**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], Qing Li[5], 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]Department of Plant Biology, Microbial and Plant Genomics Institute, University of Minnesota, Saint Paul, MN 55108, USA

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

To understand the variation in genomic patterning of DNA methylation we compared methylomes of 34 diverse angiosperm species. By analyzing whole-genome bisulfite sequencing data in a phylogenetic context it becomes clear that there is extensive variation throughout angiosperms in gene body DNA methylation, euchromatic silencing of transposons and repeats, as well as silencing of heterochromatic transposons. The Brassicaceae have 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. These results show that DNA methylation patterns are broadly 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 (Vaughn, Tanurdzic et al. 2007, Eichten, Swanson-Wagner et al. 2011, Schmitz, Schultz et al. 2013), histones and histone modifications (Filleton, Chuffart et al. 2015), 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(Colome-Tatche, Cortijo et al. 2012, Jones 2012, Schmitz and Zhang 2014). Significant amounts of epigenomic diversity are explained by genetic variation ( Bell, Pai et al. 2011, Eichten, Swanson-Wagner et al. 2011, Eichten, Briskine et al. 2013, Regulski, Lu et al. 2013, Schmitz, He et al. 2013, Schmitz, Schultz et al. 2013, Dubin, Zhang et al. 2015), however, a large portion remains unexplained and in some cases these variants arise independently of genetic variation and are thus defined as “epigenetic” (Becker, Hagmann et al. 2011, Eichten, Swanson-Wagner et al. 2011, Schmitz, Schultz et al. 2011, Eichten, Briskine et al. 2013, Regulski, Lu et al. 2013, Schmitz, He et al. 2013). Moreover, epigenetic variants are heritable and also lead to phenotypic variation (Cubas, Vincent et al. 1999, Thompson, Tor et al. 1999, Manning, Tor et al. 2006, Silveira, Trontin et al. 2013). 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 (Arabidopsis Genome 2000, Flowers and Purugganan 2008). 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 (Lane, Niederhuth et al. 2014).

In plants, cytosine methylation occurs in three sequence contexts; CG, CHG, and CHH (H = A, T, or C), are under control by distinct mechanisms (Niederhuth and Schmitz 2014). Methylation at CG (mCG) and CHG (mCHG) sites is typically symmetrical across the Watson and Crick strands (Finnegan, Genger et al. 1998). mCG is maintained by the methyltransferase MET1, which is recruited to hemi-methylated CG sites and methylates the opposing strand (Finnegan, Peacock et al. 1996, Bostick, Kim et al. 2007), whereas mCHG is maintained by the plant specific CHROMOMETHYLASE 3 (CMT3) (Lindroth, Cao et al. 2001), and is strongly associated with dimethylation of lysine 9 on histone 3 (H3K9me2) (Du, Zhong et al. 2012). The BAH and CHROMO domains of CMT3 bind to H3K9me2, leading to methylation of CHG sites (Du, Zhong et al. 2012). In turn, the histone methyltransferases KRYPTONITE (KYP), and Su(var)3-9 homologue 5 (SUVH5) and SUVH6 recognize methylated DNA and methylate H3K9 (Du, Johnson et al. 2014), leading to a self-reinforcing loop (Du, Johnson et al. 2015). Asymmetrical methylation of CHH sites (mCHH) is established and maintained by another member of the CMT family, CMT2 (Zemach, Kim et al. 2013, Stroud, Do et al. 2014). CMT2, like CMT3, also contains BAH and CHROMO domains and methylates CHH in H3K9me2 regions (Zemach, Kim et al. 2013, Stroud, Do et al. 2014).

Additionally, all three sequence contexts are methylated *de novo* via RNA-directed DNA methylation (RdDM) (Law and Jacobsen 2010). Short-interfering 24 nucleotide (nt) RNAs (siRNAs) guide the *de novo* methyltransferase DOMAINS REARRANGED METHYLTRANSFERASE 2 (DRM2) to target sites (Cao and Jacobsen 2002, Cao and Jacobsen 2002). The targets of CMT2 and RdDM are often complementary, as CMT2 in *A. thaliana* primarily methylate regions of deep heterochromatin, such as transposons bodies ( Zemach, Kim et al. 2013). 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 ( Gent, Ellis et al. 2013, Zemach, Kim et al. 2013, Stroud, Do et al. 2014). 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 (Gent, Ellis et al. 2013, Li, Gent et al. 2015). 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 (Vaughn, Tanurdzic et al. 2007, Schmitz, Schultz et al. 2011, Schmitz, Schultz et al. 2013, Stroud, Greenberg et al. 2013, Dubin, Zhang et al. 2015). Whole genome bisulfite sequencing (WGBS) is particularly powerful, as it reveals genome-wide single nucleotide resolution of DNA methylation (Cokus, Feng et al. 2008, Lister, O'Malley et al. 2008, Lister and Ecker 2009). WGBS has been used to sequence an increasing number of plant methylomes, ranging from model plants like *A. thaliana* (Cokus, Feng et al. 2008, Lister, O'Malley et al. 2008) to economically important

crops like *Z. mays* (Eichten, Swanson-Wagner et al. 2011, Gent, Ellis et al. 2013, Regulski, Lu et al. 2013, Li, Eichten et al. 2014). This has enabled a new field of comparative epigenomics, which places methylation within an evolutionary context (Feng, Cokus et al. 2010, Zemach, McDaniel et al. 2010, Takuno and Gaut 2012, Takuno and Gaut 2013). 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 (Takuno, Ran et al. 2016) but does not provide a full view of changes in the patterns of context-specific methylation at different types of genomic regions (Seymour, Koenig et al. 2014).

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 suggests that both genomic, life history, and mechanistic differences between species contribute to this variation.

## Results

### Genome-wide DNA methylation variation across angiosperms

We compared single-base resolution methylomes of 34 angiosperm species that have genome assemblies ( Sato, Nakamura et al. 2008, van Bakel, Stout et al. 2011, Goodstein, Shu et al. 2012, Dohm, Minoche et al. 2014) (**Table S1**). MethylC-seq (Lister, O'Malley et al. 2008, Urich, Nery et al. 2015) was used to sequence 26 species and an additional eight species with previously published methylomes were downloaded and reanalyzed (Schmitz, Schultz et al. 2011, Amborella Genome 2013, Gent, Ellis et al. 2013, Schmitz, He et al. 2013, Stroud, Ding et al. 2013, Zhong, Fei et al. 2013, Seymour, Koenig et al. 2014). Different metrics were used to make comparisons at a whole-genome level. The genome-wide weighted methylation level (Schultz, Schmitz et al. 2012) combines data from the number of instances of methylated cytosine sites relative to all sequenced cytosine sites, giving a single value for each context that can be compared across 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 the distribution of methylation levels at individual methylated sites and indicates within a population of cells, the proportion that are methylated (**Figure 1E-G**). 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 S1,2**). The asymmetry of CHG sites are further quantified (**Figure S3**)based on symmetric methylation levels previously determined from *A. thaliana cmt3* mutants (Bewick, Ji et al. 2016). Per-site methylation and symmetry provide information into how well methylation is maintained and how ubiquitously the sites are methylated across cell types within sequenced tissues (Willing, Rawat et al. 2015).

These data revealed that DNA methylation is a major base modification in flowering plant genomes. However, there is extensive variation between species, suggesting differences in activity amongst known DNA methylation pathways and potentially undiscovered DNA methylation pathways. Within each species, mCG had the highest levels of methylation genome-wide (**Figure 1A, Table S2**). Between species, levels ranged as much as three fold, from a low of ~30.5% in *A. thaliana* to a high of ~92.5% in *Beta vulgaris.* . 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*, all have the lowest levels of per-site mCHG methylation (**Figure 1F**). Furthermore, symmetrical mCHG sites have a wider range of methylation levels and increased asymmetry, whereas, non-

Brassicaceae species have very highly methylated symmetrical sites (**Figure S2,3B**), suggesting that the CMT3 pathway is less effective in Brassicaceae genomes. This is further evidenced by *E. salsugineum*, with the lowest mCHG levels (**Figure 1b**), is a natural *cmt3* mutant whereas *CMT3* is under relaxed selection in other Brassicaceae (Bewick, Ji et al. 2016, Bewick, Niederhuth et al. 2016). Methylation of CG sites is also less well maintained in the Brassicaceae, with *Capsella rubella* showing the lowest levels of per-site mCG methylation (**Figure 1E, S1**).

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 (Zemach, Kim et al. 2013), 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 (McKey, Elias et al.

2010). 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 (Kitimu, Taylor et al. 2015), so has altered expression of methyltransferases due to micropropagation in *Fragaria* x *ananassa* (strawberry) (Chang, Zhang et al. 2009). If clonal propagation was responsible for low mCHH, we hypothesized that going through sexual reproduction might result in increased mCHH levels, as work in *A. thaliana* suggests that mCHH is reestablished during reproduction (Slotkin, Vaughn et al. 2009, Calarco, Borges et al. 2012). To test this, 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 of sexual reproduction (**Figure S5**). As both of these results are based on one generation of sexual reproduction it may be that this is insufficient to fully observe any changes. This will require further studies of samples collected over multiple generations from matching lines that have been either clonally propagated or propagated through seed for numerous 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 (PGLS) (Martins and Hansen 1997) which takes into account the phylogenetic relationship of species and modifies the slope and intercept of the regression line (**Table S3**). This methodology minimizes the influence of outlier species and the effects of having numerous closely related species of on the slope of the regression line. Phylogenetic relationships were inferred from a species tree constructed using 50 single copy loci for use in PGLS (**Figure S6**) (Duarte, Wall et al. 2010). Correlations from PGLS were corrected for the total number of PGLS tests conducted. Indeed, positive correlations were found between mCG (p-value = $2.9 \times 10^{-3}$) and the strongest correlation for mCHG and genome size (p-value = $2.2 \times 10^{-6}$) (**Figure 2A**). No correlation was observed between mCHH and genome size after multiple testing correction (p-value > 0.05) (**Figure 2A**). A relationship between genic methylation level and genome size has been previously reported (Takuno, Ran et al. 2016). We found that within coding sequences (CDS) methylation levels were correlated with genome size for both mCHG (p-value = $5 \times 10^{-6}$) and mCHH (p-value = $1.4 \times 10^{-5}$), but not for mCG (p-value > 0.18) (**Figure 2B**).

The highest levels of DNA methylation are typically found in centromeres and pericentromeric regions (Cokus, Feng et al. 2008, Lister, O'Malley et al. 2008, Seymour, Koenig et al. 2014). The distributions of methylation at chromosomal levels were examined in 100kb sliding windows (**Figure 2C, 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 Materials and 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 (Zemach, Kim et al.

2013). 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 (Bewick, Ji et al. 2016). 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* (Li, Eichten et al. 2014). 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 = $5 \times 10^{-5}$) and mCHG levels (p-value = $7.5 \times 10^{-3}$) (**Figure 3A, Table S3**). This likely also explains the correlation of methylation with genome size, as large genomes often have more repetitive elements (Flavell, Bennett et al. 1974, Bennetzen, Ma et al. 2005). 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 (Zemach, Li et al. 2005, Stroud, Do et al. 2014). Although coding sequence (CDS) mCHG and mCHH correlates with genome size (**Figure 2B)**, only CDS mCHG correlated with the total number of repeats (p-value = $3.8 \times 10^{-2}$) (**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 = $2.2 \times 10^{-3}$, mCHG p-value = $3.6 \times 10^{-4}$, mCHH p-value = $2 \times 10^{-4}$) (**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 (Law and Jacobsen 2010), whereas genes that are only mCG methylated within the gene body are often constitutively expressed genes (Tran, Henikoff et al. 2005, Zhang, Yazaki et al. 2006, Zilberman, Gehring et al. 2007). We classified genes using a modified version of the binomial test described by Takuno and Gaut (Takuno and Gaut 2012) into one of four categories: CG gene body methylated (hereafter gbM), mCHG, mCHH, and unmethylated (UM) (**Figure S13, 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 (Takuno and Gaut 2013). 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 4B**), *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) (Zhang, Yazaki et al. 2006, Zilberman, Gehring et al. 2007). gbM genes identified in most species show this same

trend and even have comparable levels of methylation (**Figure 4C, S14**). 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, respectively 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 (Zhang, Yazaki et al. 2006, Zilberman, Gehring et al. 2007), gbM genes are more highly expressed as compared to UM and nonCG (mCHG and mCHH) genes (**Figure 4D, S15**). 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 4D, S15**) 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 (Takuno and Gaut 2012, Takuno and Gaut 2013). 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* (Takuno and Gaut 2012) and between *B. distachyon* and *O. sativa* (Takuno and Gaut 2013) 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 (Bender and Fink 1995, Cubas, Vincent et al. 1999, Manning, Tor et al. 2006, Martin, Troadec et al. 2009, Durand, Bouche et al. 2012). 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 (Schmitz, He et al. 2013, Schmitz, Schultz et al. 2013). 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 (West, Li et al. 2014). 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, S14**). Both mCHG and mCHH genes are associated with reduced expression levels (**Figure 5B, S15**). 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. This pattern was not observed in a second *C. sativus* sample and will require further study to understand its cause (**Figure S16**). 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 S17**). 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 (Seymour, Koenig et al. 2014). However, we did observe some conservation based on gene ontology (GO). The same GO terms were often enriched in multiple species (**Figure S18, 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 (Stegemann, Hartmann et al. 2003, Roark, Hui et al. 2010). 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 (Huang, Grunheit et al. 2005).

**mCHH islands**

In *Z. mays,* high mCHH is enriched in the upstream and downstream regions of highly expressed genes and are termed mCHH islands (Gent, Ellis et al. 2013, Li, Gent et al. 2015). 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 S19A, 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 S19B, 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 = $7.9 \times 10^{-6}$, downstream p-value < $8.1 \times 10^{-6}$) (**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 (Gent, Ellis et al. 2013, Li, Gent et al. 2015). This is true of several species, such as *P. persica* and all the Poaceae (**Figure 6C, S20**). However, many species showed no significant association between mCHH islands and gene expression (**Figure 6C, S20**). 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, S21**), marking a boundary of euchromatin and heterochromatin (Li, Gent et al. 2015). 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 S21**). 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 (Bewick, Ji et al. 2016). 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 (Ong-Abdullah, Ordway et al. 2015), and understanding the effects of clonal propagation will require 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* (Takuno, Ran et al. 2016) and *Selaginella moellendorffii* (Zemach, McDaniel et al. 2010). 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* (Silveira, Trontin et al. 2013). 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 (Seymour, Koenig et al. 2014, Takuno, Ran et al. 2016) 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.

## Materials and methods

### MethylC-seq and analysis

In plants, DNA methylation is highly stable between tissues and across generations (Schmitz, Schultz et al. 2011, Seymour, Koenig et al. 2014), showing little variation between replicates. DNA was isolated from leaf tissue and MethylC-seq libraries for each species were prepared as previously described (Urich, Nery et al. 2015). Previously published datasets were obtained from public databases and reanalyzed (Tuskan, Difazio et al. 2006, Schmitz, Schultz et al. 2011, Amborella Genome 2013, Gent, Ellis et al. 2013, Schmitz, He et al. 2013, Stroud, Ding et al. 2013, Zhong, Fei et al. 2013, Seymour, Koenig et al. 2014, Bewick, Ji et al. 2016). Genome sequences and annotations for most species were downloaded from Phytozome 10.1 (http://phytozome.jgi.doe.gov/pz/portal.html) ( Goodstein, Shu et al. 2012). The *L. japonicus* genome was downloaded from the *Lotus japonicus* Sequencing Project (http://www.kazusa.or.jp/lotus/) (Sato, Nakamura et al. 2008), the *B. vulgaris* genome was downloaded from *Beta vulgaris* Resource (bvseq.molgen.mpg.de/) (Dohm, Minoche et al. 2014), and the *C. sativa* genome from *C. sativa* (*Cannabis*) Genome Browser Gateway (http://genome.ccbr.utoronto.ca/cgi-bin/hgGateway) (van Bakel, Stout et al. 2011). 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 (Slater and Birney 2005) and the best hits retained.

Sequencing data for each species was aligned to their respective genome (Supplementary Table 1) (Arabidopsis Genome 2000, Jaillon, Aury et al. 2007, Ouyang,

Zhu et al. 2007, Sato, Nakamura et al. 2008, Paterson, Bowers et al. 2009, Schnable, Ware et al. 2009, Chan, Crabtree et al. 2010, International Brachypodium 2010, Schmutz, Cannon et al. 2010, Velasco, Zharkikh et al. 2010, Hu, Pattyn et al. 2011, Shulaev, Sargent et al. 2011, van Bakel, Stout et al. 2011, Young, Debelle et al. 2011, Goodstein, Shu et al. 2012, Paterson, Wendel et al. 2012, Prochnik, Marri et al. 2012, Tomato Genome 2012, Amborella Genome 2013, Hellsten, Wright et al. 2013, International Peach Genome, Verde et al. 2013, Motamayor, Mockaitis et al. 2013, Slotte, Hazzouri et al. 2013, Yang, Jarvis et al. 2013, Dohm, Minoche et al. 2014, Myburg, Grattapaglia et al. 2014, Parkin, Koh et al. 2014, Schmutz, McClean et al. 2014, Wu, Prochnik et al. 2014) and methylated sites called using previously described methods (Schultz, He et al. 2015). In brief, reads were trimmed for adapters and quality using Cutadapt (Martin 2011) and then mapped to both a converted forward strand (all cytosines to thymines) and converted reverse strand (all guanines to adenines) using bowtie (Langmead 2010). 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 (Supplementary Table 1). 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 (Bouckaert, Heled et al. 2014) on a set of 50 previously identified single copy loci (Duarte, Wall et al. 2010). Protein sequences were aligned using PASTA (Mirarab, Nguyen et al. 2015) and converted into codon alignments using custom Perl scripts. Gblocks (Castresana 2000) was used to identify conserved stretches of amino acids and then passed to JModelTest2 (Guindon and Gascuel 2003, Darriba, Taboada et al. 2012) 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 (Schultz, Schmitz et al. 2012). To determine per-site methylation levels, the weighted methylation for each cystosine with at least 3 reads of coverage was calculated and this distribution plotted. Symmetry plots were constructed by identifying paired symmetrical cytosines that had sequencing coverage and plotting the per-site methylation level of the cytosine on the Watson strand against the per-site methylation level of the Crick strand. An *A. thaliana cmt3* mutant was used to empirically determine the per-site methylation level at which symmetrical methylation disappeared (Bewick, Ji et al. 2016) as 40%. Methylated symmetrical pairs above this level were considered to be symmetrically methylated, while those below as asymmetrical.  Correlations between methylation levels, genome sizes, and gene numbers were done in R and corrected for phylogenetic signal using the APE ( Paradis,

Claude et al. 2004), phytools (Revell 2012), and NLME packages assuming a model of Brownian motion. In total, 22 comparisons were conducted (Supplementary Table 3) 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 (Quinlan and Hall 2010) 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 (Quinlan and Hall 2010) 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 (Schmitz, Schultz et al. 2011, Amborella Genome 2013, International Peach Genome, Verde et al. 2013, Stroud, Ding et al. 2013, Chavez Montes, de Fatima Rosas-Cardenas et al. 2014, Wang, Yuan et al. 2015). The small RNA toolkit from the UEA computational Biology lab was used to trim and clean the reads (Stocks, Moxon et al. 2012). For trimming, 8 bp of the 3' adapter was trimmed. Trimmed and cleaned reads were aligned using PatMan allowing for zero mismatches (Prufer, Stenzel et al. 2008). BedTools (Quinlan and Hall 2010) 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 (Takuno and Gaut 2012) (**Figure S13, Table S4**). 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) (Goodstein, Shu et al. 2012). GO term enrichment was performed using the parentCHILD algorithm (Grossmann, Bauer et al. 2007) 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 (Goodstein, Shu et al. 2012) 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 (Goodstein, Shu et al. 2012) 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 (Edgar 2004), 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) (Yang and Nielsen 2000) model in the program PAML for pairwise sequence comparison was used to estimate synonymous and non-synonymous substitution rates, and adaptive evolution (*dS*, *dN*, and ω, respectively) (Yang 1997). 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 (Schmitz, Schultz et al. 2011, Brown, Kroon et al. 2012, Chodavarapu, Feng et al. 2012, Perazzolli, Moretto et al. 2012, Tomato Genome 2012, Amborella Genome 2013, International Peach Genome, Verde et al. 2013, Schmitz, He et al. 2013, Seymour, Koenig et al. 2014, Tang, Dong et al. 2014, Livingstone, Royaert et al. 2015, Wang, Yuan et al. 2015, Bewick, Ji et al. 2016) 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 (Kim, Pertea et al. 2013) 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 (Trapnell, Roberts et al. 2012) supplied with a reference GFF.

**mCHH islands**

mCHH islands were identified for both upstream and downstream regions as previously described (Li, Gent et al. 2015). 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.

**Author Contributions**

Conceptualization, C.E.N., A.J.B.and R.J.S.; Performed Experiments, C.E.N., N.A.R., K.D.K., A.R., J.T.P., and R.J.S.; Data Analysis, C.E.N., A.J.B., L.J., M.S.A.; Writing – Original Draft, C.E.N; Writing – Review & Editing, C.E.N., A.J.B,. S.A.J., N.M.S., and R.J.S.; Resources, Q.L., J.M.B., J.A.U., C.E., J.S., J.G., S.A.J., N.M.S.

**Competing Financial Interests**

The authors declare no competing interests.

## References


Amborella Genome, P. (2013). "The Amborella genome and the evolution of flowering plants." Science **342**(6165): 1241089.

Arabidopsis Genome, I. (2000). "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana." Nature **408**(6814): 796-815.

Becker, C., J. Hagmann, J. Muller, D. Koenig, O. Stegle, K. Borgwardt and D. Weigel (2011). "Spontaneous epigenetic variation in the Arabidopsis thaliana methylome." Nature **480**(7376): 245-249.

Bell, J. T., A. A. Pai, J. K. Pickrell, D. J. Gaffney, R. Pique-Regi, J. F. Degner, Y. Gilad and J. K. Pritchard (2011). "DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines." Genome Biol **12**(1): R10.

Bender, J. and G. R. Fink (1995). "Epigenetic Control of an Endogenous Gene Family Is Revealed by a Novel Blue Fluorescent Mutant of Arabidopsis." Cell **83**(5): 725-734.

Bennetzen, J. L., J. Ma and K. M. Devos (2005). "Mechanisms of recent genome size variation in flowering plants." Ann Bot **95**(1): 127-132.

Bewick, A. J., L. Ji, C. E. Niederhuth, E.-M. Willing, B. T. Hofmeister, X. Shi, L. Wang, Z. Lu, N. A. Rohr, B. Hartwig, C. Kiefer, R. B. Deal, J. Schmutz, J. Grimwood, H. Stroud, S. E. Jacobsen, K. Schneeberger, X. Zhang and R. J. Schmitz (2016). "On the Origin and Evolutionary Consequences of Gene Body DNA Methylation." bioRxiv.

Bewick, A. J., C. E. Niederhuth, N. A. Rohr, P. T. Griffin, J. Leebens-Mack and R. J. Schmitz (2016). "The evolution of CHROMOMETHYLTRANSFERASES and gene body DNA methylation in plants." bioRxiv.

Bostick, M., J. K. Kim, P. O. Esteve, A. Clark, S. Pradhan and S. E. Jacobsen (2007). "UHRF1 plays a role in maintaining DNA methylation in mammalian cells." Science **317**(5845): 1760-1764.

Bouckaert, R., J. Heled, D. Kuhnert, T. Vaughan, C. H. Wu, D. Xie, M. A. Suchard, A. Rambaut and A. J. Drummond (2014). "BEAST 2: a software platform for Bayesian evolutionary analysis." PLoS Comput Biol **10**(4): e1003537.

Brown, A. P., J. T. Kroon, D. Swarbreck, M. Febrer, T. R. Larson, I. A. Graham, M. Caccamo and A. R. Slabas (2012). "Tissue-specific whole transcriptome sequencing in castor, directed at understanding triacylglycerol lipid biosynthetic pathways." PLoS One **7**(2): e30100.

Calarco, J. P., F. Borges, M. T. Donoghue, F. Van Ex, P. E. Jullien, T. Lopes, R. Gardner, F. Berger, J. A. Feijo, J. D. Becker and R. A. Martienssen (2012). "Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA." Cell **151**(1): 194-205.

Cao, X. and S. E. Jacobsen (2002). "Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes." Proc Natl Acad Sci U S A **99 Suppl 4**: 16491-16498.

Cao, X. and S. E. Jacobsen (2002). "Role of the Arabidopsis DRM Methyltransferases in De Novo DNA Methylation and Gene Silencing." Current Biology **12**(13): 1138-1144.

Castresana, J. (2000). "Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis." Molecular Biology and Evolution **17**(4): 540-552.

Chan, A. P., J. Crabtree, Q. Zhao, H. Lorenzi, J. Orvis, D. Puiu, A. Melake-Berhan, K. M. Jones, J. Redman, G. Chen, E. B. Cahoon, M. Gedil, M. Stanke, B. J. Haas, J. R. Wortman, C. M. Fraser-Liggett, J. Ravel and P. D. Rabinowicz (2010). "Draft genome sequence of the oilseed species Ricinus communis." Nat Biotechnol **28**(9): 951-956.

Chang, L., Z. Zhang, B. Han, H. Li, H. Dai, P. He and H. Tian (2009). "Isolation of DNA-methyltransferase genes from strawberry (Fragaria x ananassa Duch.) and their expression in relation to micropropagation." Plant Cell Rep **28**(9): 1373-1384.

Chavez Montes, R. A., F. de Fatima Rosas-Cardenas, E. De Paoli, M. Accerbi, L. A. Rymarquis, G. Mahalingam, N. Marsch-Martinez, B. C. Meyers, P. J. Green and S. de Folter (2014). "Sample sequencing of vascular plants demonstrates widespread conservation and divergence of microRNAs." Nat Commun **5**: 3722.

Chodavarapu, R. K., S. Feng, B. Ding, S. A. Simon, D. Lopez, Y. Jia, G. L. Wang, B. C. Meyers, S. E. Jacobsen and M. Pellegrini (2012). "Transcriptome and methylome interactions in rice hybrids." Proc Natl Acad Sci U S A **109**(30): 12040-12045.

Cokus, S. J., S. Feng, X. Zhang, Z. Chen, B. Merriman, C. D. Haudenschild, S. Pradhan, S. F. Nelson, M. Pellegrini and S. E. Jacobsen (2008). "Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning." Nature **452**(7184): 215-219.

Colome-Tatche, M., S. Cortijo, R. Wardenaar, L. Morgado, B. Lahouze, A. Sarazin, M. Etcheverry, A. Martin, S. H. Feng, E. Duvernois-Berthet, K. Labadie, P. Wincker, S. E. Jacobsen, R. C. Jansen, V. Colot and F. Johannes (2012). "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 **109**(40): 16240-16245.

Cubas, P., C. Vincent and E. Coen (1999). "An epigenetic mutation responsible for natural variation in floral symmetry." Nature **401**(6749): 157-161.

Darriba, D., G. L. Taboada, R. Doallo and D. Posada (2012). "jModelTest 2: more models, new heuristics and parallel computing." Nat Methods **9**(8): 772.

Dohm, J. C., A. E. Minoche, D. Holtgrawe, S. Capella-Gutierrez, F. Zakrzewski, H. Tafer, O. Rupp, T. R. Sorensen, R. Stracke, R. Reinhardt, A. Goesmann, T. Kraft, B. Schulz, P. F. Stadler, T. Schmidt, T. Gabaldon, H. Lehrach, B. Weisshaar and H. Himmelbauer (2014). "The genome of the recently domesticated crop plant sugar beet (Beta vulgaris)." Nature **505**(7484): 546-549.

Du, J., L. M. Johnson, M. Groth, S. Feng, C. J. Hale, S. Li, A. A. Vashisht, J. Gallego-Bartolome, J. A. Wohlschlegel, D. J. Patel and S. E. Jacobsen (2014). "Mechanism of DNA methylation-directed histone methylation by KRYPTONITE." Mol Cell **55**(3): 495-504.

Du, J., L. M. Johnson, S. E. Jacobsen and D. J. Patel (2015). "DNA methylation pathways and their crosstalk with histone methylation." Nat Rev Mol Cell Biol **16**(9): 519-532.

Du, J., X. Zhong, Y. V. Bernatavichute, H. Stroud, S. Feng, E. Caro, A. A. Vashisht, J. Terragni, H. G. Chin, A. Tu, J. Hetzel, J. A. Wohlschlegel, S. Pradhan, D. J. Patel and S. E. Jacobsen (2012). "Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants." Cell **151**(1): 167-180.

Duarte, J. M., P. K. Wall, P. P. Edger, L. L. Landherr, H. Ma, J. C. Pires, J. Leebens-Mack and C. W. dePamphilis (2010). "Identification of shared single copy nuclear genes in Arabidopsis,

Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels." Bmc Evolutionary Biology **10**.

Dubin, M. J., P. Zhang, D. Meng, M. S. Remigereau, E. J. Osborne, F. Paolo Casale, P. Drewe, A. Kahles, G. Jean, B. Vilhjalmsson, J. Jagoda, S. Irez, V. Voronin, Q. Song, Q. Long, G. Ratsch, O. Stegle, R. M. Clark and M. Nordborg (2015). "DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation." Elife **4**: e05255.

Durand, S., N. Bouche, E. Perez Strand, O. Loudet and C. Camilleri (2012). "Rapid establishment of genetic incompatibility through natural epigenetic variation." Curr Biol **22**(4): 326-331.

Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res **32**(5): 1792-1797.

Eichten, S. R., R. Briskine, J. Song, Q. Li, R. Swanson-Wagner, P. J. Hermanson, A. J. Waters, E. Starr, P. T. West, P. Tiffin, C. L. Myers, M. W. Vaughn and N. M. Springer (2013). "Epigenetic and genetic influences on DNA methylation variation in maize populations." Plant Cell **25**(8): 2783-2797.

Eichten, S. R., R. A. Swanson-Wagner, J. C. Schnable, A. J. Waters, P. J. Hermanson, S. Liu, C. T. Yeh, Y. Jia, K. Gendler, M. Freeling, P. S. Schnable, M. W. Vaughn and N. M. Springer (2011). "Heritable epigenetic variation among maize inbreds." PLoS Genet **7**(11): e1002372.

Feng, S., S. J. Cokus, X. Zhang, P. Y. Chen, M. Bostick, M. G. Goll, J. Hetzel, J. Jain, S. H. Strauss, M. E. Halpern, C. Ukomadu, K. C. Sadler, S. Pradhan, M. Pellegrini and S. E. Jacobsen (2010). "Conservation and divergence of methylation patterning in plants and animals." Proc Natl Acad Sci U S A **107**(19): 8689-8694.

Filleton, F., F. Chuffart, M. Nagarajan, H. Bottin-Duplus and G. Yvert (2015). "The complex pattern of epigenomic variation between natural yeast strains at single-nucleosome resolution." Epigenetics Chromatin **8**: 26.

Finnegan, E. J., R. K. Genger, W. J. Peacock and E. S. Dennis (1998). "DNA Methylation in Plants." Annu Rev Plant Physiol Plant Mol Biol **49**: 223-247.

Finnegan, E. J., W. J. Peacock and E. S. Dennis (1996). "Reduced DNA methylation in Arabidopsis thaliana results in abnormal plant development." Proc Natl Acad Sci U S A **93**(16): 8449-8454.

Flavell, R. B., M. D. Bennett, J. B. Smith and D. B. Smith (1974). "Genome size and the proportion of repeated nucleotide sequence DNA in plants." Biochem Genet **12**(4): 257-269.

Flowers, J. M. and M. D. Purugganan (2008). "The evolution of plant genomes: scaling up from a population perspective." Curr Opin Genet Dev **18**(6): 565-570.

Gent, J. I., N. A. Ellis, L. Guo, A. E. Harkess, Y. Yao, X. Zhang and R. K. Dawe (2013). "CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize." Genome Res **23**(4): 628-637.

Goodstein, D. M., S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam and D. S. Rokhsar (2012). "Phytozome: a comparative platform for green plant genomics." Nucleic Acids Res **40**(Database issue): D1178-1186.

Grossmann, S., S. Bauer, P. N. Robinson and M. Vingron (2007). "Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis." Bioinformatics **23**(22): 3024-3031.

Guindon, S. and O. Gascuel (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." Syst Biol **52**(5): 696-704.

Hellsten, U., K. M. Wright, J. Jenkins, S. Shu, Y. Yuan, S. R. Wessler, J. Schmutz, J. H. Willis and D. S. Rokhsar (2013). "Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing." Proc Natl Acad Sci U S A **110**(48): 19478-19482.

Hu, T. T., P. Pattyn, E. G. Bakker, J. Cao, J. F. Cheng, R. M. Clark, N. Fahlgren, J. A. Fawcett, J. Grimwood, H. Gundlach, G. Haberer, J. D. Hollister, S. Ossowski, R. P. Ottilar, A. A. Salamov, K. Schneeberger, M. Spannagl, X. Wang, L. Yang, M. E. Nasrallah, J. Bergelson, J. C. Carrington, B. S. Gaut, J. Schmutz, K. F. Mayer, Y. Van de Peer, I. V. Grigoriev, M. Nordborg, D. Weigel and Y. L. Guo (2011). "The Arabidopsis lyrata genome sequence and the basis of rapid genome size change." Nat Genet **43**(5): 476-481.

Huang, C. Y., N. Grunheit, N. Ahmadinejad, J. N. Timmis and W. Martin (2005). "Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes." Plant Physiol **138**(3): 1723-1733.

International Brachypodium, I. (2010). "Genome sequencing and analysis of the model grass Brachypodium distachyon." Nature **463**(7282): 763-768.

International Peach Genome, I., I. Verde, A. G. Abbott, S. Scalabrin, S. Jung, S. Shu, F. Marroni, T. Zhebentyayeva, M. T. Dettori, J. Grimwood, F. Cattonaro, A. Zuccolo, L. Rossini, J. Jenkins, E. Vendramin, L. A. Meisel, V. Decroocq, B. Sosinski, S. Prochnik, T. Mitros, A. Policriti, G. Cipriani, L. Dondini, S. Ficklin, D. M. Goodstein, P. Xuan, C. Del Fabbro, V. Aramini, D. Copetti, S. Gonzalez, D. S. Horner, R. Falchi, S. Lucas, E. Mica, J. Maldonado, B. Lazzari, D. Bielenberg, R. Pirona, M. Miculan, A. Barakat, R. Testolin, A. Stella, S. Tartarini, P. Tonutti, P.

Arus, A. Orellana, C. Wells, D. Main, G. Vizzotto, H. Silva, F. Salamini, J. Schmutz, M. Morgante and D. S. Rokhsar (2013). "The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution." Nat Genet **45**(5): 487-494.

Jaillon, O., J. M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyere, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M. E. Pe, G. Valle, M. Morgante, M. Caboche, A. F. Adam-Blondon, J. Weissenbach, F. Quetier, P. Wincker and C. French-Italian Public Consortium for Grapevine Genome (2007). "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla." Nature **449**(7161): 463-467.

Jones, P. A. (2012). "Functions of DNA methylation: islands, start sites, gene bodies and beyond." Nat Rev Genet **13**(7): 484-492.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley and S. L. Salzberg (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." Genome Biol **14**(4): R36.

Kitimu, S. R., J. Taylor, T. J. March, F. Tairo, M. J. Wilkinson and C. M. Rodríguez López (2015). "Meristem micropropagation of cassava (Manihot esculenta) evokes genome-wide changes in DNA methylation." Frontiers in Plant Science **6**.

Lane, A. K., C. E. Niederhuth, L. Ji and R. J. Schmitz (2014). "pENCODE: a plant encyclopedia of DNA elements." Annu Rev Genet **48**: 49-70.

Langmead, B. (2010). "Aligning short sequencing reads with Bowtie." Curr Protoc Bioinformatics **Chapter 11**: Unit 11 17.

Law, J. A. and S. E. Jacobsen (2010). "Establishing, maintaining and modifying DNA methylation patterns in plants and animals." Nat Rev Genet **11**(3): 204-220.

Li, Q., S. R. Eichten, P. J. Hermanson, V. M. Zaunbrecher, J. Song, J. Wendt, H. Rosenbaum, T. F. Madzima, A. E. Sloan, J. Huang, D. L. Burgess, T. A. Richmond, K. M. McGinnis, R. B. Meeley, O. N. Danilevskaya, M. W. Vaughn, S. M. Kaeppler, J. A. Jeddeloh and N. M. Springer (2014). "Genetic perturbation of the maize methylome." Plant Cell **26**(12): 4602-4616.

Li, Q., J. I. Gent, G. Zynda, J. Song, I. Makarevitch, C. D. Hirsch, C. N. Hirsch, R. K. Dawe, T. F. Madzima, K. M. McGinnis, D. Lisch, R. J. Schmitz, M. W. Vaughn and N. M. Springer (2015). "RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome." Proc Natl Acad Sci U S A.

Lindroth, A. M., X. Cao, J. P. Jackson, D. Zilberman, C. M. McCallum, S. Henikoff and S. E. Jacobsen (2001). "Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation." Science **292**(5524): 2077-2080.

Lister, R. and J. R. Ecker (2009). "Finding the fifth base: genome-wide sequencing of cytosine methylation." Genome Res **19**(6): 959-966.

Lister, R., R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar and J. R. Ecker (2008). "Highly integrated single-base resolution maps of the epigenome in Arabidopsis." Cell **133**(3): 523-536.

Livingstone, D., S. Royaert, C. Stack, K. Mockaitis, G. May, A. Farmer, C. Saski, R. Schnell, D. Kuhn and J. C. Motamayor (2015). "Making a chocolate chip: development and evaluation of a 6K SNP array for Theobroma cacao." DNA Res **22**(4): 279-291.

Manning, K., M. Tor, M. Poole, Y. Hong, A. J. Thompson, G. J. King, J. J. Giovannoni and G. B. Seymour (2006). "A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening." Nat Genet **38**(8): 948-952.

Martin, A., C. Troadec, A. Boualem, M. Rajab, R. Fernandez, H. Morin, M. Pitrat, C. Dogimont and A. Bendahmane (2009). "A transposon-induced epigenetic change leads to sex determination in melon." Nature **461**(7267): 1135-1138.

Martin, M. (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads." EMBnet. journal **17**(1): pp. 10-12.

Martins, E. P. and T. F. Hansen (1997). "Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data." American Naturalist **149**(4): 646-667.

McKey, D., M. Elias, B. Pujol and A. Duputie (2010). "The evolutionary ecology of clonally propagated domesticated plants." New Phytol **186**(2): 318-332.

Mirarab, S., N. Nguyen, S. Guo, L. S. Wang, J. Kim and T. Warnow (2015). "PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences." J Comput Biol **22**(5): 377-386.

Motamayor, J. C., K. Mockaitis, J. Schmutz, N. Haiminen, D. Livingstone, 3rd, O. Cornejo, S. D. Findley, P. Zheng, F. Utro, S. Royaert, C. Saski, J. Jenkins, R. Podicheti, M. Zhao, B. E. Scheffler, J. C. Stack, F. A. Feltus, G. M. Mustiga, F. Amores, W. Phillips, J. P. Marelli, G. D.

May, H. Shapiro, J. Ma, C. D. Bustamante, R. J. Schnell, D. Main, D. Gilbert, L. Parida and D. N. Kuhn (2013). "The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color." Genome Biol **14**(6): r53.

Myburg, A. A., D. Grattapaglia, G. A. Tuskan, U. Hellsten, R. D. Hayes, J. Grimwood, J. Jenkins, E. Lindquist, H. Tice, D. Bauer, D. M. Goodstein, I. Dubchak, A. Poliakov, E. Mizrachi, A. R. Kullan, S. G. Hussey, D. Pinard, K. van der Merwe, P. Singh, I. van Jaarsveld, O. B. Silva-Junior, R. C. Togawa, M. R. Pappas, D. A. Faria, C. P. Sansaloni, C. D. Petroli, X. Yang, P. Ranjan, T. J. Tschaplinski, C. Y. Ye, T. Li, L. Sterck, K. Vanneste, F. Murat, M. Soler, H. S. Clemente, N. Saidi, H. Cassan-Wang, C. Dunand, C. A. Hefer, E. Bornberg-Bauer, A. R. Kersting, K. Vining, V. Amarasinghe, M. Ranik, S. Naithani, J. Elser, A. E. Boyd, A. Liston, J. W. Spatafora, P. Dharmwardhana, R. Raja, C. Sullivan, E. Romanel, M. Alves-Ferreira, C. Kulheim, W. Foley, V. Carocha, J. Paiva, D. Kudrna, S. H. Brommonschenkel, G. Pasquali, M. Byrne, P. Rigault, J. Tibbits, A. Spokevicius, R. C. Jones, D. A. Steane, R. E. Vaillancourt, B. M. Potts, F. Joubert, K. Barry, G. J. Pappas, S. H. Strauss, P. Jaiswal, J. Grima-Pettenati, J. Salse, Y. Van de Peer, D. S. Rokhsar and J. Schmutz (2014). "The genome of Eucalyptus grandis." Nature **510**(7505): 356-362.

Niederhuth, C. E. and R. J. Schmitz (2014). "Covering your bases: inheritance of DNA methylation in plant genomes." Mol Plant **7**(3): 472-480.

Ong-Abdullah, M., J. M. Ordway, N. Jiang, S. E. Ooi, S. Y. Kok, N. Sarpan, N. Azimi, A. T. Hashim, Z. Ishak, S. K. Rosli, F. A. Malike, N. A. Bakar, M. Marjuni, N. Abdullah, Z. Yaakub, M. D. Amiruddin, R. Nookiah, R. Singh, E. L. Low, K. L. Chan, N. Azizi, S. W. Smith, B. Bacher, M. A. Budiman, A. Van Brunt, C. Wischmeyer, M. Beil, M. Hogan, N. Lakey, C. C. Lim,

X. Arulandoo, C. K. Wong, C. N. Choo, W. C. Wong, Y. Y. Kwan, S. S. Alwee, R. Sambanthamurthi and R. A. Martienssen (2015). "Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm." Nature.

Ouyang, S., W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, F. Thibaud-Nissen, R. L. Malek, Y. Lee, L. Zheng, J. Orvis, B. Haas, J. Wortman and C. R. Buell (2007). "The TIGR Rice Genome Annotation Resource: improvements and new features." Nucleic Acids Res **35**(Database issue): D883-887.

Paradis, E., J. Claude and K. Strimmer (2004). "APE: Analyses of Phylogenetics and Evolution in R language." Bioinformatics **20**(2): 289-290.

Parkin, I. A., C. Koh, H. Tang, S. J. Robinson, S. Kagale, W. E. Clarke, C. D. Town, J. Nixon, V. Krishnakumar, S. L. Bidwell, F. Denoeud, H. Belcram, M. G. Links, J. Just, C. Clarke, T. Bender, T. Huebert, A. S. Mason, J. C. Pires, G. Barker, J. Moore, P. G. Walley, S. Manoli, J. Batley, D. Edwards, M. N. Nelson, X. Wang, A. H. Paterson, G. King, I. Bancroft, B. Chalhoub and A. G. Sharpe (2014). "Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea." Genome Biol **15**(6): R77.

Paterson, A. H., J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, A. Poliakov, J. Schmutz, M. Spannagl, H. Tang, X. Wang, T. Wicker, A. K. Bharti, J. Chapman, F. A. Feltus, U. Gowik, I. V. Grigoriev, E. Lyons, C. A. Maher, M. Martis, A. Narechania, R. P. Otillar, B. W. Penning, A. A. Salamov, Y. Wang, L. Zhang, N. C. Carpita, M. Freeling, A. R. Gingle, C. T. Hash, B. Keller, P. Klein, S. Kresovich, M. C. McCann, R. Ming, D. G. Peterson, R. Mehboob ur, D. Ware, P. Westhoff, K. F. Mayer, J. Messing and D.

S. Rokhsar (2009). "The Sorghum bicolor genome and the diversification of grasses." Nature **457**(7229): 551-556.

Paterson, A. H., J. F. Wendel, H. Gundlach, H. Guo, J. Jenkins, D. Jin, D. Llewellyn, K. C. Showmaker, S. Shu, J. Udall, M. J. Yoo, R. Byers, W. Chen, A. Doron-Faigenboim, M. V. Duke, L. Gong, J. Grimwood, C. Grover, K. Grupp, G. Hu, T. H. Lee, J. Li, L. Lin, T. Liu, B. S. Marler, J. T. Page, A. W. Roberts, E. Romanel, W. S. Sanders, E. Szadkowski, X. Tan, H. Tang, C. Xu, J. Wang, Z. Wang, D. Zhang, L. Zhang, H. Ashrafi, F. Bedon, J. E. Bowers, C. L. Brubaker, P. W. Chee, S. Das, A. R. Gingle, C. H. Haigler, D. Harker, L. V. Hoffmann, R. Hovav, D. C. Jones, C. Lemke, S. Mansoor, M. ur Rahman, L. N. Rainville, A. Rambani, U. K. Reddy, J. K. Rong, Y. Saranga, B. E. Scheffler, J. A. Scheffler, D. M. Stelly, B. A. Triplett, A. Van Deynze, M. F. Vaslin, V. N. Waghmare, S. A. Walford, R. J. Wright, E. A. Zaki, T. Zhang, E. S. Dennis, K. F. Mayer, D. G. Peterson, D. S. Rokhsar, X. Wang and J. Schmutz (2012). "Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres." Nature **492**(7429): 423-427.

Perazzolli, M., M. Moretto, P. Fontana, A. Ferrarini, R. Velasco, C. Moser, M. Delledonne and I. Pertot (2012). "Downy mildew resistance induced by Trichoderma harzianum T39 in susceptible grapevines partially mimics transcriptional changes of resistant genotypes." BMC Genomics **13**: 660.

Prochnik, S., P. R. Marri, B. Desany, P. D. Rabinowicz, C. Kodira, M. Mohiuddin, F. Rodriguez, C. Fauquet, J. Tohme, T. Harkins, D. S. Rokhsar and S. Rounsley (2012). "The Cassava Genome: Current Progress, Future Directions." Trop Plant Biol **5**(1): 88-94.

Prufer, K., U. Stenzel, M. Dannemann, R. E. Green, M. Lachmann and J. Kelso (2008). "PatMaN: rapid alignment of short sequences to large databases." Bioinformatics **24**(13): 1530-1531.

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.

Regulski, M., Z. Lu, J. Kendall, M. T. Donoghue, J. Reinders, V. Llaca, S. Deschamps, A. Smith, D. Levy, W. R. McCombie, S. Tingey, A. Rafalski, J. Hicks, D. Ware and R. A. Martienssen (2013). "The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA." Genome Res **23**(10): 1651-1662.

Revell, L. J. (2012). "phytools: an R package for phylogenetic comparative biology (and other things)." Methods in Ecology and Evolution **3**(2): 217-223.

Roark, L. M., A. Y. Hui, L. Donnelly, J. A. Birchler and K. J. Newton (2010). "Recent and frequent insertions of chloroplast DNA into maize nuclear chromosomes." Cytogenet Genome Res **129**(1-3): 17-23.

Sato, S., Y. Nakamura, T. Kaneko, E. Asamizu, T. Kato, M. Nakao, S. Sasamoto, A. Watanabe, A. Ono, K. Kawashima, T. Fujishiro, M. Katoh, M. Kohara, Y. Kishida, C. Minami, S. Nakayama, N. Nakazaki, Y. Shimizu, S. Shinpo, C. Takahashi, T. Wada, M. Yamada, N. Ohmido, M. Hayashi, K. Fukui, T. Baba, T. Nakamichi, H. Mori and S. Tabata (2008). "Genome structure of the legume, Lotus japonicus." DNA Res **15**(4): 227-239.

Schmitz, R. J., Y. He, O. Valdes-Lopez, S. M. Khan, T. Joshi, M. A. Urich, J. R. Nery, B. Diers, D. Xu, G. Stacey and J. R. Ecker (2013). "Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population." Genome Res **23**(10): 1663-1674.

Schmitz, R. J., M. D. Schultz, M. G. Lewsey, R. C. O'Malley, M. A. Urich, O. Libiger, N. J. Schork and J. R. Ecker (2011). "Transgenerational epigenetic instability is a source of novel methylation variants." Science **334**(6054): 369-373.

Schmitz, R. J., M. D. Schultz, M. A. Urich, J. R. Nery, M. Pelizzola, O. Libiger, A. Alix, R. B. McCosh, H. Chen, N. J. Schork and J. R. Ecker (2013). "Patterns of population epigenomic diversity." Nature **495**(7440): 193-198.

Schmitz, R. J. and X. Zhang (2014). "Decoding the Epigenomes of Herbaceous Plants." **69**: 247-277.

Schmutz, J., S. B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, D. L. Hyten, Q. Song, J. J. Thelen, J. Cheng, D. Xu, U. Hellsten, G. D. May, Y. Yu, T. Sakurai, T. Umezawa, M. K. Bhattacharyya, D. Sandhu, B. Valliyodan, E. Lindquist, M. Peto, D. Grant, S. Shu, D. Goodstein, K. Barry, M. Futrell-Griggs, B. Abernathy, J. Du, Z. Tian, L. Zhu, N. Gill, T. Joshi, M. Libault, A. Sethuraman, X. C. Zhang, K. Shinozaki, H. T. Nguyen, R. A. Wing, P. Cregan, J. Specht, J. Grimwood, D. Rokhsar, G. Stacey, R. C. Shoemaker and S. A. Jackson (2010). "Genome sequence of the palaeopolyploid soybean." Nature **463**(7278): 178-183.

Schmutz, J., P. E. McClean, S. Mamidi, G. A. Wu, S. B. Cannon, J. Grimwood, J. Jenkins, S. Shu, Q. Song, C. Chavarro, M. Torres-Torres, V. Geffroy, S. M. Moghaddam, D. Gao, B. Abernathy, K. Barry, M. Blair, M. A. Brick, M. Chovatia, P. Gepts, D. M. Goodstein, M. Gonzales, U. Hellsten, D. L. Hyten, G. Jia, J. D. Kelly, D. Kudrna, R. Lee, M. M. Richard, P. N. Miklas, J. M. Osorno, J. Rodrigues, V. Thareau, C. A. Urrea, M. Wang, Y. Yu, M. Zhang, R. A. Wing, P. B. Cregan, D. S. Rokhsar and S. A. Jackson (2014). "A reference genome for common bean and genome-wide analysis of dual domestications." Nat Genet **46**(7): 707-713.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, K. Ying, C. T. Yeh, S. J. Emrich, Y. Jia, A. Kalyanaraman, A. P. Hsia, W. B. Barbazuk, R. S. Baucom, T. P. Brutnell, N. C. Carpita, C. Chaparro, J. M. Chia, J. M. Deragon, J. C. Estill, Y. Fu, J. A. Jeddeloh, Y. Han, H. Lee, P. Li, D. R. Lisch, S. Liu, Z. Liu, D. H. Nagel, M. C. McCann, P. SanMiguel, A. M. Myers, D. Nettleton, J. Nguyen, B. W. Penning, L. Ponnala, K. L. Schneider, D. C. Schwartz, A. Sharma, C. Soderlund, N. M. Springer, Q. Sun, H. Wang, M. Waterman, R. Westerman, T. K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J. L. Bennetzen, R. K. Dawe, J. Jiang, N. Jiang, G. G. Presting, S. R. Wessler, S. Aluru, R. A. Martienssen, S. W. Clifton, W. R. McCombie, R. A. Wing and R. K. Wilson (2009). "The B73 maize genome: complexity, diversity, and dynamics." Science **326**(5956): 1112-1115.

Schultz, M. D., Y. He, J. W. Whitaker, M. Hariharan, E. A. Mukamel, D. Leung, N. Rajagopal, J. R. Nery, M. A. Urich, H. Chen, S. Lin, Y. Lin, I. Jung, A. D. Schmitt, S. Selvaraj, B. Ren, T. J. Sejnowski, W. Wang and J. R. Ecker (2015). "Human body epigenome maps reveal noncanonical DNA methylation variation." Nature.

Schultz, M. D., R. J. Schmitz and J. R. Ecker (2012). "'Leveling' the playing field for analyses of single-base resolution DNA methylomes." Trends Genet **28**(12): 583-585.

Seymour, D. K., D. Koenig, J. Hagmann, C. Becker and D. Weigel (2014). "Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization." PLoS Genet **10**(11): e1004785.

Shulaev, V., D. J. Sargent, R. N. Crowhurst, T. C. Mockler, O. Folkerts, A. L. Delcher, P. Jaiswal, K. Mockaitis, A. Liston, S. P. Mane, P. Burns, T. M. Davis, J. P. Slovin, N. Bassil, R. P. Hellens, C. Evans, T. Harkins, C. Kodira, B. Desany, O. R. Crasta, R. V. Jensen, A. C. Allan, T. P. Michael, J. C. Setubal, J. M. Celton, D. J. Rees, K. P. Williams, S. H. Holt, J. J. Ruiz Rojas, M. Chatterjee, B. Liu, H. Silva, L. Meisel, A. Adato, S. A. Filichkin, M. Troggio, R. Viola, T. L. Ashman, H. Wang, P. Dharmawardhana, J. Elser, R. Raja, H. D. Priest, D. W. Bryant, Jr., S. E. Fox, S. A. Givan, L. J. Wilhelm, S. Naithani, A. Christoffels, D. Y. Salama, J. Carter, E. Lopez Girona, A. Zdepski, W. Wang, R. A. Kerstetter, W. Schwab, S. S. Korban, J. Davik, A. Monfort, B. Denoyes-Rothan, P. Arus, R. Mittler, B. Flinn, A. Aharoni, J. L. Bennetzen, S. L. Salzberg, A. W. Dickerman, R. Velasco, M. Borodovsky, R. E. Veilleux and K. M. Folta (2011). "The genome of woodland strawberry (Fragaria vesca)." Nat Genet **43**(2): 109-116.

Silveira, A. B., C. Trontin, S. Cortijo, J. Barau, L. E. Del Bem, O. Loudet, V. Colot and M. Vincentz (2013). "Extensive natural epigenetic variation at a de novo originated gene." PLoS Genet **9**(4): e1003437.

Slater, G. S. and E. Birney (2005). "Automated generation of heuristics for biological sequence comparison." BMC Bioinformatics **6**: 31.

Slotkin, R. K., M. Vaughn, F. Borges, M. Tanurdzic, J. D. Becker, J. A. Feijo and R. A. Martienssen (2009). "Epigenetic reprogramming and small RNA silencing of transposable elements in pollen." Cell **136**(3): 461-472.

Slotte, T., K. M. Hazzouri, J. A. Agren, D. Koenig, F. Maumus, Y. L. Guo, K. Steige, A. E. Platts, J. S. Escobar, L. K. Newman, W. Wang, T. Mandakova, E. Vello, L. M. Smith, S. R. Henz, J. Steffen, S. Takuno, Y. Brandvain, G. Coop, P. Andolfatto, T. T. Hu, M. Blanchette, R. M. Clark, H. Quesneville, M. Nordborg, B. S. Gaut, M. A. Lysak, J. Jenkins, J. Grimwood, J. Chapman, S. Prochnik, S. Shu, D. Rokhsar, J. Schmutz, D. Weigel and S. I. Wright (2013). "The Capsella rubella genome and the genomic consequences of rapid mating system evolution." Nat Genet **45**(7): 831-835.

Stegemann, S., S. Hartmann, S. Ruf and R. Bock (2003). "High-frequency gene transfer from the chloroplast genome to the nucleus." Proc Natl Acad Sci U S A **100**(15): 8828-8833.

Stocks, M. B., S. Moxon, D. Mapleson, H. C. Woolfenden, I. Mohorianu, L. Folkes, F. Schwach, T. Dalmay and V. Moulton (2012). "The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets." Bioinformatics **28**(15): 2059-2061.

Stroud, H., B. Ding, S. A. Simon, S. Feng, M. Bellizzi, M. Pellegrini, G. L. Wang, B. C. Meyers and S. E. Jacobsen (2013). "Plants regenerated from tissue culture contain stable epigenome changes in rice." Elife **2**: e00354.

Stroud, H., T. Do, J. Du, X. Zhong, S. Feng, L. Johnson, D. J. Patel and S. E. Jacobsen (2014). "Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis." Nat Struct Mol Biol **21**(1): 64-72.

Stroud, H., M. V. Greenberg, S. Feng, Y. V. Bernatavichute and S. E. Jacobsen (2013). "Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome." Cell **152**(1-2): 352-364.

Takuno, S. and B. S. Gaut (2012). "Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly." Mol Biol Evol **29**(1): 219-227.

Takuno, S. and B. S. Gaut (2013). "Gene body methylation is conserved between plant orthologs and is of evolutionary consequence." Proc Natl Acad Sci U S A **110**(5): 1797-1802.

Takuno, S., J.-H. Ran and B. S. Gaut (2016). "Evolutionary patterns of genic DNA methylation vary across land plants." Nature Plants **2**(2): 15222.

Tang, S., Y. Dong, D. Liang, Z. Zhang, C.-Y. Ye, P. Shuai, X. Han, Y. Zhao, W. Yin and X. Xia (2014). "Analysis of the Drought Stress-Responsive Transcriptome of Black Cottonwood (Populus trichocarpa) Using Deep RNA Sequencing." Plant Molecular Biology Reporter **33**(3): 424-438.

Thompson, A. J., M. Tor, C. S. Barry, J. Vrebalov, C. Orfila, M. C. Jarvis, J. J. Giovannoni, D. Grierson and G. B. Seymour (1999). "Molecular and genetic characterization of a novel pleiotropic tomato-ripening mutant." Plant Physiol **120**(2): 383-390.

Tomato Genome, C. (2012). "The tomato genome sequence provides insights into fleshy fruit evolution." Nature **485**(7400): 635-641.

Tran, R. K., J. G. Henikoff, D. Zilberman, R. F. Ditt, S. E. Jacobsen and S. Henikoff (2005). "DNA methylation profiling identifies CG methylation clusters in Arabidopsis genes." Curr Biol **15**(2): 154-159.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn and L. Pachter (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." Nat Protoc **7**(3): 562-578.

Tuskan, G. A., S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, L. Sterck, A. Aerts, R. R. Bhalerao, R. P. Bhalerao, D. Blaudez, W. Boerjan, A. Brun, A. Brunner, V. Busov, M. Campbell, J. Carlson, M. Chalot, J. Chapman, G. L. Chen, D. Cooper, P. M. Coutinho, J. Couturier, S. Covert, Q. Cronk, R. Cunningham, J. Davis, S. Degroeve, A. Dejardin, C. Depamphilis, J. Detter, B. Dirks, I. Dubchak, S. Duplessis, J. Ehlting, B. Ellis, K. Gendler, D. Goodstein, M. Gribskov, J. Grimwood, A. Groover, L. Gunter, B. Hamberger, B. Heinze, Y. Helariutta, B. Henrissat, D. Holligan, R. Holt, W. Huang, N. Islam-Faridi, S. Jones, M. Jones-Rhoades, R. Jorgensen, C. Joshi, J. Kangasjarvi, J. Karlsson, C. Kelleher, R. Kirkpatrick, M. Kirst, A. Kohler, U. Kalluri, F. Larimer, J. Leebens-Mack, J. C. Leple, P. Locascio, Y. Lou, S. Lucas, F. Martin, B. Montanini, C. Napoli, D. R. Nelson, C. Nelson, K. Nieminen, O. Nilsson, V. Pereda, G. Peter, R. Philippe, G. Pilate, A. Poliakov, J. Razumovskaya, P. Richardson, C. Rinaldi, K. Ritland, P. Rouze, D. Ryaboy, J. Schmutz, J. Schrader, B. Segerman, H. Shin, A. Siddiqui, F. Sterky, A. Terry, C. J. Tsai, E. Uberbacher, P. Unneberg, J. Vahala, K. Wall, S. Wessler, G. Yang, T. Yin, C. Douglas, M.

Marra, G. Sandberg, Y. Van de Peer and D. Rokhsar (2006). "The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)." Science **313**(5793): 1596-1604.

Urich, M. A., J. R. Nery, R. Lister, R. J. Schmitz and J. R. Ecker (2015). "MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing." Nat Protoc **10**(3): 475-483.

van Bakel, H., J. M. Stout, A. G. Cote, C. M. Tallon, A. G. Sharpe, T. R. Hughes and J. E. Page (2011). "The draft genome and transcriptome of Cannabis sativa." Genome Biol **12**(10): R102.

Vaughn, M. W., M. Tanurdzic, Z. Lippman, H. Jiang, R. Carrasquillo, P. D. Rabinowicz, N. Dedhia, W. R. McCombie, N. Agier, A. Bulski, V. Colot, R. W. Doerge and R. A. Martienssen (2007). "Epigenetic natural variation in Arabidopsis thaliana." PLoS Biol **5**(7): e174.

Velasco, R., A. Zharkikh, J. Affourtit, A. Dhingra, A. Cestaro, A. Kalyanaraman, P. Fontana, S. K. Bhatnagar, M. Troggio, D. Pruss, S. Salvi, M. Pindo, P. Baldi, S. Castelletti, M. Cavaiuolo, G. Coppola, F. Costa, V. Cova, A. Dal Ri, V. Goremykin, M. Komjanc, S. Longhi, P. Magnago, G. Malacarne, M. Malnoy, D. Micheletti, M. Moretto, M. Perazzolli, A. Si-Ammour, S. Vezzulli, E. Zini, G. Eldredge, L. M. Fitzgerald, N. Gutin, J. Lanchbury, T. Macalma, J. T. Mitchell, J. Reid, B. Wardell, C. Kodira, Z. Chen, B. Desany, F. Niazi, M. Palmer, T. Koepke, D. Jiwan, S. Schaeffer, V. Krishnan, C. Wu, V. T. Chu, S. T. King, J. Vick, Q. Tao, A. Mraz, A. Stormo, K. Stormo, R. Bogden, D. Ederle, A. Stella, A. Vecchietti, M. M. Kater, S. Masiero, P. Lasserre, Y. Lespinasse, A. C. Allan, V. Bus, D. Chagne, R. N. Crowhurst, A. P. Gleave, E. Lavezzo, J. A. Fawcett, S. Proost, P. Rouze, L. Sterck, S. Toppo, B. Lazzari, R. P. Hellens, C. E. Durel, A. Gutin, R. E. Bumgarner, S. E. Gardiner, M. Skolnick, M. Egholm, Y. Van de Peer, F. Salamini and R. Viola (2010). "The genome of the domesticated apple (Malus x domestica Borkh.)." Nat Genet **42**(10): 833-839.

Wang, M., D. Yuan, L. Tu, W. Gao, Y. He, H. Hu, P. Wang, N. Liu, K. Lindsey and X. Zhang (2015). "Long noncoding RNAs and their proposed functions in fibre development of cotton (Gossypium spp.)." New Phytol **207**(4): 1181-1197.

West, P. T., Q. Li, L. Ji, S. R. Eichten, J. Song, M. W. Vaughn, R. J. Schmitz and N. M. Springer (2014). "Genomic distribution of H3K9me2 and DNA methylation in a maize genome." PLoS One **9**(8): e105267.

Willing, E.-M., V. Rawat, T. Mandáková, F. Maumus, G. V. James, K. J. V. Nordström, C. Becker, N. Warthmann, C. Chica, B. Szarzynska, M. Zytnicki, M. C. Albani, C. Kiefer, S. Bergonzi, L. Castaings, J. L. Mateos, M. C. Berns, N. Bujdoso, T. Piofczyk, L. de Lorenzo, C. Barrero-Sicilia, I. Mateos, M. Piednoël, J. Hagmann, R. Chen-Min-Tao, R. Iglesias-Fernández, S. C. Schuster, C. Alonso-Blanco, F. Roudier, P. Carbonero, J. Paz-Ares, S. J. Davis, A. Pecinka, H. Quesneville, V. Colot, M. A. Lysak, D. Weigel, G. Coupland and K. Schneeberger (2015). "Genome expansion of Arabis alpina linked with retrotransposition and reduced symmetric DNA methylation." Nature Plants **1**(2): 14023.

Wu, G. A., S. Prochnik, J. Jenkins, J. Salse, U. Hellsten, F. Murat, X. Perrier, M. Ruiz, S. Scalabrin, J. Terol, M. A. Takita, K. Labadie, J. Poulain, A. Couloux, K. Jabbari, F. Cattonaro, C. Del Fabbro, S. Pinosio, A. Zuccolo, J. Chapman, J. Grimwood, F. R. Tadeo, L. H. Estornell, J. V. Munoz-Sanz, V. Ibanez, A. Herrero-Ortega, P. Aleza, J. Perez-Perez, D. Ramon, D. Brunel, F. Luro, C. Chen, W. G. Farmerie, B. Desany, C. Kodira, M. Mohiuddin, T. Harkins, K. Fredrikson, P. Burns, A. Lomsadze, M. Borodovsky, G. Reforgiato, J. Freitas-Astua, F. Quetier, L. Navarro, M. Roose, P. Wincker, J. Schmutz, M. Morgante, M. A. Machado, M. Talon, O. Jaillon, P. Ollitrault, F. Gmitter and D. Rokhsar (2014). "Sequencing of diverse mandarin, pummelo and

orange genomes reveals complex history of admixture during citrus domestication." Nat Biotechnol **32**(7): 656-662.

Yang, R., D. E. Jarvis, H. Chen, M. A. Beilstein, J. Grimwood, J. Jenkins, S. Shu, S. Prochnik, M. Xin, C. Ma, J. Schmutz, R. A. Wing, T. Mitchell-Olds, K. S. Schumaker and X. Wang (2013). "The Reference Genome of the Halophytic Plant Eutrema salsugineum." Front Plant Sci **4**: 46.

Yang, Z. (1997). "PAML: a program package for phylogenetic analysis by maximum likelihood." Bioinformatics **13**(5): 555-556.

Yang, Z. and R. Nielsen (2000). "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models." Mol Biol Evol **17**(1): 32-43.

Young, N. D., F. Debelle, G. E. Oldroyd, R. Geurts, S. B. Cannon, M. K. Udvardi, V. A. Benedito, K. F. Mayer, J. Gouzy, H. Schoof, Y. Van de Peer, S. Proost, D. R. Cook, B. C. Meyers, M. Spannagl, F. Cheung, S. De Mita, V. Krishnakumar, H. Gundlach, S. Zhou, J. Mudge, A. K. Bharti, J. D. Murray, M. A. Naoumkina, B. Rosen, K. A. Silverstein, H. Tang, S. Rombauts, P. X. Zhao, P. Zhou, V. Barbe, P. Bardou, M. Bechner, A. Bellec, A. Berger, H. Berges, S. Bidwell, T. Bisseling, N. Choisne, A. Couloux, R. Denny, S. Deshpande, X. Dai, J. J. Doyle, A. M. Dudez, A. D. Farmer, S. Fouteau, C. Franken, C. Gibelin, J. Gish, S. Goldstein, A. J. Gonzalez, P. J. Green, A. Hallab, M. Hartog, A. Hua, S. J. Humphray, D. H. Jeong, Y. Jing, A. Jocker, S. M. Kenton, D. J. Kim, K. Klee, H. Lai, C. Lang, S. Lin, S. L. Macmil, G. Magdelenat, L. Matthews, J. McCorrison, E. L. Monaghan, J. H. Mun, F. Z. Najar, C. Nicholson, C. Noirot, M. O'Bleness, C. R. Paule, J. Poulain, F. Prion, B. Qin, C. Qu, E. F. Retzel, C. Riddle, E. Sallet, S. Samain, N. Samson, I. Sanders, O. Saurat, C. Scarpelli, T. Schiex, B. Segurens, A. J. Severin, D. J. Sherrier, R. Shi, S. Sims, S. R. Singer, S. Sinharoy, L. Sterck, A. Viollet, B. B. Wang, K. Wang, M. Wang,

X. Wang, J. Warfsmann, J. Weissenbach, D. D. White, J. D. White, G. B. Wiley, P. Wincker, Y. Xing, L. Yang, Z. Yao, F. Ying, J. Zhai, L. Zhou, A. Zuber, J. Denarie, R. A. Dixon, G. D. May, D. C. Schwartz, J. Rogers, F. Quetier, C. D. Town and B. A. Roe (2011). "The Medicago genome provides insight into the evolution of rhizobial symbioses." Nature **480**(7378): 520-524.

Zemach, A., M. Y. Kim, P. H. Hsieh, D. Coleman-Derr, L. Eshed-Williams, K. Thao, S. L. Harmer and D. Zilberman (2013). "The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin." Cell **153**(1): 193-205.

Zemach, A., Y. Li, B. Wayburn, H. Ben-Meir, V. Kiss, Y. Avivi, V. Kalchenko, S. E. Jacobsen and G. Grafi (2005). "DDM1 binds Arabidopsis methyl-CpG binding domain proteins and affects their subnuclear localization." Plant Cell **17**(5): 1549-1558.

Zemach, A., I. E. McDaniel, P. Silva and D. Zilberman (2010). "Genome-wide evolutionary analysis of eukaryotic DNA methylation." Science **328**(5980): 916-919.

Zhang, X., J. Yazaki, A. Sundaresan, S. Cokus, S. W. Chan, H. Chen, I. R. Henderson, P. Shinn, M. Pellegrini, S. E. Jacobsen and J. R. Ecker (2006). "Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis." Cell **126**(6): 1189-1201.

Zhong, S., Z. Fei, Y. R. Chen, Y. Zheng, M. Huang, J. Vrebalov, R. McQuinn, N. Gapper, B. Liu, J. Xiang, Y. Shao and J. J. Giovannoni (2013). "Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening." Nat Biotechnol **31**(2): 154-159.

Zilberman, D., M. Gehring, R. K. Tran, T. Ballinger and S. Henikoff (2007). "Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription." Nat Genet **39**(1): 61-69.

**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.

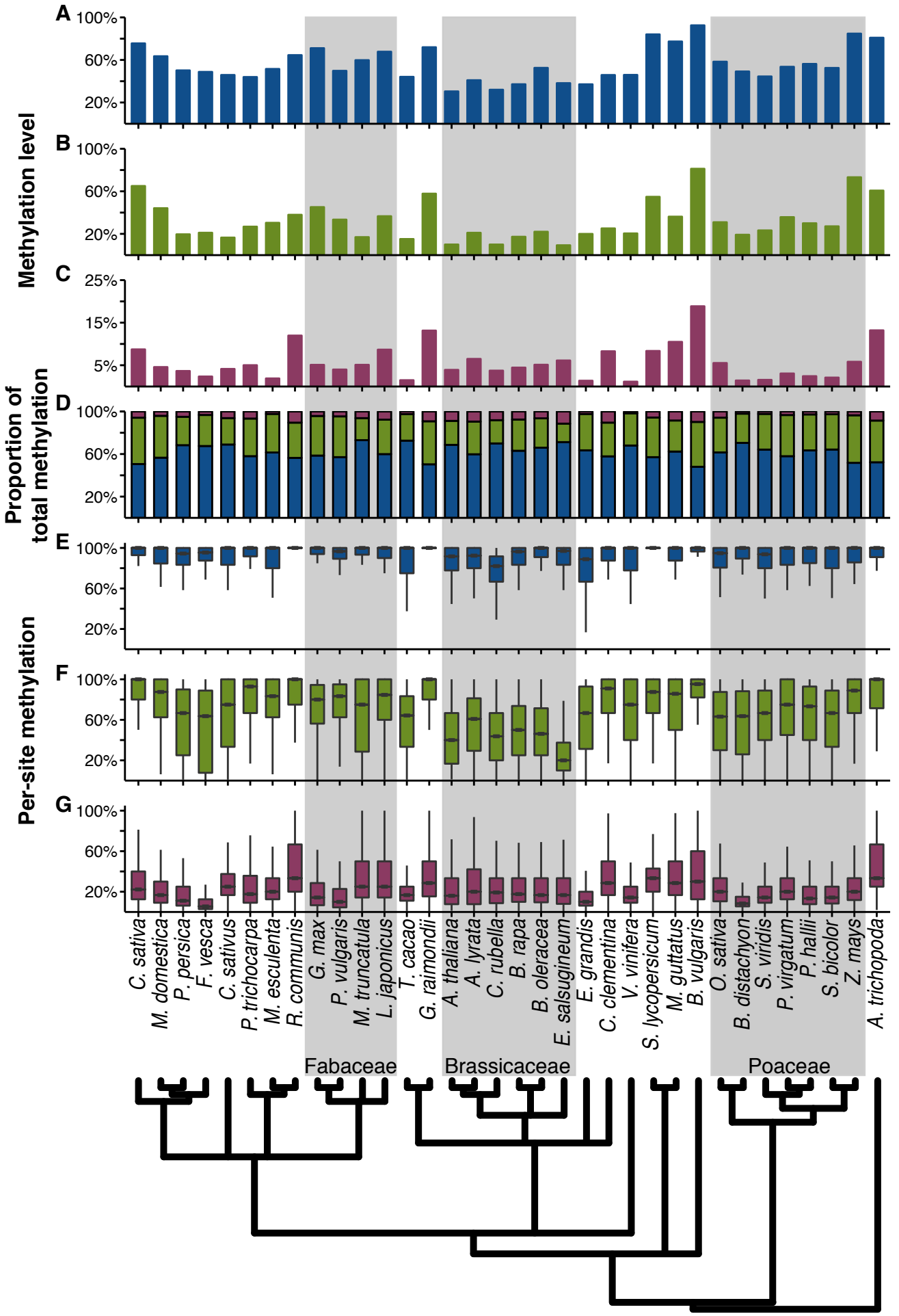

**Figure 2: (A)** Genome-wide methylation levels are correlated to genome size for mCG (blue) and mCHG (green), but not for mCHH (maroon) Significant relationships indicated.
**(B)** Coding region (CDS) methylation levels is not correlated to genome size for mCG (blue), but is for mCHG (green), and mCHH (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).

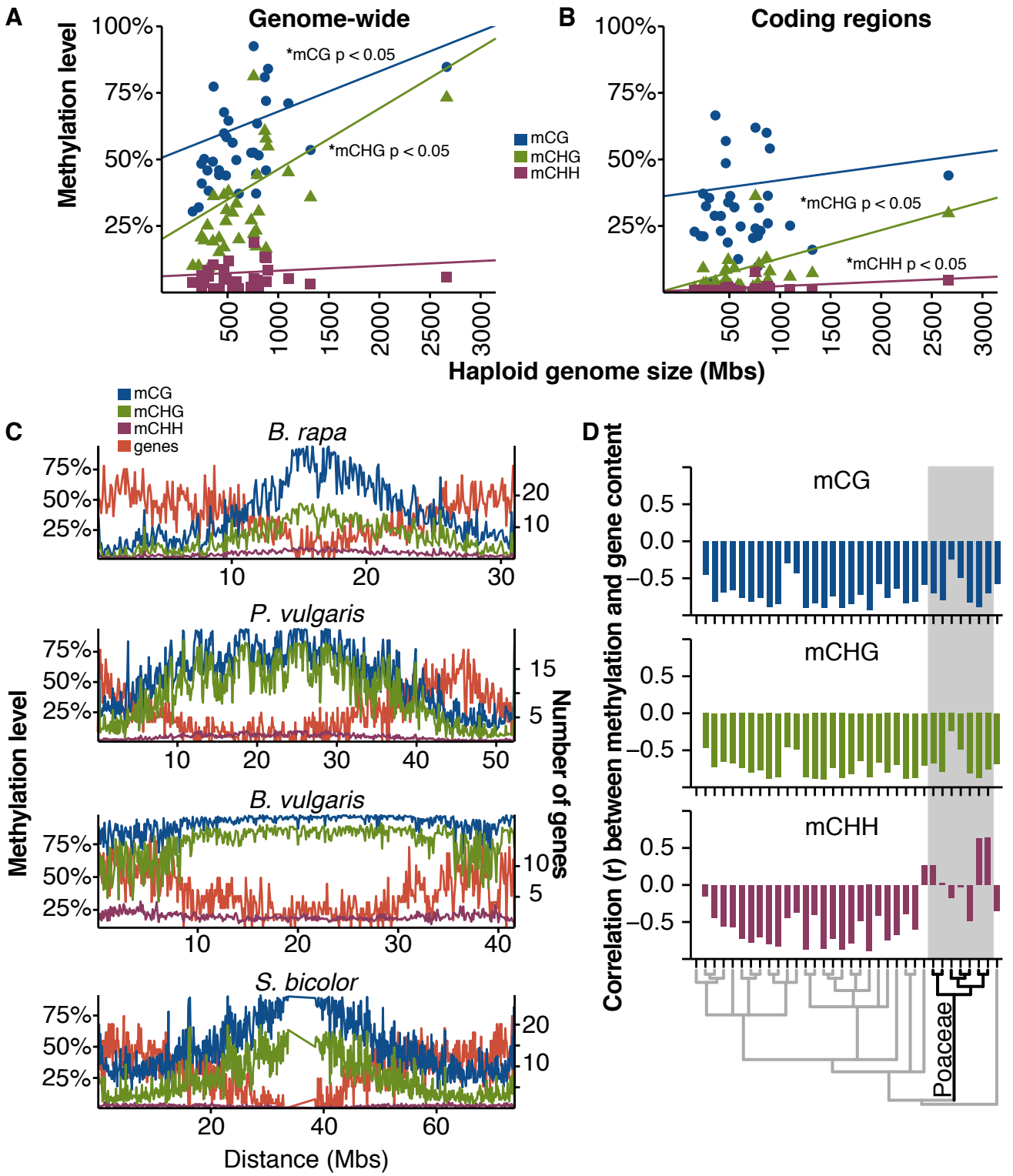

**Figure 3: (A)** Genome-wide methylation levels were correlated with repeat number for mCG (blue) and mCHG (green), but not for mCHH (maroon). Significant relationships indicated. **(B)** Distribution of methylation levels for repeats in each species. **(C)** Patterns of methylation upstream, across,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

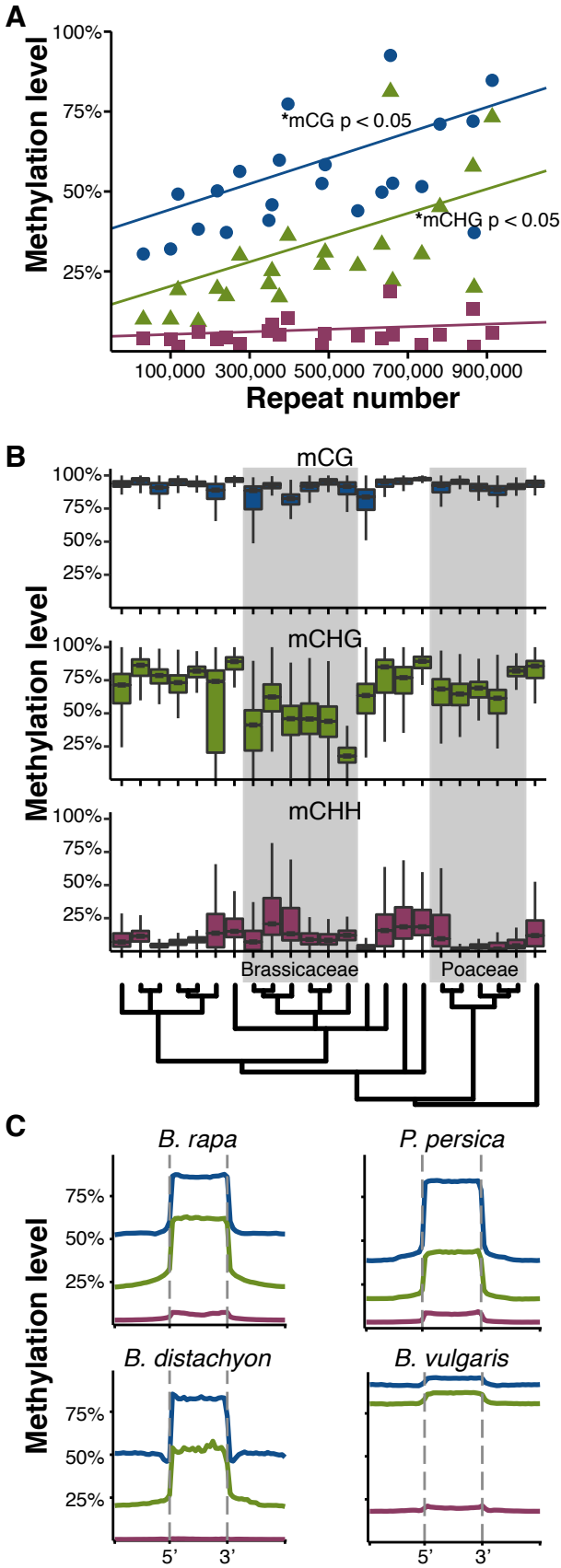

**Figure 4: (A)** Heatmap showing methylation state of orthologous genes (horizontal axis) to *A. thaliana* for each species (vertical axis). Species are organized according to phylogenetic relationship. **(B)** Percentage of genes in each species that are gbM (mCG only in coding sequences). The Brassicacea 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.

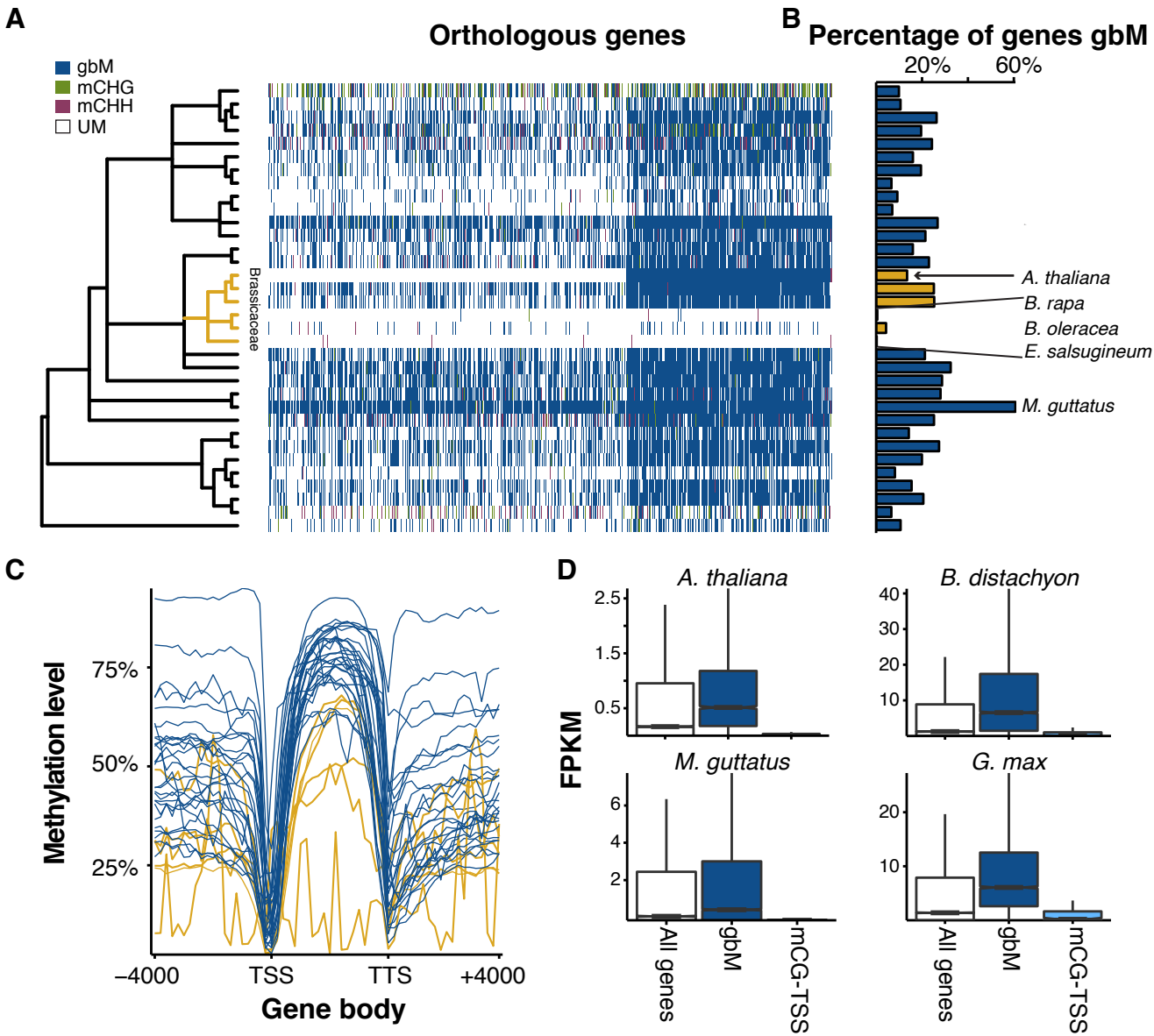

**Figure 5: (A)** Methylation levels for mCG (blue) (mCG only in coding sequences), mCHG (green) (mCG and mCHG in coding sequences), and mCHH (maroon) (mCG, mCHG, and mCHH in coding sequences) were plotted upstream, across, and downstream of mCHG and mCHH genes. **(B)** Gene expression of mCHG and mCHH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

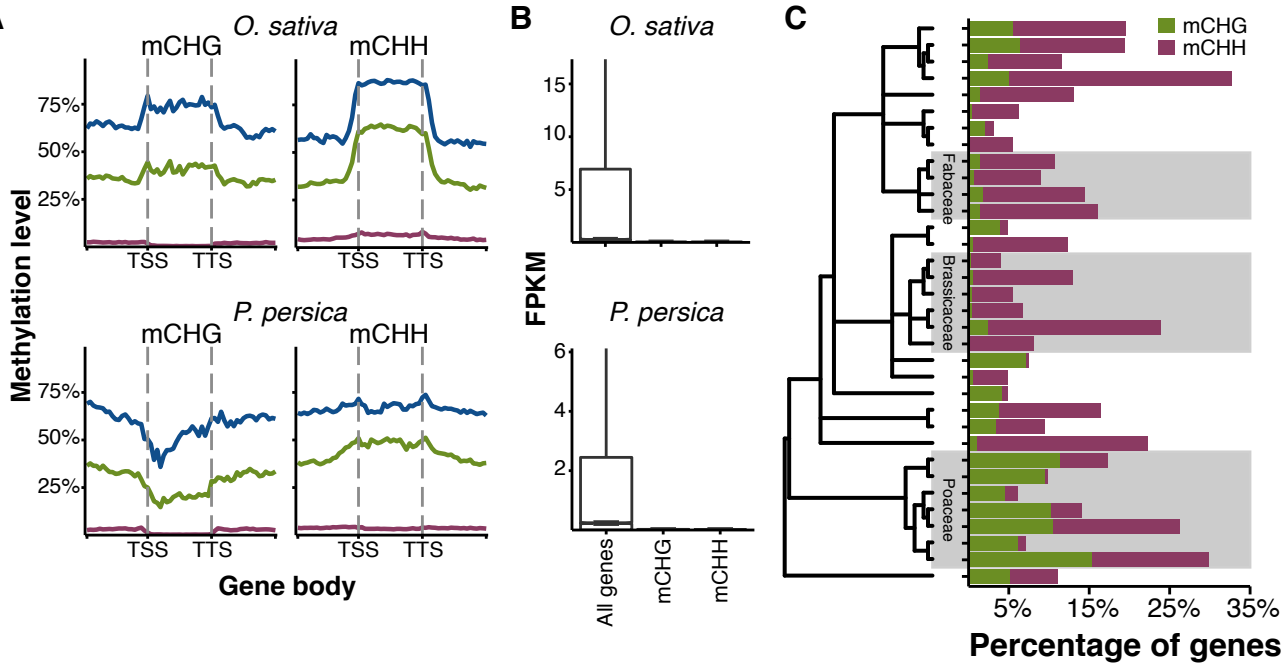
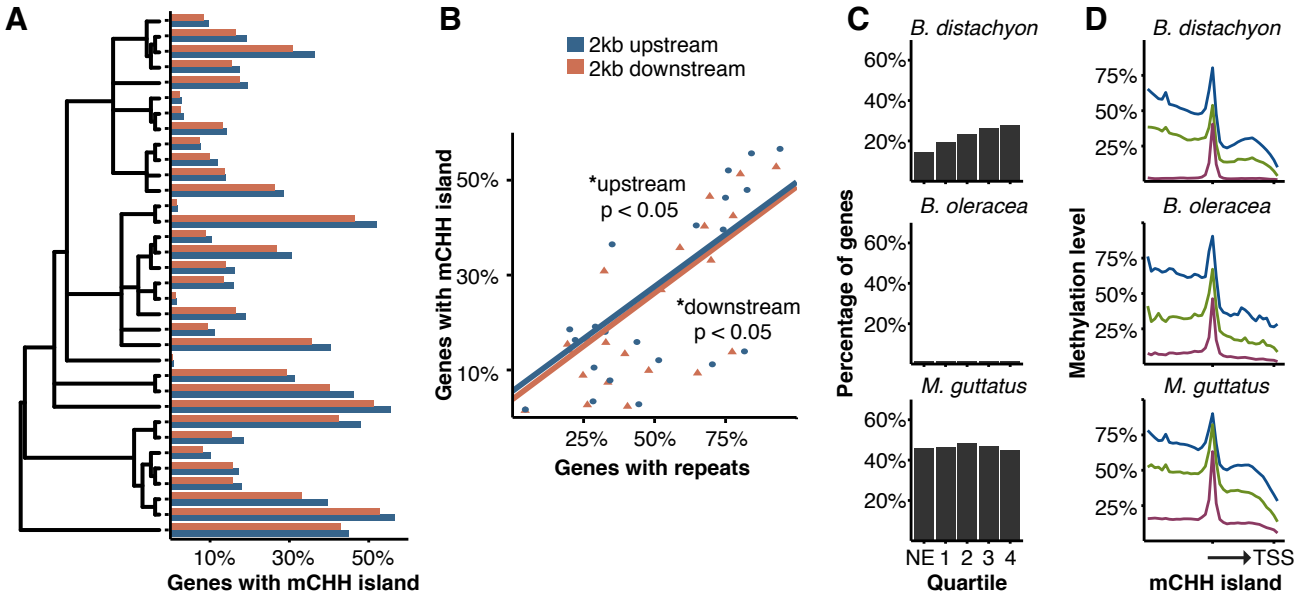
**Figure 6: (A)** Percentage of genes with mCHH islands 2kb upstream or downstream. **(B)** Upstream and downstream mCHH islands are correlated with upstream and downstream repeats (respectively). Significant relationships indicated. **(C)** Association of upstream mCHH islands with gene expression. 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.