Niederhuth et al. Genome Biology (2016) 17:194
DOI 10.1186/s13059-016-1059-0

Genome Biology

RESEARCH Open Access

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*]

Abstract

Background: DNA methylation is an important feature of plant epigenomes, involved in the formation of heterochromatin and affecting gene expression. Extensive variation of DNA methylation patterns within a species has been uncovered from studies of natural variation. However, the extent to which DNA methylation varies between flowering plant species is still unclear. To understand the variation in genomic patterning of DNA methylation across flowering plant species, we compared single base resolution DNA methylomes of 34 diverse angiosperm species.

Results: 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. Furthermore, low levels of CHH methylation are observed in a number of species, especially in clonally propagated species.

Conclusions: These results reveal the extent of variation in DNA methylation in angiosperms and show that DNA methylation patterns are broadly a reflection of the evolutionary and life histories of plant species.

Background

Biological diversity is established at multiple levels. Historically this has focused on studying the contribution of genetic variation. However, epigenetic variations manifested in the form of DNA methylation [1–3], histones and histone modifications [4], which together make up the epigenome, might also contribute to biological diversity. These components are integral to proper regulation of many aspects of the genome; including chromatin structure, transposon silencing, regulation of gene expression, and recombination [5–8]. Significant amounts of epigenomic diversity are explained by genetic variation [2, 3, 9–13], however, a large portion remains unexplained and in some cases these variants arise independently of genetic variation and are thus defined as "epigenetic" [2, 10–12, 14, 15]. Moreover, epigenetic variants can be heritable and also lead to phenotypic variation [16–19]. 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, when most plant genomes are much larger [20, 21]. 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 [22].

In plants, cytosine methylation occurs in three sequence contexts; CG, CHG, and CHH (H = A, T, or C), and are under control by distinct mechanisms [23].

* Correspondence: schmitz@uga.edu
†Equal contributors
[1]Department of Genetics, University of Georgia, 120 East Green Street, Athens, GA 30602, USA
Full list of author information is available at the end of the article

Methylation at CG (mCG) and CHG (mCHG) sites is typically symmetrical across the Watson and Crick strands [24]. mCG is maintained by methyltransferase 1 (MET1), which is recruited to hemi-methylated CG sites and methylates the opposing strand [25, 26], whereas mCHG is maintained by the plant specific chromomethylase 3 (CMT3) [27], and is strongly associated with dimethylation of lysine 9 on histone 3 (H3K9me2) [28]. The BAH and CHROMO domains of CMT3 bind to H3K9me2, leading to methylation of CHG sites [28]. In turn, the histone methyltransferases kryptonite (KYP), and Su(var)3-9 homologue 5 (SUVH5) and SUVH6 recognize methylated DNA and methylate H3K9 [29], leading to a self-reinforcing loop [30]. Asymmetrical methylation of CHH sites (mCHH) is established and maintained by another member of the CMT family, CMT2 [31, 32]. CMT2, like CMT3, also contains BAH and CHROMO domains and methylates CHH in H3K9me2 regions [31, 32]. Additionally, all three sequence contexts are methylated de novo via RNA-directed DNA methylation (RdDM) [33]. Short-interfering 24 nucleotide (nt) RNAs (siRNAs) guide the de novo methyltransferase domains rearranged methyltransferase 2 (DRM2) to target sites [34, 35]. The targets of CMT2 and RdDM are often complementary, as CMT2 in *A. thaliana* primarily methylate regions of deep heterochromatin, such as transposons bodies [31]. 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 [31, 32, 36] The mCHH islands in *Zea mays* are associated with upstream and downstream of more highly expressed genes where they might function to prevent transcription of neighboring transposons [36, 37]. 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 DNA methylation across entire genomes and within species [1, 3, 13, 15, 38]. Whole-genome bisulfite sequencing (WGBS) is particularly powerful, as it reveals genome-wide single nucleotide resolution of DNA methylation [39–41]. WGBS has been used to sequence an increasing number of plant methylomes, ranging from model plants like *A. thaliana* [39, 40] to economically important crops like *Z. mays* [2, 11, 36, 42]. This has enabled a new field of comparative epigenomics, which places DNA methylation within an evolutionary context [43–46]. The use of WGBS together with de novo transcript assemblies has provided an opportunity to monitor the changes in DNA methylation of gene bodies among species [47] but does not provide a full view of changes in the patterns of context-specific DNA methylation at different types of genomic regions [48].

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 lower mCHG levels and lower numbers or even losses of CG gene body methylation (gbM) when compared to other species. 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 from the leaves of 34 angiosperm species that have genome assemblies [49–52] (Additional file 1: Table S1). MethylC-seq [40, 53] was used to sequence 26 species and an additional eight species with previously published methylomes were downloaded and reanalyzed [12, 15, 36, 48, 54–56]. Different metrics were used to make comparisons at a whole-genome level. The genome-wide weighted DNA methylation level [57] 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 (Fig. 1a–c). The proportion that each DNA methylation context makes up of all DNA methylation indicates the predominance of specific DNA methylation pathways (Fig. 1d). The per-site DNA methylation level is the distribution of DNA methylation levels at individual methylated sites and indicates within a population of cells, the proportion that are methylated (Fig. 1e–g, Additional file 1: Figure S1). Symmetry is a comparison of per-site DNA methylation levels at cytosines on the Watson versus the Crick strand for the symmetrical CG and CHG contexts (Additional file 1: Figures S2 and S3). As CMT3 is responsible for maintaining the symmetrical DNA methylation of CHG sites [27], we can use *A. thaliana cmt3* mutants to establish thresholds with which to identify sites as symmetrical or asymmetrical and [58] quantify the asymmetry of mCHG sites (Additional file 1: Figure S4). Per-site DNA methylation and symmetry provide information into how well DNA methylation is maintained and how ubiquitously the sites are methylated across cell types within sequenced tissues [59].

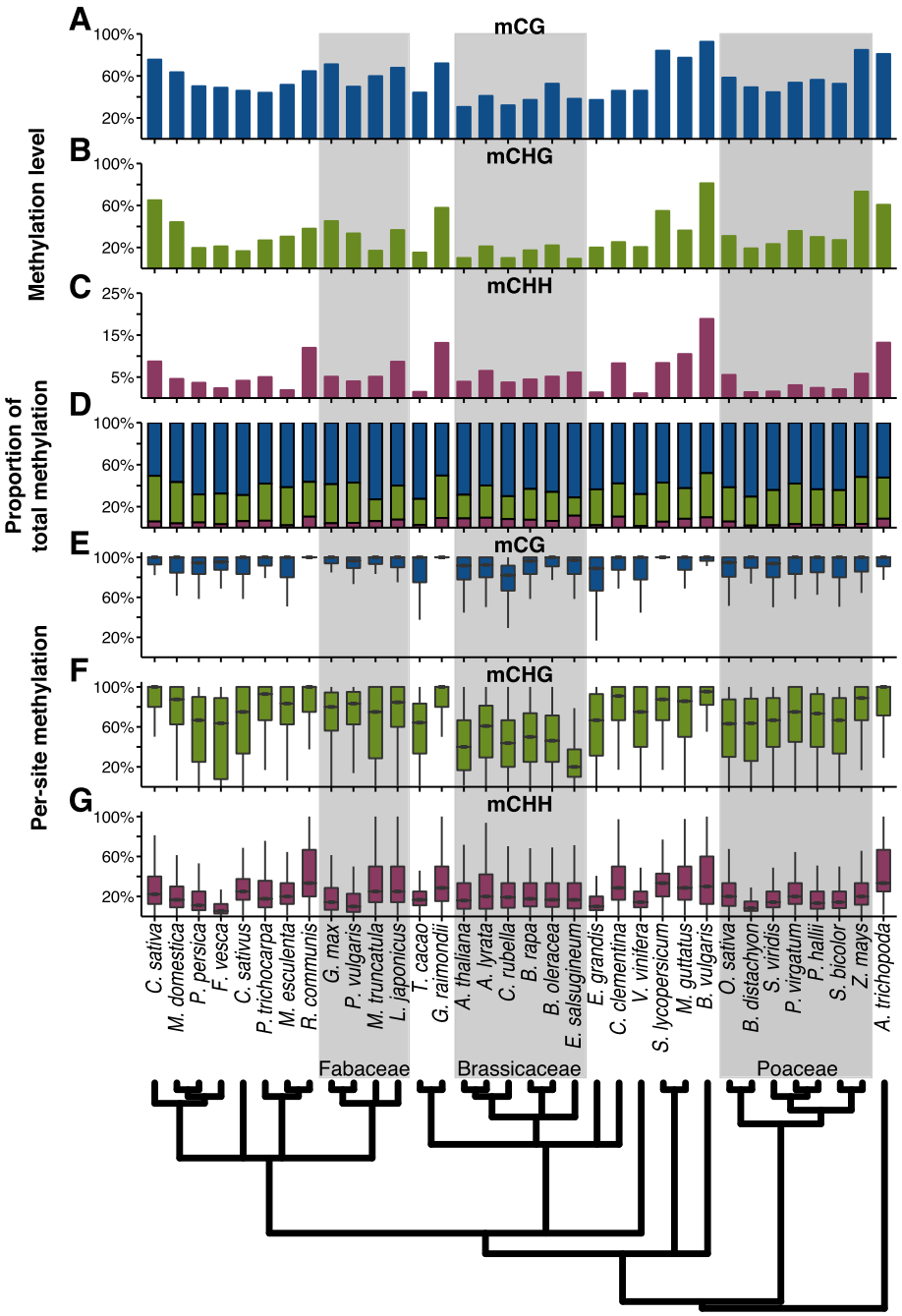

Fig. 1 Genome-wide methylation levels for **a** mCG, **b** mCHG, and **c** mCHH. **d** Using the genome-wide methylation levels, the proportion that each context contributes towards the total methylation (mC) was calculated. **e** The distribution of per-site methylation levels for mCG, **f** mCHG, and **g** mCHH. Species are organized according to their phylogenetic relationship

There is extensive variation between species. Within each species, mCG had the highest levels of DNA methylation genome-wide (Fig. 1a, Additional file 2: 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 approximately eight-fold between species, from only ~9.3 % in *Eutrema salsugineum* to ~81.2 % in *B. vulgaris* (Fig. 1b, Additional file 2: 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 (Fig. 1c, Additional file 2:

Table S2). The lowest mCHH level was found in *Vitis vinifera* with only ~1.1 % mCHH. mCG is the most predominant type of DNA methylation making up the largest proportion of the total DNA methylation in all examined species (Fig. 1d). *B. vulgaris* was a notable outlier, having the highest levels of DNA methylation in all contexts and having particularly high mCHH levels. The between-species variation observed was much greater than within species variation, when compared to *A. thaliana* accessions from the 1001 Epigenomes Project (Additional file 1: Figure S5) [60]. 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 and hypotheses regarding the evolution of DNA methylation pathways across flowering plants. In general, the Brassicaceae (mustard) family, which includes *A. thaliana*, has lower median levels of per-site mCHG methylation when compared to other species (Fig. 1f). Furthermore, symmetrical mCHG sites have a wider range of DNA methylation levels and increased asymmetry, whereas non-Brassicaceae species have very highly methylated symmetrical sites (Additional file 1: Figures S3 and S4b), suggesting that the CMT3 pathway is less effective in Brassicaceae genomes or that it operates in a cell-specific manner. This is further evidenced by *E. salsugineum*, with the lowest mCHG levels (Fig. 1b), which is a natural *cmt3* mutant, whereas *CMT3* is under relaxed selection in other Brassicaceae [61, 62]. Methylation of CG sites is also less well maintained in the Brassicaceae, with *Capsella rubella* showing the lower levels of per-site mCG methylation (Fig. 1e, Additional file 1: Figure S1).

Within the Fabaceae (legume) family, *Glycine max* and *Phaseolus vulgaris*, 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 (Fig. 1c and g). The Poaceae (grass) family, in general, have much lower levels of mCHH (~1.4–5.8 %), both in terms of total DNA 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 some of the lowest of all species, whereas others like *Oryza sativa* and *Z. mays* have levels comparable to *A. thaliana*. In *Z. mays*, CMT2 has been lost [31], 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*, *Eucalyptus grandis*. No causal factor based on examined genomic features or examined DNA methylation pathways was identified; however, these plants are commonly propagated via clonal methods [63]. Among non-Poaceae species, the six lowest mCHH levels were found in species with histories of clonal propagation (Additional file 1: Figure S6). Effects of micropropagation on DNA methylation in *M. esculenta* using DNA methylation-sensitive amplified polymorphisms have been observed before [64], so has altered expression of methyltransferases due to micropropagation in *Fragaria* x *ananassa* (common garden strawberry) [65]. If repeated rounds of clonal propagation were responsible for low mCHH, we hypothesized that going through a single round of sexual reproduction might result in increased mCHH levels, as work in *A. thaliana* suggests that mCHH is re-established during reproduction [66, 67]. To test this hypothesis, we examined a DNA methylome of a parental *M. esculenta* plant that had 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 micro-propagated 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 (Additional file 1: Figure S7). As both of these results are based on one generation of sexual reproduction, it may be that this is insufficient to fully restore DNA methylation or that clonal propagation is not causal for the low levels of mCHH observed. 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 in the copy number for repetitive elements. The majority of changes in genome size among the species we examined are 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) [68] which takes into account the phylogenetic relationship and non-independence of species as

more closely related species are more alike (Additional file 1: Table S3). Phylogenetic relationships were inferred from a species tree constructed using 50 single copy loci for use in PGLS (Additional file 1: Figure S8) [69]. A previous report had found a relationship between total methylation and genome size, but did not take into account the sequence context of that methylation [70]. Positive correlations were found between mCG and genome size (p value = 2.9×10^{-3}) and between mCHG and genome size (p value = 2.2×10^{-6}) (Fig. 2a), but no correlation was found with mCHH and genome size (Fig. 2a). This dataset was limited to one larger genome greater than 2 Gb, *Z. mays*, so we tested the effect that this had on the results. After removal of *Z. mays*, genome-wide mCHG methylation remained correlated with genome size, whereas mCG and mCHH showed no correlation (Additional file 1: Figure S9).

Similarly, a relationship between genic methylation level and genome size in plants has also been previously reported [47]. We found that within coding sequences (CDS) methylation levels were correlated with genome size for both mCHG (p value = 5×10^{-6}) and mCHH (p value = 1.4×10^{-5}), but in contrast found no correlation for mCG (p value > 0.18) (Fig. 2b). This prior study included many non-Angiosperm species and a limited set of Angiosperms and had also found that the correlation with mCG disappeared after removal of non-Angiosperm species [47]. Our observed correlations between CDS methylation and genome size were strongly driven by the large genome of *Z. mays*, and after its removal, no correlation was observed for any methylation context (Additional file 1: Figure S9). These results and those others [47, 70] suggest that the relationship between DNA methylation, both across the genome and within genes, and genome size is still not fully resolved and will require more extensive studies to resolve.

The highest levels of DNA methylation are typically found in centromeres and pericentromeric regions [39, 40, 48]. The distributions of DNA methylation at chromosomal levels were examined in 100 kb sliding windows (Fig. 2c, Additional file 1: Figure S10). The number of genes per window was used as a proxy to differentiate euchromatin and heterochromatin. Both mCG and mCHG have negative correlations between DNA methylation level and gene number, indicating that these two DNA methylation types are mostly found in gene-poor heterochromatic regions (Fig. 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 clade (Fig. 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 "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 (Additional file 3: Figure S11). For most species, regions of higher DNA methylation are often smaller in size, with regions of low or intermediate DNA methylation being larger (Additional file 3: Figure S12). More small RNAs, in particular 24 nt siRNAs map to regions of higher mCHH methylation (Additional file 3: Figure S13) and these regions of high 24 nt siRNAs tend to be smaller in size (Additional file 3: Figure S14). This may be because RdDM is primarily found on the edges of transposons whereas other mechanisms predominate in regions of deep heterochromatin [31]. Using these results, we can make inferences into the architecture of the methylome.

mCHG and mCHH regions are more variable in both size and DNA methylation levels than mCG regions, as little variability in mCG regions was found between species (Additional file 3: Figure S11). For mCHG regions, the Brassicaceae differed the most having lower DNA methylation levels and *E. salsugineum* the lowest. This fits with *E. salsugineum* being a *cmt3* mutant and RdDM likely being responsible for residual mCHG [62]. However, the sizes of these regions are similar to other species, indicating that this has not resulted in fragmentation of these regions (Additional file 3: Figure S11). 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 DNA methylation in contrast to *M. truncatula* and *L. japonicus* (Additional file 3: Figure S11). As these lower methylated mCHH regions are larger in size (Additional file 3: Figure S12) and less targeted by 24 nt siRNAs (Additional file 3: Figure S13), 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* (Additional file 2: Table S2). In the Poaceae, we also find that mCHH regions are more highly methylated, even though genome-wide, mCHH levels are lower (Additional file 3: Figure S11). This indicates that much of the mCHH in these genomes comes from smaller regions targeted by RdDM (Additional file 3: Figures S12 and S13), which is supported by RdDM mutants in *Z. mays* [42]. In contrast, previously discussed species like *M. esculenta*, *T. cacao*, and *V. vinifera* had mCHH regions of both low DNA methylation and small

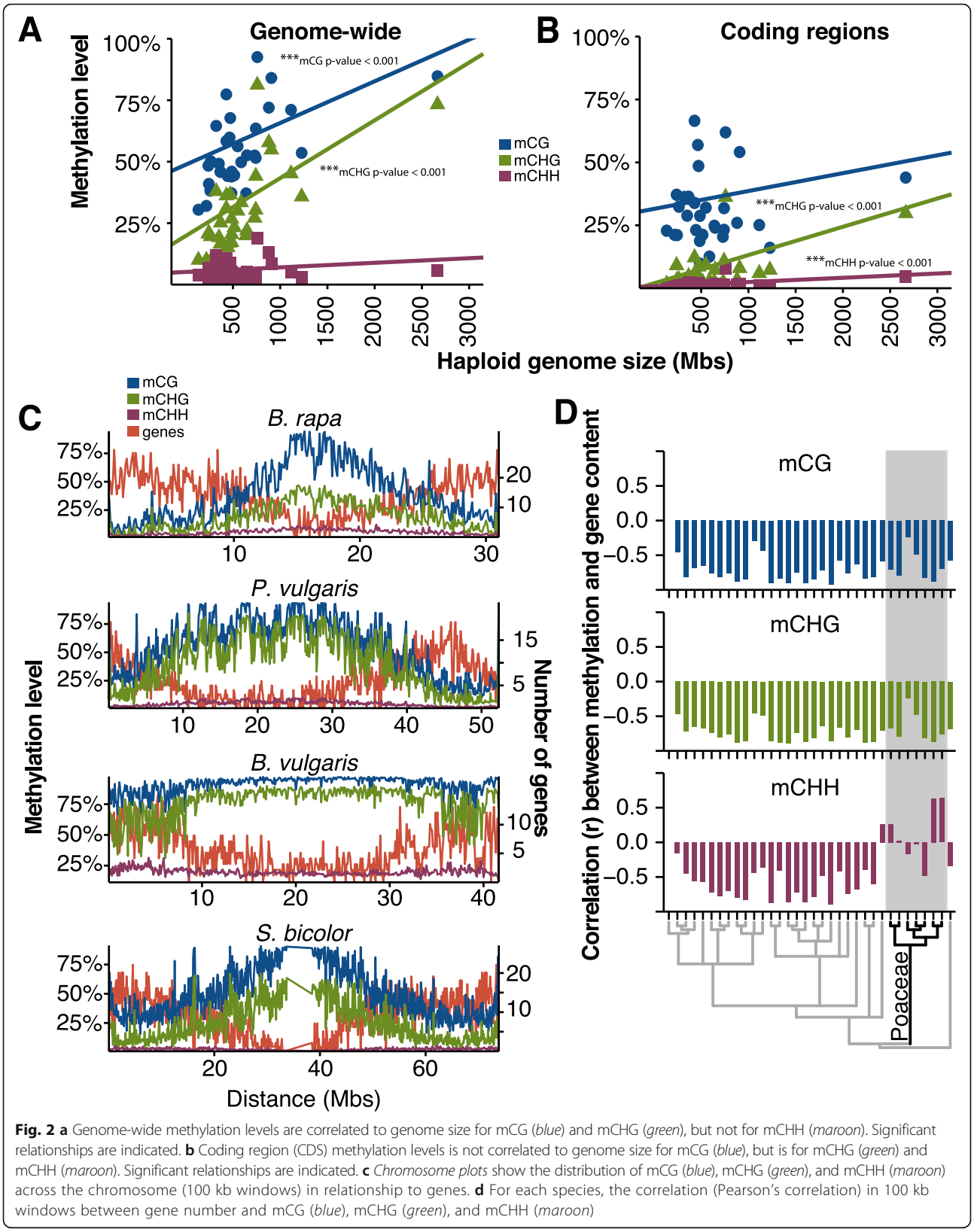

Fig. 2 **a** Genome-wide methylation levels are correlated to genome size for mCG (*blue*) and mCHG (*green*), but not for mCHH (*maroon*). Significant relationships are indicated. **b** Coding region (CDS) methylation levels is not correlated to genome size for mCG (*blue*), but is for mCHG (*green*)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

size which could indicate that effect of all mCHH pathways have been limited in these species (Additional file 3: Figure S12 and S13).

DNA methylation of repeats

Genome-wide mCG and mCHG levels are related to the proliferation of repetitive elements. The extent which heterochromatin and repeats are represented among the genomes studied does vary with the completeness of the assembled genomes. Despite this, however, correlations were found between repeat number and mCG (p value = 3.0×10^{-2}) and mCHG levels (p value = 4.9×10^{-4}) (Fig. 3a, Additional file 1: Table S3). This likely explains the correlation of DNA methylation with genome size, as large genomes often have more repetitive elements [71, 72]. No such correlation between mCHH levels and repeat numbers was found (p value = 1) (Fig. 3a). This was unexpected given that mCHH is generally associated with repetitive sequences in many plant species [32, 73]. Both CDS mCHG and mCHH correlated with the total number of repeats (p value = 8.7×10^{-3}, p value = 1.5×10^{-2}, respectively), but CDS mCG did not (p value = 1) (Additional file 3: Figure S15A). CDS mCHG and mCHH were also correlated with the presence of repeats within gene bodies (exons, introns, and untranslated regions: mCHG p value = 1.6×10^{-3}, mCHH p value = 2.0×10^{-3}), whereas mCG was not (p value = 1) (Additional files 1 and 3: Table S3 and Figure S15b). Plotting the percentage of genes containing repeats against the total number of repeats showed a relationship between the percentage of repeat content in genes and total number of repeats (p value = 2.4×10^{-6}) (Additional file 3: Figure S15C). After *Z. mays*, *B. vulgaris* has the highest percentage of genes containing repeats, much more so than expected given the total repeat content. This may explain in part why it has the highest CDS methylation levels.

Considerable variation exists in DNA methylation patterns within repeats. Across all species, repeats were heavily methylated at CG sequences, but were more variable in CHG and CHH methylation (Fig. 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 DNA methylation sequence contexts as compared to outside the repeat (Fig. 3c, Additional file 3: S16). 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.

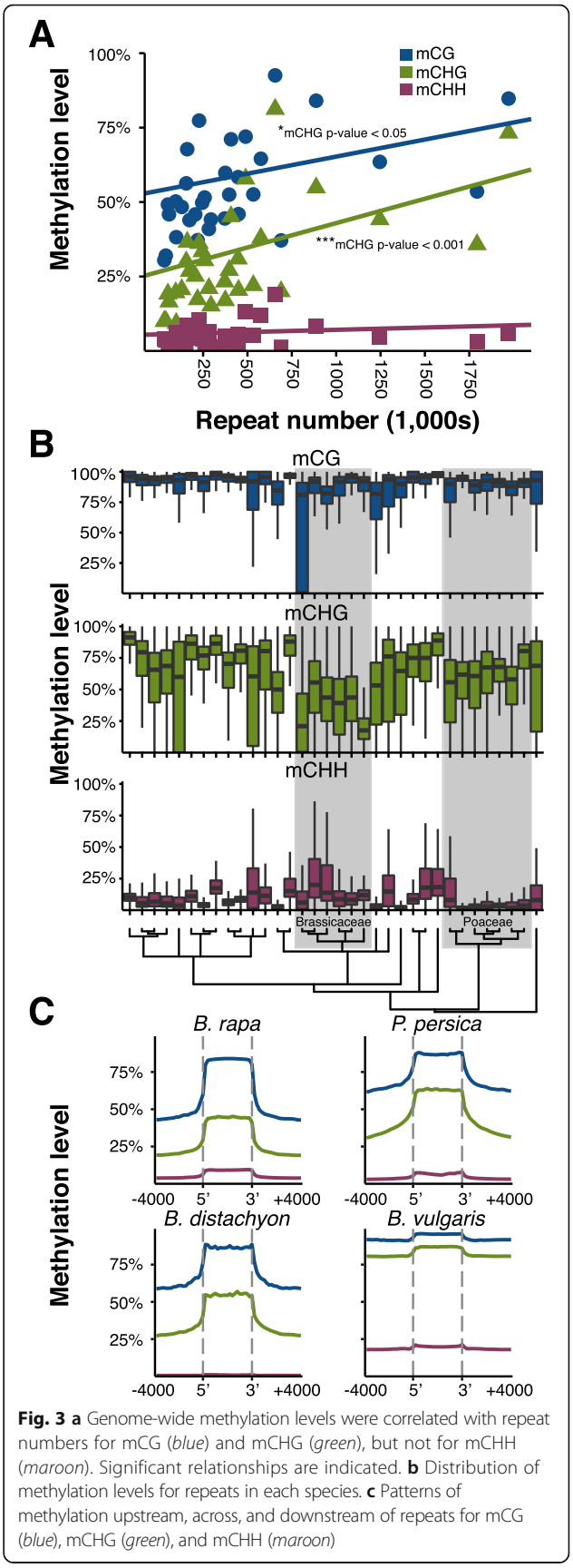

Fig. 3 **a** Genome-wide methylation levels were correlated with repeat numbers for mCG (*blue*) and mCHG (*green*), but not for mCHH (*maroon*). Significant relationships are indicated. **b** Distribution of methylation levels for repeats in each species. **c** Patterns of methylation upstream, across, and downstream of repeats for mCG (*blue*), mCHG (*green*), and mCHH (*maroon*)

CG gene body methylation

DNA methylation within genes in all three contexts is associated with suppressed gene expression [33], whereas genes that are only mCG methylated within the gene body are often constitutively expressed genes [74–76]. We classified genes using a modified version of the binomial test described by Takuno and Gaut [45] into one of four categories: CG gene body methylated (hereafter gbM), mCHG, mCHH, and unmethylated (UM) (Additional files 3, 4, and 5: Figure S17 and Table S4). This approach enables a consistent and statistically based classification of genes, but cannot fully capture finer details such as the pattern of methylation. GbM genes are methylated at CG sites, but not at CHG or CHH. Non-CG contexts are often coincident with mCG, for example RdDM regions are methylated in all three contexts. We further classified non-CG methylated genes as mCHG genes (mCHG and mCG, no mCHH) or mCHH genes (mCHH, mCHG, and mCG). Genes with insignificant amounts of DNA methylation were classified as unmethylated.

Between species, the DNA methylation status of gbM can be conserved across orthologs [46]. The DNA methylation state of orthologous genes across all species was compared using *A. thaliana* as an anchor (Fig. 4a). *A. lyrata* and *C. rubella* are the most closely related to *A. thaliana* and also have the greatest conservation of DNA 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 (Additional file 2: Table S2). However, when the percentage of gbM genes in the genome is taken into account (Fig. 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 DNA 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 DNA methylation around the transcriptional start site (TSS), increasing mCG throughout the gene body and a sharp decrease at the transcriptional termination site (TTS) [75, 76]. GbM genes identified in most species show this same trend and even have comparable levels of DNA methylation (Fig. 4c, Additional file 3: Figure S18). 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 DNA methylation levels, respectively, in gbM genes; and *E. salsugineum* shows no canonical gbM having only a few genes that passed statistical tests for having enrichment of mCG in gene bodies. As has previously been found [75, 76], gbM genes are more highly expressed as compared to UM and non-CG (mCHG and mCHH) genes (Fig. 4d, Additional file 3: Figure S19). The exception to this is *E. salsugineum* where the few genes that showed statistically significant amounts of mCG have almost no expression, supporting that they are not truly gbM genes but instead statistical anomalies associated with high numbers of statistical tests. 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 (Fig. 4d, Additional file 3: Figure S19) 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 have previously been reported to evolve more slowly [45, 46]. 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 (Additional files 3 and 6: Tables S5 and S6). With the exception of *E. salsugineum*, which lacks canonical gbM, these genes were longer and had more exons than UM genes (Additional files 3 and 6: Tables S5 and S6). 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* [45] and between *B. distachyon* and *O. sativa* [46] are more slowly evolving than UM orthologs. We verified this result for *A. thaliana* and *A. lyrata*. Within dicots, this result remains over short evolutionary distances, but it breaks down over greater distances with gbM genes typically evolving at equivalent rates as UM and, in some cases, faster rates (Additional files 3 and 6:

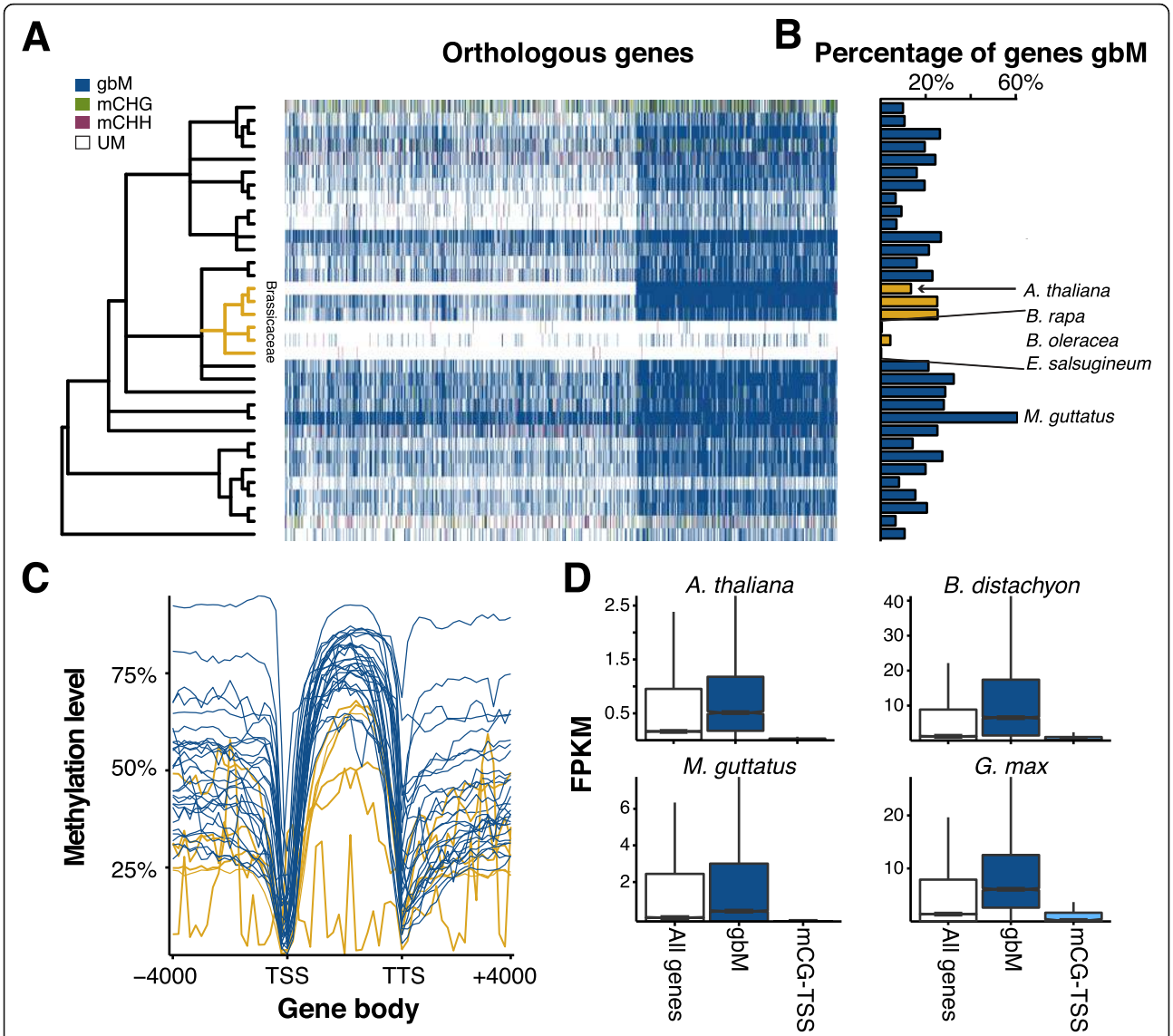

Fig. 4 **a** *Heatmap* showing methylation state of orthologous genes (horizontal axis) to *A. thaliana* for each species (*vertical axis*). Species are organized according to phylogenetic relationship. **b** Percentage of genes in each species that are gbM (mCG only in coding sequences). The Brassicaceae 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, whereas mCG over the TSS (mCG-TSS) has reduced gene expression

Tables S5 and S6). Between *B. distachyon* and *O. sativa*, and across the Poaceae, we found the opposite result. GbM genes typically were evolving at faster rates (Additional files 3 and 6: Tables S5 and S6). To increase the robustness of our analyses across such diverse species, we incorporated several differences in our methods and choice of molecular evolution model, which could account for these discrepancies (see "Methods"). Why gbM genes are evolving at faster rates in the Poaceae is unknown and future studies will be needed to resolve this.

Non-CG methylated genes

Non-CG methylation exists within genes and is known to suppress gene expression [16, 18, 77–79]. Differences in annotation quality could lead to some transposons being misannotated as genes and thus as targets of non-CG methylation. However, work in both *A. thaliana* and *G. max* have shown that some percentage of protein-coding genes do indeed contain non-CG methylation [3, 12]. 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 [80]. Based on this difference in DNA methylation, mCHG and mCHH genes were maintained as separate categories (Additional files 4 and 5: Table S4). The DNA methylation profiles of mCHG and mCHH genes often resembled that of repeats (Fig. 5a, Additional file 3: Figure S18). Both mCHG and mCHH genes are associated with reduced expression levels (Fig. 5b, Additional file 3: Figure S19). As mCHG methylation is present in mCHH genes, this may indicate that mCHG alone is sufficient for reduced gene expression. It was also observed that *Cucumis sativus* has an unusual pattern of mCHH in many highly expressed genes, although this pattern was not observed in a second *C. sativus* sample and will require further study to understand the basis for this difference (Additional file 3: Figure S20). The number of genes possessing non-CG types of DNA methylation ranged from as low as ~3 % of genes (*M. esculenta*) to as high as ~32 % of genes (*F. vesca*) (Fig. 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 non-CG genes.

Unlike gbM genes, there was no conservation of DNA methylation status across orthologs of mCHG and mCHH genes (Additional file 3: Figure S21). For many non-CG 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 non-CG methylation between orthologs within the Brassicaceae [48]. However, we did observe some conservation based on gene ontology (GO). The same GO terms were often enriched in multiple species (Additional files 3 and 7: Figure S22 and Table S7). The most commonly enriched terms were involved in proteolysis, cell death, and defense responses; these processes 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 [81, 82]. 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 [83].

Transposable element insertions near or within genes can be one cause of non-CG methylated genes. To test

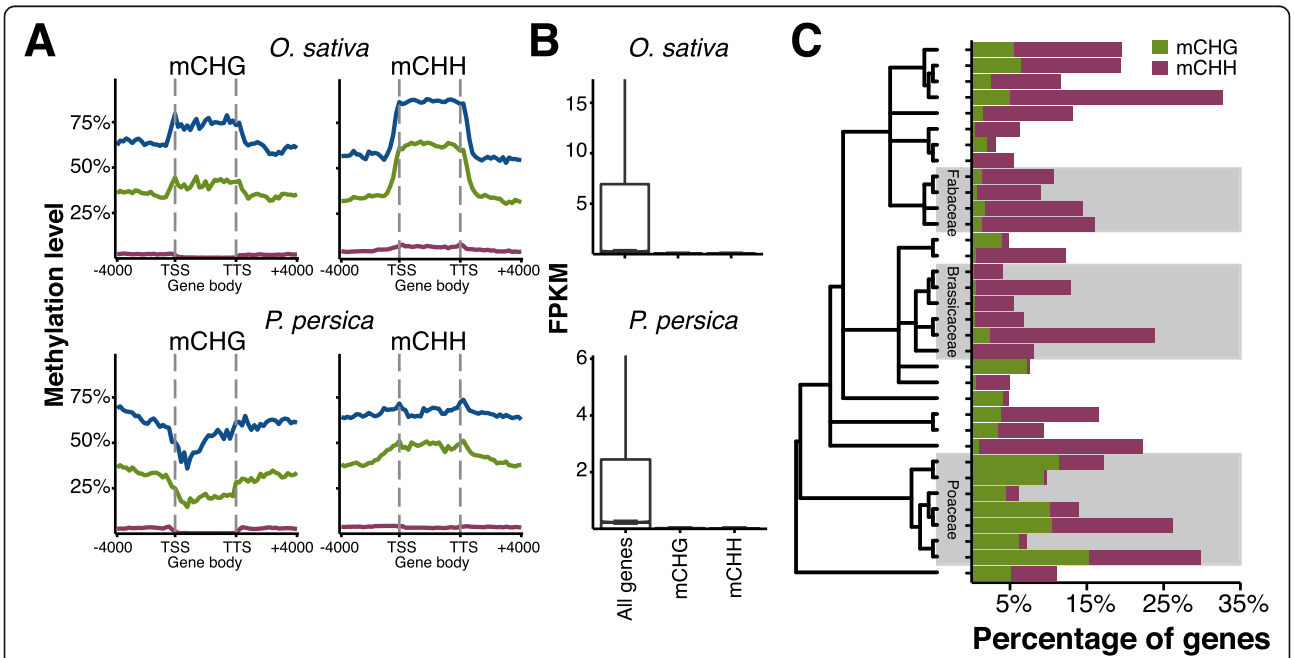

Fig. 5 **a** Methylation levels for mCG (*blue*) (mCG only in coding sequences), mCHG (*green*) (mCG and mCHG in coding sequences), and mCHH (*maroon*) (mCG, mCHG, and mCHH in coding sequences) were plotted upstream, across, and downstream of mCHG and mCHH genes. **b** Gene expression of mCHG and mCHH genes vs. all genes. **c** The percentage of mCHG and mCHH genes per species. Species are arranged by phylogenetic relationship

this, we looked for enrichment of TEs upstream, within, and downstream of gbM, mCHG, and mCHH genes (Additional file 3: Figure S23). For the majority of species, TEs were indeed enriched near or within non-CG methylated genes. There were exceptions, however, as in both *S. lycopersicum* and *Z. mays*, there was no enrichment of TEs associated with these genes. Surprisingly, there was enrichment for TEs associated with gbM genes in many species. In large genomes like *Z. mays*, nearly every gene is associated with a TE in some way, indicating that the presence of an associated TE alone is not the only cause of non-CG methylation within genes.

Non-coding sequences and regulatory regions

Outside of the gene body, DNA methylation might have an impact on gene expression through the DNA methylation of neighboring transcription factor binding sites (TFBS) or other regulatory elements. To date, there is limited in vivo evidence of such effects in plants, although the recent example of *repressor of silencing 1* (*ROS1*) hints at this possibility [84, 85]. In vitro evidence also supports the possibility of DNA methylation inhibiting and in some cases, promoting, transcription factor binding [86]. Conserved non-coding sequences contain many important regulatory elements, including TFBS [87, 88]. We identified CNS regions for a sample of species across the phylogeny and plotted DNA methylation levels (Fig. 6a, Additional file 3: Figure S24). DNA methylation in all three contexts was depleted across these regions, compared to outside. Locations of CNS regions were defined as either proximal (within 1 kbps), distal (>1 kbps), within untranslated regions (UTR), or within introns. Similar patterns were observed for CNS regions whether they were located proximally or distally to a gene (Additional file 3: Figure S24). UTR and intronic CNS sequences do show elevated levels of mCG in comparison, which might result from elevated mCG levels across the gene bodies of gbM genes. In *Z. mays*, high mCHH is enriched in the upstream and downstream regions of highly expressed genes and are termed mCHH islands [36, 37]. We identified mCHH islands 2 kb upstream and downstream of annotated genes for each species, finding that the percentage of genes with such regions varied considerably across species (Fig. 6b). Although some species other than *Z. mays* also show an association between mCHH islands and gene expression, many showed no such association, indicating no universal causal relationship between the two (Fig. 6c, Additional file 3: Figure S25). 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 (Fig. 6d, Additional file 3: Figure S26) [37]. However, this difference in DNA methylation level is much less pronounced in most other species as compared to *Z. mays* (Additional file 3: Figure S26). It is thought that these differences in DNA methylation on proximal versus distal sides of mCHH islands mark euchromatin-heterochromatin boundaries [37]. Indeed, mCHH

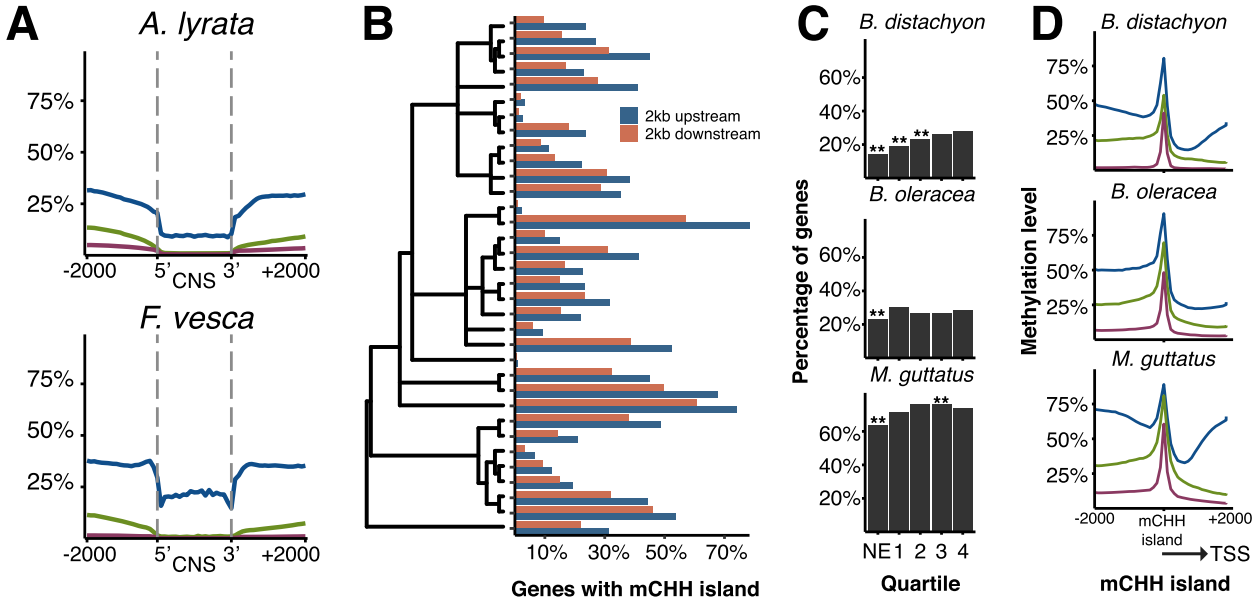

Fig. 6 a Patterns of methylation across conserved non-coding sequences (CNS) for mCG (*blue*), mCHG (*green*), and mCHH (*maroon*). **b** Percentage of genes with mCHH islands 2 kb upstream or downstream. **c** Association of upstream mCHH islands with gene expression. Genes are divided into not-expressed (NE) and quartiles of increasing expression. ** indicates a difference in proportion from the fourth quartile at $p < 0.01$. **d** Patterns of upstream mCHH islands for mCG (*blue*), mCHG (*green*), and mCHH (*maroon*)

islands are often associated with transposons [36, 37], however, there was no correlation found between the total number of repeats in the genome and the number of genes with mCHH islands (Additional files 1 and 3: Table S3 and Figure S27a). When correlated to the percentage of genes with repeats 2 kb upstream or downstream, both upstream and downstream mCHH islands are correlated (upstream p value = 4.5×10^{-5}, downstream p value = 9.3×10^{-4}) (Additional files 1 and 3: Table S3 and Figure S27b). While there was a correlation between the total repeat content and the percentage of upstream and downstream repeats (upstream p value = 1.3×10^{-2}, downstream p value = 1.6×10^{-3}), there were numerous outlying species which may explain the lack of correlation between mCHH islands and total repeat content (Additional files 1 and 3: Table S3 and Figure S27c). This supports a hypothesis that transposon distribution as opposed to transposon load alone is critical in shaping the epigenome.

Discussion and conclusions

We present the methylomes of 34 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 DNA methylation and distribution of DNA methylation, with the greatest variation being observed in non-CG contexts. The Brassicaceae show overall lower mCHG levels and reduced numbers of gbM genes, leading to a complete loss in *E. salsugineum*, that is associated with loss of CMT3 [62]. 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 tend to have lower mCHH levels, suggesting a potential effect. Epigenetic variation induced by propagation techniques can be of agricultural and economic importance [89], and understanding the effects of clonal propagation will require future studies over multiple generations. Evaluation of per-site DNA methylation levels, methylated regions, their structure, and association with small RNAs suggests 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 DNA methylation level and distribution with impacts on DNA 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 *Selaginella moellendorffii* (lycophyte) [44], *Physcomitrella patens* (moss) [44], and *Marchantia polymorpha* (liverwort) [47]. That it has also been lost in the angiosperm *E. salsugineum* indicates that it is dispensable over evolutionary time [58]. Non-CG methylation shows no conservation at the level of individual genes, which indicates that it is gained and lost in a lineage specific manner. It is an open question as to the evolutionary origins of non-CG methylation within genes. This type of DNA methylation within in genes is typically associated with the presence of upstream or downstream TEs. However, species like *Z. mays* are an exception, with nearly every gene associated with TEs, suggesting that other causes might also exist. Many non-CG genes lack orthologous genes, which could indicate a preferential targeting of de novo genes, as in the case of the *qua-quine starch* (*QQS*) gene in *A. thaliana* [19]. 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 conserved non-coding sequences, and mCHH islands are also examined. CNS regions show depletion of all type of DNA methylation, hinting that DNA methylation may have an inhibitory role at regulatory regions. While mCHH islands 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 DNA 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 [47, 48] and now shows that there is still much to be learned about underlying causes of variation in this molecular trait. Due to its role in gene expression and its potential to vary independently of genetic variation, understanding these causes will be necessary to a more complete understanding of the role of DNA methylation underlying biological diversity.

Methods

MethylC-seq and analysis

In plants, DNA methylation is highly stable between tissues and across generations [15, 48], showing little variation between replicates. DNA was isolated from leaf tissue and MethylC-seq libraries for each species were prepared as previously described [53]. Previously published datasets were obtained from public databases and reanalyzed [12, 15, 36, 48, 54–56, 62, 90]. Genome

sequences and annotations for most species were downloaded from Phytozome 10.1 (http://phytozome.jgi.doe.gov/pz/portal.html) [51]. The *L. japonicus* genome was downloaded from the *Lotus japonicus* Sequencing Project (http://www.kazusa.or.jp/lotus/) [49], the *B. vulgaris* genome was downloaded from the *Beta vulgaris* Resource (bvseq.molgen.mpg.de/) [52], and the *C. sativa* genome from the *C. sativa* (*Cannabis*) Genome Browser Gateway (http://genome.ccbr.utoronto.ca/cgi-bin/hgGateway) [50]. As gene annotations for *S. viridis* were not available, gene models from the closely related *S. italica* were mapped onto the *S. viridis* genome using Exonerate [91] and the best hits retained. As repeat annotations were unavailable for 12 of the species studied, RepeatMasker [92] was used to annotate repetitive elements and transposons using plant repetitive element sequences downloaded from Repbase [93] and *A. thaliana* transposable element sequences [20].

Sequencing data for each species was aligned to their respective genome (Additional file 1: Table S1) [20, 49–52, 54, 94–116] and methylated sites called using previously described methods [117]. In brief, reads were trimmed for adapters and quality using Cutadapt [118] and then mapped to both a converted forward strand (all cytosines to thymines) and converted reverse strand (all guanines to adenines) using bowtie [119]. 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 (Additional file 1: Table S1). Cytosines were called as methylated using a binomial test using the non-conversion rate as the expected probability followed by multiple testing correction using Benjamini–Hochberg false discovery rate (FDR). A minimum of three reads mapping to a site was required to call a site as methylated. Data are available at the Plant Methylome DB http://schmitzlab.genetics.uga.edu/plantmethylomes.

Phylogenetic tree

A species tree was constructed using BEAST2 [120] on a set of 50 previously identified single copy loci [69]. Protein sequences were aligned using PASTA [121] and converted into codon alignments using custom Perl scripts. Gblocks [122] was used to identify conserved stretches of amino acids and then passed to JModelTest2 [123, 124] 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 [57]. To determine per-site methylation levels, the weighted methylation for each cytosine with at least 3 reads of coverage was calculated and this distribution plotted. Symmetry plots were constructed by identifying paired symmetrical cytosines with 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 [62] at 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 [125], phytools [126], and NLME packages assuming a model of Brownian motion. In total, 22 comparisons were conducted (Additional file 1: Table S3) and a p value < 0.05 after Bonferroni correction. Distribution of methylation levels and genes across chromosomes was conducted by dividing the genome into 100 kb windows, sliding every 50 kb using BedTools [127] 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 [127] and custom scripts. For each context, only methylated sites in that respective context were considered and used to define the region. The genome was divided into 25 bp windows and all windows that contained at least one methylated cytosine in the context of interest were retained. Windows were merged if they were within 100 bp of each other. 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-sequencing (RNA-seq) datasets for additional species

were downloaded from GEO and the SRA and reanalyzed [15, 54, 55, 101, 128, 129]. The small RNA toolkit from the UEA computational Biology lab was used to trim and clean the reads [130]. For trimming, 8 bp of the 3' adapter was trimmed. Trimmed and cleaned reads were aligned using PatMan allowing for zero mismatches [131]. BedTools [127] and custom scripts were used to calculate overlap with mCHH regions.

Gene-level analyses

Genes were classified as gbM, mCHG, or mCHH by applying a binomial test to the number of methylated sites in a gene [45] (Additional files 3, 4 and 5: Figure S17 and 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 applied to each gene for each sequence context and q-values calculated by adjusting p values by Benjamini–Hochberg FDR. Genes were classified as gbM 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.

TEs mapping to within 2000 bp upstream, within, or 2000 bp downstream of a gene were identified using BedTools. GbM, mCHG, and mCHH genes were then tested for enrichment of TEs upstream, within, or downstream using Fisher's exact test against the background of all the genes in a genome. GO terms for each gene were downloaded from phytozome 10.1 (http://phytozome.jgi.doe.gov/pz/portal.html) [51]. GO term enrichment was performed using the parentCHILD algorithm [132] with the F-statistic as implemented in the topGO module in R. Multiple testing correction was then applied using the Benjamini–Hochberg procedure. 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 [51] 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 [51] 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 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 *O. 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 [133], 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. gbM-gbM and UM-UM). The *yn00* (Yang-Neilsen) [134] model in the program PAML for pairwise sequence comparison was used to estimate synonymous substitution rates, non-synonymous substitution rates, and adaptive evolution (dS, dN, and ω, respectively) [135]. 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 gbM and UM methylated genes.

RNA-seq mapping and analysis

RNA-seq datasets [12, 15, 48, 54, 62, 101, 109, 129, 136–140] 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 [141] supplied with a reference genome feature file (GFF) with the following arguments -I 50000 –b2-very-sensitive –b2-D 50 (Additional file 1: Table S1). Transcripts were then quantified using Cufflinks v2.2.1 [142] supplied with a reference GFF.

Conserved non-coding sequences

The CNS Discovery Pipeline 3.0 [88] was used to call conserved non-coding sequences through pair-wise comparison of closely related species (*A. lyrata-A. thaliana*, *B. distachyon-O. sativa*, *F. vesca-P. persica*, *G. raimondii-T. cacao*, *M. esculenta-P. trichocarpa*). As the genomes of some species analyzed here are as yet unpublished, we restricted our analysis to a representative subset of species with published genomes taken from across the phylogeny. CNS regions were defined as 5' distal, 5' proximal, intronic, 3' proximal, and 3' distal by the CNS Discovery Pipeline 3.0. Coordinates for CNS regions were extracted and methylation levels calculated across 2 kb upstream, across the CNS, and 2 kb downstream. BED files of called CNS regions are available at GitHub (https://github.com/chadn737/Widespread-natural-variation-of-DNA-methylation-within-angiosperms).

mCHH islands

mCHH islands were identified for both upstream and downstream regions as previously described [37]. Briefly, methylation levels were determined for 100 bp windows across the genome. Windows of 25 % or greater mCHH with at least five methylated CHH sites, were identified 2 kb upstream and downstream of genes. Genes with more missing data in more than half the neighboring windows were removed. Methylation levels were then plotted centered on the window of highest mCHH, extending 2 kb in both directions. Genes associated with mCHH islands were categorized as non-expressed (NE) or divided into one of four quartiles based on their expression level. Differences in the proportions of each expression quartiles were determined in a pair-wise manner using prop.test in R with p value < 0.01 [37].

Additional files

Additional file 1: Supplemental Tables and Figures. **Table S1**, **Table S3**, and **Figures S1–10**. (PDF 16677 kb)

Additional file 2: Table S2. Trait and feature information for each species. (XLSX 15 kb)

Additional file 3: Supplemental Table and Figures. **Table S5** and **Figures S11-27**

Additional file 4: Table S4. Classification of gene methylation status: part A species *A. thaliana* through *L. japonicus*. (XLSX 10029 kb)

Additional file 5: Table S4. Classification of gene methylation status: part B species *M. domestica* through *Z. mays*. (XLSX 10996 kb)

Additional file 6: Table S6. Summary statistics for evolutionary analysis of gbM genes. (XLSX 31 kb)

Additional file 7: Table S7. GO term enrichment for each gene class and the species for which that GO term is enriched in. (XLSX 68 kb)

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. We would also like to thank the Georgia Genomics Facility and Georgia Advanced Computing Resource Center (GACRC) for technical support, particularly Dr. Shanho Tsai and Yecheng Huang of the GACRC for their efforts.

Funding

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 JS. 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 –1339194), by the Office of the Vice President of Research at UGA, and by The Pew Charitable Trusts to RJS. CEN was supported by a NSF postdoctoral fellowship (IOS – 1402183).

Availability of data and materials

Genome browsers for all methylation data used in this paper are located at Plant Methylation DB (http://schmitzlab.genetics.uga.edu/plantmethylomes). Sequence data for MethylC-seq, RNA-seq, and small RNA-seq are located at the Gene Expression Omnibus, accession GSE79526. Code and other relevant data is available on GitHub (https://github.com/chadn737/Widespread-natural-variation-of-DNA-methylation-within-angiosperms.git).

Authors' contributions

Conceptualization: CEN, AJB, and RJS; Performed experiments: CEN, NAR, KDK, AR, JTP, and RJS; Data Analysis: CEN, AJB, LJ, and MSA; Writing – Original Draft: CEN; Writing – Review and Editing: CEN, AJB, SAJ, NMS, and RJS; Resources: QL, JMB, JAU, CE, JS, JG, SAJ, and NMS. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval

Ethics approval was not needed for this study.

Author details

[1]Department of Genetics, University of Georgia, 120 East Green Street, 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, UT 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, CA, USA.

Received: 2 June 2016 Accepted: 9 September 2016
Published online: 27 September 2016

References

1. Vaughn MW, Tanurdzic M, Lippman Z, Jiang H, Carrasquillo R, Rabinowicz PD, et al. Epigenetic natural variation in Arabidopsis thaliana. PLoS Biol. 2007;5(7):e174. doi:10.1371/journal.pbio.0050174.
2. Eichten SR, Swanson-Wagner RA, Schnable JC, Waters AJ, Hermanson PJ, Liu S, et al. Heritable epigenetic variation among maize inbreds. PLoS Genet. 2011;7(11):e1002372. doi:10.1371/journal.pgen.1002372.
3. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, et al. Patterns of population epigenomic diversity. Nature. 2013;495(7440):193–8. doi:10.1038/nature11968.
4. Filleton F, Chuffart F, Nagarajan M, Bottin-Duplus H, Yvert G. The complex pattern of epigenomic variation between natural yeast strains at single-nucleosome resolution. Epigenetics Chromatin. 2015;8:26. doi:10.1186/s13072-015-0019-3.
5. Colome-Tatche M, Cortijo S, Wardenaar R, Morgado L, Lahouze B, Sarazin A, et al. Features of the Arabidopsis recombination landscape resulting from the combined loss of sequence variation and DNA methylation. Proc Natl Acad Sci U S A. 2012;109(40):16240–5. doi:10.1073/Pnas.1212955109.
6. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13(7):484–92. doi:10.1038/nrg3230.
7. Schmitz RJ, Zhang X. Decoding the epigenomes of herbaceous plants. Genomes of Herbaceous Land Plants. 2014;69:247–77. doi: 10.1016/b978-0-12-417163-3.00010-x.
8. Yelina NE, Lambing C, Hardcastle TJ, Zhao X, Santos B, Henderson IR. DNA methylation epigenetically silences crossover hot spots and controls chromosomal domains of meiotic recombination in Arabidopsis. Genes Dev. 2015;29(20):2183–202. doi:10.1101/gad.270876.115.
9. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011;12(1):R10. doi:10.1186/gb-2011-12-1-r10.
10. Eichten SR, Briskine R, Song J, Li Q, Swanson-Wagner R, Hermanson PJ, et al. Epigenetic and genetic influences on DNA methylation variation in maize populations. Plant Cell. 2013;25(8):2783–97. doi:10.1105/tpc.113.114793.
11. Regulski M, Lu Z, Kendall J, Donoghue MT, Reinders J, Llaca V, et al. The maize methylome influences mRNA splice sites and reveals widespread paramutation-like switches guided by small RNA. Genome Res. 2013;23(10):1651–62. doi:10.1101/gr.153510.112.
12. Schmitz RJ, He Y, Valdes-Lopez O, Khan SM, Joshi T, Urich MA, et al. Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. Genome Res. 2013;23(10):1663–74. doi:10.1101/gr.152538.112.
13. Dubin MJ, Zhang P, Meng D, Remigereau MS, Osborne EJ, Paolo Casale F, et al. DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. eLife. 2015;4:e05255. doi:10.7554/eLife.05255.
14. Becker C, Hagmann J, Muller J, Koenig D, Stegle O, Borgwardt K, et al. Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. Nature. 2011;480(7376):245–9. doi:10.1038/nature10555.
15. Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, et al. Transgenerational epigenetic instability is a source of novel methylation variants. Science. 2011;334(6054):369–73. doi:10.1126/science.1212959.
16. Cubas P, Vincent C, Coen E. An epigenetic mutation responsible for natural variation in floral symmetry. Nature. 1999;401(6749):157–61. doi:10.1038/43657.
17. Thompson AJ, Tor M, Barry CS, Vrebalov J, Orfila C, Jarvis MC, et al. Molecular and genetic characterization of a novel pleiotropic tomato-ripening mutant. Plant Physiol. 1999;120(2):383–90.
18. Manning K, Tor M, Poole M, Hong Y, Thompson AJ, King GJ, et al. A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. Nat Genet. 2006;38(8):948–52. doi:10.1038/ng1841.
19. Silveira AB, Trontin C, Cortijo S, Barau J, Del Bem LE, Loudet O, et al. Extensive natural epigenetic variation at a de novo originated gene. PLoS Genet. 2013;9(4):e1003437. doi:10.1371/journal.pgen.1003437.
20. Arabidopsis Genome I. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408(6814):796–815. doi:10.1038/35048692.
21. Flowers JM, Purugganan MD. The evolution of plant genomes: scaling up from a population perspective. Curr Opin Genet Dev. 2008;18(6):565–70. doi:10.1016/j.gde.2008.11.005.
22. Lane AK, Niederhuth CE, Ji L, Schmitz RJ. pENCODE: a plant encyclopedia of DNA elements. Annu Rev Genet. 2014;48:49–70. doi:10.1146/annurev-genet-120213-092443.
23. Niederhuth CE, Schmitz RJ. Covering your bases: inheritance of DNA methylation in plant genomes. Mol Plant. 2014;7(3):472–80. doi:10.1093/mp/sst165.
24. Finnegan EJ, Genger RK, Peacock WJ, Dennis ES. DNA Methylation in Plants. Annu Rev Plant Physiol Plant Mol Biol. 1998;49:223–47. doi:10.1146/annurev.arplant.49.1.223.
25. Finnegan EJ, Peacock WJ, Dennis ES. Reduced DNA methylation in Arabidopsis thaliana results in abnormal plant development. Proc Natl Acad Sci U S A. 1996;93(16):8449–54.
26. Bostick M, Kim JK, Esteve PO, Clark A, Pradhan S, Jacobsen SE. UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science. 2007;317(5845):1760–4. doi:10.1126/science.1147939.
27. Lindroth AM, Cao X, Jackson JP, Zilberman D, McCallum CM, Henikoff S, et al. Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. Science. 2001;292(5524):2077–80. doi:10.1126/science.1059745.
28. Du J, Zhong X, Bernatavichute YV, Stroud H, Feng S, Caro E, et al. Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. Cell. 2012;151(1):167–80. doi:10.1016/j.cell.2012.07.034.
29. Du J, Johnson LM, Groth M, Feng S, Hale CJ, Li S, et al. Mechanism of DNA methylation-directed histone methylation by KRYPTONITE. Mol Cell. 2014;55(3):495–504. doi:10.1016/j.molcel.2014.06.009.
30. Du J, Johnson LM, Jacobsen SE, Patel DJ. DNA methylation pathways and their crosstalk with histone methylation. Nat Rev Mol Cell Biol. 2015;16(9):519–32. doi:10.1038/nrm4043.
31. Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, Thao K, et al. The Arabidopsis nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell. 2013;153(1):193–205. doi:10.1016/j.cell.2013.02.033.
32. Stroud H, Do T, Du J, Zhong X, Feng S, Johnson L, et al. Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. Nat Struct Mol Biol. 2014;21(1):64–72. doi:10.1038/nsmb.2735.
33. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010;11(3):204–20. doi:10.1038/nrg2719.
34. Cao X, Jacobsen SE. Locus-specific control of asymmetric and CpNpG methylation by the DRM and CMT3 methyltransferase genes. Proc Natl Acad Sci U S A. 2002;99(4):16491–8. doi:10.1073/pnas.162371599.
35. Cao X, Jacobsen SE. Role of the Arabidopsis DRM methyltransferases in de novo DNA methylation and gene silencing. Curr Biol. 2002;12(13):1138–44. doi:10.1016/s0960-9822(02)00925-9.
36. Gent JI, Ellis NA, Guo L, Harkess AE, Yao Y, Zhang X, et al. CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. Genome Res. 2013;23(4):628–37. doi:10.1101/gr.146985.112.
37. Li Q, Gent JI, Zynda G, Song J, Makarevitch I, Hirsch CD, et al. RNA-directed DNA methylation enforces boundaries between heterochromatin and euchromatin in the maize genome. Proc Natl Acad Sci U S A. 2015. doi:10.1073/pnas.1514680112.
38. Stroud H, Greenberg MV, Feng S, Bernatavichute YV, Jacobsen SE. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell. 2013;152(1-2):352–64. doi:10.1016/j.cell.2012.10.054.
39. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature. 2008;452(7184):215–9. doi:10.1038/nature06745.
40. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008;133(3):523–36. doi:10.1016/j.cell.2008.03.029.
41. Lister R, Ecker JR. Finding the fifth base: genome-wide sequencing of cytosine methylation. Genome Res. 2009;19(6):959–66. doi:10.1101/gr.083451.108.
42. Li Q, Eichten SR, Hermanson PJ, Zaunbrecher VM, Song J, Wendt J, et al. Genetic perturbation of the maize methylome. Plant Cell. 2014;26(12):4602–16. doi:10.1105/tpc.114.133140.
43. Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, et al. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A. 2010;107(19):8689–94. doi:10.1073/pnas.1002720107.

44. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science. 2010;328(5980):916–9. doi:10.1126/science.1186366.
45. Takuno S, Gaut BS. Body-methylated genes in Arabidopsis thaliana are functionally important and evolve slowly. Mol Biol Evol. 2012;29(1):219–27. doi:10.1093/molbev/msr188.
46. Takuno S, Gaut BS. Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. Proc Natl Acad Sci U S A. 2013;110(5):1797–802. doi:10.1073/pnas.1215380110.
47. Takuno S, Ran J-H, Gaut BS. Evolutionary patterns of genic DNA methylation vary across land plants. Nat Plants. 2016;2(2):15222. doi:10.1038/nplants.2015.222.
48. Seymour DK, Koenig D, Hagmann J, Becker C, Weigel D. Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. PLoS Genet. 2014;10(11):e1004785. doi:10.1371/journal.pgen.1004785.
49. Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, et al. Genome structure of the legume, Lotus japonicus. DNA Res. 2008;15(4):227–39. doi:10.1093/dnares/dsn008.
50. van Bakel H, Stout JM, Cote AG, Tallon CM, Sharpe AG, Hughes TR, et al. The draft genome and transcriptome of Cannabis sativa. Genome Biol. 2011;12(10):R102. doi:10.1186/gb-2011-12-10-r102.
51. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40(Database issue):D1178–86. doi:10.1093/nar/gkr944.
52. Dohm JC, Minoche AE, Holtgrawe D, Capella-Gutierrez S, Zakrzewski F, Tafer H, et al. The genome of the recently domesticated crop plant sugar beet (Beta vulgaris). Nature. 2014;505(7484):546–9. doi:10.1038/nature12817.
53. Urich MA, Nery JR, Lister R, Schmitz RJ, Ecker JR. MethylC-seq library preparation for base-resolution whole-genome bisulfite sequencing. Nat Protoc. 2015;10(3):475–83. doi:10.1038/nprot.2014.114.
54. Amborella GP. The Amborella genome and the evolution of flowering plants. Science. 2013;342(6165):1241089. doi:10.1126/science.1241089.
55. Stroud H, Ding B, Simon SA, Feng S, Bellizzi M, Pellegrini M, et al. Plants regenerated from tissue culture contain stable epigenome changes in rice. eLife. 2013;2:e00354. doi:10.7554/eLife.00354.
56. Zhong S, Fei Z, Chen YR, Zheng Y, Huang M, Vrebalov J, et al. Single-base resolution methylomes of tomato fruit development reveal epigenome modifications associated with ripening. Nat Biotechnol. 2013;31(2):154–9. doi:10.1038/nbt.2462.
57. Schultz MD, Schmitz RJ, Ecker JR. 'Leveling' the playing field for analyses of single-base resolution DNA methylomes. Trends Genet. 2012;28(12):583–5. doi:10.1016/j.tig.2012.10.012.
58. Bewick AJ, Ji L, Niederhuth CE, Willing E-M, Hofmeister BT, Shi X, et al. On the origin and evolutionary consequences of gene body DNA methylation. Proc Natl Acad Sci. 2016:201604666. doi: 10.1073/pnas.1604666113.
59. Willing E-M, Rawat V, Mandáková T, Maumus F, James GV, Nordström KJV, et al. Genome expansion of Arabis alpina linked with retrotransposition and reduced symmetric DNA methylation. Nat Plants. 2015;1(2):14023. doi:10.1038/nplants.2014.23.
60. Kawakatsu T, Huang S-shan C, Jupe F, Sasaki E, Schmitz Robert J, Urich Mark A, et al. Epigenomic diversity in a global collection of Arabidopsis thaliana accessions. Cell. 2016;166(2):492–505. doi:10.1016/j.cell.2016.06.044.
61. Bewick AJ, Niederhuth CE, Rohr NA, Griffin PT, Leebens-Mack J, Schmitz RJ. The evolution of CHROMOMETHYLTRANSFERASES and gene body DNA methylation in plants. bioRxiv. 2016. doi:10.1101/054924.
62. Bewick AJ, Ji L, Niederhuth CE, Willing E-M, Hofmeister BT, Shi X, et al. On the origin and evolutionary consequences of gene body DNA methylation. bioRxiv. 2016. doi:10.1101/045542.
63. McKey D, Elias M, Pujol B, Duputie A. The evolutionary ecology of clonally propagated domesticated plants. New Phytol. 2010;186(2):318–32. doi:10.1111/j.1469-8137.2010.03210.x.
64. Kitimu SR, Taylor J, March TJ, Tairo F, Wilkinson MJ, Rodríguez López CM. Meristem micropropagation of cassava (Manihot esculenta) evokes genome-wide changes in DNA methylation. Front Plant Sci. 2015;6:590. doi:10.3389/fpls.2015.00590.
65. Chang L, Zhang Z, Han B, Li H, Dai H, He P, et al. Isolation of DNA-methyltransferase genes from strawberry (Fragaria x ananassa Duch.) and their expression in relation to micropropagation. Plant Cell Rep. 2009;28(9):1373–84. doi:10.1007/s00299-009-0737-8.
66. Slotkin RK, Vaughn M, Borges F, Tanurdzic M, Becker JD, Feijo JA, et al. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell. 2009;136(3):461–72. doi:10.1016/j.cell.2008.12.038.
67. Calarco JP, Borges F, Donoghue MT, Van Ex F, Jullien PE, Lopes T, et al. Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. Cell. 2012;151(1):194–205. doi:10.1016/j.cell.2012.09.001.
68. Martins EP, Hansen TF. Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. Am Nat. 1997;149(4):646–67. doi:10.1086/286013.
69. Duarte JM, Wall PK, Edger PP, Landherr LL, Ma H, Pires JC, et al. Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. BMC Evol Biol. 2010;10:61. doi:10.1186/1471-2148-10-61.
70. Alonso C, Perez R, Bazaga P, Herrera CM. Global DNA cytosine methylation as an evolving trait: phylogenetic signal and correlated evolution with genome size in angiosperms. Front Genet. 2015;6:4. doi:10.3389/fgene.2015.00004.
71. Flavell RB, Bennett MD, Smith JB, Smith DB. Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem Genet. 1974;12(4):257–69.
72. Bennetzen JL, Ma J, Devos KM. Mechanisms of recent genome size variation in flowering plants. Ann Bot. 2005;95(1):127–32. doi:10.1093/aob/mci008.
73. Zemach A, Li Y, Wayburn B, Ben-Meir H, Kiss V, Avivi Y, et al. DDM1 binds Arabidopsis methyl-CpG binding domain proteins and affects their subnuclear localization. Plant Cell. 2005;17(5):1549–58. doi:10.1105/tpc.105.031567.
74. Tran RK, Henikoff JG, Zilberman D, Ditt RF, Jacobsen SE, Henikoff S. DNA methylation profiling identifies CG methylation clusters in Arabidopsis genes. Curr Biol. 2005;15(2):154–9. doi:10.1016/j.cub.2005.01.008.
75. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H, et al. Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell. 2006;126(6):1189–201. doi:10.1016/j.cell.2006.08.003.
76. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet. 2007;39(1):61–9. doi:10.1038/ng1929.
77. Bender J, Fink GR. Epigenetic control of an endogenous gene family is revealed by a novel blue fluorescent mutant of Arabidopsis. Cell. 1995;83(5):725–34. doi:10.1016/0092-8674(95)90185-X.
78. Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, et al. A transposon-induced epigenetic change leads to sex determination in melon. Nature. 2009;461(7267):1135–8. doi:10.1038/nature08498.
79. Durand S, Bouche N, Perez Strand E, Loudet O, Camilleri C. Rapid establishment of genetic incompatibility through natural epigenetic variation. Curr Biol. 2012;22(4):326–31. doi:10.1016/j.cub.2011.12.054.
80. West PT, Li Q, Ji L, Eichten SR, Song J, Vaughn MW, et al. Genomic distribution of H3K9me2 and DNA methylation in a maize genome. PLoS One. 2014;9(8):e105267. doi:10.1371/journal.pone.0105267.
81. Stegemann S, Hartmann S, Ruf S, Bock R. High-frequency gene transfer from the chloroplast genome to the nucleus. Proc Natl Acad Sci U S A. 2003;100(15):8828–33. doi:10.1073/pnas.1430924100.
82. Roark LM, Hui AY, Donnelly L, Birchler JA, Newton KJ. Recent and frequent insertions of chloroplast DNA into maize nuclear chromosomes. Cytogenet Genome Res. 2010;129(1-3):17–23. doi:10.1159/000312724.
83. Huang CY, Grunheit N, Ahmadinejad N, Timmis JN, Martin W. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiol. 2005;138(3):1723–33. doi:10.1104/pp.105.060327.
84. Lei M, Zhang H, Julian R, Tang K, Xie S, Zhu JK. Regulatory link between DNA methylation and active demethylation in Arabidopsis. Proc Natl Acad Sci U S A. 2015;112:3553–7. doi:10.1073/pnas.1502279112.
85. Williams BP, Pignatta D, Henikoff S, Gehring M. Methylation-sensitive expression of a DNA demethylase gene serves as an epigenetic rheostat. PLoS Genet. 2015;11(3):e1005142. doi:10.1371/journal.pgen.1005142.
86. O'Malley RC, Huang SS, Song L, Lewsey MG, Bartlett A, Nery JR, et al. Cistrome and epicistrome features shape the regulatory DNA landscape. Cell. 2016;165(5):1280–92. doi:10.1016/j.cell.2016.04.038.
87. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet. 2013;45(8):891–8. doi:10.1038/ng.2684.
88. Turco G, Schnable JC, Pedersen B, Freeling M. Automated conserved non-coding sequence (CNS) discovery reveals differences in gene content and promoter evolution among grasses. Front Plant Sci. 2013;4:170. doi:10.3389/fpls.2013.00170.
89. Ong-Abdullah M, Ordway JM, Jiang N, Ooi SE, Kok SY, Sarpan N, et al. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. Nature. 2015;525:533–7. doi:10.1038/nature15365.

90. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, et al. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science. 2006;313(5793):1596–604. doi:10.1126/science.1128691.
91. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31. doi:10.1186/1471-2105-6-31.
92. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. http://www.repeatmasker.org/. 2013-2015. Accessed 7 Jul 2016.
93. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11. doi:10.1186/s13100-015-0041-9.
94. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449(7161):463–7. doi:10.1038/nature06148.
95. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, et al. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res. 2007;35(Database issue):D883–7. doi:10.1093/nar/gkl976.
96. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The Sorghum bicolor genome and the diversification of grasses. Nature. 2009;457(7229):551–6. doi:10.1038/nature07723.
97. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J, Jin D, et al. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature. 2012;492(7429):423–7. doi:10.1038/nature11798.
98. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009; 326(5956):1112–5. doi:10.1126/science.1178534.
99. Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, et al. Draft genome sequence of the oilseed species Ricinus communis. Nat Biotechnol. 2010; 28(9):951–6. doi:10.1038/nbt.1674.
100. International Brachypodium I. Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature. 2010;463(7282):763–8. doi:10.1038/nature08747.
101. International Peach Genome I, Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, et al. The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet. 2013;45(5):487–94. doi:10.1038/ng.2586.
102. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463(7278):178–83. doi:10.1038/nature08670.
103. Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, et al. A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet. 2014;46(7):707–13. doi:10.1038/ng.3008.
104. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, et al. The genome of the domesticated apple (Malus x domestica Borkh.). Nat Genet. 2010;42(10):833–9. doi:10.1038/ng.654.
105. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, et al. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet. 2011;43(5):476–81. doi:10.1038/ng.807.
106. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, et al. The genome of woodland strawberry (Fragaria vesca). Nat Genet. 2011;43(2):109–16. doi:10.1038/ng.740.
107. Young ND, Debelle F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, et al. The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature. 2011;480(7378):520–4. doi:10.1038/nature10625.
108. Prochnik S, Marri PR, Desany B, Rabinowicz PD, Kodira C, Mohiuddin M, et al. The cassava genome: current progress, future directions. Trop Plant Biol. 2012;5(1):88–94. doi:10.1007/s12042-011-9088-z.
109. Tomato Genome C. The tomato genome sequence provides insights into fleshy fruit evolution. Nature. 2012;485(7400):635–41. doi:10.1038/nature11119.
110. Hellsten U, Wright KM, Jenkins J, Shu S, Yuan Y, Wessler SR, et al. Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. Proc Natl Acad Sci U S A. 2013;110(48):19478–82. doi:10.1073/pnas.1319032110.
111. Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Livingstone 3rd D, Cornejo O, et al. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol. 2013;14(6):r53. doi:10.1186/gb-2013-14-6-r53.
112. Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, Guo YL, et al. The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat Genet. 2013;45(7):831–5. doi:10.1038/ng.2669.
113. Yang R, Jarvis DE, Chen H, Beilstein MA, Grimwood J, Jenkins J, et al. The reference genome of the halophytic plant Eutrema salsugineum. Front Plant Sci. 2013;4:46. doi:10.3389/fpls.2013.00046.
114. Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, et al. The genome of Eucalyptus grandis. Nature. 2014;510(7505):356–62. doi:10.1038/nature13308.
115. Parkin IA, Koh C, Tang H, Robinson SJ, Kagale S, Clarke WE, et al. Transcriptome and methylome profiling reveals relics of genome dominance in the mesopolyploid Brassica oleracea. Genome Biol. 2014;15(6):R77. doi:10.1186/gb-2014-15-6-r77.
116. Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, Murat F, et al. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Nat Biotechnol. 2014;32(7):656–62. doi:10.1038/nbt.2906.
117. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015;523:212–6. doi:10.1038/nature14465.
118. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–2.
119. Langmead B. Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics. 2010;Chapter 11:Unit 11 7. doi: 10.1002/0471250953.bi1107s32.
120. Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2014;10(4):e1003537. doi:10.1371/journal.pcbi.1003537.
121. Mirarab S, Nguyen N, Guo S, Wang LS, Kim J, Warnow T. PASTA: Ultra-large multiple sequence alignment for nucleotide and amino-acid sequences. J Comput Biol. 2015;22(5):377–86. doi:10.1089/cmb.2014.0156.
122. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17(4):540–52.
123. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003;52(5):696–704. doi:10.1080/10635150390235520.
124. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012;9(8):772. doi:10.1038/nmeth.2109.
125. Paradis E, Claude J, Strimmer K. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20(2):289–90. doi:10.1093/bioinformatics/btg412.
126. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 2012;3(2):217–23. doi:10.1111/j.2041-210X.2011.00169.x.
127. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2. doi:10.1093/bioinformatics/btq033.
128. Chavez Montes RA, de Fatima R-CF, De Paoli E, Accerbi M, Rymarquis LA, Mahalingam G, et al. Sample sequencing of vascular plants demonstrates widespread conservation and divergence of microRNAs. Nat Commun. 2014;5:3722. doi:10.1038/ncomms4722.
129. Wang M, Yuan D, Tu L, Gao W, He Y, Hu H, et al. Long noncoding RNAs and their proposed functions in fibre development of cotton (Gossypium spp.). New Phytol. 2015;207(4):1181–97. doi:10.1111/nph.13429.
130. Stocks MB, Moxon S, Mapleson D, Woolfenden HC, Mohorianu I, Folkes L, et al. The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. Bioinformatics. 2012;28(15):2059–61. doi:10.1093/bioinformatics/bts311.
131. Prufer K, Stenzel U, Dannemann M, Green RE, Lachmann M, Kelso J. PatMaN: rapid alignment of short sequences to large databases. Bioinformatics. 2008;24(13):1530–1. doi:10.1093/bioinformatics/btn223.
132. Grossmann S, Bauer S, Robinson PN, Vingron M. Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. Bioinformatics. 2007;23(22):3024–31. doi:10.1093/bioinformatics/btm440.
133. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7. doi:10.1093/nar/gkh340.
134. Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol. 2000;17(1):32–43.
135. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics. 1997;13(5):555–6. doi:10.1093/bioinformatics/13.5.555.
136. Brown AP, Kroon JT, Swarbreck D, Febrer M, Larson TR, Graham IA, et al. Tissue-specific whole transcriptome sequencing in castor, directed at understanding triacylglycerol lipid biosynthetic pathways. PLoS One. 2012; 7(2):e30100. doi:10.1371/journal.pone.0030100.

137. Chodavarapu RK, Feng S, Ding B, Simon SA, Lopez D, Jia Y, et al. Transcriptome and methylome interactions in rice hybrids. Proc Natl Acad Sci U S A. 2012; 109(30):12040–5. doi:10.1073/pnas.1209297109.
138. Perazzolli M, Moretto M, Fontana P, Ferrarini A, Velasco R, Moser C, et al. Downy mildew resistance induced by Trichoderma harzianum T39 in susceptible grapevines partially mimics transcriptional changes of resistant genotypes. BMC Genomics. 2012;13:660. doi:10.1186/1471-2164-13-660.
139. Tang S, Dong Y, Liang D, Zhang Z, Ye C-Y, Shuai P, et al. Analysis of the drought stress-responsive transcriptome of black cottonwood (Populus trichocarpa) using deep RNA sequencing. Plant Mol Biol Report. 2014;33(3): 424–38. doi:10.1007/s11105-014-0759-4.
140. Livingstone D, Royaert S, Stack C, Mockaitis K, May G, Farmer A, et al. Making a chocolate chip: development and evaluation of a 6K SNP array for Theobroma cacao. DNA Res. 2015;22(4):279–91. doi:10.1093/dnares/dsv009.
141. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36. doi:10.1186/gb-2013-14-4-r36.
142. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012;7(3):562–78. doi:10.1038/nprot.2012.016.