

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

3
4 **RUNNING TITLE:** CMT gene family in plants

5
6 Adam J. Bewick^{1*}, Chad E. Niederhuth¹, Nicolas A. Rohr¹, Patrick T. Griffin¹, Jim
7 Leebens-Mack², Robert J. Schmitz¹

8
9 ¹Department of Genetics, University of Georgia, Athens, GA 30602, USA

10 ²Department of Plant Biology, University of Georgia, Athens, GA 30602, USA

11
12 **CO-CORRESPONDING AUTHORS:** Adam J. Bewick, bewickaj@uga.edu and
13 Robert J. Schmitz, schmitz@uga.edu

14
15 **KEY WORDS:** CHROMOMETHYLTRANSFERASE, Phylogenetics, DNA
16 methylation, WGBS

17
18 **WORD COUNT:** ≈4582

19
20 **TABLE COUNT:** 0 main, 1 SI

21
22 **FIGURE COUNT:** 5 main, 5 SI

23
24

25 ABSTRACT

26

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

41

42 INTRODUCTION

43

44 DNA methylation is an important chromatin modification that protects the genome
45 from selfish genetic elements, is sometimes important for proper gene
46 expression, and is involved in genome stability. In plants, DNA methylation is
47 found at cytosines in three sequence contexts: CG, CHG, and CHH (H is any
48 base, but G). Establishment and maintenance of DNA methylation at these three
49 sequence contexts is performed by a suite of distinct *de novo* and maintenance
50 DNA methyltransferases, respectively. CHROMOMETHYLTRANSFERASES
51 (CMTs) are an important class of plant-specific DNA methylation maintenance
52 enzymes, which are characterized by the presence of a CHROMATIN
53 ORGANISATION MODIFIER (CHROMO) domain between the cytosine
54 methyltransferase catalytic motifs I and IV¹. Identification, expression and
55 functional characterization of CMTs have been extensively performed in the
56 model plant *Arabidopsis thaliana*²⁻⁴ and in the model grass species *Zea mays*⁵.
57 Homologous CMTs have been identified in other flowering plants⁶⁻⁸, the moss
58 *Physcomitrella patens*, the lycophyte *Selaginella moellendorffii*, and the green
59 algae *Chlorella* NC64A and *Volvox carter*⁹. There are three CMT genes encoded
60 in *A. thaliana*: CMT1, CMT2, and CMT3^{2,10-12}. CMT1 is the least studied of the
61 three chromomethyltransferases as a handful of *A. thaliana* accessions contain
62 an EVELKNIEVEL retroelement insertion or a frameshift mutation truncating the
63 protein, which suggested that CMT1 is nonessential¹⁰. The majority of DNA
64 methylation present in pericentromeric regions of the genome composed of long
65 transposable elements is targeted by a CMT2-dependent pathway^{4,3}. Allelic
66 variation at CMT2 has been shown to alter genome-wide levels of CHH DNA
67 methylation, and plastic alleles of CMT2 may play a role in adaptation to
68 temperature¹³⁻¹⁵. Whereas DNA methylation at CHG sites is often maintained by

69 CMT3 through a reinforcing loop with histone H3 lysine 9 di-methylation
70 (H3K9me2) catalyzed by the KRYPTONITE (KYP)/SUVH4 lysine
71 methyltransferase^{5,16}.

72 A large number of plant genes exclusively contain CG DNA methylation in
73 the transcribed region and a depletion of CG DNA methylation from both the
74 transcriptional start and stop sites (referred to as “gene body DNA methylation”,
75 or “gbM”)¹⁷⁻¹⁹. The penetrance of this DNA methylation pattern is strong, and can
76 be observed without extracting gbM genes from metaplots²⁰. The current model
77 for the evolution of gbM relies on rare transcription-coupled
78 incorporation/methylation of histone H3K9me2 in gene bodies with subsequent
79 failure of INCREASED IN BONSAI METHYLATION 1 (IBM1) to de-methylate
80 H3K9me2²¹. This provides a substrate for CMT3 to bind and methylate DNA and
81 through an unknown mechanism leads to CG DNA methylation, which is
82 maintained over evolutionary timescales by CMT3 and through DNA replication
83 by MET1. Methylated DNA then provides a substrate for KYP and related family
84 members, which increases the rate at which H3K9 is methylated^{21,22}. Finally,
85 gene body methylation spreads throughout the gene over evolutionary time²¹.
86 Support for this model is evident in the eudicot species *Eutrema salsugineum*
87 and *Conringia planisiliqua* (family Brassicaceae), which have independently lost
88 CMT3 and gbM²¹. The loss of CMT3 in these species phenocopies the reduced
89 levels of CHG DNA methylation found in the *A. thaliana cmt3* mutants²³. Closely
90 related Brassicaceae species have reduced levels of CHG DNA methylation on a
91 per cytosine basis and have reduced gbM relative to other species, but still
92 possess CMT3^{20,21}, which indicates changes at the molecular level may have
93 disrupted function of CMT3.

94 Previous phylogenetic studies have proposed that CMT1 and CMT3 are
95 more closely related to each other than to CMT2, and that ZMET2 and ZMET5
96 proteins are more closely related to CMT3 than to CMT1 or CMT2⁶, and the
97 placement of non-seed plant CMTs more closely related to CMT3²⁴. However,
98 these studies were not focused on resolving phylogenetic relationships within the
99 CMT gene family, but rather relationships of CMTs between a handful of species.
100 These studies have without question laid the groundwork to understand CMT-
101 dependent DNA methylation pathways and patterns in plants. However, the
102 massive increase in transcriptome data for a broad sampling of plant species
103 together with advancements in sequence alignment and phylogenetic inference
104 algorithms have made it possible to incorporate thousands of sequences into a
105 single phylogeny, allowing for a more complete understanding of the CMT gene
106 family. A comprehensive plant CMT gene phylogeny has implications for
107 mechanistic understanding of DNA methylation across the plant kingdom.
108 Additionally, advancements in high throughput screening of DNA methylation has
109 made it possible to get accurately estimate genome-wide levels from species
110 without a sequenced reference assembly²⁵. Understanding the evolutionary
111 relationships of CMT proteins is foundational for inferring the evolutionary origins,
112 maintenance, and consequences of genome-wide DNA methylation and gbM.

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

129

130 RESULTS AND DISCUSSION

131

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

152 The relationships among CMTs suggests that CMT2 and CMT ϵ lineages
153 arose from a duplication event within a common ancestor of all flowering plant
154 species (Fig. 1C). Relationships among non-angiosperm CMTs largely
155 recapitulate species relationships³¹ but the existence of two distinct clades for
156 both gymnosperms and ferns suggesting gene duplication and loss events early

157 in vascular plant evolution. The non-angiosperm CMT protein-coding genes
158 analyzed here include those previously identified in the lycophyte *S.*
159 *moellendorffii*⁹ and the moss *P. patens*^{9,33}. CMTs identified in a few green algae
160 from the 1KP sequencing data also belong to this clade (Fig. S1). However,
161 previously identified CMT proteins in *Chlamydomonas reinhardtii*, *Chlorella*
162 NC64A and *Volvox carteri* were not included in the CMT gene family clade
163 because they lacked the CHROMO and other domains typically associated with
164 CMT proteins (Fig. S2B), and *C. reinhardtii* and *V. carteri* sequences are more
165 homologous to METHYTRANSFERASE 1 (MET1) than other green algae CMT
166 (Table S1). The greatly increased sampling of non-angiosperm CMTs defines the
167 understanding of relationships of CMTs in early land plants and plants in
168 general^{9,24,33,34}.

169 Further diversification of CMT proteins occurred in the eudicots; a second
170 duplication gave rise to what is now called CMT1 and CMT3 (Fig. 1C). Thus,
171 CMT1 and CMT3 are eudicot-specific, and the monocot and magnoliid CMTs
172 (CMT $\epsilon^{\text{monocot+magnoliid}}$) are sister to both CMT1 and CMT3 (Fig. 1C). CMT1, CMT3
173 and CMT $\epsilon^{\text{monocot+magnoliid}}$ share a common ancestor with basal angiosperms
174 (CMT $\epsilon^{\text{basal angiosperm}}$). This and the previously mentioned angiosperm-specific
175 duplication events may have coincided with the ancestral angiosperm whole
176 genome duplication (WGD) event or the eudicot WGD event, respectively³⁵. Not
177 all eudicots possess CMT1, CMT2 and CMT3, but rather species exhibit CMT
178 gene content ranging from zero to three of these homologs, suggesting multiple
179 independent losses and fluidity of the functionality of these proteins (Fig. S3 and
180 Table S1).

181 Functional differences among these paralogs has been characterized in *A.*
182 *thaliana*; CMT3 is functional and is required for the maintenance of DNA
183 methylation at CHG sites, however allelic diversity of CMT1 is high, which
184 suggests alleles are segregating neutrally or under relaxed selection in
185 populations and may not serve a function in the maintenance of CHG DNA
186 methylation. Additionally, CMT1 does not compensate CMT3 function *in vivo* in
187 *A. thaliana* $\Delta cmt3$ ²³. The expression and persistence of CMT1 in numerous other
188 eudicots raises the possibility of functional divergence and convergence over the
189 evolutionary history of eudicots. For example, CMT1 may have evolved a novel
190 function (neofunctionalization), CMT1 and CMT3 may both be required to achieve
191 the same phenotype (subfunctionalization) or CMT1 may have a redundant
192 function to that of CMT3 (redundancy). The latter possibility does not seem to be
193 the case for naturally occurring $\Delta cmt3$ species, *E. salsugineum*, as CMT1 does
194 not recover levels of CHG DNA methylation as expected²¹. However,
195 subfunctionalization of CMT1 in *E. salsugineum* seems plausible since DNA
196 methylation at CHG sites still occurs at low levels²¹. The exact fate of CMT1 and
197 interplay between these two paralogs in shaping the epigenome remains
198 unknown.

199 The *Zea mays*-specific ZMET2 and ZMET5, and closely related CMT
200 proteins in other monocots, monocots/commelinids, and magnoliids form a well-

201 supported monophyletic clade with bootstrap support of 95%
202 ($\text{CMT}\epsilon^{\text{monocot+magnoliid}}$, Fig. 1). The inclusion of magnoliids in the
203 $\text{CMT}\epsilon^{\text{monocot+magnoliid}}$ clade is interesting because monophyletic support of the
204 magnoliidae as basal to all monocots has been hypothesized^{36,37}. Akin to
205 eudicots, monocots possess combinations of $\text{CMT}\epsilon^{\text{monocot+magnoliid}}$ and CMT2 (Fig.
206 S4). For example, the model grass species *Z. mays* has loss CMT2, whereas the
207 closely related species *Sorghum bicolor* possess both $\text{CMT}\epsilon^{\text{monocot+magnoliid}}$ and
208 CMT2 (Fig. S3). $\text{CMT}\epsilon^{\text{monocot+magnoliid}}$ is not strictly homologous to CMT3, and
209 represents a unique monophyletic group that is co-orthologous to CMT1 and
210 CMT3. However, ZMET2 is functionally orthologous to CMT3 and maintains DNA
211 methylation at CHG sites⁵. But, unlike CMT3, ZMET2 is associated with DNA
212 methylation at CHH sites within some loci³⁸. Given the inclusion of monocot and
213 magnoliid species the monophyletic $\text{CMT}\epsilon^{\text{monocot+magnoliid}}$ clade, this dual-function
214 is expected to be present in other monocot species, and magnoliid species.

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

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

241
242 **Non-neutral evolution of CMT contributes to the epigenome of flowering**
243 **plants.** Gene body methylation can be found in all angiosperms investigated to
244 date, with the exception of a couple species in the Brassicaceae^{20,21}. Shared by

245 all angiosperms are CMTs belonging to the CMT ϵ clade (Fig. 1), suggesting the
246 evolution of this clade is tightly linked to the evolution of gbM in plants. For
247 example, recent work identified independent events that lead to the loss of CMT3
248 in *E. salsugineum* and *C. planisiliqua*, which is linked to the loss of gbM over
249 evolutionary time²¹. Peculiarly, closely related species with CMT3 – *Brassica*
250 *oleracea*, *Brassica rapa* and *Schrenkiella parvulum* (Fig. 3 Clade B) – have
251 reduced numbers of gbM loci compared to a sister clade – *A. thaliana*,
252 *Arabidopsis lyrata* and *Capsella rubella* (Fig. 3 Clade C) – and other eudicot
253 species²⁰. At the molecular level, CMT3 has evolved under less selective
254 constraint – measured as dN/dS (ω) – in the Brassicaceae (Clade A) ($\omega=0.175$)
255 compared to 162 eudicot species ($\omega=0.0961$), and with reduced apparent
256 constraint in the clade containing *B. oleracea*, *B. rapa* and *S. parvulum*
257 ($\omega=0.241$) compared to the clade containing *A. thaliana*, *A. lyrata* and *C. rubella*
258 ($\omega=0.164$) (Fig. 3). A hypothesis of positive selection was not preferred to
259 contribute to the increased rates of ω in either clade (Fig. 3). Relaxed selective
260 constraint may be associated with decreased gbM in some members of the
261 Brassicaceae. The severity of these substitutions within Brassicaceae clades A,
262 B and C (Fig. 3) effects the respective epigenomes differently, and substitutions
263 at CMT3 have progressed over evolutionary time from compromising levels of
264 CHG DNA methylation (Clades A and C) to the number of gbM loci (Clade B).

265 It is conceivable that additional independent losses or non-neutral
266 evolution of CMT ϵ has shaped the epigenome of other species of plants. Several
267 eudicot species of plants possess truncated annotations of CMT3 or CMT3 was
268 absent from the assembled transcriptomes. These species often possessed low
269 levels (e.g., *Kalanchoe tomentosa*) of CG DNA methylation in gene bodies (Table
270 S1; Fig. S4). However, other species missing CMT3 showed similar patterns of
271 DNA methylation to species with CMT3 (e.g., *S. dulcificum* and *A. thaliana*),
272 which may represent false-negatives (Table S1; Fig. S4). Alternatively, not
273 enough time has passed since the loss of CMT3 to effect levels of CG DNA
274 methylation within gene bodies. Co-orthologous proteins to CMT3, including
275 CMT $\epsilon^{\text{monocot+magnoliid}}$ and CMT $\epsilon^{\text{basal angiosperm}}$, are similar to CMT3 functionally⁵ and
276 at the amino acid level to CMT3 (e.g., ZMET2 shares 417/915 amino acid sites),
277 thus similar phenotypic consequences are expected for naturally occurring
278 CMT $\epsilon^{\text{monocot+magnoliid}}$ and CMT $\epsilon^{\text{basal angiosperm}}$ mutants. The monocot *Dioscorea*
279 *elephantipes* is missing CMT $\epsilon^{\text{monocot+magnoliid}}$ from its assembled transcriptome,
280 and has low levels of CG DNA methylation within gene bodies (Table S1).
281 However, this observation is from only a couple dozen assembled transcripts.

282
283 **CG DNA methylation within gene bodies of non-flowering plants is not**
284 **typical of gbM patterns in angiosperms.** CG DNA methylation within gene
285 bodies can be found in most taxonomic groups with lycophyte, moss and
286 liverwort as the exceptions (Fig. 4A). Similar to what has been documented in
287 angiosperms²⁰ there exists substantial variation of DNA methylation across non-
288 flowering plants, and within taxonomic groups of non-flowering plants (Fig. 4A;

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

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

329

330 **CONCLUSION**

331

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

341 We hypothesize that the angiosperm-specific duplication ≥ 236 MYA³⁵
342 gave rise to what is now CMT2 and CMT ϵ , and gbM in angiosperms.
343 Furthermore, the eudicot-specific duplication event (≥ 134 MYA³⁵) gave rise to the
344 eudicot-specific proteins CMT1 and CMT3. Eudicot species without CMT3 have
345 loss of gbM, and currently at least three independent losses have been
346 supported with phylogenetic analyses and sequencing data. Non-neutral
347 evolution at CMT3 has played an important role in shaping the epigenome of the
348 Brassicaceae. Reduced selective constraint at CMT3 has reduced levels of CHG
349 DNA methylation in the Brassicaceae, and further reductions in *B. oleracea*, *B.*
350 *rapa*, and *S. parvulum* have reduced levels of gbM and the number of gbM loci.
351 Similarly, corresponding phenotypic consequences for naturally occurring
352 mutants of co-orthologous proteins to CMT3, CMT ϵ ^{monocot+magnoliid}, has been
353 observed in the monocot *D. elephantipes*. Also, it seems plausible that the loss of
354 CMT would lead to reductions in the histone modification H3K9me2, and this
355 could subsequently lead to genomic structural consequences. The majority of
356 non-flowering plants do not possess signatures of DNA methylation within gene
357 bodies that is typical of species with gbM. Convergent evolution of gene body
358 methylation, through a CMT mechanism, may be present in species outside of
359 flowering plants, however support for this hypothesis remains sparse.

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

371

372 MATERIALS AND METHODS

373

374 1KP sequencing, transcriptome assembling and orthogrouping.

375 Transcriptome data from the One Thousand Plants (1KP) Consortium for a total

376 of 1329 species were included in this analysis, including both amino acid and
377 coding sequence (Table S1). Additionally, gene annotations from 20 additional
378 species – *A. lyrata*, *Br. distachyon*, *B. oleracea*, *B. rapa*, *Cit. clementina*, *Ca.*
379 *rubella*, *Can. sativa*, *C. sativus*, *E. salsugineum*, *F. vesca*, *G. max*, *Go. raimondii*,
380 *L. japonicus*, *M. domestica*, *Me. truncatula*, *Pan. hallii*, *Pan. virgatum*, *R.*
381 *communis*, *Se. viridis*, and *Z. mays* – from Phytozome were included. The CMT
382 gene family was extracted from the previously compiled orthogroupings using the
383 *A. thaliana* gene identifier for CMT1, CMT2 and CMT3. This orthogroup
384 determined by the 1KP Consortium included all three *A. thaliana* CMT proteins,
385 and a total of 5383 sequences. Sequences from species downloaded from
386 Phytozome, that were not included in sequences generated by 1KP, were
387 included to the gene family through reciprocal best BLAST with *A. thaliana*
388 CMT1, CMT2 and CMT3. In total the CMT gene family included 5449 sequences
389 from 1043 species. We used the protein structure of *A. thaliana* as a reference to
390 filter the sequences found within the CMT gene family. Sequences were retained
391 if the included the same base pfam domains as *A. thaliana* – CHROMO
392 (CHRromatin Organisation MOdifier) domain, BAH domain, and C-5 cytosine-
393 specific DNA methylase – as identified by Interproscan⁴³. These filtered
394 sequences represent a set of high-confident, functional, ideal CMT proteins,
395 which included 773 sequences from 432 species, and were used for phylogenetic
396 analyses.

397
398 **Phylogeny construction.** To estimate the gene tree for the CMT sequences, a
399 series of alignment and phylogenetic estimation steps were conducted. An initial
400 protein alignment was carried out using Pasta with the default settings⁴⁴. The
401 resulting alignment was back-translated using the CDS sequence into in-frame
402 codons using a custom Perl script. A phylogeny was estimated by RAxML⁴⁵ (-m
403 GTRGAMMA) with 1000 rapid bootstrap replicates using the in-framed aligned
404 sequences, and with only the first and second codon positions. Long branches
405 can effect parameter estimation for the substitution model, which can in turn
406 degrade phylogenetic signal. Therefore, phylogenies were constructed with and
407 without green algae species, and were rooted to the green algae clade or
408 liverworts, respectively. The species *Balanophora fungosa* has been reported to
409 have a high substitution rate, which can also produce long branches, and was
410 removed prior to phylogenetic analyses.

411
412 **Codon analysis.** Similar methodology as described above was used to construct
413 phylogenetic trees for testing hypotheses on the rates of evolution in a
414 phylogenetic context. However, the program Gblocks⁴⁶ was used to identify
415 conserved amino acids in codon alignments. The parameters for Gblocks were
416 kept at the default settings, except we allowed for %50 gapped positions. The
417 program Phylogenetic Analysis by Maximum Likelihood (PAML)⁴⁷ was used to
418 test branches (branch test) and sites along branches (branch-site test) for
419 deviations from the background rate of molecular evolution (dN/dS ; ω) and for

420 deviations from the neutral expectation, respectively. Branches tested and a
421 summary of each test can be found in Table S1.

422

423 **MethylC-seq.** MethylC-seq libraries were prepared according to the following
424 protocol⁴⁸. Prior to mapping each transcript was searched for the longest open
425 reading frame from all six possible frames, and only transcripts beginning with a
426 start codon and ending with one of the three stop codons were kept. All
427 sequencing data for each species was aligned to their respective transcriptome
428 or species within the same genus using the methylpy pipeline⁴⁹, which can be
429 found in Table S1. Weighted methylation was calculated for each sequence
430 context (CG, CHG and CHH) by dividing the total number of aligned methylated
431 reads by the total number of methylated plus un-methylated reads. Since, per site
432 sequencing coverage was low – on average ~1X – subsequent binomial tests
433 could not be performed to bin genes as gbM²⁰. To investigate the affect of low
434 coverage we compared levels of DNA methylation of 1X randomly sampled
435 MethylC-seq reads to actual levels for 33 angiosperm species and *Chlorella*
436 *NC64A*^{9,20}. On average DNA methylation levels determined from 1X sequencing
437 coverage are 1.28, 2.63, and 2.17 times higher at CG, CHG and CHH sequence
438 contexts compared to the actual levels within coding regions. These values were
439 used to adjust levels of DNA methylation within coding regions for species of
440 plants sequenced in this study. Genome-wide levels of DNA methylation were
441 estimated using *FAST^mC* and the *plant* model²⁵.

442

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

458

459 **ACKNOWLEDGEMENTS**

460

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

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

469

470 **FIGURE LEGENDS**

471

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

494

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

506

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

520
521 **Figure 4. Gene-body methylation (gbM) is unique to flowering plants.** (A)
522 DNA methylation at CG, CHG, and CHH sites within gene bodies can be found at
523 the majority of species investigated. DNA methylation within gene bodies at any
524 sequence context was not observed in lycophyte, moss and liverwort. Variation of
525 DNA methylation levels within gene bodies at all sequence contexts is high
526 across all land plants, and within major taxonomic groups. CG DNA methylation
527 levels are typically higher than CHG, followed by CHH. However, levels of CG
528 and CHG DNA methylation within genes are similar in several gymnosperms and
529 ferns, and the ratio of CG:CHG is significantly lower (T-test) in gymnosperms and
530 ferns compared to flowering plants. (B) Levels of CG and CHG DNA methylation
531 are highly correlated in non-flowering plants compared to flowering plants. (C)
532 Also, CG and CHG DNA methylation tend to mirror one another throughout the
533 gene bodies of non-flowering plants as opposed to flowering plants. Together
534 these results suggest a relationship between mechanism(s) involved in
535 methylating cytosines at CG and CHG sites within gene bodies of non-flowering
536 plants. Cladogram was obtained from Open Tree of Life⁵⁰.

537
538 **Figure 5. Distribution of CG DNA methylation within gene bodies of non-**
539 **flowering plants is not indicative of gbM.** Levels of CG DNA methylation
540 within gene bodies of species known to possess gbM loci are exemplified by a
541 normal distribution-like pattern with depletions at the TSS and TTS (Eudicot,
542 Monocot, and Basal angiosperm). However, non-flowering plants, which are
543 suspected not to contain gbM loci based on the evolution of CMT ϵ , do not exhibit
544 this type of pattern (Gymnosperm, Fern, Lycophyte, Moss, Liverwort, and Green
545 algae). All non-flowering plants have a spike in CG DNA methylation at the TTS.
546 Gymnosperms possess a methyl-type that shares characteristics to species with,
547 and without gbM: a normal distribution-like pattern of CG DNA methylation with
548 depletions at the TSS, and a spike at the TTS. Error bars represent standard
549 error of the mean.

550

551 **SUPPLEMENTAL INFORMATION**

552

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

561

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

577

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

590

591 **Figure S4. Genome-wide levels of DNA methylation.** DNA methylation levels
592 estimated by alignment to a reference genome or predicted by *FAST^mC²⁵*.
593 Cladogram was obtained from Open Tree of Life⁵⁰.

594

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

598

599 REFERENCES

600

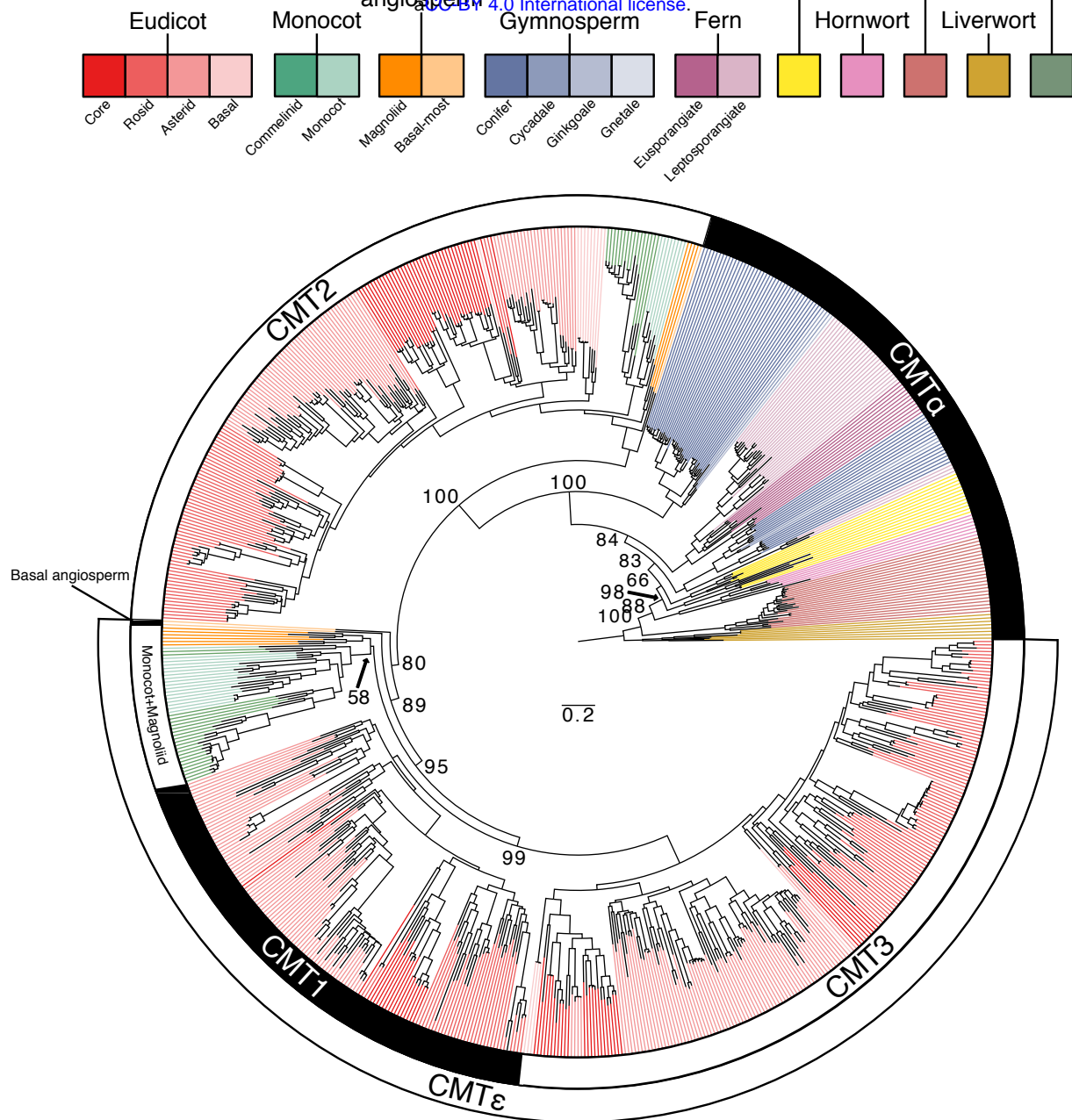
- 601 1. Bartee, L., Malagnac, F. & Bender, J. *Arabidopsis* cmt3 chromomethylase
602 mutations block non-CG methylation and silencing of an endogenous gene.
603 *Genes Dev* 15, 1753–8 (2001).
- 604 2. Jackson, J. P., Lindroth, A. M., Cao, X. & Jacobsen, S. E. Control of CpNpG
605 DNA methylation by the KRYPTONITE histone H3 methyltransferase. *Nature*
606 416, 556–560 (2002).
- 607 3. Zemach, A. et al. The *Arabidopsis* nucleosome remodeler DDM1 allows DNA
608 methyltransferases to access H1-containing heterochromatin. *Cell* 153, 193–
609 205 (2013).
- 610 4. Stroud, H. et al. Non-CG methylation patterns shape the epigenetic landscape
611 in *Arabidopsis*. *Nat Struct Mol Biol* 21, 64–72 (2014).
- 612 5. Du, J. et al. Dual binding of chromomethylase domains to H3K9me2-
613 containing nucleosomes directs DNA methylation in plants. *Cell* 151, 167–180
614 (2012).
- 615 6. Papa, C. M., Springer, N. M., Muszynski, M. G., Meeley, R. & Kaeppler, S. M.
616 Maize chromomethylase *Zea* methyltransferase2 Is required for CpNpG
617 methylation. *The Plant Cell* 13, 1919–1928 (2001).
- 618 7. Hou, P. Q. et al. Functional characterization of *Nicotiana benthamiana*
619 chromomethylase 3 in developmental programs by virus-induced gene
620 silencing. *Physiologia Plantarum* 150, 119–132 (2013).
- 621 8. Garg, R., Kumari, R., Tiwari, S. & Goyal, S. Genomic survey, gene expression
622 analysis and structural modeling suggest diverse roles of DNA
623 methyltransferases in legumes. *PLoS One* 9, e88947 (2014).
- 624 9. Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide
625 evolutionary analysis of eukaryotic DNA methylation. *Science* 328, 916–919
626 (2010).
- 627 10. Henikoff, S. & Comai, L. A DNA methyltransferase homolog with a
628 chromodomain exists in multiple polymorphic forms in *Arabidopsis*. *Genetics*
629 149, 307–318 (1998).
- 630 11. Finnegan, E. J. & Kovac, K. A. Plant DNA methyltransferases. *Plant*
631 *Molecular Biology* 43, 189–201 (2000).
- 632 12. McCallum, C. M., Comai, L., Greene, E. A. & Henikoff, S. Targeted screening
633 for induced mutations. *Nature Biotechnology* 18, 455–457 (2000).
- 634 13. Shen, X. et al. Natural CMT2 variation is associated with genome-wide
635 methylation changes and temperature seasonality. *PLoS Genet* 10, e1004842
636 (2014).
- 637 14. Bewick, A. J. & Schmitz, R. J. Epigenetics in the wild. *eLife* DOI:
638 10.7554/eLife.07808 (2015).

- 639 15. Dubin, M. J. et al. DNA methylation in *Arabidopsis* has a genetic basis and
640 shows evidence of local adaptation. eLife DOI: 10.7554/eLife.05255 (2015).
- 641 16. Du, J. et al. Mechanism of DNA methylation-directed histone methylation by
642 KRYPTONITE. Molecular Cell 55, 495–504 (2014).
- 643 17. Tran, R. K. et al. DNA methylation profiling identifies CG methylation clusters
644 in *Arabidopsis* genes. Curr Biol 15, 154–9 (2005).
- 645 18. Zhang, X. et al. Genome-wide high-resolution mapping and functional
646 analysis of DNA methylation in *Arabidopsis*. Cell 126, 1189–201 (2006).
- 647 19. Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T. & Henikoff, S.
648 Genomewide analysis of *Arabidopsis thaliana* DNA methylation uncovers an
649 interdependence between methylation and transcription. Nat Genet 39, 61–9
650 (2007).
- 651 20. Niederhuth, C. E., Bewick, A. J. et al. Widespread natural variation of DNA
652 methylation within angiosperms. biorxiv doi: <http://dx.doi.org/10.1101/045880>
653 (2016).
- 654 21. Bewick, A. J., Ji, L., Niedtherhuth, C. E., Willing, E. et al. On the origin and
655 evolutionary consequences of gene body DNA methylation. biorxiv doi:
656 <http://dx.doi.org/10.1101/045542> (2016).
- 657 22. Inagaki, S. & Kakutani T. What triggers differential DNA methylation of genes
658 and TEs: contribution of body methylation? Cold Spring Harb Symp Quant
659 Biol. 77, 155-160 (2012).
- 660 23. Stroud, H., Greenber, M. V. C., Feng, S., Bernatavichute, Y. V. & Jacobsen,
661 S. E. Comprehensive analysis of silencing mutants reveals complex
662 regulation of the Arabidopsis methylome. Cell 152, 352–364 (2013).
- 663 24. Noy-Malka, C. et al. A single CMT methyltransferase homolog is involved in
664 CHG DNA methylation and development of *Physcomitrella patens*. Plant Mol
665 Biol 84, 719–35 (2014).
- 666 25. Bewick, A. J. et al. FAST^mC: a suite of predictive models for non-reference-
667 based estimations of DNA methylation. G3 6, 447–452 (2015).
- 668 26. Kenrick, P. & Crane, P. R. The origin and early evolution of plants on land.
669 Nature 389, 33–39 (1997).
- 670 27. Wellman, C. H., Osterloff, P. L. & Mohiuddin, U. Fragments of the earliest
671 land plants. Nature 425, 282–285 (2003).
- 672 28. Steemans, P. et al. Origin and radiation of the earliest vascular land plants.
673 Science 324, 353 (2009).
- 674 29. Rubinstein, C. V., Gerrienne, P., de la Puente, G. S., Astini, R. A. &
675 Steemans, P. Early Middle Ordovician evidence for land plants in Argentina
676 (eastern Gondwana). New Phytologist 188, 365–369 (2010).
- 677 30. Bhattacharya, D. & Medlin, L. Algal phylogeny and the origin of land plants.
678 Plant Physiology 116, 9–15 (1998).
- 679 31. Stiller, J. W. & Hall, B. D. The origin of red algae: Implications for plastid
680 evolution. PNAS 94, 4520–4525 (1997).
- 681 32. Wickett, N. J. et al. Phylotranscriptomic analysis of the origin and early
682 diversification of land plants. PNAS 111, E4859–68 (2014).

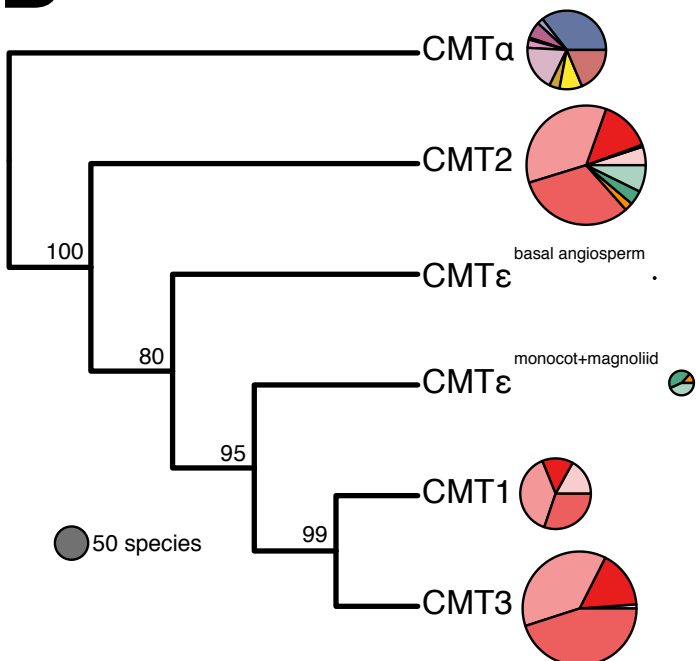
- 683 33. Malik, G., Dangwal, M., Kapoor, S. & Kapoor, M. Role of DNA methylation in
684 growth and differentiation in *Physcomitrella patens* and characterization of
685 cytosine DNA methyltransferases. *FEBS J* 279, 4081–94 (2012).
- 686 34. Feng, S. et al. Conservation and divergence of methylation patterning in
687 plants and animals. *PNAS* 107, 8689–8694 (2010).
- 688 35. Jiao, Y. et al. Ancestral polyploidy in seed plants and angiosperms. *Nature*
689 473, 97–100 (2011).
- 690 36. Qiu, Y. et al. Phylogenetic analyses of basal angiosperms based on nine
691 plastid, mitochondrial, and nuclear Genes. *International Journal of Plant*
692 *Sciences* 166, 815–842 (2005).
- 693 37. Cai, Z. et al. Complete plastid genome sequences of *Drimys*, *Liriodendron*,
694 and *Piper*: implications for the phylogenetic relationships of magnoliids. *BMC*
695 *Evolutionary Biology* 6, doi:10.1186/1471-2148-6-77 (2006).
- 696 38. Li, Q. et al. Genetic perturbation of the maize methylome. *Plant Cell* 26,
697 4602–16 (2014).
- 698 39. Casas-Mollano, J. A., Jeong, B., Xu, J., Moriyama, H. & Cerutti, H. The
699 MUT9p kinase phosphorylates histone H3 threonine 3 and is necessary for
700 heritable epigenetic silencing in *Chlamydomonas*. *PNAS* 105, 6486–6491
701 (2008).
- 702 40. Wang, Z. et al. Osmotic stress induces phosphorylation of histone H3 at
703 threonine 3 in pericentromeric regions of *Arabidopsis thaliana*. *PNAS* 112,
704 8487–8492 (2015).
- 705 41. Takuno, S., Ran, J.-H. & Gaut, B. S. Evolutionary patterns of genic DNA
706 methylation vary across land plants. *Nature Plants* 15222,
707 doi:10.1038/nplants.2015.222 (2016).
- 708 42. Lister, R. et al. Human DNA methylomes at base resolution show widespread
709 epigenomic differences. *Nature* 462, 315–22 (2009).
- 710 43. Jones, P. et al. InterProScan 5: genome-scale protein function classification.
711 *Bioinformatics* 30, 1236–40 (2014).
- 712 44. Mirarab, S. et al. PASTA: Ultra-Large Multiple Sequence Alignment for
713 Nucleotide and Amino-Acid Sequences. *J Comput Biol* 22, 377–86 (2015).
- 714 45. Stamatakis, A. RAxML Version 8: A tool for phylogenetic analysis and
715 postanalysis of large phylogenies. *Bioinformatics* 30, 1312–1313 (2014).
- 716 46. Castresana, J. Selection of conserved blocks from multiple alignments for
717 their use in phylogenetic analysis. *Molecular Biology and Evolution* 17, 540–
718 552 (2000).
- 719 47. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular*
720 *Biology and Evolution* 24, 1586–1591 (2007).
- 721 48. Urich, M. A., Nery, J. R., Lister, R., Schmitz, R. J. & Ecker, J. R. MethylCseq
722 library preparation for base-resolution whole-genome bisulfite sequencing.
723 *Nature Protocols* 10, 475–483 (2015).
- 724 49. Schultz, M. D. et al. Human body epigenome maps reveal noncanonical DNA
725 methylation variation. *Nature* 523, 212–6 (2015).

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

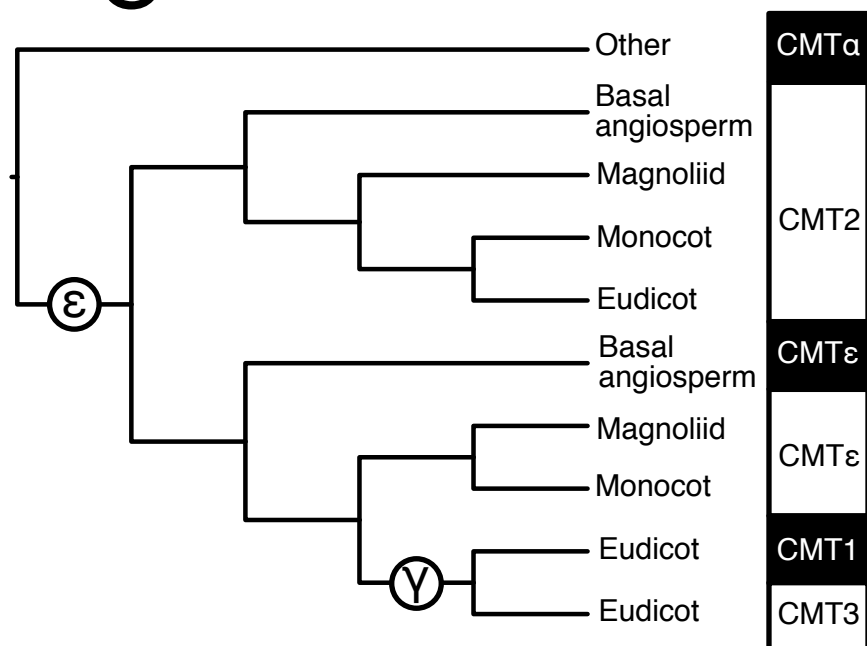
A



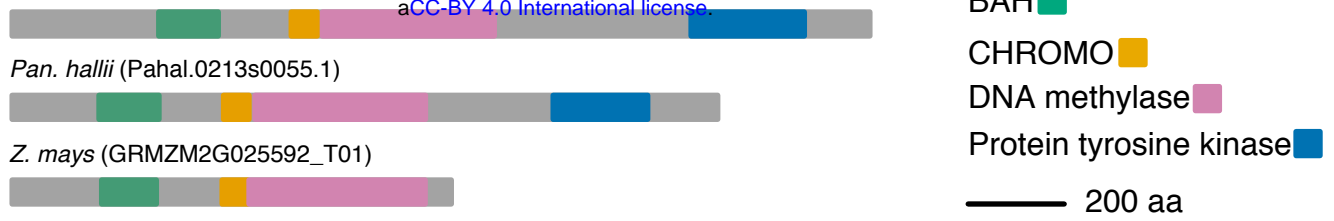
B



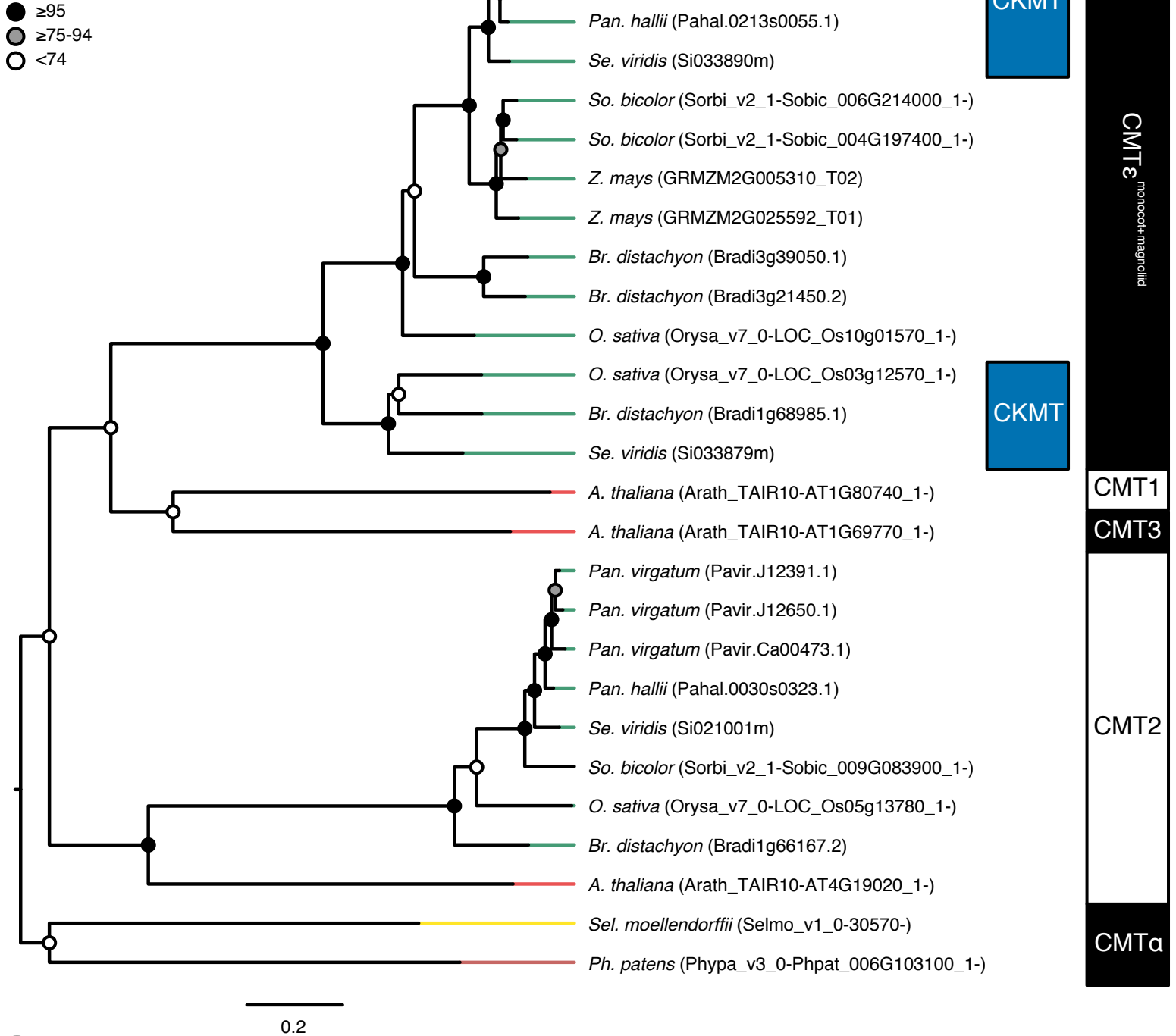
C



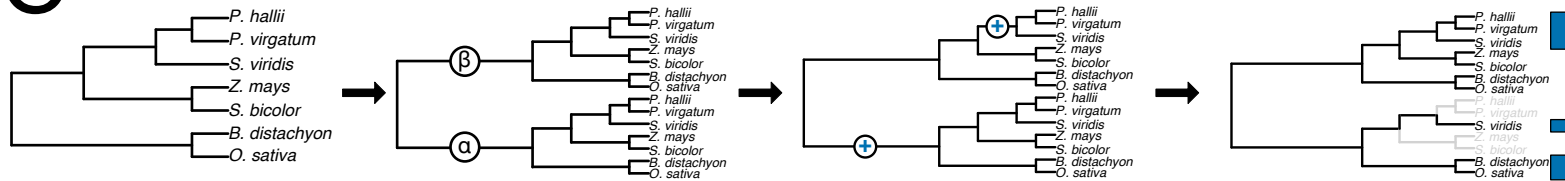
A

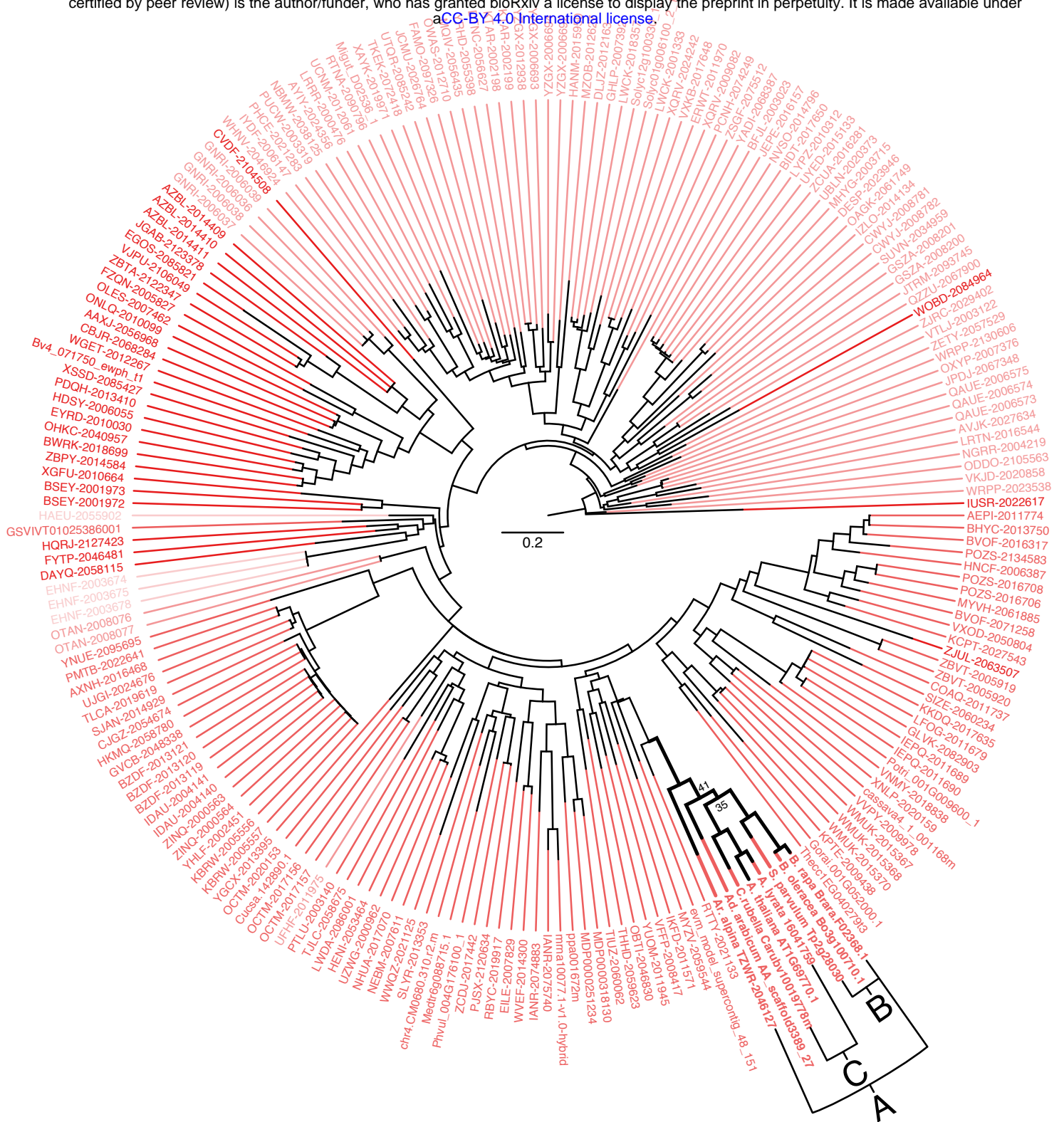


B



C





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)

