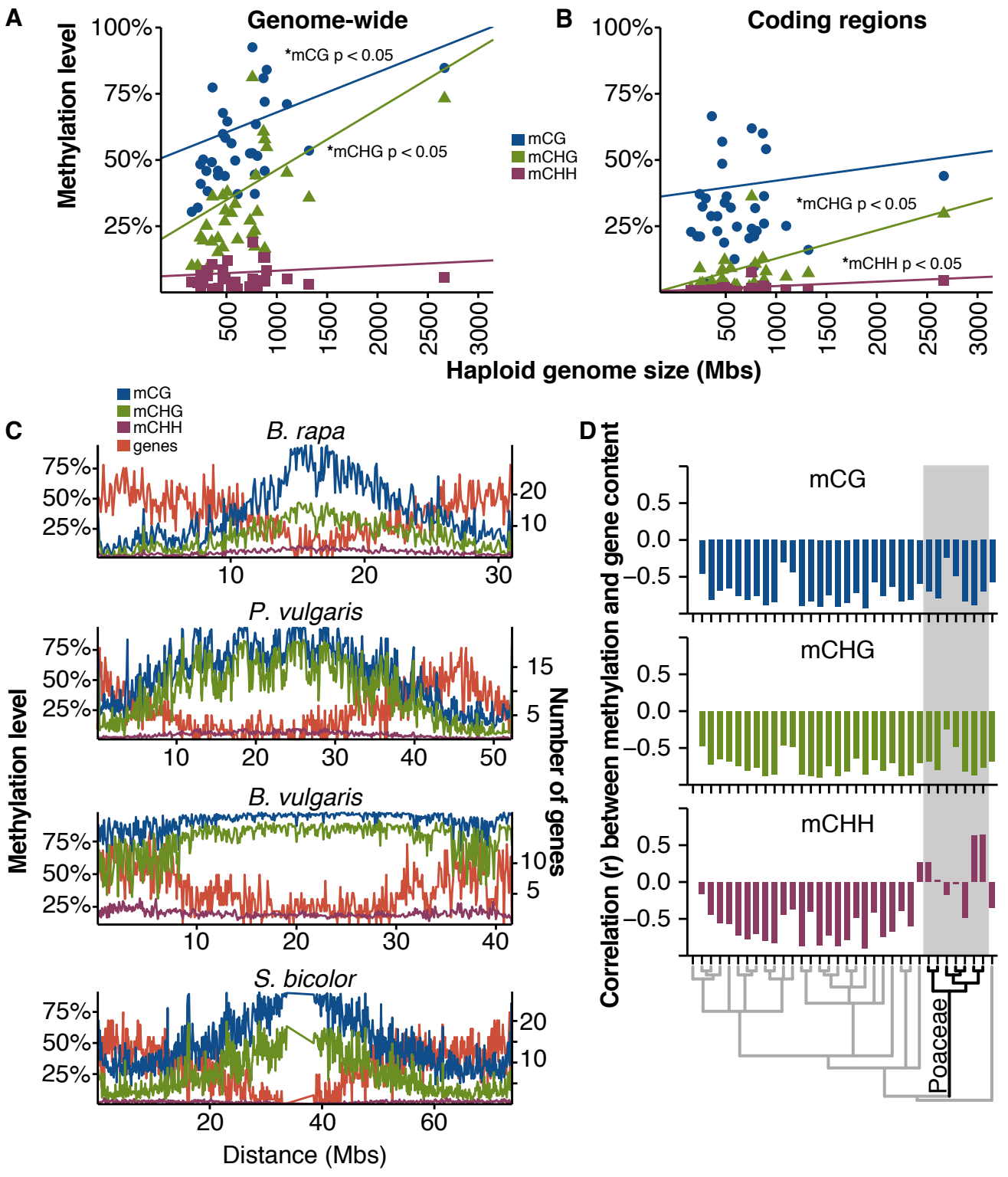
22. Niederhuth, C.E. and R.J. Schmitz, Covering your bases: inheritance of DNA methylation in plant genomes. Mol Plant, 2014. **7**(3): p. 472-80.
23. Finnegan, E.J., et al., *DNA Methylation in Plants.* Annu Rev Plant Physiol Plant Mol Biol, 1998. **49**: p. 223-247.
24. Finnegan, E.J., W.J. Peacock, 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.
25. Bostick, M., et al., UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science, 2007. **317**(5845): p. 1760-4.
26. Lindroth, A.M., et al., Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. Science, 2001. **292**(5524): p. 2077-80.
27. 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.
28. Du, J., et al., Mechanism of DNA methylation-directed histone methylation by KRYPTONITE. Mol Cell, 2014. **55**(3): p. 495-504.
29. Du, J., et al., DNA methylation pathways and their crosstalk with histone methylation. Nat Rev Mol Cell Biol, 2015. **16**(9): p. 519-32.
30. Zemach, A., et al., The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell, 2013. **153**(1): p. 193-205.
31. Stroud, H., et al., Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. Nat Struct Mol Biol, 2014. **21**(1): p. 64-72.
32. 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.
33. Cao, X. and S.E. Jacobsen, Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes. Proc Natl Acad Sci U S A, 2002. **99 Suppl 4**: p. 16491-8.
34. Cao, X. and S.E. Jacobsen, Role of the Arabidopsis DRM Methyltransferases in De Novo DNA Methylation and Gene Silencing. Current Biology, 2002. **12**(13): p. 1138-1144.
35. Gent, J.I., et al., CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. Genome Res, 2013. **23**(4): p. 628-37.
36. Li, Q., et al., RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. Proc Natl Acad Sci U S A, 2015.
37. Stroud, H., et al., Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell, 2013. **152**(1-2): p. 352-64.
38. Cokus, S.J., et al., Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature, 2008. **452**(7184): p. 215-9.
39. Lister, R., et al., Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell, 2008. **133**(3): p. 523-36.
40. Lister, R. and J.R. Ecker, Finding the fifth base: genome-wide sequencing of cytosine methylation. Genome Res, 2009. **19**(6): p. 959-66.
41. Li, Q., et al., Genetic perturbation of the maize methylome. Plant Cell, 2014. **26**(12): p. 4602-16.
42. 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.
43. Zemach, A., et al., Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science, 2010. **328**(5980): p. 916-9.
44. 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.

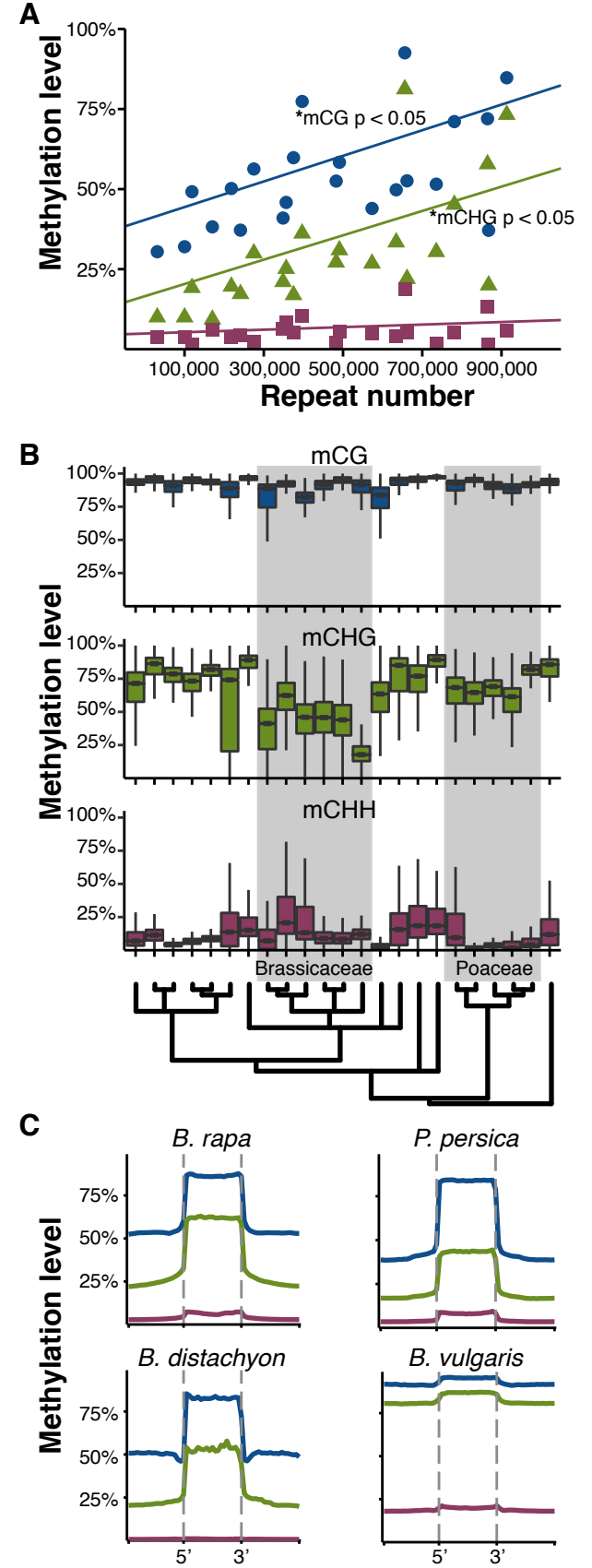
45. Takuno, S. and B.S. Gaut, Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. Proc Natl Acad Sci U S A, 2013. **110**(5): p. 1797-802.
46. Takuno, S., J.-H. Ran, and B.S. Gaut, Evolutionary patterns of genic DNA methylation vary across land plants. Nature Plants, 2016. **2**(2): p. 15222.
47. 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.
48. Urich, M.A., et al., MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nat Protoc, 2015. **10**(3): p. 475-83.
49. Amborella Genome, P., The Amborella genome and the evolution of flowering plants. Science, 2013. **342**(6165): p. 1241089.
50. Stroud, H., et al., Plants regenerated from tissue culture contain stable epigenome changes in rice. Elife, 2013. **2**: p. e00354.
51. Zhong, S., et al., Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. Nat Biotechnol, 2013. **31**(2): p. 154-9.
52. Schultz, M.D., R.J. Schmitz, and J.R. Ecker, 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. Trends Genet, 2012. **28**(12): p. 583-5.
53. Willing, E.-M., et al., Genome expansion of Arabis alpina linked with retrotransposition and reduced symmetric DNA methylation. Nature Plants, 2015. **1**(2): p. 14023.
54. Bewick, A.J., et al., On the Origin and Evolutionary Consequences of Gene Body DNA Methylation. bioRxiv, 2016.
55. McKey, D., et al., The evolutionary ecology of clonally propagated domesticated plants. New Phytol, 2010. **186**(2): p. 318-32.
56. Kitimu, S.R., et al., Meristem micropropagation of cassava (Manihot esculenta) evokes genome-wide changes in DNA methylation. Frontiers in Plant Science, 2015. **6**.
57. Chang, L., et al., Isolation of DNA-methyltransferase genes from strawberry (Fragaria x ananassa Duch.) and their expression in relation to micropropagation. Plant Cell Rep, 2009. **28**(9): p. 1373-84.
58. Martins, E.P. and T.F. Hansen, Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. American Naturalist, 1997. **149**(4): p. 646-667.
59. Duarte, J.M., et al., Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. Bmc Evolutionary Biology, 2010. **10**.
60. Flavell, R.B., et al., Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem Genet, 1974. **12**(4): p. 257-69.
61. Bennetzen, J.L., J. Ma, and K.M. Devos, Mechanisms of recent genome size variation in flowering plants. Ann Bot, 2005. **95**(1): p. 127-32.
62. Tran, R.K., et al., DNA methylation profiling identifies CG methylation clusters in Arabidopsis genes. Curr Biol, 2005. **15**(2): p. 154-9.
63. Zhang, X., et al., Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell, 2006. **126**(6): p. 1189-201.
64. 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.
65. 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. 725-734.
66. Martin, A., et al., A transposon-induced epigenetic change leads to sex determination in melon. Nature, 2009. **461**(7267): p. 1135-8.

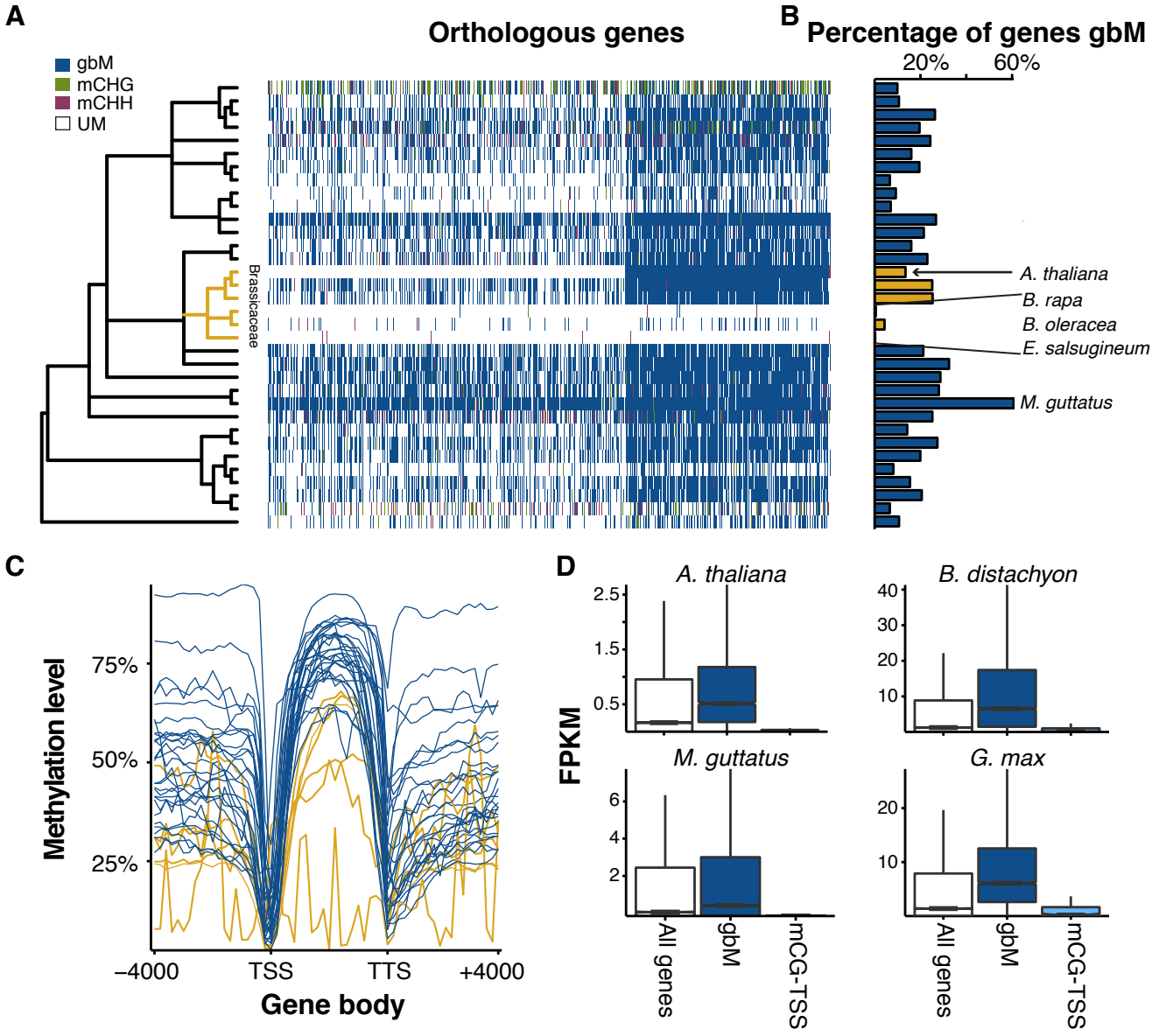
67. Durand, S., et al., Rapid establishment of genetic incompatibility through natural epigenetic variation. Curr Biol, 2012. **22**(4): p. 326-31.
68. West, P.T., et al., Genomic distribution of H3K9me2 and DNA methylation in a maize genome. PLoS One, 2014. **9**(8): p. e105267.
69. Stegemann, S., et al., High-frequency gene transfer from the chloroplast genome to the nucleus. Proc Natl Acad Sci U S A, 2003. **100**(15): p. 8828-33.
70. Roark, L.M., et al., Recent and frequent insertions of chloroplast DNA into maize nuclear chromosomes. Cytogenet Genome Res, 2010. **129**(1-3): p. 17-23.
71. Huang, C.Y., et al., Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiol, 2005. **138**(3): p. 1723-33.
72. Ong-Abdullah, M., et al., Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature, 2015.
73. Goodstein, D.M., et al., Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res, 2012. **40**(Database issue): p. D1178-86.
74. Sato, S., et al., Genome structure of the legume, Lotus japonicus. DNA Res, 2008. **15**(4): p. 227-39.
75. Dohm, J.C., et al., The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature, 2014. **505**(7484): p. 546-9.
76. van Bakel, H., et al., The draft genome and transcriptome of Cannabis sativa. Genome Biol, 2011. **12**(10): p. R102.
77. Slater, G.S. and E. Birney, Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics, 2005. **6**: p. 31.
78. Schultz, M.D., et al., Human body epigenome maps reveal noncanonical DNA methylation variation. Nature, 2015.
79. Martin, M., Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 2011. **17**(1): p. pp. 10-12.
80. Langmead, B., Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics, 2010. **Chapter 11**: p. Unit 11 7.
81. Bouckaert, R., et al., BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol, 2014. **10**(4): p. e1003537.
82. Mirarab, S., et al., PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. J Comput Biol, 2015. **22**(5): p. 377-86.
83. Castresana, J., Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution, 2000. **17**(4): p. 540-552.
84. Guindon, S. and O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol, 2003. **52**(5): p. 696-704.
85. Darriba, D., et al., jModelTest 2: more models, new heuristics and parallel computing. Nat Methods, 2012. **9**(8): p. 772.
86. Paradis, E., J. Claude, and K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics, 2004. **20**(2): p. 289-290.
87. Revell, L.J., phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution, 2012. **3**(2): p. 217-223.
88. Quinlan, A.R. and I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 2010. **26**(6): p. 841-2.
89. International Peach Genome, I., et al., The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet, 2013. **45**(5): p. 487-94.

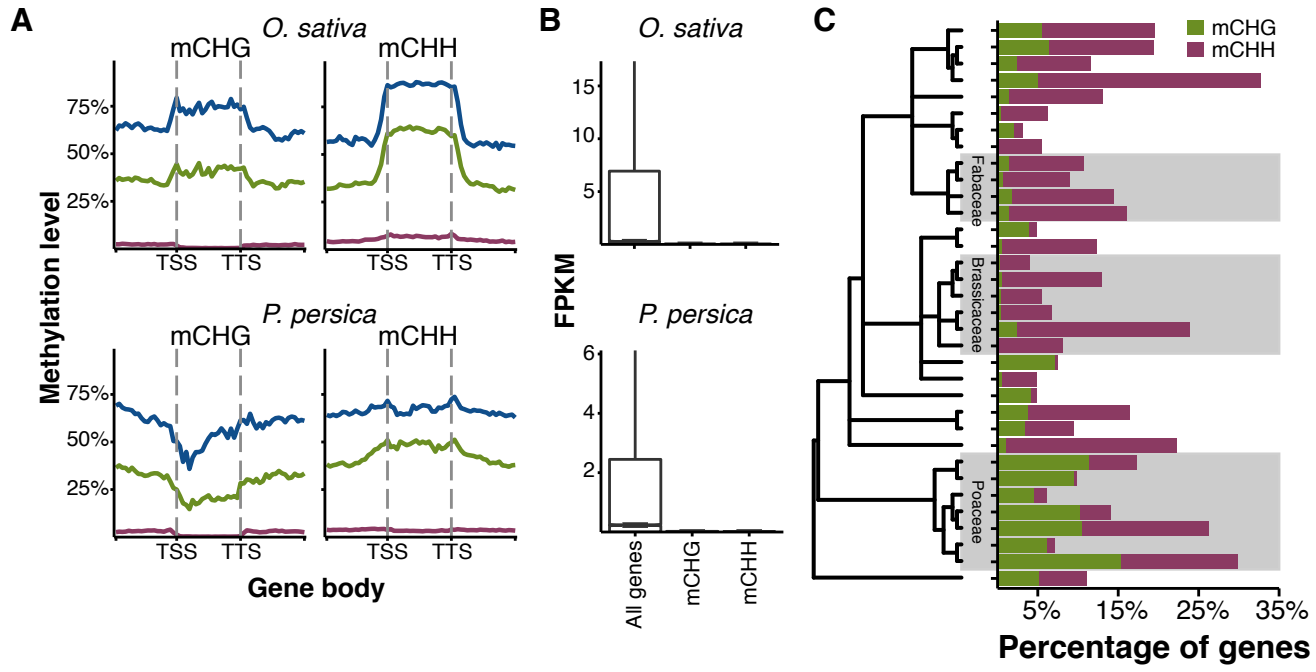
90. Chavez Montes, R.A., et al., Sample sequencing of vascular plants demonstrates widespread conservation and divergence of microRNAs. Nat Commun, 2014. **5**: p. 3722.
91. Wang, M., et al., Long noncoding RNAs and their proposed functions in fibre development of cotton (Gossypium spp.). New Phytol, 2015. **207**(4): p. 1181-97.
92. Stocks, M.B., et al., The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. Bioinformatics, 2012. **28**(15): p. 2059-61.
93. Prufer, K., et al., PatMaN: rapid alignment of short sequences to large databases. Bioinformatics, 2008. **24**(13): p. 1530-1531.
94. Grossmann, S., et al., Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. Bioinformatics, 2007. **23**(22): p. 3024-31.
95. Yang, Z. and R. Nielsen, Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol, 2000. **17**(1): p. 32-43.
96. Yang, Z., PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics, 1997. **13**(5): p. 555-556.
97. Brown, A.P., et al., Tissue-specific whole transcriptome sequencing in castor, directed at understanding triacylglycerol lipid biosynthetic pathways. PLoS One, 2012. **7**(2): p. e30100.
98. Chodavarapu, R.K., et al., Transcriptome and methylome interactions in rice hybrids. Proc Natl Acad Sci U S A, 2012. **109**(30): p. 12040-5.
99. Perazzolli, M., et al., Downy mildew resistance induced by Trichoderma harzianum T39 in susceptible grapevines partially mimics transcriptional changes of resistant genotypes. BMC Genomics, 2012. **13**: p. 660.
100. Tomato Genome, C., The tomato genome sequence provides insights into fleshy fruit evolution. Nature, 2012. **485**(7400): p. 635-41.
101. Tang, S., et al., Analysis of the Drought Stress-Responsive Transcriptome of Black Cottonwood (Populus trichocarpa) Using Deep RNA Sequencing. Plant Molecular Biology Reporter, 2014. **33**(3): p. 424-438.
102. Livingstone, D., et al., Making a chocolate chip: development and evaluation of a 6K SNP array for Theobroma cacao. DNA Res, 2015. **22**(4): p. 279-91.
103. Kim, D., et al., TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol, 2013. **14**(4): p. R36.
104. Trapnell, C., et al., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc, 2012. **7**(3): p. 562-78.

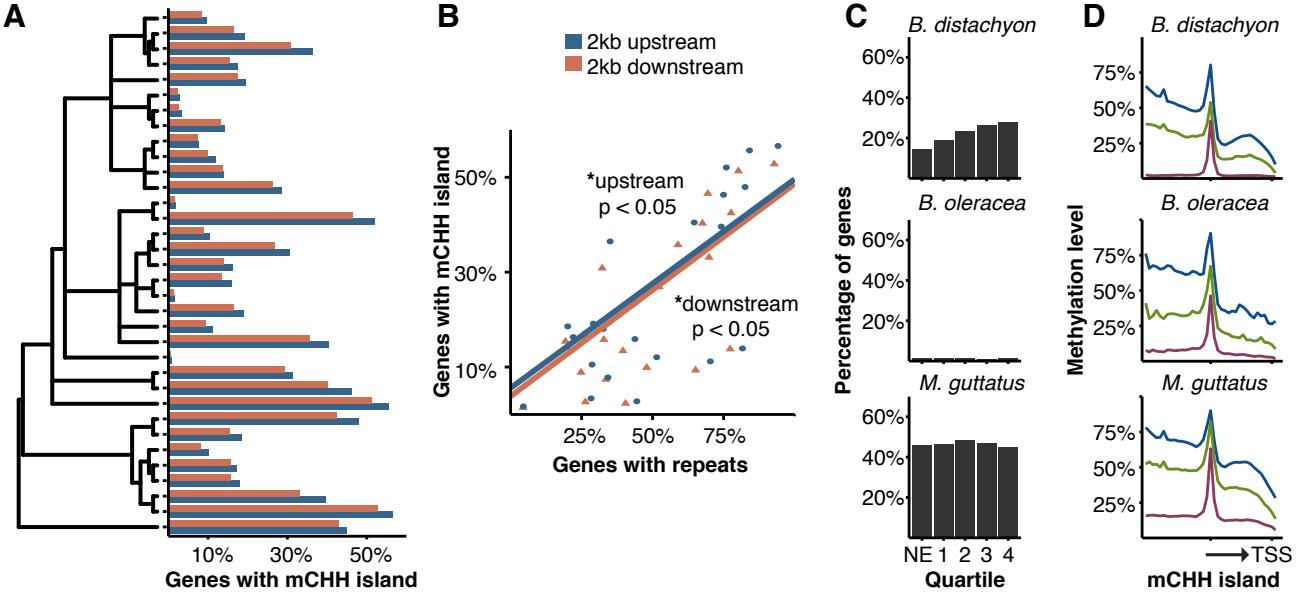

Methylation level
Proportion of total methylation
Per-site methylation
A
B
C
D
E
F
G
100%
60%
20%
25%
15%
5%
C. sativa
M. domestica
P. persica
F. vesca
C. sativus
P. trichocarpa
M. esculenta
R. communis
G. max
P. vulgaris
M. truncatula
L. japonicus
T. cacao
G. raimondii
A. thaliana
A. lyrata
C. rubella
B. rapa
B. oleracea
E. salsugineum
E. grandis
C. clementina
V. vinifera
S. lycopersicum
M. guttatus
B. vulgaris
O. sativa
B. distachyon
S. viridis
P. virgatum
P. hallii
S. bicolor
Z. mays
A. trichopoda
Fabaceae
Brassicaceae
Poaceae


A
Genome-wide
Methylation level
100%
75%
50%
25%
*mCG p < 0.05
*mCHG p < 0.05
mCG
mCHG
mCHH
500
1000
1500
2000
2500
3000
B
Coding regions
100%
75%
50%
25%
*mCHG p < 0.05
*mCHH p < 0.05
500
1000
1500
2000
2500
3000
Haploid genome size (Mbs)
C
mCG
mCHG
mCHH
genes
B. rapa
P. vulgaris
B. vulgaris
S. bicolor
75%
50%
25%
Methylation level
Number of genes
Distance (Mbs)
D
mCG
mCHG
mCHH
0.5
0.0
−0.5
Correlation (r) between methylation and gene content
Poaceae


A
Methylation level
100%
75%
50%
25%
*mCG p < 0.05
*mCHG p < 0.05
100,000
300,000
500,000
700,000
900,000
Repeat number
B
Methylation level
mCG
mCHG
mCHH
100%
75%
50%
25%
Brassicaceae
Poaceae
C
Methylation level
B. rapa
P. persica
B. distachyon
B. vulgaris
75%
50%
25%
5'
3'


A
Orthologous genes
gbM
mCHG
mCHH
UM
Brassicaceae
B
Percentage of genes gbM
20%
60%
A. thaliana
B. rapa
B. oleracea
E. salsugineum
M. guttatus
C
Methylation level
75%
50%
25%
–4000
TSS
TTS
+4000
Gene body
D
FPKM
A. thaliana
B. distachyon
M. guttatus
G. max
All genes
gbM
mCG-TSS


A
O. sativa
mCHG
mCHH
75%
50%
25%
TSS
TTS
P. persica
Methylation level
Gene body
B
FPKM
15
10
5
6
4
2
All genes
mCHG
mCHH
C
Fabaceae
Brassicaceae
Poaceae
5%
15%
25%
35%
Percentage of genes


A
10%
30%
50%
Genes with mCHH island
B
2kb upstream
2kb downstream
*upstream
p < 0.05
*downstream
p < 0.05
Genes with mCHH island
50%
30%
10%
25%
50%
75%
Genes with repeats
C
B. distachyon
B. oleracea
M. guttatus
60%
40%
20%
Percentage of genes
NE 1 2 3 4
Quartile
D
B. distachyon
B. oleracea
M. guttatus
75%
50%
25%
Methylation level
TSS
mCHH island