

1 **TITLE:** The evolution of CHROMOMETHYLASES and gene body DNA
2 methylation in plants

3

4 **RUNNING TITLE:** CMT gene family in plants

5

6 Adam J. Bewick¹, Chad E. Niederhuth¹, Lexiang Ji², Nicholas A. Rohr¹, Patrick
7 T. Griffin¹, Jim Leebens-Mack³, Robert J. Schmitz¹

8

9 ¹Department of Genetics, University of Georgia, Athens, GA 30602, USA

10 ²Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

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

12

13 **CORRESPONDING AUTHOR:** Robert J. Schmitz, schmitz@uga.edu

14

15 **KEY WORDS:** CHROMOMETHYLASE, Phylogenetics, DNA methylation, WGBS

16

17 **ABSTRACT**

18

19 **Background.** The evolution of gene body methylation (gbM), its origins and its
20 functional consequences are poorly understood. By pairing the largest collection
21 of transcriptomes (>1000) and methylomes (77) across Viridiplantae we provide
22 novel insights into the evolution of gbM and its relationship to
23 CHROMOMETHYLASE (CMT) proteins.

24

25 **Results.** CMTs are evolutionary conserved DNA methyltransferases in
26 Viridiplantae. Duplication events gave rise to what are now referred to as CMT1,
27 2 and 3. Independent losses of CMT1, 2 and 3 in eudicots, CMT2 and ZMET in
28 monocots and monocots/commelinids, variation in copy number and non-neutral
29 evolution suggests overlapping or fluid functional evolution of this gene family.
30 DNA methylation within genes is widespread and is found in all major taxonomic
31 groups of Viridiplantae investigated. Genes enriched with methylated CGs (mCG)

32 were also identified in species sister to angiosperms. The proportion of genes
33 and DNA methylation patterns associated with gbM are restricted to angiosperms
34 with a functional CMT3 or ortholog. However, mCG-enriched genes in the
35 gymnosperm *Pinus taeda* shared some similarities with gbM genes in *Amborella*
36 *trichopoda*. Additionally, gymnosperms and ferns share a CMT homolog closely
37 related to CMT2 and 3. Hence, the dependency of gbM on a CMT most likely
38 extends to all angiosperms and possibly gymnosperms and ferns.

39

40 **Conclusions.** The resulting gene family phylogeny of CMT transcripts from the
41 most diverse sampling of plants to date redefines our understanding of CMT
42 evolution and its evolutionary consequences on DNA methylation. Future,
43 functional tests of homologous and paralogous CMTs will uncover novel roles
44 and consequences to the epigenome.

45

46 BACKGROUND

47

48 DNA methylation is an important chromatin modification that protects the genome
49 from selfish genetic elements, is important for proper gene expression, and is
50 involved in genome stability. In plants, DNA methylation is found at cytosines (C)
51 in three sequence contexts: CG, CHG, and CHH (H is any nucleotide, but G). A
52 suite of distinct *de novo* and maintenance DNA methyltransferases establish and
53 maintain DNA methylation at these three sequence contexts, respectively.
54 CHROMOMETHYLASES (CMTs) are an important class of plant-specific DNA
55 methylation enzymes, which are characterized by the presence of a CHRomatins
56 Organisation MOdifier (CHROMO) domain between the cytosine
57 methyltransferase catalytic motifs I and IV [1]. Identification, expression, and
58 functional characterization of CMTs have been extensively performed in the
59 model plant *Arabidopsis thaliana* [2, 3, 4] and in the model grass species *Zea*
60 *mays* [5, 6, 7].

61 There are three CMT genes encoded in the *A. thaliana* genome: CMT1,
62 CMT2, and CMT3 [2, 8, 9, 10]. CMT1 is the least studied of the three CMTs as a

63 handful of *A. thaliana* accessions contain an Evelknievel retroelement insertion or
64 a frameshift mutation truncating the protein, which suggested that CMT1 is
65 nonessential [8]. The majority of DNA methylation at CHH sites (mCHH) at long
66 transposable elements in pericentromeric regions of the genome is targeted by a
67 CMT2-dependent pathway [3, 4]. Allelic variation at CMT2 has been shown to
68 alter genome-wide levels of CHH DNA methylation (mCHH), and plastic alleles of
69 CMT2 may play a role in adaptation to temperature [11, 12 ,13]. In contrast, DNA
70 methylation at CHG (mCHG) sites is often maintained by CMT3 through a
71 reinforcing loop with histone H3 lysine 9 di-methylation (H3K9me2) catalyzed by
72 the KRYPTONITE (KYP)/SU(VAR)3-9 HOMOLOG 4 (SUVH4), SUVH5 and
73 SUVH6 lysine methyltransferases [2, 6, 14, 15]. In *Z. mays*, ZMET2 is a
74 functional homolog of CMT3 and catalyzes the maintenance of mCHG [5, 6, 7]. A
75 paralog of ZMET2, ZMET5, contributes to the maintenance of mCHG to a lesser
76 degree in *Z. mays* [5, 7]. Homologous CMTs have been identified in other
77 flowering plants (angiosperms) [16, 17, 18, 19]: the moss *Physcomitrella patens*,
78 the lycophyte *Selaginella moellendorffii*, and the green algae *Chlorella* sp.
79 NC64A and *Volvox carteri* [20]. The function of CMTs in species sister to
80 angiosperms (flowering plants) is poorly understood. However, in at least *P.*
81 *patens* a CMT protein contributes to mCHG [21].

82 A large number of genes in angiosperms exclusively contain CG DNA
83 methylation (mCG) in the transcribed region and a depletion of mCG from both
84 the transcriptional start and stop sites (referred to as “gene body DNA
85 methylation”, or “gbM”) [22, 23, 24, 25]. GbM genes are generally constitutively
86 expressed, evolutionarily conserved, and typically longer than un-methylated
87 genes [25, 26, 27]. How gbM is established and subsequently maintained is
88 unclear. However, recently it was discovered that CMT3 has been independently
89 lost in two angiosperm species belonging to the Brassicaceae family of plants
90 and this coincides with the loss of gbM [19, 25]. Furthermore, *A. thaliana* and
91 closely related Brassicaceae species have reduced levels of mCHG on a per
92 cytosine basis, but still posses CMT3 [19, 25], which indicates changes at the
93 molecular level may have disrupted function of CMT3. This has led to a

94 hypothesis that the evolution of gbM is linked to incorporation/methylation of
95 histone H3 lysine-9 di-methylation (H3K9me2) in gene bodies with subsequent
96 failure of INCREASED IN BONSAI METHYLATION 1 (IBM1) to de-methylate
97 H3K9me2 [19, 28]. This provides a substrate for CMT3 to bind and methylate
98 DNA, and through an unknown mechanism leads to mCG. MCG is maintained
99 over evolutionary timescales by the CMT3-dependent mechanism and during
100 DNA replication by the maintenance DNA METHYTRANSFERASE 1 (MET1).
101 Methylated DNA then provides a substrate for binding by KRYPTONITE (KYP)
102 and related family members through their SRA domains, which increases the rate
103 at which H3K9 is di-methylated [29]. Finally, mCG spreads throughout the gene
104 over evolutionary time [19]. A similar model was previously proposed, which links
105 gene body mCG with transcription, mCHG, and IBM1 activity [28].

106 Previous phylogenetic studies have proposed that CMT1 and CMT3 are
107 more closely related to each other than to CMT2, and that ZMET2 and ZMET5
108 proteins are more closely related to CMT3 than to CMT1 or CMT2 [5], and the
109 placement of non-seed plant CMTs more closely related to CMT3 [21]. However,
110 these studies were not focused on resolving phylogenetic relationships within the
111 CMT gene family, but rather relationships of CMTs between a handful of species.
112 These studies have without question laid the groundwork to understand CMT-
113 dependent DNA methylation pathways and patterns in plants. However, the
114 massive increase in transcriptome data from a broad sampling of plant species
115 together with advancements in sequence alignment and phylogenetic inference
116 algorithms have made it possible to incorporate thousands of sequences into a
117 single phylogeny, allowing for a more complete understanding of the CMT gene
118 family. Understanding the evolutionary relationships of CMT proteins is
119 foundational for inferring the evolutionary origins, maintenance, and
120 consequences of genome-wide DNA methylation and gbM.

121 Here we investigate phylogenetic relationships of CMTs at a much larger
122 evolutionary timescale using data generate from the 1KP Consortium
123 (www.onekp.com). In the present study we have analyzed 771 mRNA transcripts
124 and annotated genomes, identified as belonging to the CMT gene family, from an

125 extensive taxonomic sampling of 443 different species including eudicots (basal,
126 core, rosid, and asterid), basal angiosperms, monocots and
127 monocots/commelinid, magnoliids, gymnosperms (conifers, Cycadales,
128 Ginkgoales), monilophytes (ferns and fern allies), lycophytes, bryophytes
129 (mosses, liverworts and hornworts), and green algae. CMT homologs identified
130 across Viridiplantae (land plants and green algae) indicate that CMT genes
131 originated prior to the origin of Embryophyta (land plants) (\geq 480 MYA) [30, 31,
132 32, 33]. In addition, phylogenetic relationships suggests at least two duplication
133 events occurred within the angiosperm lineage giving rise to the CMT1, CMT2,
134 and CMT3 gene clades. In the light of CMT evolution we explored patterns of
135 genomic and genic DNA methylation levels in 77 species of Viridiplantae,
136 revealing diversity of the epigenome within and between major taxonomic
137 groups, and the evolution of gbM in association with the origin of CMT3 and
138 orthologous sequences.

139

140 RESULTS

141

142 **The origins of CHROMOMETHYLASES.** CMTs are found in most major
143 taxonomic groups of land plants and some algae: eudicots, basal angiosperms,
144 monocots and commelinids, magnoliids, gymnosperms, ferns, lycophytes,
145 mosses, liverworts, hornworts, and green algae (Fig. 1a and Table S1). CMTs
146 were not identified in transcriptome data sets for species sister to Viridiplantae
147 including those belonging to Glauco phyta, red algae, Dinophyceae, Chromista,
148 and Euglenozoa. CMTs were identified in a few green algae species: *Picocystis*
149 *salinarum*, *Cylindrocystis* sp., and *Cylindrocystis brebissonii*. Additionally,
150 functional CMTs – based on presence/absence of characterized protein domains
151 – were not identified from three species within the gymnosperm order Gnetales.
152 A transcript with a CHROMO and C-5 cytosine-specific DNA methylase domain
153 was identified in *Welwitschia mirabilis* (Gnetales), but this transcript did not
154 include a Bromo Adjacent Homology (BAH) domain. The BAH domain is an
155 interaction surface that is required to capture H3K9me2, and mutations that

156 abolish this interaction causes a failure of a CMT protein (i.e., ZMET2) binding to
157 nucleosomes and a complete loss of activity *in vivo* [6]. Therefore, although a
158 partial sequence is present, it might represent a nonfunctional allele of a CMT.
159 Alternatively, it might represent an incomplete transcript generated during
160 sequencing and assembling of the transcriptome. Overall, the presence of CMT
161 homologs across Viridiplantae and their absence from sister taxonomic groups
162 suggest CMT evolved following the divergence of green algae [35, 35].

163 The relationships among CMTs suggest that CMT2 and the clade
164 containing CMT1, CMT3 and ZMET arose from a duplication event at the base of
165 all angiosperms (Fig. 1b). This duplication event might have coincided with event
166 ϵ , the ancestral angiosperm whole genome duplication (WGD) event [36].
167 Relationships among clades sister to angiosperm CMTs largely recapitulate
168 species relationships (Fig. 1a) [34, 37]. However, CMTs in gymnosperms and
169 ferns are paraphyletic (Fig. 1a). Similarly, these homologous sequences might
170 have been derived from a WGD (i.e., ζ , the ancestral seed plant WGD), with one
171 paralog being the ancestor to CMT1, CMT2 and CMT3, and ZMET [36].
172 Previously identified CMTs in *S. moellendorffii* [20] and *P. patens* [20, 38] were
173 identified, which are sister to clades containing CMT1, CMT2, and CMT3 and
174 ZMET (Fig. 1a). CMTs previously identified in the green algae *Chlamydomonas*
175 *reinhardtii*, *Chlorella* sp. NC64A and *Volvox carteri* were excluded from
176 phylogenetic analysis because they lacked the CHROMO and other domains
177 typically associated with CMT proteins (Figure S1). Furthermore, based on
178 percent amino acid identity *C. reinhardtii* and *V. carteri* CMT sequences are
179 homologous to MET1 (Table S2). Similar to *S. moellendorffii* and *P. patens* CMT
180 sequences, green algae CMT sequences are sister to clades containing CMT1,
181 CMT2, and CMT3 and ZMET (Figure S2). The increase taxonomic sampling re-
182 defines relationships of CMTs in early-diverged land plants and in Viridiplantae in
183 general [18, 20, 21, 38, 39].

184 Further diversification of CMT proteins occurred in eudicots. CMT1 and
185 CMT3 clades contain only sequences from eudicots (Fig. 1a and b). This
186 relationship supports the hypothesis that CMT1 and CMT3 arose from a

187 duplication event shared by all eudicots. Thus, CMT1 and CMT3 might be the
188 result of the γ WGD event at the base of eudicots [36]. Synteny between CMT1
189 and CMT3, despite ~125 million years of divergence, further supports this
190 hypothesis (Figure S3a). Not all eudicots possess CMT1, CMT2, and CMT3, but
191 rather exhibit CMT gene content ranging from zero to three (Figure S4a). Also,
192 many species possess multiple copies of CMT1, CMT2, or CMT3. The
193 presence/absence of CMTs might represent differences in transcriptome
194 sequencing coverage or spatial and temporal divergence of expression.
195 However, CMT2, CMT3, and homologous proteins have functions in methylating
196 a significant number of non-CG sites throughout the entire genome and thus are
197 broadly expressed in *A. thaliana*, *Z. mays* and other species [8, 6, 18, 25].
198 Additionally, eudicot species with sequenced and assembled genomes show
199 variation in the presence/absence and copy number of CMTs. Hence the type of
200 tissue(s) used in transcriptome sequencing (www.onekp.com) would have limited
201 biases against CMTs, suggesting that the variation reflects presence/absence at
202 a genetic level.

203 The *Z. mays* in-paralogs ZMET2 and ZMET5, and closely related CMTs in
204 other monocots, commelinids, and magnoliids form a well-supported
205 monophyletic clade (Fig. 1a and Figure S3b and c). In addition to *Z. mays*, in-
206 paralogs were identified in *Sorghum bicolor* and *Brachypodium distachyon* (Fig.
207 1a and Figure S3b). Relationships of *S. bicolor* and *Z. mays* ZMETs differed
208 between gene and amino acid derived phylogenies (Fig. 1a and Figure S3b).
209 However, synteny between paralogs of both species supports two independent
210 duplications (Figure S3c). Also, paralogous ZMETs are shared across species
211 (Figure S3b). These shared paralogs might have originated from a Poaceae-
212 specific duplication event, which was followed by losses in some species. The
213 contribution of each paralog to DNA methylation and other chromatin
214 modifications remains unknown at this time.

215 Akin to eudicots, monocots and monocots/commelinids possess
216 combinations of ZMET and CMT2 (Figure S4b). For example, the model grass
217 species *Z. mays* has lost CMT2, whereas the closely related species *S. bicolor*

218 possess both ZMET and CMT2 (Table S1). ZMET is not strictly homologous to
219 CMT3, and represents a unique monophyletic group that is sister to both CMT1
220 and CMT3. However, ZMET2 is functionally homologous to CMT3 and maintains
221 DNA methylation at CHG sites [6, 7]. Unlike CMT3, ZMET2 is associated with
222 DNA methylation at CHH sites within some loci [7]. Given the inclusion of
223 monocot and magnoliid species in the monophyletic ZMET clade, this dual-
224 function is expected to be present in other monocot species, and in magnoliid
225 species.

226 Overall, these redefined CMT clades, and monophyletic clades of broad
227 taxonomic groups, are well supported (Fig. 1a). Thus, the identification of novel
228 CMTs in magnoliids, gymnosperms, lycophytes, hornworts, liverworts,
229 bryophytes, and green algae pushes the timing of evolution of CMT, and
230 potentially certain mechanisms maintaining mCHG and mCHH, prior to the origin
231 of Embryophyta (\geq 480 million years ago [MYA]) [30, 31, 32, 33].

232

233 **Reduced selective constraint of CMT3 in the Brassicaceae affects gbM.**
234 Recent work has described the DNA methylomes of 34 angiosperms, revealing
235 extensive variation across this group of plants [25]. This variation was
236 characterized in terms of levels of DNA methylation, and number of DNA
237 methylated genes. DNA methylation levels describe variation within a population
238 of cells. Although ascribing genes as DNA methylated relies on levels of DNA
239 methylation, this metric provides insights into the predominant DNA methylation
240 pathway and expected relationship to genic characteristics [25, 26, 27]. The
241 genetic underpinnings of this variation are not well understood, but some light
242 has been shed through investigating DNA methylation within the Brassicaceae
243 [19]. The Brassicaceae have reduced levels of genomic and genic levels of
244 mCG, genome-wide per-site levels of mCHG, and numbers of gbM genes [19,
245 25]. In at least *E. salsugineum* and *C. planisiliqua* this reduction in levels of DNA
246 methylation and numbers of gbM genes has been attributed to the loss of the
247 CMT3 [19]. However, closely related species with CMT3 – *Brassica oleracea*,
248 *Brassica rapa* and *Schrenkiella parvula* – have reduced levels of gbM and

249 numbers of gbM genes compared to the sister clade of *A. thaliana*, *Arabidopsis*
250 *lyrata*, and *Capsella rubella* (Fig. 2a and b) and overall to other eudicots [19, 25].
251 Although CMT3 is present, changes at the sequence level, including the
252 evolution of deleterious or functionally null alleles, could disrupt function to
253 varying degrees. At the sequence level, CMT3 has evolved at a higher rate of
254 molecular evolution – measured as $dN/dS (\omega)$ – in the Brassicaceae ($\omega=0.175$)
255 compared to 162 eudicots ($\omega=0.097$), with further increases in the clade
256 containing *B. oleracea*, *B. rapa* and *S. parvula* ($\omega=0.241$) compared to the clade
257 containing *A. thaliana*, *A. lyrata* and *C. rubella* ($\omega=0.164$) (Fig. 2). A low
258 background rate of molecular evolution suggests purifying selection acting to
259 maintain low allelic variation across eudicots. Conversely, increased rates of
260 molecular evolution can be a consequence of positive selection. However, a
261 hypothesis of positive selection was not preferred to contribute to the increased
262 rates of ω in either Brassicaceae clade (Table S3). Alternatively, relaxed
263 selective constraint possibly resulted in an increased ω , which might have
264 introduced null alleles ultimately affecting function of CMT3, and in turn, affecting
265 levels of DNA methylation and numbers of gbM genes. The consequence of the
266 higher rates of molecular evolution in the clade containing *Brassica spp.* and *S.*
267 *parvula* relative to all eudicots and other Brassicaceae are correlated with an
268 exacerbated reduction in the numbers of gbM loci and their methylation levels,
269 which suggests unique substitutions between clades or a quantitative affect to an
270 increase in the number of substitutions. However, at least some substitutions
271 affecting function are shared between Brassicaceae clades because both have
272 reductions in per-site levels of mCHG [19].

273

274 **Divergence of DNA methylation patterns within gene bodies during**
275 **Viridiplantae evolution.** Levels and distributions of DNA methylation within gene
276 bodies are variable across Viridiplantae. Levels of mCG range from ~2% in *S.*
277 *moellendorffii* to ~86% in *Chlorella sp.* NC64A (Fig. 3a). Other plant species fall
278 between these two extremes (Fig. 4a) [25]. *Beta vulgaris* remains distinct among
279 angiosperms and Viridiplantae with respect to levels of DNA methylation at all

280 sequence contexts (Fig. 3a). Similarly, *Z. mays* is distinct among monocots and
281 monocots/commelinids (Fig. 3a). Gymnosperms and ferns possess similar levels
282 of mCG to mCHG within gene bodies and levels of mCHG qualitatively parallel
283 those of mCG similar to observations in recently published study (Fig. 3a and
284 Figure S5) [40]. A similar pattern is observed in *Z. mays*. However, this pattern is
285 not shared by other monocots/commelinids [25]. High levels of mCHG is
286 common across the gymnosperms and ferns investigated in this study, and tends
287 to be higher than levels observed in angiosperms (Fig. 3a and Figure S5) [25].
288 DNA methylation at CG, CHG and CHH sites within gene bodies was detected in
289 the liverwort *Marchantia polymorpha* (Fig. 3a). Furthermore, DNA methylation at
290 CG sites was not detected in the *P. patens* when all genes are considered (Fig.
291 3a). Overall, increased taxonomic sampling has revealed natural variation
292 between and within groups of Viridiplantae.

293 Despite the presence of mCG within the gene bodies of angiosperms,
294 gymnosperms, ferns, lycophytes, liverworts, and green algae; the distributions
295 across gene bodies differ (Fig. 3b). In angiosperms (eudicots, commelinids,
296 monocots and basal angiosperms) CG DNA methylation is depleted at the
297 transcriptional start and termination sites (TSS and TTS, respectively), and
298 gradually increases towards the center of the gene body (Fig. 3b). In the basal
299 angiosperm *Amborella trichopoda*, levels of mCG decline sharply prior to the TTS
300 (Fig. 3b). Similar to angiosperms, mCG is reduced at the TSS relative to the
301 gene body in the gymnosperm *Pinus taeda* (Fig. 4b). However, mCG is not
302 reduced at the TTS (Fig. 3b). Additionally, DNA methylation at non-CG (mCHG
303 and mCHH) sites is not reduced at the TSS and TTS. Little difference in mCG,
304 mCHG, and mCHH within gene bodies, and upstream and downstream regions
305 are observed in *S. moellendorffii* (Fig. 3b). Additionally, mCG, mCHG, and
306 mCHH are not excluded from the TSS and TTS (Fig. 3b). As opposed to
307 angiosperms and gymnosperms (*P. taeda*), mCG in *M. polymorpha* decreases
308 towards the center of the gene body (Fig. 3b). This distribution also occurs for
309 methylation at non-CG sites in *M. polymorpha* (Fig. 3b). Additionally, *M.*
310 *polymorpha* has distinctive high levels of mCG, mCHG, and mCHH surrounding

311 the TSS and TTS (Fig. 3b). Finally, in *Chlorella* sp. NC64A, mCG is enriched at
312 near 100% across the entire gene body (Fig. 3b).

313 The presence of mCG within gene bodies indicates that a gene could
314 possess gbM. However, other types of DNA methylated genes contain high
315 levels of mCG [25], thus enrichment tests were performed to identify genes that
316 are significantly enriched for mCG and depleted of non-CG methylation (i.e., gbM
317 genes). Genes matching this DNA methylation enrichment profile were identified
318 in species sister to angiosperms: gymnosperms, lycophytes, liverworts, mosses
319 and green algae (Fig. 3b). The proportion of genes within each genome or subset
320 of the genome (*P. taeda*) was small compared to angiosperms with gbM (Fig.
321 3b). Furthermore, the number of gbM genes was comparable to angiosperms
322 without gbM, which suggests these identified genes are the result of statistical
323 noise (Figure S6a). This is most likely the case for lycophytes, liverworts, mosses
324 and green algae, since the levels of mCG within genes bodies is highly skewed
325 (Figure S7). Additionally, the distribution of mCG and non-CG methylation across
326 the gene body is unlike the distribution of gbM genes (Fig. 3b) [25]. However, the
327 gymnosperm *P. taeda* shares some similarities to gbM genes of the basal
328 angiosperm *A. trichopoda* (Fig. 3b). Hence, mCG-enriched genes identified in
329 gymnosperms, lycophytes, liverworts, mosses and green algae are most likely
330 not gbM.

331

332 **Correlated evolution of CMT3 and the histone de-methylase IBM1 in**
333 **angiosperms.** The exact mechanisms by which genes are targeted to become
334 gbM and the establishment of DNA methylation at CG sites is currently unknown.
335 One proposed possibility is the failure of IBM1 to remove H3K9me2 within genes.
336 This would provide the necessary substrate for CMT3 to associate with
337 nucleosomes in genes. Due to the tight association between CMT3 and IBM1
338 (and SUVH4/5/6) these proteins might have evolved together. Resolution of
339 phylogenetic relationships supports monophyly of IBM1 and orthologous
340 sequences that is unique to angiosperms (Fig. 4 and Figure S8). Furthermore,
341 high levels of mCHG and/or similar levels of mCHG to mCG in gymnosperms,

342 ferns, *S. moellendorffii* (lycophyte), *M. polymorpha* (liverwort) and *P. patens*
343 (moss) compared to angiosperms suggests a functionally homologous histone
344 de-methylase is not present in these taxonomic groups and species. The
345 absence of IBM1 in the basal-most angiosperm *A. trichopoda* and similarities of
346 DNA methylation distribution between gbM genes and *P. taeda* mCG-enriched
347 genes further supports a role of CMT3 and IBM1 in maintenance of mCG within
348 gene bodies. Unlike CMT3 and IBM1, histone methylases SUVH4 and SUVH5/6
349 are common to all taxonomic groups investigated, which suggests common
350 ancestry and shared functions of transposon silencing (Fig. 4 and Figures S8 and
351 S9) [22, 41, 42, 43]. However, a Brassicaceae-specific duplication event gave
352 rise to SUVH5 and SUVH6, and other Viridiplantae possess a homologous
353 SUVH5/6 (hSUVH5/6) (Figure S9b and c). Additionally, a duplication event
354 shared by all monocots and monocots/commelinids generated paralogous
355 hSUVH5/6, and additionally duplication event occurred in the Poaceae (Figure
356 S9d). The duplication event that gave rise to ZMET paralogs in the Poaceae
357 might have also generated the paralogous hSUVH5/6. The diversity in levels and
358 patterns of DNA methylation within gene bodies suggests corresponding
359 changes in function of DNA methyltransferases and/or histone de-methylases
360 during Viridiplantae divergence. Furthermore, monophyletic, angiosperm-specific
361 clades of a gbM-dependent CMT and IBM1 suggest co-evolution of proteins
362 involved in the gbM pathway.

363

364 **DISCUSSION**

365

366 CMTs are conserved DNA methyltransferases across Viridiplantae. Evolutionary
367 phenomenon and forces have shaped the relationships of CMTs, which have
368 most likely contributed to functional divergence among and within taxonomic
369 groups of Viridiplantae. Duplication events have contributed to the unique
370 relationships of CMTs, and have given rise to clade-, family- and species-specific
371 CMTs. This includes the eudicot-specific CMT1 and CMT3, paralogous CMTs
372 within monocots/commelinids (ZMETs), and the *Z. mays*-specific ZMET2 and

373 ZMET5. The paralogous CMT1 and CMT3, and orthologous CMTs in monocots,
374 monocots/commelinids, magnoliids and basal angiosperms form a superclade
375 that is sister to CMT2. Homologous CMTs in gymnosperms and ferns are
376 paraphyletic, and clades are sister to all CMTs – including CMT1, CMT2, CMT3
377 and ZMET – in angiosperms. CMTs have been shown to maintain methylation at
378 CHG sites (CMT3 and ZMET5, and hCMT β in *P. patens*) and methylate CHH
379 sites within deep heterochromatin (CMT2) [2, 4, 6, 11, 14, 15, 20, 21], whereas
380 CMT1 is nonfunctional in at least *A. thaliana* accessions [8]. However, recent
381 work has provided evidence for the role of CMT3 in the evolution of mCG within
382 gene bodies, and specifically gbM, within angiosperms [19]. Additionally, non-
383 neutral evolution of CMT3 can affect levels of genome-wide mCHG and within
384 gene body mCG, and the number of gbM genes. Hence, functional divergence
385 following duplication might be more widespread [44], and the exact fate of
386 paralogous CMTs and interplay between paralogs in shaping the epigenome
387 remain unknown at this time.

388 DNA methylation within genes is common in Viridiplantae. However,
389 certain classes of DNA methylated genes might be specific to certain taxonomic
390 groups within the Viridiplantae. GbM is a functionally enigmatic class of DNA
391 methylated genes, which is characterized by an enrichment of mCG and
392 depletion of non-mCG within transcribed regions and depletion of DNA
393 methylation from all sequence contexts at the TSS and TTS. These genes are
394 typically constitutively expressed, evolutionary conserved, housekeeping genes,
395 which compose a distinct proportion of protein coding genes [19, 25, 26, 27, 45].
396 GbM genes have been mostly studied in angiosperms and evidence for the
397 existence of this class of DNA methylated gene outside of angiosperms is limited
398 [40]. However, in the present study, genes matching the DNA methylation profile
399 of gbM genes – enrichment of mCG and depletion of non-mCG – were identified
400 in taxonomic groups sister to angiosperms: gymnosperms, lycophytes, liverworts,
401 mosses and green algae. It is unclear if these genes are gbM genes in light of
402 findings in angiosperms [19, 25, 26, 27]. For example, the low proportion of
403 mCG-enriched genes supports the absence of gbM in gymnosperms, lycophytes,

404 liverworts, mosses and green algae. Additionally, the distribution of mCG among
405 all genes and across the gene body of mCG-enriched genes supports the
406 absence of gbM in lycophytes, liverworts, mosses and green algae. However,
407 similar distributions of mCG between gbM genes in the basal angiosperm *A.*
408 *trichopoda* and mCG-enriched genes in the gymnosperm *P. taeda* are observed,
409 which support the presence of gbM in this species and possibly other
410 gymnosperms. Also, a small proportion of mCG-enriched genes in gymnosperms
411 are homologous to gbM genes in *A. thaliana* (Figure S6b). With that being said,
412 sequence conservation is not the most robust indicator of gbM [25]. GbM genes
413 compose a unique class of genes with predictable characteristics [19, 25, 26, 27].
414 Through comparison of mCG-enriched genes identified in early diverging
415 Viridiplantae to angiosperms with and without gbM, there is stronger support that
416 this epigenetic feature is unique to angiosperms. However, future work including
417 deeper WGBS and RNA-seq, and additional and improved genome assemblies –
418 especially for gymnosperms and ferns – will undoubtedly contribute to our
419 understanding of the evolution of gbM.

420 GbM is dependent on the CHG maintenance methyltransferase CMT3 or
421 an orthologous CMT in angiosperms. Support for the dependency of gbM on
422 CMT3 comes from the naturally occurring $\Delta cmt3$ mutants *E. salsugineum* and *C.*
423 *planisilqua*, which is correlated with the lack of gbM genes [19, 25]. The
424 independent loss of CMT3 has also affected mCHG with low overall and per-site
425 levels recorded for these species [25]. Both species belong to the Brassicaceae
426 family, and other species within this family show reduced numbers of gbM genes
427 compared to other eudicot and angiosperm species [25]. Although CMT3 is
428 present in these species, relaxed selective constraint might have introduced
429 alleles which functionally compromise CMT3 resulting in decreased per-site
430 levels of mCHG and the number of gbM genes [25]. The functional compromises
431 of CMT3 non-neutral evolution are shared and have diverged between clades of
432 Brassicaceae, respectively, which might reflect shared ancestry between clades
433 and the unique evolutionary history of each clade. Furthermore, more relaxed
434 selective constraint – as in the *Brassica* spp. and *S. parvula* – is correlated with a

435 more severe phenotype relative to the other Brassicaceae clade. The
436 dependency of gbM on a CMT protein might extend into other taxonomic groups
437 of plants. Phylogenetic relationships of CMTs found in Viridiplantae and the
438 location of *A. thaliana* CMTs support a eudicot-specific, monophyletic CMT3
439 clade. The CMT3 clade is part of a superclade, which includes a monophyletic
440 clade of monocot (ZMET) and magnoliid CMTs, and a CMT from the basal
441 angiosperm *A. trichopoda*. Thus, the CMT-dependent gbM pathway might be
442 specific to angiosperms. However, a homologous, closely related CMT in
443 gymnosperms and ferns (i.e., hCMT α) might have a similar function. It is
444 conceivable that other proteins and chromatin modifications that interact with
445 CMTs and non-CG methylation are important for the evolution of gbM, and thus
446 have evolved together. Specifically, IBM1 that de-methylates H3K9me2 and
447 SUVH4/5/6 that binds to H3K9me2 and methylates CHG sites both act upstream
448 of CMT3. One proposed model for the evolution of gbM requires failure of IBM1
449 and rare mis-incorporation of H3K9me2, which initiates mCHG by SUVH4/5/6
450 and maintenance by CMT3 [19, 28]. IBM1 shares similar patterns and taxonomic
451 diversity as CMT3 and orthologous CMTs involved in gbM. Also, unlike most
452 angiosperms investigated to date – with *A. trichopoda* as the exception –
453 gymnosperms and ferns do not possess an IBM1 ortholog, hence IBM1 might be
454 important for the distribution of mCG within gene bodies. Furthermore, the lack of
455 IBM1 in *A. trichopoda* and *P. taeda* might explain some similarities shared
456 between gbM genes and mCG-enriched genes with respect to the deposition of
457 mCG, respectively. However, the exact relationship between gbM and IBM1 is
458 unknown and similarities in underlying nucleotide composition of genes might
459 affect distribution of mCG. Overall, the patterns of DNA methylation within gene
460 bodies and the phylogenetic relationships of CMTs support a CMT3 and
461 orthologous CMT-dependent mechanism for the maintenance of gbM in
462 angiosperms, which is stochastically initiated by IBM1.

463

464 **CONCLUSIONS**

465

466 In summary, we present the most comprehensive CMT gene-family phylogeny to
467 date. CMTs are ancient proteins that evolved prior to the diversification of
468 Embryophyta. A shared function of CMTs is the maintenance of DNA methylation
469 at non-CG sites, which has been essential for DNA methylation at long
470 transposable elements in the pericentromeric regions of the genome [6, 14, 15].
471 However, CMTs in some species of eudicots have been shown to be important
472 for mCG within gbM genes [19]. Refined relationships between CMT1, CMT2,
473 CMT3, ZMET, and other homologous CMT clades have shed light on current
474 models for the evolution of gbM, and provided a framework for further research
475 on the role of CMTs in establishment and maintenance of DNA methylation and
476 histone modifications. Patterns of DNA methylation within gene bodies have
477 diverged between Viridiplantae. Other taxonomic groups do not share the pattern
478 of mCG associated with gbM genes in the majority of angiosperms, which further
479 supports specificity of gbM in angiosperms. However, genic DNA methylation
480 commonalities between angiosperms and other taxonomic groups were
481 identified. DNA methylation within gene bodies and its consequences of or
482 relationship to expression and other genic features has been extensively studied
483 in angiosperms [25] and shifting focus to other taxonomic groups of plants for
484 deep methylome analyses will aid in understanding the shared consequences of
485 genic DNA methylation. Understanding the evolution of additional chromatin
486 modifiers will undoubtedly unravel the epigenome and reveal unique
487 undiscovered mechanisms.

488

489 METHODS

490

491 **1KP sequencing, transcriptome assembling and orthogrouping.** The One
492 Thousand Plants (1KP) Consortium includes assembled transcriptomes and
493 predicted protein coding sequences from a total of 1329 species of plants (Table
494 S1). Additionally, gene annotations from 24 additional species – *Arabidopsis*
495 *lyrata*, *Brachypodium distachyon*, *Brassica oleracea*, *Brassica rapa*, *Citrus*
496 *clementina*, *Capsella rubella*, *Cannabis sativa*, *Cucumis sativus*, *Eutrema*

497 *salsuginosum*, *Fragaria vesca*, *Glycine max*, *Gossypium raimondii*, *Lotus*
498 *japonicus*, *Malus domestica*, *Marchantia polymorpha*, *Medicago truncatula*,
499 *Panicum hallii*, *Panicum virgatum*, *Pinus taeda*, *Physcomitrella patens*, *Ricinus*
500 *communis*, *Setaria viridis*, *Selaginella moellendorffii*, and *Zea mays* – were
501 included (<https://phytozome.jgi.doe.gov/pz/portal.html>) and
502 (<http://pinegenome.org/pinerefseq/>). The CMT gene family was extracted from the
503 previously compiled 1KP orthogroupings using the *A. thaliana* gene identifier for
504 CMT1, CMT2 and CMT3. A single orthogroup determined by the 1KP
505 Consortium included all three *A. thaliana* CMT proteins, and a total of 5383
506 sequences. Sequences from species downloaded from Phytozome, that were not
507 included in sequences generated by 1KP, were included to the gene family
508 through reciprocal best BLAST with *A. thaliana* CMT1, CMT2 and CMT3. In total
509 the CMT gene family included 5449 sequences from 1043 species. We used the
510 protein structure of *A. thaliana* as a reference to filter the sequences found within
511 the CMT gene family. Sequences were retained if they included the same base
512 PFAM domains as *A. thaliana* – CHROMO, BAH, and C-5 cytosine-specific DNA
513 methylase domains – as identified by Interproscan [46]. These filtered sequences
514 represent a set of high-confident, functional, ideal CMT proteins, which included
515 771 sequences from 432 species, and were used for phylogenetic analyses.
516

517 **Phylogeny construction.** To estimate the gene tree for the CMT sequences, a
518 series of alignment and phylogenetic estimation steps were conducted. An initial
519 protein alignment was carried out using Pasta with the default settings [47]. The
520 resulting alignment was back-translated using the coding sequence (CDS) into
521 an in-frame codon alignment. A phylogeny was estimated by RAxML [48] (-m
522 GTRGAMMA) with 1000 rapid bootstrap replicates using the in-frame alignment,
523 and with only the first and second codon positions. Long branches can effect
524 parameter estimation for the substitution model, which can in turn degrade
525 phylogenetic signal. Therefore, phylogenies were constructed with and without
526 green algae species, and were rooted to the green algae clade or liverworts,
527 respectively. The species *Balanophora fungosa* has been reported to have a high

528 substitution rate, which can also produce long branches, and was removed prior
529 to phylogenetic analyses. Identical workflows were used for jumonji (jmjC)
530 domain-containing (i.e., IBM1), SUVH4, and SUVH5/6 gene families.

531

532 **Codon analysis.** Similar methodology as described above was used to construct
533 phylogenetic trees for testing hypotheses on the rates of evolution in a
534 phylogenetic context. However, the program Gblocks [49] was used to identify
535 conserved codons. The parameters for Gblocks were kept at the default settings,
536 except allowing for 50% gapped positions. The program Phylogenetic Analysis
537 by Maximum Likelihood (PAML) [50] was used to test branches (branch test) and
538 sites along branches (branch-site test) for deviations from the background rate of
539 molecular evolution (ω) and for deviations from the neutral expectation,
540 respectively. Branches tested and a summary of each test can be found in Table
541 S3.

542

543 **MethylC-seq.** MethylC-seq libraries were prepared according to the following
544 protocol [51]. For *A. thaliana*, *A. trichopoda*, *Chlorella* sp. NC64A, *M.*
545 *polymorpha*, *P. patens*, *P. taeda*, *S. moellendorffii*, and *Z. mays* reads were
546 mapped to the respective genome assemblies. *P. taeda* has a large genome
547 assembly of ~23 Gbp divided among ~14k scaffolds
548 (http://dendrome.ucdavis.edu/ftp/Genome_Data/genome/pinerefseq/Pita/v1.01/R_EADME.txt). Due to computational limitations imposed by the large genome size
549 only 4 Gbp of the *P. taeda* genome assembly was used for mapping, which
550 includes 2411 (27%) of the high quality gene models. Prior to mapping for
551 species with only transcriptomes each transcript was searched for the longest
552 open reading frame from all six possible frames, and only transcripts beginning
553 with a start codon and ending with one of the three stop codons were kept. All
554 sequencing data for each species was aligned to their respective transcriptome
555 or species within the same genus using the methylpy pipeline [52]. All MethylC-
556 seq data used in this study can be found in Tables S4 and S5. Weighted
557 methylation was calculated for each sequence context (CG, CHG and CHH) by
558

559 dividing the total number of aligned methylated reads by the total number of
560 methylated plus un-methylated reads. Since, per site sequencing coverage was
561 low – on average ~1 \times – subsequent binomial tests could not be performed for the
562 majority of species to bin genes as gbM [25]. To investigate the affect of low
563 coverage we compared levels of DNA methylation of 1 \times randomly sampled
564 MethylC-seq reads to actual levels for 32 angiosperm species, *S. moellendorffii*
565 (lycophyte), *M. polymorpha* (liverwort) and *Chlorella sp.* NC64A (green algae)
566 [19, 20, 25, 40]. Specifically, a linear model was constructed between deep (x)
567 and 1 \times (y) sequencing coverage, which was then used to extrapolate levels of
568 DNA methylation and 95% confidence intervals (CI) from low sequence coverage
569 species (Figure S10 and Tables S5).

570

571 **Genic DNA methylation analyses and metaplots.** DNA methylation was
572 estimated as weighted DNA methylation, which is the total number of aligned
573 DNA methylated reads divide by the total number of methylated plus un-
574 methylated reads. This metric of DNA methylation was estimated for each
575 sequence context within coding regions. For *P. taeda* only high quality gene
576 models were used, since low quality models cannot distinguish between
577 pseudogenes and true protein coding genes. For genic metaplots, the gene body
578 – start to stop codon – was divided into 20 windows. Additionally, for species with
579 assembled and annotated genomes regions 1000 or 4000 bp upstream and
580 downstream were divided into 20 windows. Weighted DNA methylation was
581 calculated for each window. The mean weighted methylation for each window
582 was then calculated for all genes and plotted in R v3.2.4 ([https://www.r-
583 project.org/](https://www.r-project.org/)).

584

585 **mCG-enrichment test.** Sequence context enrichment for each gene was
586 determined through a binomial test followed by Benjamini–Hochberg FDR [25,
587 26]. A context-specific background level of DNA methylation determined from the
588 coding sequence was used as a threshold in determining significance. Genes
589 were classified as mCG-enriched/gbM if they had reads mapping to at least 10

590 CG sites and a q-value <0.05 for mCG, and a q-value >0.05 for mCHG and
591 mCHH.

592

593 DECLARATIONS

594

595 **Acknowledgements.** We thank Nathan Springer for comments and discussions.
596 Kevin Tarner (UGA greenhouse), and Michael Wenzel and Ron Determann
597 (Atlanta Botanical Gardens) for plant tissue. Brigitte T. Hofmeister for
598 establishment and maintenance of the genome browser. We also thank the
599 Georgia Genomics Facility (GGF) for sequencing. Computational resources were
600 provided by the Georgia Advanced Computing Resource Center (GACRC). We
601 thank Gane Ka-Shu Wong and the 1000 Plants initiative (1KP, onekp.com) for
602 advanced access to transcript assemblies.

603

604 **Funding.** The work conducted by the National Science Foundation (NSF) (MCB–
605 1402183) and by The Pew Charitable Trusts to R.J.S.

606

607 **Availability of data and materials.** Genome browsers for all methylation data
608 used in this paper are located at Plant Methylation DB
609 (<http://schmitzlab.genetics.uga.edu/plantmethylomes>). Sequence data for
610 MethylC-seq are located at the Gene Expression Omnibus, accession
611 GSE81702.

612

613 **Authors' contributions.** Conceptualization: AJB, and RJS; Performed
614 experiments: AJB, CEN, LJ, and NAR; Data Analysis: AJB, CEN, and LJ; Writing
615 – Original Draft: AJB; Writing – Review and Editing: AJB, JL-M, and RJS;
616 Resources: JL-M, and RJS. All authors read and approved the final manuscript.

617

618 **Competing interests.** The authors declare that they have no competing
619 interests.

620

621 **Ethics approval.** Ethics approval was not needed for this study.

622

623 **FIGURE LEGENDS**

624

625 **Fig. 1. Phylogenetic relationships of CMTs across Embryophyta.** **a**, CMTs
626 are separated into four monophyletic clades based on bootstrap support and the
627 relationship of *A. thaliana* CMTs: (i) the gbM-dependent CMT superclade with
628 subclades CMT1, CMT3, ZMET and *A. trichopoda*; (ii) CMT2 and; (iii)
629 homologous (hCMT) α and β . CMT1 and CMT3 clades only contain eudicot
630 species of plants suggesting a eudicot-specific duplication event that occurred
631 after the divergence of eudicots from monocots and monocots/commelinids.
632 Sister to CMT1 and CMT3 is the monophyletic group ZMET, which contains
633 monocots, monocots/commelinids, and magnoliids. CMT2 is sister to CMT1 and
634 CMT3. Lastly, the polyphyletic hCMT clades are sister to all previously
635 mentioned clades. HCMT α is sister to CMT2 and the CMT superclade and
636 contains gymnosperm and ferns. HCMT β contains gymnosperms, ferns and
637 other early diverging land plants. **b**, A collapsed CMT gene family tree showing
638 the seven clades described in **a**. Pie charts represent species diversity within
639 each clade, and are scaled to the number of species. Two duplication events
640 shared by all angiosperms (ε) and eudicots (\square) gave rise to what is now referred
641 to as CMT1, CMT2 and CMT3. These duplication events correspond to what was
642 reported by Jiao et al. (2011). Values at nodes in **a** and **b** represent bootstrap
643 support from 1000 replicates, and **a** was rooted to the clade containing all
644 liverwort species.

645

646 **Fig. 2. Non-neutral evolution of CMT3 in the Brassicaceae is correlated with**
647 **reduced levels of genic mCG and numbers of gbM loci.** **a**, Distribution of
648 mCG upstream, downstream and within gene bodies of Brassicaceae species
649 and outgroup species *Prunus persica*. MCG levels within gene bodies of
650 Brassicaceae species are within the bottom 38% of 34 angiosperms. Data used
651 represents a subset of that previously published by [19] and [25]. TSS:

652 transcriptional start site; and TTS: transcriptional termination site. **b**, Similarly the
653 number of gbM genes within the genome of Brassicaceae species are within the
654 bottom 15% of 34 angiosperms. The size of the circle corresponds to the number
655 of gbM genes within each genome. Data used represents a subset of that
656 previously published by [19] and [25]. **c**, Changes at the amino acid level of
657 CMT3 is correlated to reduced genic levels of DNA methylation and number of
658 gbM genes in the Brassicaceae. An overall higher rate of molecular evolution
659 measured as the number of non-synonymous substitutions per non-synonymous
660 site divided by the number of synonymous substitutions per synonymous site (ω)
661 was detected in the Brassicaceae. Also, a higher rate ratio of ω was detected in
662 the Brassicaceae clade containing *B. rapa* and closely related species compared
663 to the clade containing *A. thaliana* and closely related species. The higher rate
664 ratio in the Brassicaceae, compared the background branches, was not attributed
665 to positive selection.

666

667 **Fig. 3. Variation in levels of DNA methylation within gene bodies across**
668 **Viridiplantae.** **a**, DNA methylation at CG, CHG, and CHH sites within gene
669 bodies can be found at the majority of species investigated. Variation of DNA
670 methylation levels within gene bodies at all sequence contexts is high across all
671 land plants, and within major taxonomic groups. mCG levels are typically higher
672 than mCHG, followed by mCHH. However, levels of mCG and mCHG within
673 genes are similar in gymnosperms and ferns. Error bars represent 95%
674 confidence intervals for species with low sequencing coverage. Cladogram was
675 generated from Open Tree of Life [53]. **b**, The distribution of DNA methylation
676 within genes (all [dashed lines] and mCG-enriched/gbM [solid lines]) has
677 diverged among taxonomic groups of Viridiplantae represented by specific
678 species. Based on the distribution of DNA methylation, and number of mCG-
679 enriched genes, gbM is specific to angiosperms. However, mCG-enriched genes
680 in *P. taeda* share some DNA methylation characteristics to *A. trichopoda*.
681 However, other characteristics associated with gbM genes remains unknown at
682 this time for mCG-enriched genes in gymnosperms and other early diverging

683 Viridiplantae. The yellow-highlighted line represents the average from 100
684 random sampling of 100 gbM genes in angiosperms and was used to assess
685 biases in numbers of mCG-enriched genes identified. NCR: non-conversion rate;
686 TSS: transcriptional start site; and TTS: transcriptional termination site.

687

688 **Fig. 4. Presence/absence (+/-) of genes likely involved in the evolution of**
689 **gbM and heterochromatin formation for various taxonomic groups of**
690 **Viridiplantae.** Families (orthogroups) of gbM- and heterochromatin-related
691 genes are taxonomically diverse. However, after phylogenetic resolution, clades
692 containing proteins of known function in *A. thaliana* are less diverse. Specifically,
693 the CMT3 and orthologous genes (ZMET2 and ZMET5, and *A. trichopoda*
694 CMT3), and IBM1 are angiosperm-specific. Other clades – SUVH4 and
695 homologous SUVH5/6 (hSUVH5/6) – are more taxonomically diverse, which
696 might relate to universal functions in heterochromatin formation.

697

698 SUPPLEMENTAL INFORMATION

699

700 **Figure S1. CMT proteins in green algae (*C. reinhardtii*, *Chlorella NC64A*,**
701 **and *V. carteri*) might represent misidentified homologs. a,** A midpoint rooted
702 gene tree constructed from a subset of species and green algae using protein
703 sequences. Previously identified CMT homologs in *C. reinhardtii*, *Chlorella*
704 NC64A, and *V. carteri* (JGI accession ids 190580, 52630, and 94056,
705 respectively) have low amino acid sequence similarity to *A. thaliana* CMT
706 compared to other green algae species (Table S1), which is reflected in long
707 branches, especially for *C. reinhardtii* and *V. carteri*. Values on branches are raw
708 branch lengths represented as amino acid substitutions per amino acid site. **b,**
709 Protein structure of previously identified CMT homologs in *C. reinhardtii*,
710 *Chlorella NC64A*, and *V. carteri* and those identified in green algae from the 1KP
711 dataset. Reported CMTs in *C. reinhardtii* and *Chlorella NC64A* do not contain
712 CHROMO domains, and the homolog in *V. carteri* does not contain any

713 recognizable PFAM domains, however BAH, CHROMO and a DNA methylase
714 domain can all be identified in green algae CMT homologs from the 1KP dataset.
715

716 **Figure S2. Phylogenetic relationships among CMTs in Viridiplantae.** CMTs
717 are separated into four monophyletic clades based on bootstrap support and the
718 relationship of *A. thaliana* CMTs: (i) the gbM-dependent CMT superclade with
719 subclades CMT1, CMT3, ZMET and *A. trichopoda*; (ii) CMT2 and; (iii)
720 homologous (hCMT) α and β. Values at nodes in represent bootstrap support
721 from 1000 replicates, and the tree was rooted to the clade containing all green
722 algae species.

723
724 **Figure S3. Syntenic relationships support a Whole Genome Duplication**
725 **(WGD) event giving rise to CMT1 and CMT3 in eudicots and ZMET paralogs.**
726 **a**, Synteny was determined using CoGe's GEvo program, and is indicated by
727 connected blocks. Synteny is more pronounced in some eudicots over others,
728 which suggests sequence divergence following the shared WGD placed at the
729 base of all eudicots [36]. **b**, Phylogenetic relationships of ZMETs in the Poaceae
730 suggest WGD events are shared by several species and are species-specific as
731 is the case for ZMET2 and ZMET5 in *Z. mays*. Colors following the tip labels
732 indicate clades of paralogous ZMETs. **c**, Similarly to eudicots, WGD is supported
733 by synteny upstream and downstream of ZMET paralogs.

734
735 **Figure S4. Presence and absence of CMTs and ZMETs in eudicots, and**
736 **monocots and monocots/commelinids, respectively.** **a**, Eudicot (basal, core,
737 rosid, and asterid) species of plants possess different combinations of CMT1,
738 CMT2, and CMT3. CMT3 was potentially loss from 46/262 (18%), and CMT1 is
739 found in 106/262 (40%) of eudicot species sequenced by the 1KP Consortium.
740 Species without CMT3 are predicted to have significantly reduced levels of gbM
741 loci compared to eudicot species with CMT3. The presence of CMT1 in
742 numerous species suggests a yet to be determined functional role of CMT1 in
743 DNA methylation and/or chromatin modification. **b**, Similarly to eudicots,

744 monocots and monocots/commelinids have different combinations of CMT2 and
745 ZMET, which may reflect differences in genome structure, and DNA methylation
746 and chromatin modification patterns.

747

748 **Figure S5. Metagene plots of DNA methylation across gene bodies.** DNA
749 methylation levels within all full-length coding sequences or transcripts for
750 additional species used in this study.

751

752 **Figure S6. MCG-enriched genes in species sister to angiosperms are rare**
753 **and not strongly conserved.** **a**, The proportion of mCG-enriched genes are
754 variable across Embryophyta. However, lowest levels are seen in species null for
755 CMT3 and that possess a non-orthologous CMT3 (white circles). Additionally,
756 species that possess a CMT3 that has experienced elevated rates of evolution
757 (ω) have a lower proportion of mCG-enriched genes (gray circles). **b**, The
758 majority of mCG-enriched genes are orthologous to non-mCG-enriched genes in
759 *A. thaliana* or have no hits to an *A. thaliana* gene based on an e-value of $\leq 1E-06$.
760 However, *P. taeda* is an exception, which suggests some of the mCG-enriched
761 genes are conserved to gbM genes in *A. thaliana*.

762

763 **Figure S7. MCG in genes of species sister to angiosperms are biased**
764 **towards extreme low or high levels.** Distributions of mCG across all genes with
765 sufficient coverage for species with sequenced genomes (see Methods).

766

767 **Figure S8. Jumonji (jmjC) domain-containing gene family phylogeny.** The
768 jmjC domain-containing family contains five monophyletic clades based on the
769 location of *A. thaliana* genes. Only angiosperm sequences can be found within
770 the clade containing *A. thaliana* IBM1. Scale bar represents nucleotide
771 substitutions per site.

772

773 **Figure S9. SUVH4 and SUVH5/6 gene family phylogenies.** **a**, SUVH4 gene
774 family approximately recapitulate species relationships, and angiosperm-specific

775 monophyletic clades are not observed based on bootstrap support and the
776 placement of *A. thaliana* SUVH4. **b**, However, Brassicaceae-specific
777 monophyletic clades delineate SUVH5 and SUVH6, hence a homologous
778 SUVH5/6 (hSUVH5/6) sequence is found in other Embryophyta. However, some
779 nodes – especially those delineating monocot and monocot/commelinid
780 hSUVH5/6 sequences – are weakly supported. **c**, Phylogenetic relationships
781 support a Brassicaceae-specific duplication event, which gave rise to SUVH5 and
782 SUVH6. **d**, Reanalyzing monocot and monocot/commelinid hSUVH5/6
783 sequences increases bootstrap support delineating two monophyletic clades.
784 This relationship is analogous to SUVH5 and SUVH6 in Brassicaceae, but
785 encompasses all monocots and monocot/commelinids. Furthermore, Poaceae-
786 specific monophyletic clades are observed within each of the monocot- and
787 monocot/commelinid-specific monophyletic clades. Phylogenetic relationships
788 support multiple duplication events in the monocots and monocot/commelinids.
789 Values at nodes represent bootstrap support and scale bar represents nucleotide
790 substitutions per site.

791

792 **Figure S10. A linear model to determine DNA methylation levels from low**
793 **sequence coverage WGBS.** A strong linear correlation is observed between
794 DNA methylation levels at CG, CHG and CHH sites determined from low,
795 subsampled and full WGBS coverage. A linear model was generated for each
796 sequence context, which was used to extrapolate levels of DNA methylation from
797 species with low WGBS coverage. Each data point represents a single plant
798 species from [19, 20, 25, 40].

799

800 **Table S1. Taxonomic, sequence, and phylogenetic summary of sequences**
801 **used in Fig. 1 and Supplementary Fig. 2.**

802

803 **Table S2. Best BLASTp hits of published green algae CMTs suggest mis-**
804 **annotation compared to green algae CMTs identified in the current study.**

805

806 **Table S3. A summary of branch and branch-site tests implemented in
807 PAML.**

808

809 **Table S4. Reduced (1x) and deep sequencing coverage estimates of DNA
810 methylation levels from 34 Viridiplantae species.**

811

812 **Table S5. DNA methylation levels of species sequenced in this study and
813 the levels predicted by a context-specific linear model.**

814

815 **REFERENCES**

816

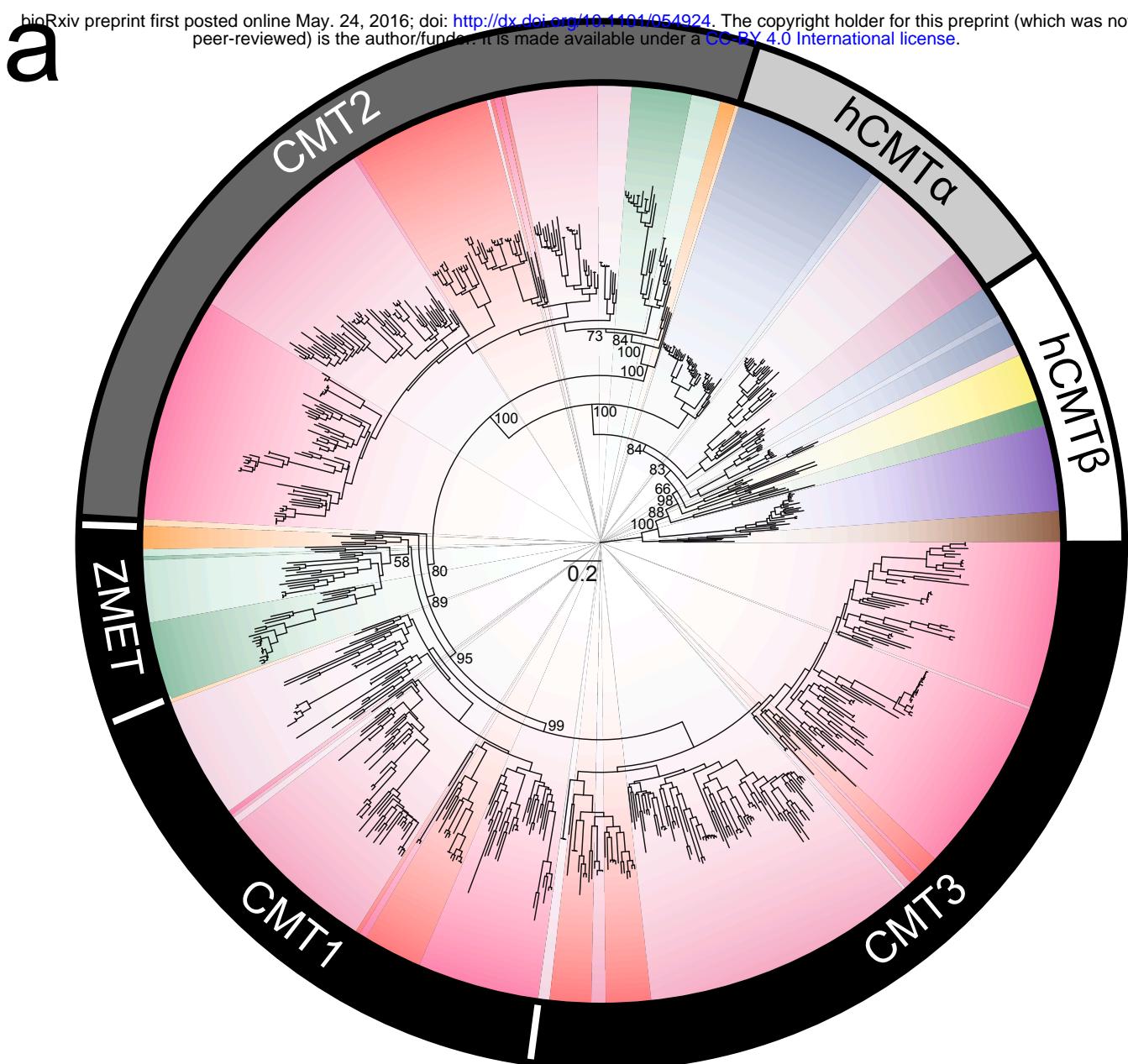
- 817 1. Bartee L, Malagnac F, Bender J. Arabidopsis cmt3 chromomethylase
818 mutations block non-CG methylation and silencing of an endogenous gene.
819 *Genes Dev.* 2001;15:1753–58.
- 820 2. Jackson JP, Lindroth AM, Cao X, Jacobsen SE. Control of CpNpG DNA
821 methylation by the KRYPTONITE histone H3 methyltransferase. *Nature.*
822 2002;416:556–60.
- 823 3. Zemach A, et al. The Arabidopsis nucleosome remodeler DDM1 allows DNA
824 methyltransferases to access H1-containing heterochromatin. *Cell.*
825 2013;153:193–205.
- 826 4. Stroud H, et al. Non-CG methylation patterns shape the epigenetic landscape
827 in Arabidopsis. *Nat. Struct. Mol. Biol.* 2014;21:64–72.
- 828 5. Papa CM, Springer NM, Muszynski MG, Meeley R, Kaepller SM. Maize
829 chromomethylase Zea methyltransferase2 is required for CpNpG methylation.
830 *Plant Cell.* 2001;13:1919–28.
- 831 6. Du J, et al. Dual binding of chromomethylase domains to H3K9me2-
832 containing nucleosomes directs DNA methylation in plants. *Cell.*
833 2012;151:167–80.
- 834 7. Li Q, et al. Genetic perturbation of the maize methylome. *Plant Cell.*
835 2014;26:4602–16.

- 836 8. Henikoff S, Comai L. A DNA methyltransferase homolog with a
837 chromodomain exists in multiple polymorphic forms in *Arabidopsis*. *Genetics*.
838 1998;149:307–18.
- 839 9. Finnegan EJ, Kovac KA. Plant DNA methyltransferases. *Plant Mol. Biol.*
840 2000;43:189–201.
- 841 10. McCallum CM, Comai L, Greene EA, Henikoff S. Targeted screening for
842 induced mutations. *Nature Biotechnol.* 2000;18:455–57.
- 843 11. Shen X, et al. Natural CMT2 variation is associated with genome-wide
844 methylation changes and temperature seasonality. *PLoS Genet.*
845 2014;10:e1004842.
- 846 12. Bewick AJ, Schmitz RJ. Epigenetics in the wild. *eLife*. 2015;
847 doi:10.7554/eLife.07808.
- 848 13. Dubin MJ, et al. DNA methylation in *Arabidopsis* has a genetic basis and
849 shows evidence of local adaptation. *eLife*. 2015; doi:10.7554/eLife.05255.
- 850 14. Du J, et al. Mechanism of DNA methylation-directed histone methylation by
851 KRYPTONITE. *Mol. Cell*. 2014;55:495–504.
- 852 15. Du J, Johnson LM, Jacobsen SE, Patel DJ. DNA methylation pathways and
853 their crosstalk with histone methylation. *Nat. Rev. Mol. Cell Biol.*
854 2015;16:519–32.
- 855 16. Hou PQ, et al. Functional characterization of *Nicotiana benthamiana*
856 chromomethylase 3 in developmental programs by virus-induced gene
857 silencing. *Physiologia Plantarum*. 2013;150:119–32.
- 858 17. Garg R, Kumari R, Tiwari S, Goyal S. Genomic survey, gene expression
859 analysis and structural modeling suggest diverse roles of DNA
860 methyltransferases in legumes. *PLoS One*. 2014;9:e88947.
- 861 18. Lin YT, Wei HM, Lu HY, Lee YI, Fu SF. Developmental- and tissue-specific
862 expression of NbCMT3-2 encoding a chromomethylase in *Nicotiana*
863 benthamiana. *Plant Cell Physiol*. 2015;56:1124–43.
- 864 19. Bewick AJ, et al. On the origin and evolutionary consequences of gene body
865 DNA methylation. *Proc. Natl Acad. Sci. USA*. 2016;113:9111–16.

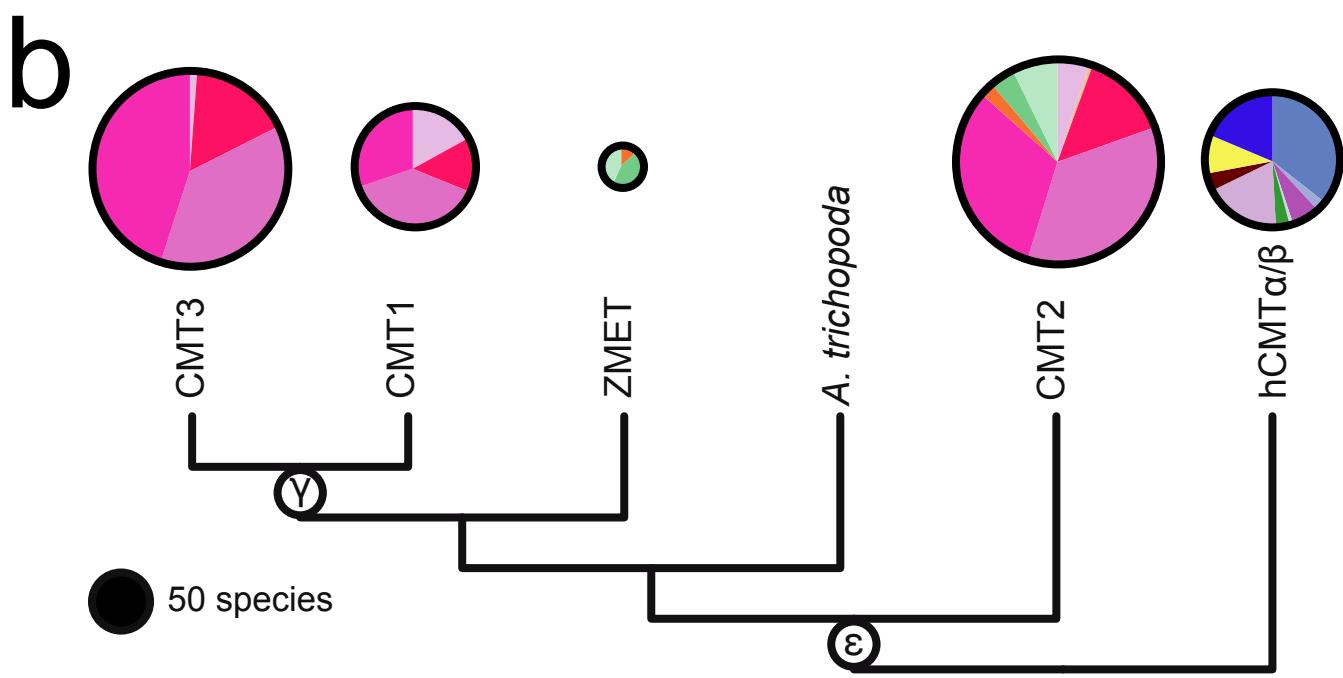
- 866 20. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary
867 analysis of eukaryotic DNA methylation. *Science*. 2010;328:916-9.
- 868 21. Noy-Malka C, et al. A single CMT methyltransferase homolog is involved in
869 CHG DNA methylation and development of *Physcomitrella patens*. *Plant Mol*
870 *Biol*. 2014;84:719–35.
- 871 22. Tran RK, et al. DNA methylation profiling identifies CG methylation clusters in
872 *Arabidopsis* genes. *Curr. Biol*. 2005;15:154–9.
- 873 23. Zhang X, et al. Genome-wide high-resolution mapping and functional analysis
874 of DNA methylation in *Arabidopsis*. *Cell*. 2006;126:1189–201.
- 875 24. Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. Genomewide
876 analysis of *Arabidopsis thaliana* DNA methylation uncovers an
877 interdependence between methylation and transcription. *Nat. Genet*.
878 2007;39:61–9.
- 879 25. Niederhuth CE, et al. Widespread natural variation of DNA methylation within
880 angiosperms. *Genome Biol*. 2016;17:194.
- 881 26. Takuno S, Gaut BS. Body-methylated genes in *Arabidopsis thaliana* are
882 functionally important and evolve slowly. *Mol. Biol. Evol*. 2012;1:219-27.
- 883 27. Takuno S, Gaut BS. Gene body methylation is conserved between plant
884 orthologs and is of evolutionary consequence. *Proc. Natl Acad. Sci. USA*.
885 2013;110:1797–802.
- 886 28. Inagaki S, Kakutani T. What triggers differential DNA methylation of genes
887 and TEs: contribution of body methylation? *Cold Spring Harb Symp Quant*
888 *Biol*. 2012;77:155–60.
- 889 29. Johnson LM, et al. SRA- and SET-domain-containing proteins link RNA
890 polymerase V occupancy to DNA methylation. *Nature*. 2014;507:124–8.
- 891 30. Kenrick P, Crane PR. The origin and early evolution of plants on land. *Nature*.
892 1997;389:33–9.
- 893 31. Wellman CH, Osterloff PL, Mohiuddin U. Fragments of the earliest land
894 plants. *Nature*. 2003;425:282–5.
- 895 32. Steemans P, et al. Origin and radiation of the earliest vascular land plants.
896 *Science*. 2009;324:353.

- 897 33. Rubinstein CV, Gerrienne P, de la Puente GS, Astini RA, Steemans P. Early
898 Middle Ordovician evidence for land plants in Argentina (eastern Gondwana).
899 *New Phytologist*. 2010;188:365–9.
- 900 34. Stiller JW, Hall BD. The origin of red algae: Implications for plastid evolution.
901 *Proc. Natl Acad. Sci. USA*. 1997;94:4520–5.
- 902 35. Bhattacharya D, Medlin L. Algal phylogeny and the origin of land plants. *Plant*
903 *Physiol.* 1998;116:9–15.
- 904 36. Jiao Y, et al. Ancestral polyploidy in seed plants and angiosperms. *Nature*.
905 2011;473:97–100.
- 906 37. Wickett NJ, et al. Phylogenomic analysis of the origin and early
907 diversification of land plants. *Proc. Natl Acad. Sci. USA*. 2014;111:E4859–68.
- 908 38. Malik G, Dangwal M, Kapoor S, Kapoor M. Role of DNA methylation in growth
909 and differentiation in *Physcomitrella patens* and characterization of cytosine
910 DNA methyltransferases. *FEBS J.* 2012;279:4081–94.
- 911 39. Feng S, et al. Conservation and divergence of methylation patterning in plants
912 and animals. *Proc. Natl Acad. Sci. USA*. 2010;107:8689–94.
- 913 40. Takuno S, Ran J-H, Gaut BS. Evolutionary patterns of genic DNA methylation
914 vary across land plants. *Nature Plants*. 2016;15222:
915 doi:10.1038/nplants.2015.222.
- 916 41. Lippman Z, May B, Yordan C, Singer T, Martienssen R. Distinct mechanisms
917 determine transposon inheritance and methylation via small interfering RNA
918 and histone modification. *PLoS Biol.* 2003;1:E67.
- 919 42. Qin FJ, Sun QW, Huang LM, Chen XS, Zhou DX. Rice SUVH histone
920 methyltransferase genes display specific functions in chromatin modification
921 and retrotransposon repression. *Mol Plant*. 2010;3:773–82.
- 922 43. Stroud H, Greenber MVC, Feng S, Bernatavichute YV, Jacobsen SE.
923 Comprehensive analysis of silencing mutants reveals complex regulation of
924 the *Arabidopsis* methylome. *Cell*. 2013;152:352–64.
- 925 44. Lynch M, Conery JS. The evolutionary fate and consequences of duplicate
926 genes. *Science*. 2000;290:1151–5.

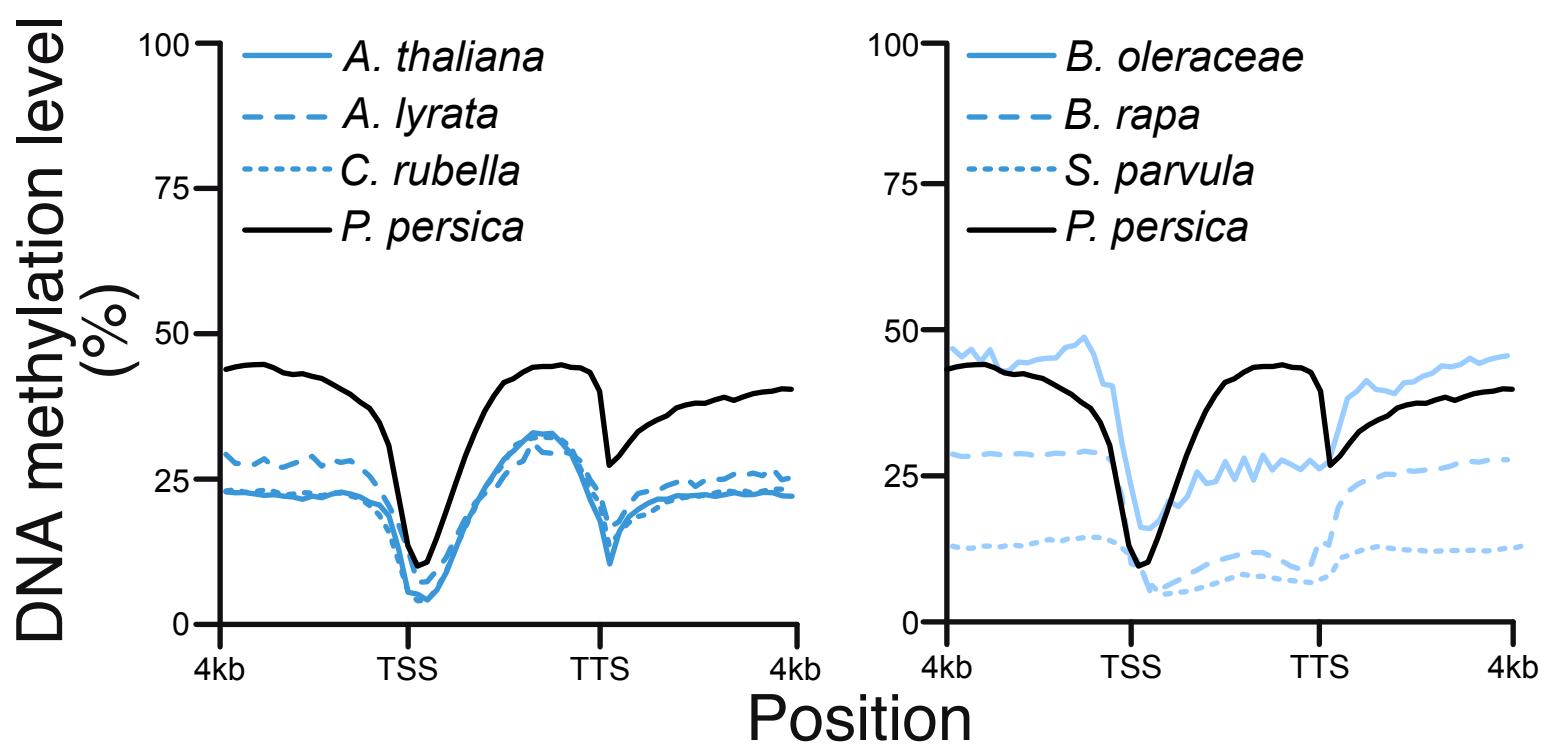
- 927 45. Aceituno FF, Moseyko N, Rhee SY, Gutiérrez RA. The rules of gene
928 expression in plants: organ identity and gene body methylation are key
929 factors for regulation of gene expression in *Arabidopsis thaliana*. *BMC*
930 *Genomics*. 2008;9: doi:10.1186/1471-2164-9-438.
- 931 46. Jones P. et al. InterProScan 5: genome-scale protein function classification.
932 *Bioinformatics*. 2014;30:1236–40.
- 933 47. Mirarab S, et al. PASTA: Ultra-Large Multiple Sequence Alignment for
934 Nucleotide and Amino-Acid Sequences. *J Comput Biol*. 2015;22:377–86.
- 935 48. Stamatakis A. RAxML Version 8: A tool for phylogenetic analysis and
936 postanalysis of large phylogenies. *Bioinformatics*. 2014;30:1312–3.
- 937 49. Castresana J. Selection of conserved blocks from multiple alignments for their
938 use in phylogenetic analysis. *Mol. Biol. Evol*. 2000;17:540–52.
- 939 50. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol.*
940 *Evol*. 1997;24:1586–91.
- 941 51. Urich, M. A., Nery, J. R., Lister, R., Schmitz, R. J. & Ecker, J. R. MethylCseq
942 library preparation for base-resolution whole-genome bisulfite sequencing.
943 *Nat. Protoc.* 2015;10:475–83.
- 944 52. Schultz MD, et al. Human body epigenome maps reveal noncanonical DNA
945 methylation variation. *Nature*. 2015;523:212–6.
- 946 53. Hinchliff CE, et al. Synthesis of phylogeny and taxonomy into a
947 comprehensive tree of life. *Proc. Natl Acad. Sci. USA*. 2015;112:12764–9.



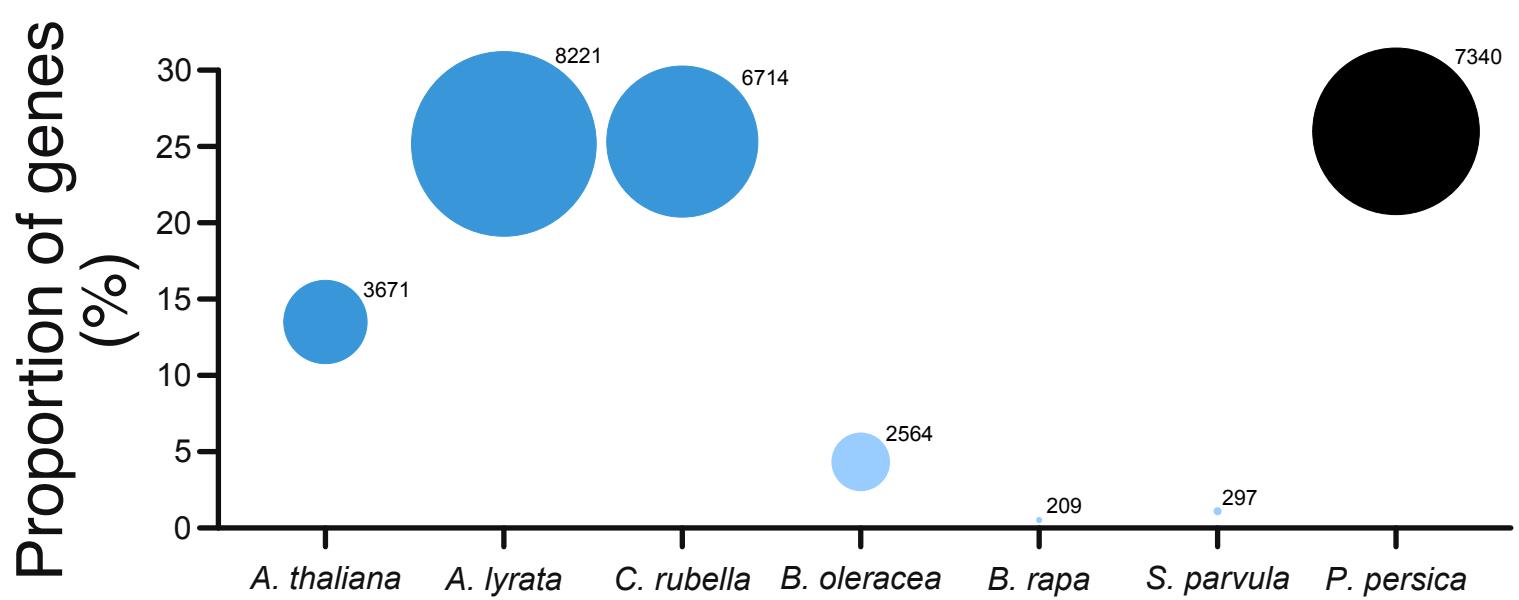
Embryophytes																	
Tracheophytes																	
Angiosperms										Gymnosperms			Ferns	Lyc.	Bryophytes		
Eudicots	Rosids	Asterids	Basal	Com.	Monocots	Mag.	Bas.-mos.	Conifers	Cyc.	Gin.	Eus.	Lep.	Lyc.	Hor.	Mosses	Liv.	
Core																	



a



b



c

