TITLE: The evolution of CHROMOMETHYLTRANSFERASES and gene body DNA methylation in plants

RUNNING TITLE: CMT gene family in plants

Adam J. Bewick[1]*, Chad E. Niederhuth[1], Nicolas A. Rohr[1], Patrick T. Griffin[1], Jim Leebens-Mack[2], Robert J. Schmitz[1]

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

CO-CORRESPONDING AUTHORS: Adam J. Bewick, bewickaj@uga.edu and Robert J. Schmitz, schmitz@uga.edu

KEY WORDS: CHROMOMETHYLTRANSFERASE, Phylogenetics, DNA methylation, WGBS

WORD COUNT: ≈4582

TABLE COUNT: 0 main, 1 SI

FIGURE COUNT: 5 main, 5 SI

ABSTRACT

The evolution of gene body methylation (gbM) and the underlying mechanism is poorly understood. By pairing the largest collection of CHROMOMETHYLTRANSFERASE (CMT) sequences (773) and methylomes (72) across land plants and green algae we provide novel insights into the evolution of gbM and its underlying mechanism. The angiosperm- and eudicot-specific whole genome duplication events gave rise to what are now referred to as CMT1, 2 and 3 lineages. CMTε, which includes the eudicot-specific CMT1 and 3, and orthologous angiosperm clades, is essential for the perpetuation of gbM in angiosperms, implying that gbM evolved at least 236 MYA. Independent losses of CMT1, 2 and 3 in eudicots, and CMT2 and CMTε$^{\text{monocot+magnoliid}}$ in monocots suggests overlapping or fluid functional evolution. The resulting gene family phylogeny of CMT transcripts from the most diverse sampling of plants to date redefines our understanding of CMT evolution and its evolutionary consequences on DNA methylation.

INTRODUCTION

DNA methylation is an important chromatin modification that protects the genome from selfish genetic elements, is sometimes important for proper gene expression, and is involved in genome stability. In plants, DNA methylation is found at cytosines in three sequence contexts: CG, CHG, and CHH (H is any base, but G). Establishment and maintenance of DNA methylation at these three sequence contexts is performed by a suite of distinct *de novo* and maintenance DNA methyltransferases, respectively. CHROMOMETHYLTRANSFERASES (CMTs) are an important class of plant-specific DNA methylation maintenance enzymes, which are characterized by the presence of a CHRromatin Organisation MOdifier (CHROMO) domain between the cytosine methyltransferase catalytic motifs I and IV[1]. Identification, expression and functional characterization of CMTs have been extensively performed in the model plant *Arabidopsis thaliana*[2-4] and in the model grass species *Zea mays*[5]. Homologous CMTs have been identified in other flowering plants[6-8], the moss *Physcomitrella patens*, the lycophyte *Selaginella moellendorffii*, and the green algae *Chlorella* NC64A and *Volvox carteri*[9]. There are three CMT genes encoded in *A. thaliana*: CMT1, CMT2, and CMT3[2,10-12]. CMT1 is the least studied of the three chromomethyltransferases as a handful of *A. thaliana* accessions contain an Evelknievel retroelement insertion or a frameshift mutation truncating the protein, which suggested that CMT1 is nonessential[10]. The majority of DNA methylation present in pericentromeric regions of the genome composed of long transposable elements is targeted by a CMT2-dependent pathway[4,3]. Allelic variation at CMT2 has been shown to alter genome-wide levels of CHH DNA methylation, and plastic alleles of CMT2 may play a role in adaptation to temperature[13-15]. Whereas DNA methylation at CHG sites is often maintained by

CMT3 through a reinforcing loop with histone H3 lysine 9 di-methylation (H3K9me2) catalyzed by the KRYPTONITE (KYP)/SUVH4 lysine methyltransferase[5,16].

A large number of plant genes exclusively contain CG DNA methylation in the transcribed region and a depletion of CG DNA methylation from both the transcriptional start and stop sites (referred to as "gene body DNA methylation", or "gbM")[17-19]. The penetrance of this DNA methylation pattern is strong, and can be observed without exctracting gbM genes from metaplots[20]. The current model for the evolution of gbM relies on rare transcription-coupled incorporation/methylation of histone H3K9me2 in gene bodies with subsequent failure of INCREASED IN BONSAI METHYLATION 1 (IBM1) to de-methylate H3K9me2[21]. This provides a substrate for CMT3 to bind and methylate DNA and through an unknown mechanism leads to CG DNA methylation, which is maintained over evolutionary timescales by CMT3 and through DNA replication by MET1. Methylated DNA then provides a substrate for KYP and related family members, which increases the rate at which H3K9 is methylated[21,22]. Finally, gene body methylation spreads throughout the gene over evolutionary time[21]. Support for this model is evident in the eudicot species *Eutrema salsugineum* and *Conringia planisiliqua* (family Brassicaceae), which have independently lost CMT3 and gbM[21]. The loss of CMT3 in these species phenocopies the reduced levels of CHG DNA methylation found in the *A. thaliana cmt3* mutants[23]. Closely related Brassicaceae species have reduced levels of CHG DNA methylation on a per cytosine basis and have reduced gbM relative to other species, but still posses CMT3[20,21], which indicates changes at the molecular level may have disrupted function of CMT3.

Previous phylogenetic studies have proposed that CMT1 and CMT3 are more closely related to each other than to CMT2, and that ZMET2 and ZMET5 proteins are more closely related to CMT3 than to CMT1 or CMT2[6], and the placement of non-seed plant CMTs more closely related to CMT3[24]. However, these studies were not focused on resolving phylogenetic relationships within the CMT gene family, but rather relationships of CMTs between a handful of species. These studies have without question laid the groundwork to understand CMT-dependent DNA methylation pathways and patterns in plants. However, the massive increase in transcriptome data for a broad sampling of plant species together with advancements in sequence alignment and phylogenetic inference algorithms have made it possible to incorporate thousands of sequences into a single phylogeny, allowing for a more complete understanding of the CMT gene family. A comprehensive plant CMT gene phylogeny has implications for mechanistic understanding of DNA methylation across the plant kingdom. Additionally, advancements in high throughput screening of DNA methylation has made it possible to get accurately estimate genome-wide levels from species without a sequenced reference assembly[25]. Understanding the evolutionary relationships of CMT proteins is foundational for inferring the evolutionary origins, maintenance, and consequences of genome-wide DNA methylation and gbM.

Here we investigate phylogenetic relationships of CMTs at a much larger evolutionary timescale using data generate from the 1KP Consortium. In the present study we have analyzed 773 mRNA transcripts from 443 different species, identified as belonging to the CMT gene family, from an extensive taxonomic sampling including eudicots (basal, core, rosid, and asterid), basal angiosperms, monocots and commelinids, magnoliids, gymnosperms (conifers, cycadales, ginkgoales), monilophytes (ferns and fern allies), lycophytes, bryophytes (mosses, liverworts and hornworts) and green algae. CMT homologs identified in all major land plant lineages and green algae, indicate that CMT genes originated prior to the origin of land plants (≥480 MYA)[26-29]. In addition, phylogenetic relationships suggests at least two duplication events occurred within the angiosperm lineage giving rise to the CMT1, CMT2, and CMT3 gene clades. In the light of CMT evolution we explored patterns of genomic and genic DNA methylation levels in 72 species of land plants, revealing diversity of the epigenome within and between major taxonomic groups, and the evolution of gbM in association with the origin of the CMTε gene clade.

RESULTS AND DISCUSSION

The origins of CHROMOMETHYLTRANSFERASES. CMT proteins are found in most major taxonomic groups of land plants and some algae: eudicots (basal, core, rosid, and asterid), basal angiosperms, monocots and commelinids, magnoliids, gymnosperms (conifers, cycadales, ginkgoales), ferns (leptosporangiates, eusporangiates), lycophytes, mosses, liverworts, hornworts and green algae (Fig. 1A and B and Table S1). CMT genes were not identified in transcriptome data sets for species outside of green plants (Viridiplantae) including species within the glaucophyta, red algae, dinophyceae, chromista, and euglenozoa, as CMT genes were only identified in a few green algae lineages. Interestingly, CMT genes were not sampled from three species within the gymnosperm order Gnetales. A transcript with CHROMO and C-5 cytosine-specific DNA methylase domains was identified in *Welwitschia mirabilis* RNA-seq data, but this transcript did not include a Bromo Adjacent Homology (BAH) domain. The BAH domain is an interaction surface that is required to capture H3K9me2, and mutations that abolish this interaction causes a failure of a CMT protein (e.g., ZMET2) binding to nucleosomes, and a complete loss of activity *in vivo*[5]. Therefore, although present, it may represent a nonfunctional copy of a CMT protein. Alternatively, it may represent an incomplete transcript. Most of the species without a CMT protein are red algae, which may have lost CMT or CMT evolved after the diversification from green algae[30,31].

The relationships among CMTs suggests that CMT2 and CMTε lineages arose from a duplication event within a common ancestor of all flowering plant species (Fig. 1C). Relationships among non-angiosperm CMTs largely recapitulate species relationships[31] but the existence of two distinct clades for both gymnosperms and ferns suggesting gene duplication and loss events early in vascular plant evolution. The non-angiosperm CMT protein-coding genes analyzed here include those previously identified in the lycophyte *S. moellendorffii*[9] and the moss *P. patens*[9,33]. CMTs identified in a few green algae from the 1KP sequencing data also belong to this clade (Fig. S1). However, previously identified CMT proteins in *Chlamydomonas reinhardtii*, *Chlorella* NC64A and *Volvox carteri* were not included in the CMT gene family clade because they lacked the CHROMO and other domains typically associated with CMT proteins (Fig. S2B), and *C. reinhardtii* and *V. carteri* sequences are more homologous to METHYTRANSFERASE 1 (MET1) than other green algae CMT (Table S1). The greatly increased sampling of non-angiosperm CMTs defines the understanding of relationships of CMTs in early land plants and plants in general[9,24,33,34].

Further diversification of CMT proteins occurred in the eudicots; a second duplication gave rise to what is now called CMT1 and CMT3 (Fig. 1C). Thus, CMT1 and CMT3 are eudicot-specific, and the monocot and magnoliid CMTs (CMTε$^{\text{monocot+magnoliid}}$) are sister to both CMT1 and CMT3 (Fig. 1C). CMT1, CMT3 and CMTε$^{\text{monocot+magnoliid}}$ share a common ancestor with basal angiosperms (CMTε$^{\text{basal angiosperm}}$). This and the previously mentioned angiosperm-specific duplication events may have coincided with the ancestral angiosperm whole genome duplication (WGD) event or the eudicot WGD event, respectively[35]. Not all eudicots possess CMT1, CMT2 and CMT3, but rather species exhibit CMT gene content ranging from zero to three of these homologs, suggesting multiple independent losses and fluidity of the functionality of these proteins (Fig. S3 and Table S1).

Functional differences among these paralogs has been characterized in *A. thaliana*; CMT3 is functional and is required for the maintenance of DNA methylation at CHG sites, however allelic diversity of CMT1 is high, which suggests alleles are segregating neutrally or under relaxed selection in populations and may not serve a function in the maintenance of CHG DNA methylation. Additionally, CMT1 does not compensate CMT3 function *in vivo* in *A. thaliana Δcmt3*[23]. The expression and persistence of CMT1 in numerous other eudicots raises the possibility of functional divergence and convergence over the evolutionary history of eudicots. For example, CMT1 may have evolved a novel function (neofunctionalization), CMT1 and CMT3 may both be required to achieve the same phenotype (subfunctionalization) or CMT1 may have a redundant function to that of CMT3 (redundancy). The latter possibility does not seem to be the case for naturally occurring *Δcmt3* species, *E. salsugineum*, as CMT1 does not recover levels of CHG DNA methylation as expected[21]. However, subfunctionalization of CMT1 in *E. salsugineum* seems plausible since DNA methylation at CHG sites still occurs at low levels[21]. The exact fate of CMT1 and interplay between these two paralogs in shaping the epigenome remains unknown.

The *Zea mays*-specific ZMET2 and ZMET5, and closely related CMT proteins in other monocots, monocots/commelinids, and magnoliids form a well-supported monophyletic clade with bootstrap support of 95% (CMTε$^{monocot+magnoliid}$; Fig. 1). The inclusion of magnoliids in the CMTε$^{monocot+magnoliid}$ clade is interesting because monophyletic support of the magnoliidae as basal to all monocots has been hypothesized[36,37]. Akin to eudicots, monocots possess combinations of CMTε$^{monocot+magnoliid}$ and CMT2 (Fig. S4). For example, the model grass species *Z. mays* has loss CMT2, whereas the closely related species *Sorghum bicolor* possess both CMTε$^{monocot+magnoliid}$ and CMT2 (Fig. S3). CMTε$^{monocot+magnoliid}$ is not strictly homologous to CMT3, and represents a unique monophyletic group that is co-orthologous to CMT1 and CMT3. However, ZMET2 is functionally orthologous to CMT3 and maintains DNA methylation at CHG sites[5]. But, unlike CMT3, ZMET2 is associated with DNA methylation at CHH sites within some loci[38]. Given the inclusion of monocot and magnoliid species the monophyletic CMTε$^{monocot+magnoliid}$ clade, this dual-function is expected to be present in other monocot species, and magnoliid species.

Several monocot/commelinid species possess a CMTε$^{monocot+magnoliid}$ protein with a kinase domain, hereafter referred to as CHROMO-kinase methyltransferase (CKMT) (Fig. 2A). These proteins are restricted to the true grasses (family Poaceae), and species *Panicum hallii*, *Panicum virgatum* and *Setaria viridis* exclusively possess CKMT proteins (Fig. 2B). The relationship among CMTε$^{monocot+magnoliid}$ and CKMT is polyphyletic, and suggests that the addition of a kinase domain was through two independent fusion and deletion events affecting different paralogous genes (Fig. 2C). One fusion event shared by all species in the family Poaceae in paralog α and other in the clade containing *Pan. hallii*, *Pan. virgatum* and *Se. viridis* in paralog β (Fig. 2C). Subsequently, two deletion events occurred in paralog α: one in the clade containing *So. bicolor* and *Z. mays* and the other in the clade containing *Pan. hallii*, *Pan. virgatum* and *Se. viridis* (Fig. 2C). Interestingly, genes with protein kinase domains are in close proximity to CKMT (Table S1). Kinase proteins have shown to be involved in histone phosphorylation in the green algae *C. reinhardtii*[39] and *A. thaliana*[40]. The kinase and DNA methyltransferase domains of CKMTs may coincide with dual functions: histone phosphorylation and DNA methylation. Further, functional exploration of CKMTs in the grasses will help identify its contribution to DNA methylation, chromatin modifications and gene expression.

Overall, these redefined CMT clades, and monophyletic clades of broad taxonomic groups, are well supported with bootstrap support of ≥80 (Fig. 1). Thus, the identification of novel CMT proteins in magnoliids, gymnosperms, bryophytes, lycophytes, liverworts, hornworts, and green algae pushes the timing of evolution of CMT, and potentially certain mechanisms maintaining CHG and CHH DNA methylation, prior to the origin of land plants (≥480)[26-29].

Non-neutral evolution of CMT contributes to the epigenome of flowering plants. Gene body methylation can be found in all angiosperms investigated to date, with the exception of a couple species in the Brassicaceae[20,21]. Shared by all angiosperms are CMTs belonging to the CMTε clade (Fig. 1), suggesting the evolution of this clade is tightly linked to the evolution of gbM in plants. For example, recent work identified independent events that lead to the loss of CMT3 in *E. salsugineum* and *C. planisiliqua*, which is linked to the loss of gbM over evolutionary time[21]. Peculiarly, closely related species with CMT3 – *Brassica oleracea*, *Brassica rapa* and *Schrenkiella parvulum* (Fig. 3 Clade B) – have reduced numbers of gbM loci compared to a sister clade – *A. thaliana*, *Arabidopsis lyrata* and *Capsella rubella* (Fig. 3 Clade C) – and other eudicot species[20]. At the molecular level, CMT3 has evolved under less selective constraint – measured as *dN/dS* (ω) – in the Brassicaceae (Clade A) (ω=0.175) compared to 162 eudicot species (ω=0.0961), and with reduced apparent constraint in the clade containing *B. oleracea*, *B. rapa* and *S. parvulum* (ω=0.241) compared to the clade containing *A. thaliana*, *A. lyrata* and *C. rubella* (ω=0.164) (Fig. 3). A hypothesis of positive selection was not preferred to contribute to the increased rates of ω in either clade (Fig. 3). Relaxed selective constraint may be associated with decreased gbM in some members of the Brassicaceae. The severity of these substitutions within Brassicaceae clades A, B and C (Fig. 3) effects the respective epigenomes differently, and substitutions at CMT3 have progressed over evolutionary time from compromising levels of CHG DNA methylation (Clades A and C) to the number of gbM loci (Clade B).

It is conceivable that additional independent losses or non-neutral evolution of CMTε has shaped the epigenome of other species of plants. Several eudicot species of plants possess truncated annotations of CMT3 or CMT3 was absent from the assembled transcriptomes. These species often possessed low levels (e.g., *Kalanchoe tomentosa*) of CG DNA methylation in gene bodies (Table S1; Fig. S4). However, other species missing CMT3 showed similar patterns of DNA methylation to species with CMT3 (e.g., *S. dulcificum* and *A. thaliana*), which may represent false-negatives (Table S1; Fig. S4). Alternatively, not enough time has passed since the loss of CMT3 to effect levels of CG DNA methylation within gene bodies. Co-orthologous proteins to CMT3, including CMTε$^{\text{monocot+magnoliid}}$ and CMTε$^{\text{basal angiosperm}}$, are similar to CMT3 functionally[5] and at the amino acid level to CMT3 (e.g., ZMET2 shares 417/915 amino acid sites), thus similar phenotypic consequences are expected for naturally occurring CMTε$^{\text{monocot+magnoliid}}$ and CMTε$^{\text{basal angiosperm}}$ mutants. The monocot *Dioscorea elephantipes* is missing CMTε$^{\text{monocot+magnoliid}}$ from its assembled transcriptome, and has low levels of CG DNA methylation within gene bodies (Table S1). However, this observation is from only a couple dozen assembled transcripts.

CG DNA methylation within gene bodies of non-flowering plants is not typical of gbM patterns in angiosperms. CG DNA methylation within gene bodies can be found in most taxonomic groups with lycophyte, moss and liverwort as the exceptions (Fig. 4A). Similar to what has been documented in angiosperms[20] there exists substantial variation of DNA methylation across non-flowering plants, and within taxonomic groups of non-flowering plants (Fig. 4A;

Fig. S4)[41]. Levels of CG DNA methylation are typically much higher than CHG DNA methylation (and CHH DNA methylation) in species with gbM (Fig. 4A)[20]. However, CG and CHG DNA methylation within gene bodies are at a similar level in non-flowering plants – CG:CHG DNA methylation levels are closer to one – especially in gymnosperms (conifer, cycadale, and gnetale) and ferns (eusporangiate and leptosporangiate) (Fig. 4A). Additionally, there is a stronger correlation of gene body levels of CG and CHG DNA methylation in non-flowering plants compared to flowering plants, which suggests a shared association of the mechanism(s) involved in methylating cytosines at these two sequence contexts (Fig. 4B). Furthermore, levels of CG and CHG DNA methylation within gene bodies of non-flowering plants tends to mirror one another (Fig. 4C), further suggesting a relationship between mechanisms that methylate cytosines at CG and CHG sites or a single mechanism that methylates cytosines at both sequence contexts. Thus, enrichment of only CG DNA methylation within gene bodies (i.e., angiosperm-like gbM) seems unlikely, or restricted to a limited number of loci in gymnosperms and other non-flowering plants.

A gradual increase in levels of CG DNA methylation towards the center of the gene body is not observed in non-flowering plants (Fig. 5). On average species belonging to these taxonomic groups – gymnosperms, ferns, lycophyte, moss, fern, and green algae – do not exhibit the typical pattern of minimal CG DNA methylation at the transcriptional start site (TSS) and transcriptional terminate site (TTS) found in species with gbM (Fig. 4B). The opposite is observed for the majority of non-flowering plants, especially at the TTS where a spike in CG DNA methylation occurs (Fig. 5). Green algae represents a different CG methyl-type within gene bodies compared to flowering and non-flowering plants, which is characterized by hypomethylation spreading towards the TTS (Fig. 5), which is more similar to metazoan and/or mammalian CG DNA methylation patterns within gene bodies[9,42]. Qualitatively, gymnosperms have a CG methyl-type that is a mesh of flowering and non-flowering plants; a gradual increase in levels of CG DNA methylation towards the center of gene body and a spike in CG DNA methylation at the TTS (Fig. 5). Two paralogous CMT proteins are present in most species of gymnosperms, and independent evolution of a CMT-dependent mechanism for the perpetuation of CG DNA methylation within gene bodies may have occurred with one paralog being opted to perform this function. Some support for this hypothesis is observed in the gnetale *Gnetum gnemon*, which does not possess a full or partial copy of CMT protein(s), and has very low levels of CG (and CHG and CHH) DNA methylation within gene bodies compared to other gymnosperms and other land plants (Fig. S5). However, further investigation is needed into the functionality of CMTs and the relationship to gbM in gymnosperms.

CONCLUSION

In summary, we present the most comprehensive CMT gene-family phylogeny to date. Refined relationships between CMT1, CMT2 and CMT3, and other CMTs have shed light on current models for the evolution of gbM, and provided a framework for further research on the role of CMT in establishment and maintenance of DNA methylation and histone modifications. CMTs are ancient proteins that evolved prior to the diversification of land plants. A shared function of CMTs is the maintenance of non-CG DNA methylation, which has been essential for DNA methylation at long transposable elements in the pericentromeric regions of the genome, and the evolution of gbM in angiosperms.

We hypothesize that the angiosperm-specific duplication ≥236 MYA[35] gave rise to what is now CMT2 and CMTε, and gbM in angiosperms. Furthermore, the eudicot-specific duplication event (≥134 MYA[35]) gave rise to the eudicot-specific proteins CMT1 and CMT3. Eudicot species without CMT3 have loss of gbM, and currently at least three independent losses have been supported with phylogenetic analyses and sequencing data. Non-neutral evolution at CMT3 has played an important role in shaping the epigenome of the Brassicaceae. Reduced selective constraint at CMT3 has reduced levels of CHG DNA methylation in the Brassicaceae, and further reductions in *B. oleracea, B. rapa*, and *S. parvulum* have reduced levels of gbM and the number of gbM loci. Similarly, corresponding phenotypic consequences for naturally occurring mutants of co-orthologous proteins to CMT3, CMTε$^{monocot+magnoliid}$, has been observed in the monocot *D. elephantipes*. Also, it seems plausible that the loss of CMT would lead to reductions in the histone modification H3K9me2, and this could subsequently lead to genomic structural consequences. The majority of non-flowering plants do not possess signatures of DNA methylation within gene bodies that is typical of species with gbM. Convergent evolution of gene body methylation, through a CMT mechanism, may be present in species outside of flowering plants, however support for this hypothesis remains sparse.

CMTs have diversified during land plant evolution, and given this diversity it seems plausible that divergent functions among homologs and paralogs has occurred. The function of some homologs have been well studied in model species of plants including *A. thaliana* and *Z. mays*, and studies using natural epigenomic variation have revealed novel functions of CMTε in the evolution of gbM. This study has opened the door for future functional studies of CMT1 in eudicots, CKMT in true grasses, and CMTα in basal land plants. Also, it is possible that convergent evolution of mechanisms leading to gbM in species outside of angiosperms could occur. These future studies are essential to understanding the role of CMTs in DNA methylation and histone modifications, and discovery of novel mechanisms.

MATERIALS AND METHODS

1KP sequencing, transcriptome assembling and orthogrouping. Transcriptome data from the One Thousand Plants (1KP) Consortium for a total of 1329 species were included in this analysis, including both amino acid and coding sequence (Table S1). Additionally, gene annotations from 20 additional species – *A. lyrata*, *Br. distachyon*, *B. oleracea*, *B. rapa*, *Cit. clementina*, *Ca. rubella*, *Can. sativa*, *C. sativus*, *E. salsugineum*, *F. vesca*, *G. max*, *Go. raimondii*, *L. japonicus*, *M. domestica*, *Me. truncatula*, *Pan. hallii*, *Pan. virgatum*, *R. communis*, *Se. viridis*, and *Z. mays* – from Phytozome were included. The CMT gene family was extracted from the previously compiled orthogroupings using the *A. thaliana* gene identifier for CMT1, CMT2 and CMT3. This orthogroup determined by the 1KP Consortium included all three *A. thaliana* CMT proteins, and a total of 5383 sequences. Sequences from species downloaded from Phytozome, that were not included in sequences generated by 1KP, were included to the gene family through reciprocal best BLAST with *A. thaliana* CMT1, CMT2 and CMT3. In total the CMT gene family included 5449 sequences from 1043 species. We used the protein structure of *A. thaliana* as a reference to filter the sequences found within the CMT gene family. Sequences were retained if the included the same base pfam domains as *A. thaliana* – CHROMO (CHRromatin Organisation MOdifier) domain, BAH domain, and C-5 cytosine-specific DNA methylase – as identified by Interproscan[43]. These filtered sequences represent a set of high-confident, functional, ideal CMT proteins, which included 773 sequences from 432 species, and were used for phylogenetic analyses.

Phylogeny construction. To estimate the gene tree for the CMT sequences, a series of alignment and phylogenetic estimation steps were conducted. An initial protein alignment was carried out using Pasta with the default settings[44]. The resulting alignment was back-translated using the CDS sequence into in-frame codons using a custom Perl script. A phylogeny was estimated by RAxML[45] (-m GTRGAMMA) with 1000 rapid bootstrap replicates using the in-framed aligned sequences, and with only the first and second codon positions. Long branches can effect parameter estimation for the substitution model, which can in turn degrade phylogenetic signal. Therefore, phylogenies were constructed with and without green algae species, and were rooted to the green algae clade or liverworts, respectively. The species *Balanophora fungosa* has been reported to have a high substitution rate, which can also produce long branches, and was removed prior to phylogenetic analyses.

Codon analysis. Similar methodology as described above was used to construct phylogenetic trees for testing hypotheses on the rates of evolution in a phylogenetic context. However, the program Gblocks[46] was used to identify conserved amino acids in codon alignments. The parameters for Gblocks were kept at the default settings, except we allowed for %50 gapped positions. The program Phylogenetic Analysis by Maximum Likelihood (PAML)[47] was used to test branches (branch test) and sites along branches (branch-site test) for deviations from the background rate of molecular evolution (*dN/dS*; ω) and for deviations from the neutral expectation, respectively. Branches tested and a summary of each test can be found in Table S1.

MethylC-seq. MethylC-seq libraries were prepared according to the following protocol[48]. Prior to mapping each transcript was searched for the longest open reading frame from all six possible frames, and only transcripts beginning with a start codon and ending with one of the three stop codons were kept. All sequencing data for each species was aligned to their respective transcriptome or species within the same genus using the methylpy pipeline[49], which can be found in Table S1. Weighted methylation was calculated for each sequence context (CG, CHG and CHH) by dividing the total number of aligned methylated reads by the total number of methylated plus un-methylated reads. Since, per site sequencing coverage was low – on average ~1X – subsequent binomial tests could not be performed to bin genes as gbM[20]. To investigate the affect of low coverage we compared levels of DNA methylation of 1X randomly sampled MethylC-seq reads to actual levels for 33 angiosperm species and Chlorella NC64A[9,20]. On average DNA methylation levels determined from 1X sequencing coverage are 1.28, 2.63, and 2.17 times higher at CG, CHG and CHH sequence contexts compared to the actual levels within coding regions. These values were used to adjust levels of DNA methylation within coding regions for species of plants sequenced in this study. Genome-wide levels of DNA methylation were estimated using *FAST^{m}C* and the *plant* model[25].

Metagene plots. The gene body – start to stop codon – was divided into 20 windows and weighted methylation levels were calculated for each window. The mean weighted methylation for each window was then calculated for all genes and plotted in R v3.2.4. Genic CHG DNA methylation often paralleled levels of CG DNA methylation in gymnosperms and ferns, thus to investigate the contribution of only CG DNA methylation, and the contribution of CG DNA methylation via gbM mechanisms, reads with CHG methylation were removed. Additionally, different levels of CG DNA methylation within genes between species can obscure the distribution and the qualitative assessment of gbM. To overcome this obstacle the level of CG DNA methylation was standardized to the bin with the highest level for each species. Thus, making the levels of CG DNA methylation within each species comparable across species. Several species were removed from the metaplot analysis due to poor mapping of MethylC-seq reads compared to other species, and these included *A. lyrata*, *Ca. rubella*, and *P. persica*.

ACKNOWLEDGEMENTS

We thank Nathan Springer for comments and discussions as well as Kevin Tarner (UGA greenhouse), and Michael Wenzel and Ron Determann (Atlanta Botanical Gardens) for plant tissue. We also thank the Georgia Genomics Facility

(GGF) for sequencing. Computational resources were provided by the Georgia Advanced Computing Resource Center (GACRC). We thank Gane Ka-Shu Wong and the 1000 Plants initiative (1KP, onekp.com) for advanced access to transcript assemblies. This work was supported by the National Science Foundation (NSF) (MCB–1402183) and by The Pew Charitable Trusts to R.J.S.

FIGURE LEGENDS

Figure 1. Phylogenetic relationships among CMTs in land plants. (A) CMTs are separated into five monophyletic clades and one polyphyletic clade based on bootstrap support and placement within a phylogenetic context: the superclade CMTε with subclades CMT1, CMT3, CMT$\varepsilon^{\text{monocot+magnoliid}}$ and CMT$\varepsilon^{\text{basal angiosperm}}$, CMT2, and CMTα. CMT1 and CMT3 clades only contain eudicot species of plants suggesting a eudicot-specific duplication event that occurred after the divergence of eudicots from monocots and monocots/commelinids. Sister to CMT1 and CMT3 is the monophyletic group CMT$\varepsilon^{\text{monocot+magnoliid}}$, which contains monocots, monocots/commelinids, and magnoliids species, and subsequently basal-angiosperms (CMT$\varepsilon^{\text{basal angiosperm}}$). CMT2 is sister to CMT1 and CMT3, and contains all major taxonomic groups of land plants and green algae (Fig. S1). Lastly, the polyphyletic CMTα is ancestral to all previously mentioned clades. This delineation is somewhat *ad hoc*, and it seems plausible that CMTε or CMT2 would include species of gymnosperms (conifers, cycadales, and gnetales) and ferns (eusporangiate and leptosporangiate). (B) A collapsed CMT gene family tree showing the six major clades. Pie charts represent species diversity within each clade, and are scaled to the number of species. (C) Two duplication events shared by all angiosperms (ε) and eudicots (Γ) gave rise to what is now referred to as CMT1, CMT2 and CMT3. Values at nodes in (A) and (B) represent bootstrap support from 1000 replicates, and (A) was rooted to the clade containing all liverwort species. The duplication events in (C) correspond to what was reported by Jiao et al. (2011).

Figure 2. CHROMO-kinase methyltransferase (CKMT) proteins are unique to true grasses (family Poaceae). (A) Several species belonging to the family Poaceae possess a CMT with a protein tyrosine kinase domain (e.g., *O. sativa* and *Pan. hallii*) compared to other Poaceae (e.g., *Z. mays*). (B) Species with CKMTs are polyphyletic within the CMT$\varepsilon^{\text{monocot+magnoliid}}$ clade. (C) Relationships among grasses with and without CKMT suggests a duplication event followed by two fusion (+) and two loss events (grey) occurred. The presence of a kinase and DNA methylase domain suggests these novel proteins could play a role in both histone 3 threonine 3 phosphorylation (H3T3ph) and DNA methylation. Shaded circles at nodes in (B) represent bootstrap support from 1000 replicates, and the tree was rooted to *Ph. patens* and *Sel. moellendorffii*.

Figure 3. Non-neutral evolution of CMT3 in the Brassicaceae. Branch and branch-site tests for violations of the neutral expectation were performed in the Brassicaceae for CMT3 (bolded). An overall higher rate of molecular evolution measured as the number of non-synonymous substitutions per non-synonymous divided by the number of synonymous substitutions per synonymous (*dN/dS* or ω) were detected in the Brassicaceae (Clade A). Also, a higher rate ratio of ω was detected in the Brassicaceae (B) clade containing B. rapa and closely related species compared to the clade (C) containing A. thaliana and closely related species. The higher rate ratio in clade B, compared the background branches, was not attributed to positive selection (chart). Therefore, relaxed selective constraint on CMT3 in the Brassicaceae may be contributing to reduced levels of CHG, and reduced levels of gbM in *B. rapa*, *B. oleracae* and *S. parvulum*.

Figure 4. Gene-body methylation (gbM) is unique to flowering plants. (A) DNA methylation at CG, CHG, and CHH sites within gene bodies can be found at the majority of species investigated. DNA methylation within gene bodies at any sequence context was not observed in lycophyte, moss and liverwort. Variation of DNA methylation levels within gene bodies at all sequence contexts is high across all land plants, and within major taxonomic groups. CG DNA methylation levels are typically higher than CHG, followed by CHH. However, levels of CG and CHG DNA methylation within genes are similar in several gymnosperms and ferns, and the ratio of CG:CHG is significantly lower (T-test) in gymnosperms and ferns compared to flowering plants. (B) Levels of CG and CHG DNA methylation are highly correlated in non-flowering plants compared to flowering plants. (C) Also, CG and CHG DNA methylation tend to mirror one another throughout the gene bodies of non-flowering plants as opposed to flowering plants. Together these results suggest a relationship between mechanism(s) involved in methylating cytosines at CG and CHG sites within gene bodies of non-flowering plants. Cladogram was obtained from Open Tree of Life[50].

Figure 5. Distribution of CG DNA methylation within gene bodies of non-flowering plants is not indicative of gbM. Levels of CG DNA methylation within gene bodies of species known to possess gbM loci are exemplified by a normal distribution-like pattern with depletions at the TSS and TTS (Eudicot, Monocot, and Basal angiosperm). However, non-flowering plants, which are suspected not to contain gbM loci based on the evolution of CMTε, do not exhibit this type of pattern (Gymnosperm, Fern, Lycophyte, Moss, Liverwort, and Green algae). All non-flowering plants have a spike in CG DNA methylation at the TTS. Gymnosperms possess a methyl-type that shares characteristics to species with, and without gbM: a normal distribution-like pattern of CG DNA methylation with depletions at the TSS, and a spike at the TTS. Error bars represent standard error of the mean.

SUPPLEMENTAL INFORMATION

Figure S1. Phylogenetic relationships among CMTs in land plants and green algae. Similarly to Fig. 1, CMTs are separated into five monophyletic clades and one polyphyletic clade based on bootstrap support and placement within a phylogenetic context: the superclade CMTε with subclades CMT1, CMT3, CMTε$^{\text{monocot+magnoliid}}$ and CMTε$^{\text{basal angiosperm}}$, CMT2, and CMTα. Green algae belong to the CMTα clade. Values at nodes in represent bootstrap support from 1000 replicates, and the tree was rooted to the clade containing all green algae species.

Figure S2. CMT proteins in green algae (*Ch. reinhardtii*, *Chlorella* NC64A, and *V. carteri*) may represent misidentified homologs. (A) A midpoint rooted gene tree constructed from a subset of species and green algae using protein sequences. Previously identified CMT homologs in *Ch. reinhardtii*, *Chlorella* NC64A, and *V. carteri* (JGI accession ids 190580, 52630, and 94056, respectively) have low amino acid sequence similarity to *A. thaliana* CMT compared to other green algae species (Table S1), which is reflected in long branches, especially for *Ch. reinhardtii* and *V. carteri*. Values on branches are raw branch lengths represented as amino acid substitutions per amino acid site. (B) Protein structure of previously identified CMT homologs in *Ch. reinhardtii*, *Chlorella* NC64A, and *V. carteri* and those identified in green algae from the 1KP dataset. Reported CMTs in *Ch. reinhardtii* and *Chlorella* NC64A do not contain CHROMO domains, and the homolog in *V. carteri* does not contain any recognizable pfam domains, however BAH, CHROMO and a DNA methylase domain can all be identified in green algae CMT homologs from the 1KP dataset.

Figure S3. Permutations of presences and absences of CMT in eudicots, and monocots and monocots/commelinids. (A) Eudicot (basal, core, rosid, and asterid) species of plants possess different combinations of CMT1, CMT2, and CMT3. CMT3 was potentially loss from 46/262 (~18%), and CMT1 is found in 106/262 (~40%) of eudicot species sequenced by the 1KP Consortium. Species without CMT3 are predicted to have significantly reduced levels of gbM loci compared to eudicot species with CMT3. The presence of CMT1 in numerous species suggests a yet to be determined functional role of CMT1 in DNA methylation and/or chromatin modification. (B) Similarly to eudicots, monocots and monocots/commelinids have different combinations of CMTε$^{\text{monocot+magnoliid}}$ and CMT2, which may reflect differences in genome structure, and DNA methylation and chromatin modification patterns.

Figure S4. Genome-wide levels of DNA methylation. DNA methylation levels estimated by alignment to a reference genome or predicted by *FAST^{m}C*[25]. Cladogram was obtained from Open Tree of Life[50].

Figure S5. Metagene plots of DNA methylation across gene bodies. DNA methylation levels within all full-length genes for additional species used in this study. All reads were used for generating these plots.

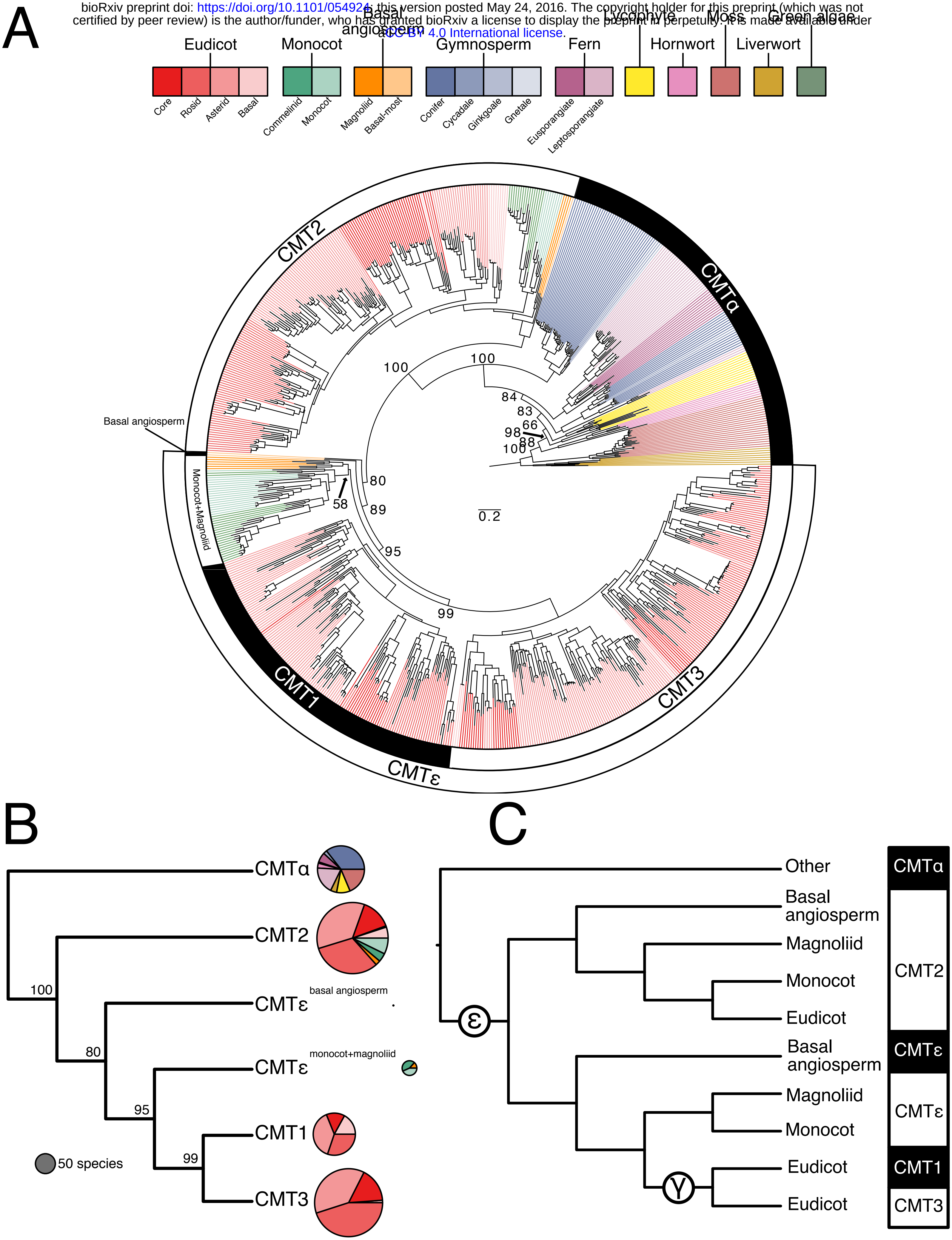
REFERENCES

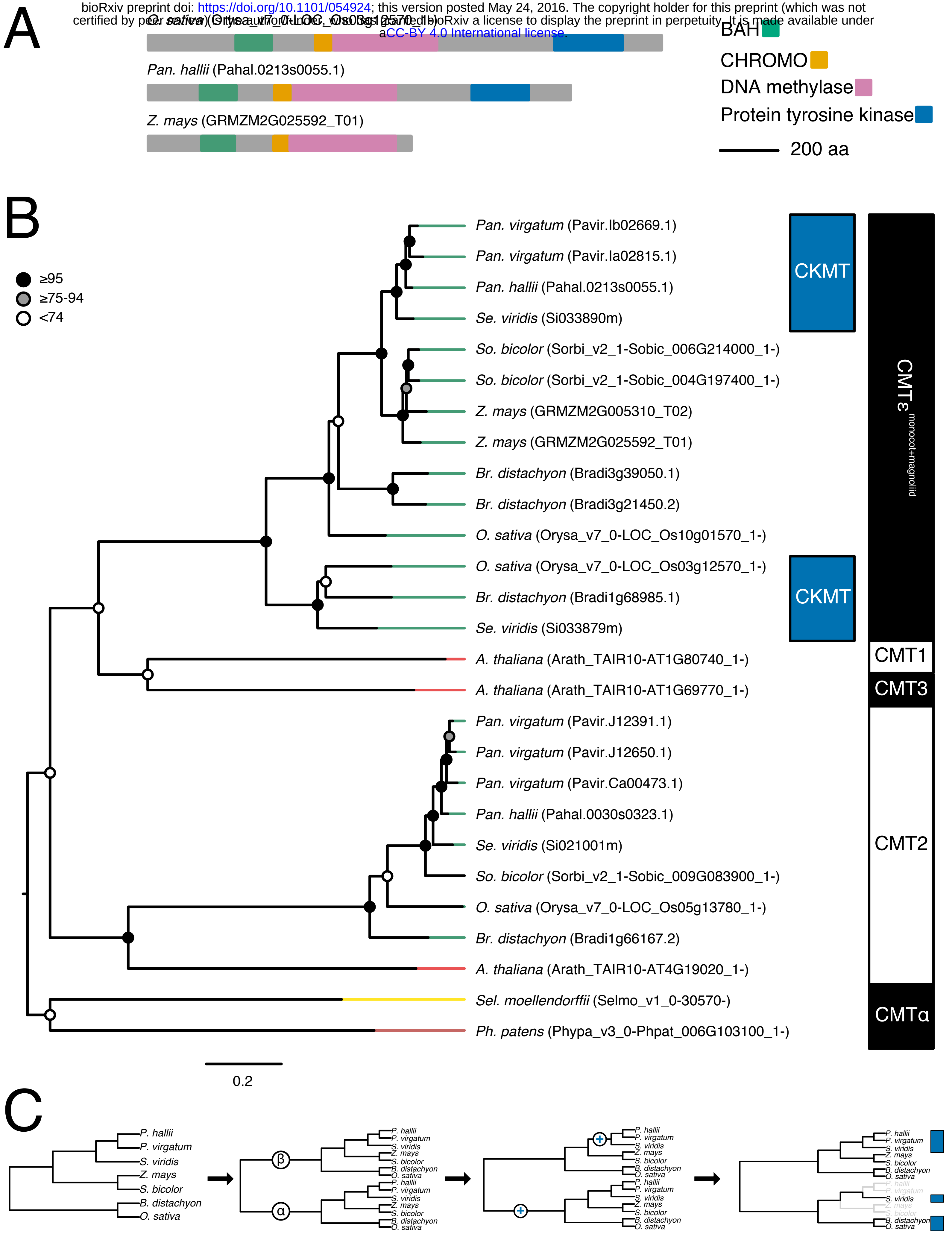
1. Bartee, L., Malagnac, F. & Bender, J. *Arabidopsis* cmt3 chromomethylase mutations block non-CG methylation and silencing of an endogenous gene. Genes Dev 15, 1753–8 (2001).
2. Jackson, J. P., Lindroth, A. M., Cao, X. & Jacobsen, S. E. Control of CpNpG DNA methylation by the KRYPTONITE histone H3 methyltransferase. Nature 416, 556–560 (2002).
3. Zemach, A. et al. The *Arabidopsis* nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell 153, 193–205 (2013).
4. Stroud, H. et al. Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*. Nat Struct Mol Biol 21, 64–72 (2014).
5. Du, J. et al. Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. Cell 151, 167–180 (2012).
6. Papa, C. M., Springer, N. M., Muszynski, M. G., Meeley, R. & Kaeppler, S. M. Maize chromomethylase *Zea* methyltransferase2 Is required for CpNpG methylation. The Plant Cell 13, 1919–1928 (2001).
7. Hou, P. Q. et al. Functional characterization of *Nicotiana benthamiana* chromomethylase 3 in developmental programs by virus-induced gene silencing. Physiologia Plantarum 150, 119–132 (2013).
8. Garg, R., Kumari, R., Tiwari, S. & Goyal, S. Genomic survey, gene expression analysis and structural modeling suggest diverse roles of DNA methyltransferases in legumes. PLoS One 9, e88947 (2014).
9. Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 328, 916–919 (2010).
10. Henikoff, S. & Comai, L. A DNA methyltransferase homolog with a chromodomain exists in multiple polymorphic forms in *Arabidopsis*. Genetics 149, 307–318 (1998).
11. Finnegan, E. J. & Kovac, K. A. Plant DNA methyltransferases. Plant Molecular Biology 43, 189–201 (2000).
12. McCallum, C. M., Comai, L., Greene, E. A. & Henikoff, S. Targeted screening for induced mutations. Nature Biotechnology 18, 455–457 (2000).
13. Shen, X. et al. Natural CMT2 variation is associated with genome-wide methylation changes and temperature seasonality. PLoS Genet 10, e1004842 (2014).
14. Bewick, A. J. & Schmitz, R. J. Epigenetics in the wild. eLife DOI: 10.7554/eLife.07808 (2015).

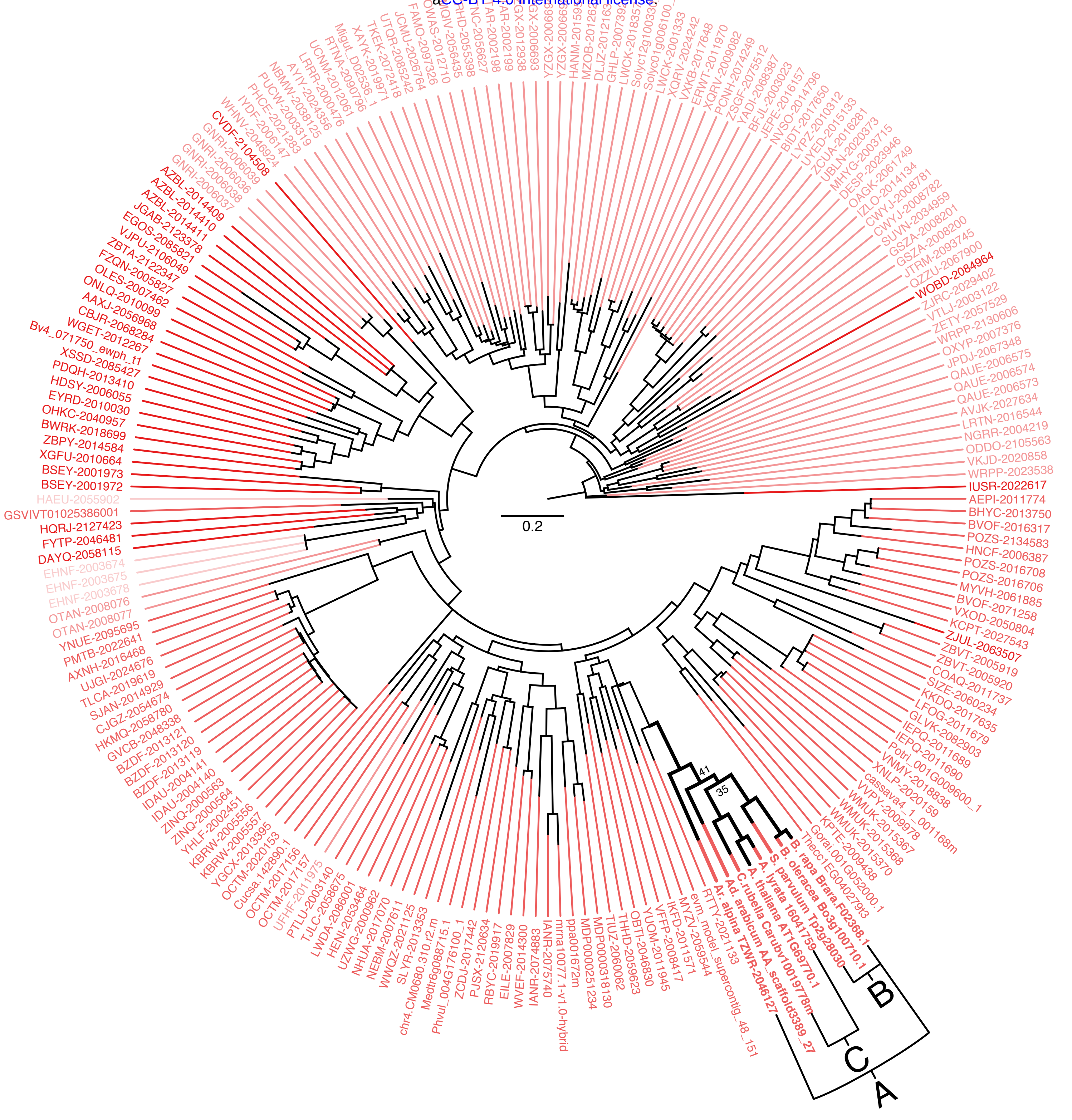
15. Dubin, M. J. et al. DNA methylation in *Arabidopsis* has a genetic basis and shows evidence of local adaptation. eLife DOI: 10.7554/eLife.05255 (2015).
16. Du, J. et al. Mechanism of DNA methylation-directed histone methylation by KRYPTONITE. Molecular Cell 55, 495–504 (2014).
17. Tran, R. K. et al. DNA methylation profiling identifies CG methylation clusters in *Arabidopsis* genes. Curr Biol 15, 154–9 (2005).
18. Zhang, X. et al. Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. Cell 126, 1189–201 (2006).
19. Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T. & Henikoff, S. Genomewide analysis of *Arabidopsis* thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat Genet 39, 61–9 (2007).
20. Niederhuth, C. E., Bewick, A. J. et al. Widespread natural variation of DNA methylation within angiosperms. biorxiv doi: http://dx.doi.org/10.1101/045880 (2016).
21. Bewick, A. J., Ji, L., Niedtherhuth, C. E., Willing, E. et al. On the origin and evolutionary consequences of gene body DNA methylation. biorxiv doi: http://dx.doi.org/10.1101/045542 (2016).
22. Inagaki, S. & Kakutani T. What triggers differential DNA methylation of genes and TEs: contribution of body methylation? Cold Spring Harb Symp Quant Biol. 77, 155-160 (2012).
23. Stroud, H., Greenber, M. V. C., Feng, S., Bernatavichute, Y. V. & Jacobsen, S. E. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell 152, 352–364 (2013).
24. Noy-Malka, C. et al. A single CMT methyltransferase homolog is involved in CHG DNA methylation and development of *Physcomitrella patens*. Plant Mol Biol 84, 719–35 (2014).
25. Bewick, A. J. et al. *FAST^{m}C*: a suite of predictive models for non-reference-based estimations of DNA methylation. G3 6, 447–452 (2015).
26. Kenrick, P. & Crane, P. R. The origin and early evolution of plants on land. Nature 389, 33–39 (1997).
27. Wellman, C. H., Osterloff, P. L. & Mohiuddin, U. Fragments of the earliest land plants. Nature 425, 282–285 (2003).
28. Steemans, P. et al. Origin and radiation of the earliest vascular land plants. Science 324, 353 (2009).
29. Rubinstein, C. V., Gerrienne, P., de la Puente, G. S., Astini, R. A. & Steemans, P. Early Middle Ordovician evidence for land plants in Argentina (eastern Gondwana). New Phytologist 188, 365–369 (2010).
30. Bhattacharya, D. & Medlin, L. Algal phylogeny and the origin of land plants. Plant Physiology 116, 9–15 (1998).
31. Stiller, J. W. & Hall, B. D. The origin of red algae: Implications for plastid evolution. PNAS 94, 4520–4525 (1997).
32. Wickett, N. J. et al. Phylotranscriptomic analysis of the origin and early diversification of land plants. PNAS 111, E4859–68 (2014).

33. Malik, G., Dangwal, M., Kapoor, S. & Kapoor, M. Role of DNA methylation in growth and differentiation in *Physcomitrella patens* and characterization of cytosine DNA methyltransferases. FEBS J 279, 4081–94 (2012).
34. Feng, S. et al. Conservation and divergence of methylation patterning in plants and animals. PNAS 107, 8689–8694 (2010).
35. Jiao, Y. et al. Ancestral polyploidy in seed plants and angiosperms. Nature 473, 97–100 (2011).
36. Qiu, Y. et al. Phylogenetic analyses of basal angiosperms based on nine plastid, mitochondrial, and nuclear Genes. International Journal of Plant Sciences 166, 815–842 (2005).
37. Cai, Z. et al. Complete plastid genome sequences of Drimys, Liriodendron, and Piper: implications for the phylogenetic relationships of magnoliids. BMC Evolutionary Biology 6, doi:10.1186/1471–2148–6–77 (2006).
38. Li, Q. et al. Genetic perturbation of the maize methylome. Plant Cell 26, 4602–16 (2014).
39. Casas-Mollano, J. A., Jeong, B., Xu, J., Moriyama, H. & Cerutti, H. The MUT9p kinase phosphorylates histone H3 threonine 3 and is necessary for heritable epigenetic silencing in *Chlamydomonas*. PNAS 105, 6486–6491 (2008).
40. Wanga, Z. et al. Osmotic stress induces phosphorylation of histone H3 at threonine 3 in pericentromeric regions of *Arabidopsis thaliana*. PNAS 112, 8487–8492 (2015).
41. Takuno, S., Ran, J.-H. & Gaut, B. S. Evolutionary patterns of genic DNA methylation vary across land plants. Nature Plants 15222, doi:10.1038/nplants.2015.222 (2016).
42. Lister, R. et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462, 315–22 (2009).
43. Jones, P. et al. InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–40 (2014).
44. Mirarab, S. et al. PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. J Comput Biol 22, 377–86 (2015).
45. Stamatakis, A. RAxML Version 8: A tool for phylogenetic analysis and postanalysis of large phylogenies. Bioinformatics 30, 1312–1313 (2014).
46. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17, 540–552 (2000).
47. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution 24, 1586–1591 (2007).
48. Urich, M. A., Nery, J. R., Lister, R., Schmitz, R. J. & Ecker, J. R. MethylCseq library preparation for base-resolution whole-genome bisulfite sequencing. Nature Protocols 10, 475–483 (2015).
49. Schultz, M. D. et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature 523, 212–6 (2015).

50. Hinchliff, C. E. et al. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. PNAS 112, 12764–9 (2015).

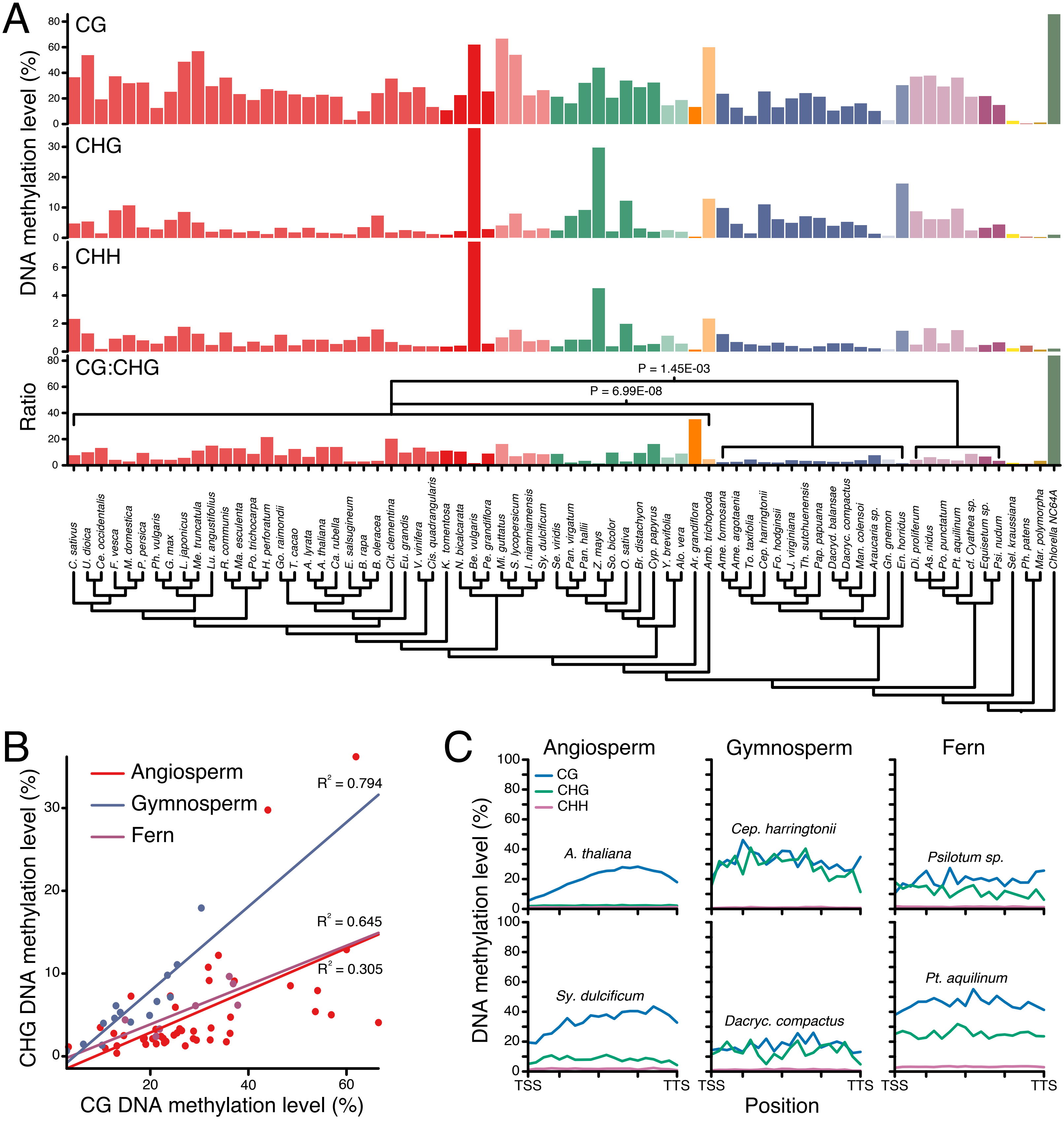

A
Eudicot
Core
Rosid
Asterid
Basal
Monocot
Commelinid
Monocot
Basal angiosperm
Magnoliid
Basal-most
Gymnosperm
Conifer
Cycadale
Ginkgoale
Gnetale
Fern
Eusporangiate
Leptosporangiate
Lycophyte
Hornwort
Moss
Liverwort
Green algae
CMT2
CMTα
Basal angiosperm
Monocot+Magnoliid
CMT1
CMTε
CMT3
100
100
84
83
66
98
88
100
80
58
89
95
99
0.2
B
CMTα
CMT2
100
basal angiosperm
CMTε
80
monocot+magnoliid
CMTε
95
CMT1
99
CMT3
50 species
C
Other
Basal angiosperm
Magnoliid
Monocot
Eudicot
Basal angiosperm
Magnoliid
Monocot
Eudicot
Eudicot
ε
γ
CMTα
CMT2
CMTε
CMTε
CMT1
CMT3

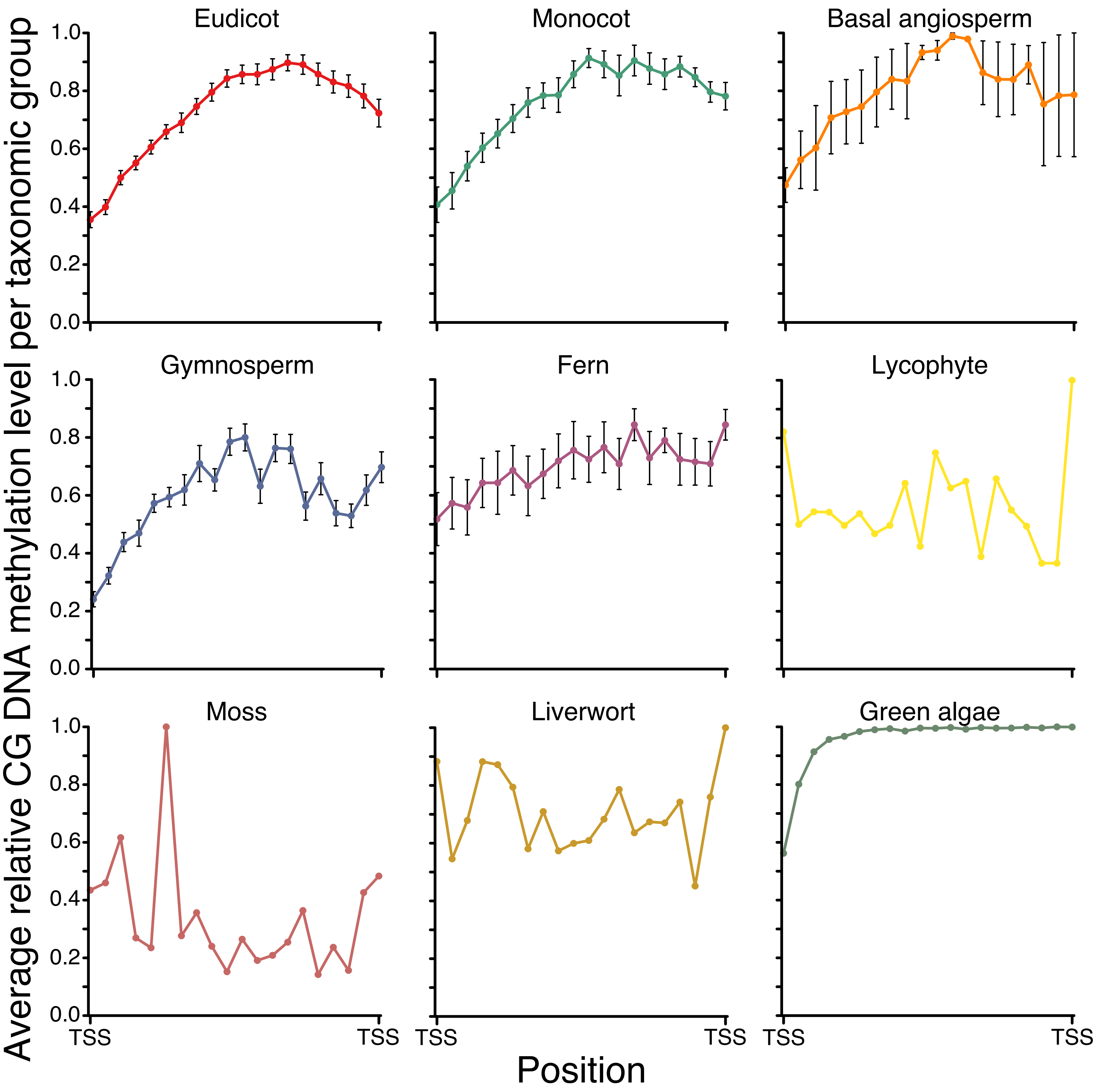

A
O. sativa (Orysa_v7_0-LOC_Os03g12570_1-)
Pan. hallii (Pahal.0213s0055.1)
Z. mays (GRMZM2G025592_T01)
BAH 🟩
CHROMO 🟧
DNA methylase 🟪
Protein tyrosine kinase 🟦
200 aa
B
≥95
≥75-94
<74
Pan. virgatum (Pavir.Ib02669.1)
Pan. virgatum (Pavir.Ia02815.1)
Pan. hallii (Pahal.0213s0055.1)
Se. viridis (Si033890m)
So. bicolor (Sorbi_v2_1-Sobic_006G214000_1-)
So. bicolor (Sorbi_v2_1-Sobic_004G197400_1-)
Z. mays (GRMZM2G005310_T02)
Z. mays (GRMZM2G025592_T01)
Br. distachyon (Bradi3g39050.1)
Br. distachyon (Bradi3g21450.2)
O. sativa (Orysa_v7_0-LOC_Os10g01570_1-)
O. sativa (Orysa_v7_0-LOC_Os03g12570_1-)
Br. distachyon (Bradi1g68985.1)
Se. viridis (Si033879m)
A. thaliana (Arath_TAIR10-AT1G80740_1-)
A. thaliana (Arath_TAIR10-AT1G69770_1-)
Pan. virgatum (Pavir.J12391.1)
Pan. virgatum (Pavir.J12650.1)
Pan. virgatum (Pavir.Ca00473.1)
Pan. hallii (Pahal.0030s0323.1)
Se. viridis (Si021001m)
So. bicolor (Sorbi_v2_1-Sobic_009G083900_1-)
O. sativa (Orysa_v7_0-LOC_Os05g13780_1-)
Br. distachyon (Bradi1g66167.2)
A. thaliana (Arath_TAIR10-AT4G19020_1-)
Sel. moellendorffii (Selmo_v1_0-30570-)
Ph. patens (Phypa_v3_0-Phpat_006G103100_1-)
CKMT
CKMT
CMTε monocot+magnoliid
CMT1
CMT3
CMT2
CMTα
0.2
C
P. hallii
P. virgatum
S. viridis
Z. mays
S. bicolor
B. distachyon
O. sativa
β
α
+
+



Model	Null -lnL	Alternative -lnL	df	p-value	ω	Summary
Branch	149273.628	149223.700	1	≤0.0001	b=0.0961 f=0.1750	Higher ω in Brassicaceae (A)
Branch	149273.628	149223.256	2	≤0.0001	b=0.0966 fB=0.2406 fC=0.1642	Higher and different ω between *Brassica* (B) and *Arabidopsis* (C) clades
Branch-site	146970.090	146970.090	1	1	0: p=0.6595, b=0.0822 f=0.0822 1: p=0.1274, b=1.0000, f=1.0000 2a: p=0.1786, b=0.0822, f=1.0000 2b: p=0.0345, b=1.0000, f=1.0000	Higher ω in *Brassica* (B) clade not due to positive selection

f: foreground; b: background; 0, 1, 2a,2b: site classes described in Yang (2007)


A
DNA methylation level (%)
CG
0
20
40
60
80
CHG
0
10
20
30
CHH
0
2
4
6
Ratio
CG:CHG
0
20
40
60
80
P = 1.45E-03
P = 6.99E-08
C. sativus
U. dioica
Ce. occidentalis
F. vesca
M. domestica
P. persica
Ph. vulgaris
G. max
L. japonicus
Me. truncatula
Lu. angustifolius
R. communis
Ma. esculenta
Po. trichocarpa
H. perforatum
Go. raimondii
T. cacao
A. lyrata
A. thaliana
Ca. rubella
E. salsugineum
B. rapa
B. oleracea
Cit. clementina
Eu. grandis
V. vinifera
Cis. quadrangularis
K. tomentosa
N. bicalcarata
Be. vulgaris
Pe. grandiflora
Mi. guttatus
S. lycopersicum
I. niamniamensis
Sy. dulcificum
Se. viridis
Pan. virgatum
Pan. hallii
Z. mays
So. bicolor
O. sativa
Br. distachyon
Cyp. papyrus
Y. brevifolia
Alo. vera
Ar. grandiflora
Amb. trichopoda
Ame. formosana
Ame. argotaenia
To. taxifolia
Cep. harringtonii
Fo. hodginsii
J. virginiana
Th. sutchuenensis
Pap. papuana
Dacryd. balansae
Dacryc. compactus
Man. colensoi
Araucaria sp.
Gn. gnemon
En. horridus
Di. proliferum
As. nidus
Po. punctatum
Pt. aquilinum
cf. Cyathea sp.
Equisetum sp.
Psi. nudum
Sel. kraussiana
Ph. patens
Mar. polymorpha
Chlorella NC64A
B
Angiosperm
Gymnosperm
Fern
$R^2 = 0.794$
$R^2 = 0.645$
$R^2 = 0.305$
CHG DNA methylation level (%)
0
10
20
30
CG DNA methylation level (%)
20
40
60
C
Angiosperm
Gymnosperm
Fern
CG
CHG
CHH
DNA methylation level (%)
0
20
40
60
80
100
A. thaliana
Cep. harringtonii
Psilotum sp.
Sy. dulcificum
Dacryc. compactus
Pt. aquilinum
TSS
TTS
Position


Eudicot
Monocot
Basal angiosperm
Gymnosperm
Fern
Lycophyte
Moss
Liverwort
Green algae
1.0
0.8
0.6
0.4
0.2
0.0
TSS
TSS
Average relative CG DNA methylation level per taxonomic group
Position