Gene duplication and rate variation in the evolution of non-photosynthetic pathways in plastids

Alissa M. Williams[1,2,*], Olivia G. Carter[1], Evan S. Forsythe[1], Hannah K. Mendoza[1], Daniel B. Sloan[1]

[1]Department of Biology, Colorado State University, Fort Collins, Colorado 80523

[2]Program in Cell and Molecular Biology, Colorado State University, Fort Collins, Colorado 80523

*corresponding author: Alissa.Williams@colostate.edu

Abstract:

While the chloroplast (plastid) is known for its role in photosynthesis, it is also involved in many other biosynthetic pathways essential for plant survival. As such, plastids contain an extensive suite of enzymes required for non-photosynthetic processes. The evolution of the associated genes has been especially dynamic in flowering plants (angiosperms), including examples of gene duplication and extensive rate variation. We examined the role of ongoing gene duplication in two key plastid enzymes, the acetyl-CoA carboxylase (ACCase) and the caseinolytic protease (Clp), responsible for fatty acid biosynthesis and protein turnover, respectively. In plants, there are two ACCase complexes—a homomeric version present in the cytosol and a heteromeric version present in the plastid. Duplications of the nuclear-encoded homomeric ACCase gene and retargeting to the plastid have been previously reported in multiple species. We find that these retargeted copies of the homomeric ACCase gene exhibit elevated rates of sequence evolution, consistent with neofunctionalization and/or relaxation of selection. The plastid Clp complex catalytic core is composed of nine paralogous proteins that arose via ancient gene duplication in the cyanobacterial/plastid lineage. We show that further gene duplication occurred more recently in the nuclear-encoded core subunits of this complex, yielding additional paralogs in many species of angiosperms. Moreover, in six of eight cases, subunits that have undergone recent duplication display increased rates of sequence evolution relative to those that have remained single copy. We also compared rate patterns between pairs of Clp core paralogs to gain insight into post-duplication evolutionary routes. These results show that gene duplication and rate variation continue to shape the plastid proteome.

23 **Introduction:**

24

25 The plastid is a dynamic proteomic environment in which key photosynthetic and non-
26 photosynthetic biochemical reactions occur. Major non-photosynthetic functions of plastids
27 include the reaction catalyzed by the acetyl-CoA carboxylase (ACCase) enzyme and protein
28 degradation performed by the caseinolytic protease (Clp) complex (Caroca et al., 2021; Green,
29 2011; Konishi et al., 1996; Nishimura et al., 2017; Nishimura and van Wijk, 2015). Both of these
30 functions are essential in plants and thus the genes involved are generally highly conserved;
31 however, these genes have undergone rapid evolution in multiple angiosperm species (Barnard-
32 Kubow et al., 2014; Erixon and Oxelman, 2008; Jansen et al., 2007; Park et al., 2017; Sloan et
33 al., 2014, 2014; Wicke et al., 2011; Williams et al., 2019, 2015; Zhang et al., 2014). While many
34 hypotheses about these patterns of accelerated evolution have been posited, the underlying
35 evolutionary mechanisms, causes, and consequences remain largely unknown.

36

37 The ACCase enzyme catalyzes the first committed step of fatty acid biosynthesis, the
38 carboxylation of acetyl-CoA to malonyl-CoA (Salie and Thelen, 2016; Sasaki and Nagano,
39 2004). This step requires four different enzyme domains—one biotin carboxylase, one biotin
40 carboxyl carrier, and two (α and β) carboxyltransferases (Salie and Thelen, 2016; Sasaki and
41 Nagano, 2004; Schulte et al., 1997). In plants, there are two forms of the ACCase enzyme. The
42 homomeric version, present in the cytosol, is encoded by a single nuclear gene (Konishi et al.,
43 1996; Konishi and Sasaki, 1994). The heteromeric version, present in the plastid, is encoded by
44 five genes in *Arabidopsis thaliana*; each functional domain is represented by a single gene
45 except for the biotin carboxyl carrier domain, which is encoded by two genes (Konishi et al.,
46 1996; Konishi and Sasaki, 1994; Salie and Thelen, 2016). Four of these genes are in the nuclear
47 genome while the fifth (*accD*) is in the plastid genome (Caroca et al., 2021; Sasaki and Nagano,
48 2004). In a few angiosperm lineages, including the Brassicaceae, Caryophyllaceae, Geraniaceae,
49 and Poaceae, there have been duplications of the homomeric ACCase gene with subsequent
50 retargeting of one copy to the plastid (Babiychuk et al., 2011; Konishi and Sasaki, 1994; Park et
51 al., 2017; Parker et al., 2014; Rockenbach et al., 2016; Schulte et al., 1997).

52

53 The Clp complex is one of the most abundant stromal proteases and degrades a variety of targets
54 (Apitz et al., 2016; Bouchnak and van Wijk, 2021; Majeran et al., 2000; Montandon et al., 2019;
55 Nishimura et al., 2017; Nishimura and van Wijk, 2015; Welsch et al., 2018). This complex
56 consists of many types of subunits. Adapters bind proteins targeted for degradation and deliver
57 them to chaperones, which use ATP to unfold the targeted proteins into the proteolytic core of
58 the complex (Nishimura and van Wijk, 2015). The core consists of 14 subunits that are encoded
59 by nine different paralogous genes (Olinares et al., 2011a; Peltier et al., 2004; Sjögren et al.,
60 2006; Stanne et al., 2007). Eight of these genes reside in the nuclear genome (*CLPP3-6, CLPR1-*
61 *4*), while the ninth is encoded in the plastid genome (*clpP1*) (Nishimura et al., 2017; Olinares et
62 al., 2011b). The ClpP subunits contain a catalytically active Ser-His-Asp triad, whereas the ClpR
63 subunits do not (Nishimura and van Wijk, 2015; Porankiewicz et al., 1999). These nine paralogs
64 are the results of gene duplications throughout cyanobacterial and plastid evolution and are
65 shared by all land plants (Olinares et al., 2011a). Ongoing gene duplication of individual
66 subunits has been noted in a handful of angiosperm lineages (Rockenbach et al., 2016; Williams
67 et al., 2021, 2019).
68
69 Thus, the evolutionary trajectory of both of these essential plastid pathways is characterized by
70 gene duplication at both ancient and recent timescales. Gene duplication is common in land
71 plants, in part due to the frequency with which whole genome duplication (polyploidization)
72 occurs in this lineage (Clark and Donoghue, 2018; De Bodt et al., 2005; del Pozo and Ramirez-
73 Parra, 2015; Flagel and Wendel, 2009; Panchy et al., 2016; Wendel et al., 2018). Nearly all
74 species of land plants have polyploidization events in their evolutionary histories (Clark and
75 Donoghue, 2018; Leebens-Mack et al., 2019; Panchy et al., 2016). Angiosperms in particular
76 seem to have a propensity for whole genome duplication; the entire clade shares an ancient
77 polyploidization event and many lineages have undergone subsequent rounds of whole genome
78 duplication (Clark and Donoghue, 2018; Panchy et al., 2016; Renny-Byfield and Wendel, 2014;
79 Soltis et al., 2009). While every gene is initially affected by whole genome duplication, only 10-
80 30% of those duplicates are maintained in the genome longer-term (Hahn, 2009; Maere et al.,
81 2005; Paterson et al., 2006). Though polyploidy is likely a main contributor to gene duplication
82 in plants, other forms of gene duplication are also prevalent (Flagel and Wendel, 2009). For
83 instance, tandem duplication has been shown to be common in both *Arabidopsis thaliana* and

84 *Oryza sativa*, where tandemly arrayed gene clusters make up 15-20% of genic content.
85 Additionally, multiple studies have shown that transposon-mediated gene duplication is
86 prevalent in plants (Flagel and Wendel, 2009; Freeling et al., 2008; Rizzon et al., 2006; Wang et
87 al., 2006).
88
89 Gene duplication is an important evolutionary process and is thought to be a major source of
90 evolutionary novelty (Hahn, 2009; Ohno, 1970; Taylor and Raes, 2004; Zhang, 2003). The most
91 common evolutionary fate of paralogs is retention of one copy and pseudogenization and loss of
92 the other copy (Lynch and Conery, 2000; Zhang, 2003; Zhang et al., 2003). However, several
93 evolutionary mechanisms have been described in which retention of both gene duplicates is
94 favored. The increased gene-dosage advantage model describes a scenario in which increased
95 amount of gene product produced by the two identical gene copies is beneficial and thus both
96 copies retain ancestral function (Hahn, 2009; Ohno, 1970; Pegueroles et al., 2013; Zhang, 2003).
97 The neofunctionalization model posits that one paralog acquires new functions while the other
98 retains ancestral functionality (Hahn, 2009; Ohno, 1970; Pegueroles et al., 2013; Zhang, 2003).
99 In the subfunctionalization model, an ancestral function is split between the two duplicates
100 (Hahn, 2009; Ohno, 1970; Pegueroles et al., 2013; Zhang, 2003), in some cases creating the
101 possibility for each paralog to optimize a subset of the ancestral function in a process known as
102 escape from adaptive conflict (Des Marais and Rausher, 2008; Huang et al., 2015; Sikosek et al.,
103 2012).
104
105 To distinguish between these evolutionary fates, many studies have employed evolutionary rate
106 comparisons (Hahn, 2009; Pegueroles et al., 2013). These comparisons involve both paralogs as
107 well as their common ancestor (Pegueroles et al., 2013). Under both the gene-dosage advantage
108 and subfunctionalization models, gene duplicates are expected to evolve at approximately the
109 same rate as each other (Pegueroles et al., 2013). The difference in evolutionary rates predicted
110 by these two models is found in comparisons to the common ancestor; with a gene-dosage
111 advantage, the expectation is that the paralogs will evolve at the same rate as the common
112 ancestor, while with subfunctionalization, the expectation is that the paralogs will evolve at an
113 increased rate relative to the common ancestor (though this assumption has been challenged)
114 (Force et al., 1999; Hahn, 2009; He and Zhang, 2005; Lynch and Force, 2000; Pegueroles et al.,

115 2013; Zhang, 2003). By contrast, under the neofunctionalization model, asymmetry between
116 evolutionary rates of paralogs is expected, where one paralog retains the ancestral evolutionary
117 rate while the other experiences rate acceleration after being freed from selective constraints
118 (Hahn, 2009; Pegueroles et al., 2013; Zhang, 2003). The proportion of paralogs with asymmetric
119 rates of evolution has been estimated at anywhere from 5% to 65% in a variety of studies
120 (Conant and Wagner, 2003; Dermitzakis and Clark, 2001; Kondrashov et al., 2002; Panchin et
121 al., 2010; Pegueroles et al., 2013; Van de Peer et al., 2001). This wide range of estimates is likely
122 due to differences in study systems, definitions and identifications of paralogs, gene types, and
123 time since duplication. Despite the varying estimates of evolutionary rate asymmetry, it is clear
124 that paralogs evolve under a mixture of evolutionary regimes.
125
126 Here, we characterize recent gene duplication events and subsequent changes in evolutionary
127 rate in ACCase and Clp core subunits. We show that ACCase genes exhibit patterns of
128 duplication, retargeting, and accelerated protein evolution consistent with neofunctionalization
129 and/or relaxed selection. Additionally, we examine duplications of nuclear-encoded plastid Clp
130 core subunits and demonstrate that duplication leads to significant changes in the rate of
131 evolution in most cases but that patterns differ across Clp subunits, meaning multiple post-
132 duplication evolutionary routes are represented across pairs of paralogs. This work provides
133 additional insights into the interplay between gene duplication and evolutionary rate in the
134 molecular evolution of plastid proteins.
135
136 **Materials and Methods:**
137
138 **Compilation and curation of *ACC* nucleotide sequences**
139
140 Previous work identified duplications of the homomeric ACCase gene *ACC* and subsequent
141 retargeting of one copy to the plastid in the angiosperm families Poaceae (Konishi and Sasaki,
142 1994; Park et al., 2017; Rockenbach et al., 2016), Brassicaceae (Babiychuk et al., 2011; Park et
143 al., 2017; Parker et al., 2014; Schulte et al., 1997), Caryophyllaceae (Rockenbach et al., 2016),
144 and Geraniaceae (Park et al., 2017). *ACC* genes were obtained for multiple species in each of
145 these families. All cytosol-targeted *ACC* genes were designated *ACC1* while all plastid-targeted

146 *ACC* genes were designated *ACC2* per established conventions (Babiychuk et al., 2011; Sasaki
147 and Nagano, 2004); thus, sharing the same identifier does not necessarily indicate orthology
148 because of the multiple independent origins of plastid-targeted *ACC2* genes. *Amborella*
149 *trichopoda*, which has a single *ACC* gene that we designated *ACC1*, was used as an outgroup.
150
151 Trimmed *ACC1* and *ACC2* coding sequences (CDSs) were obtained from Rockenbach et al.
152 (2016) for *Amborella trichopoda, Arabidopsis thaliana, Agrostemma githago*, *Silene noctiflora*,
153 *Silene paradoxa,* and *Triticum aestivum*. The trimming in Rockenbach et al (2016) was codon-
154 guided and included removal of the target peptide. *ACC1* and *ACC2* CDSs from the following
155 species were compiled using gene identifiers from Table S4 in Park et al. (2017): Geraniaceae:
156 *California macrophylla, Erodium texanum, Geranium incanum, Geranium maderense,*
157 *Geranium phaeum, Monsonia emarginata, Pelargonium cotyledonis*; Brassicaceae: *Capsella*
158 *rubella*; Poaceae: *Oryza sativa*, *Sorghum bicolor*. Duplications of *ACC* were additionally
159 identified in two Poaceae species—*Aegilops tauschii* and *Zea mays*—by performing BLAST
160 searches against these organisms on NCBI and Phytozome v13, respectively (Camacho et al.,
161 2009; Goodstein et al., 2012).
162
163 All *ACC1* and *ACC2* sequences were included in a single file and aligned using the MAFFT *einsi*
164 option (Katoh and Standley, 2013) in codon space using the *align_fasta_with_mafft_codon*
165 subroutine in the sloan.pm Perl module (https://github.com/dbsloan/perl_modules). 5′ trimming
166 was conducted according to the trimming performed in Rockenbach et al. (2016). Additional
167 trimming of poorly aligned regions was performed manually in a codon-based manner.
168
169 **Compilation and curation of Clp core subunit amino acid and nucleotide sequences**
170
171 To identify Clp core subunit amino acid sequences, a custom Python script
172 (https://github.com/alissawilliams/Gene_duplication_ACCase_Clp/scripts /local_blast5.py) was
173 used to reciprocally blast (blastp v2.2.29) *Arabidopsis thaliana* amino acid sequences against
174 predicted protein sequences from each of 22 other angiosperm species in the dataset. These 22
175 species were the same set used in Williams et al. (2019) with the exclusion of *Silene latifolia* and
176 *Silene noctiflora*, since Clp core subunit duplications have been previously studied in *Sileneae*

177 (Rockenbach et al., 2016; Williams et al., 2021, 2019). This sampling was chosen to represent
178 both the diversity of angiosperms and the range of rate variation in Clp complex evolution
179 (Williams et al., 2019; see Table S3).
180
181 Compiled amino acid sequences for each subunit were aligned using the *einsi* option in MAFFT
182 v7.222 (Katoh and Standley, 2013) and trimmed using GBLOCKS v0.91b (Castresana, 2000)
183 with parameter *-b1* set to the default value of *-b2* and parameter *-b5* set to *h*. All alignments were
184 examined manually to confirm homology. Sequences were also screened to prevent inclusion of
185 multiple splice variants from a single gene. In cases where genomic data were used, only one
186 transcript per gene was used. In cases where transcriptomic data were used, sequences were
187 eliminated when alternative splicing was obvious (i.e. inclusion of an intron where the other
188 sequence had a gap or variation only in one short piece of the transcript at either end). Catalytic
189 site status and length were determined using the amino acid sequence data.
190
191 Nucleotide sequences for each identified Clp core subunit protein sequence were compiled from
192 the corresponding CDS or transcript sequence file. For non-CDS sequences, ORFfinder
193 (Wheeler et al., 2003) was used to identify the coding sequence. Compiled CDS sequences for
194 each subunit were aligned with the MAFFT *einsi* option (Katoh and Standley, 2013) in codon
195 space as above. 5′ and 3′ end trimming was performed manually in a codon-based manner.
196
197 **Generating constraint trees for the *ACC* and Clp subunit alignments for use in PAML**
198
199 A constraint tree stipulates a fixed topology (branching order) that is used by a phylogenetic
200 program (in this case, PAML) when calculating branch lengths. To generate a constraint tree for
201 the *ACC* alignment, RAxML v8.2.12 (Stamatakis, 2014) was used on the trimmed nucleotide
202 alignment with parameters *-m = GTRGAMMA, -p = 12345, -f = a, -x = 12345,* and *-# = 100*.
203 The resultant topology confirmed that there were independent *ACC* duplications at the base of
204 each family (Park et al., 2017; Rockenbach et al., 2016).
205
206 To construct constraint trees for Clp core subunits, each trimmed amino acid alignment was
207 analyzed with ProtTest v3.4.2 (Darriba et al., 2011) to choose a model of sequence evolution.

208 The top model based on the Bayesian Information Criterion was chosen for use in PhyML v3.3
209 (Guindon et al., 2010), which was run with 1000 bootstrap replicates and 100 random starts. The
210 resultant phylogenetic trees were used to determine whether duplication events were lineage-
211 specific or shared among species in the dataset. In almost all cases, paralogs from a single
212 species were sister to one another in the trees, indicating lineage-specific duplications. There
213 were a few cases in which paralogs from a single species were not sister to one another.
214 However, given low bootstrap support and the difficulty of resolving species relationships using
215 a single gene with highly variable rates of evolution, we proceeded under the assumption that
216 these duplications were lineage-specific as well. Thus, the constraint trees for each individual
217 Clp core subunit were constructed using the known species tree (The Angiosperm Phylogeny
218 Group et al., 2016), with duplications encoded as species-specific (mapped to terminal branches
219 of the species tree).
220
221 **Running PAML for ACCase and Clp core subunit genes**
222
223 For each alignment, PAML v4.9j (Yang, 2007) was used to infer d_N/d_S values for all branches
224 using the free ratios model (model = 1) and parameters *CodonFreq = 2* and *cleandata = 0*.
225 Additionally, *model = 0* and *model = 2* runs were conducted for all alignments, again using
226 *CodonFreq = 2* and *cleandata = 0*. The *model = 0* runs forced all branches to have the same
227 d_N/d_S ratio, while the *model = 2* runs allowed different d_N/d_S values for specified groups of
228 branches.
229
230 For the *ACC* alignment, one *model = 2* run was conducted with plastid-targeted branches as the
231 foreground. The resultant tree had one d_N/d_S value for plastid-targeted (*ACC2*) branches and a
232 second d_N/d_S value for cytosol-targeted (*ACC1*) branches (including all internal pre-duplication
233 branches). This output was compared with the *model = 0* run to determine whether allowing two
234 d_N/d_S ratios (one for each of those groups) was a better fit to the data than allowing just a single
235 d_N/d_S value. For the Clp subunit alignments, *model = 2* was used twice. In the first run, all
236 terminal branches (and in the case of two subunits, internal post-duplication branches) were
237 designated as the foreground. In the second run, there were three classes of branches, where all
238 branches were categorized the same as in the first run except that post-duplication branches

239 (internal or terminal) were placed in a third category. The three-partition and two-partition
240 models were compared to determine whether allowing an additional d_N/d_S ratio for post-
241 duplication branches was a better fit to the data than just separating terminal from internal
242 branches. The models were compared using likelihood ratio tests.
243
244 For the *ACC* alignment, a branch-site test (Yang, 2007; Yang and Nielsen, 2002) was also
245 conducted to test for evidence of positive selection on branches for plastid-targeted genes, which
246 were set as the foreground branches for this analysis. A null model and an alternative model both
247 used the parameters *model = 2, NSsites = 2, CodonFreq = 2*, and *cleandata = 0*. The alternative
248 model otherwise used all default values, while the null model additionally used *fix_omega = 1*
249 and *omega = 1*. The models were compared using a likelihood ratio test.
250
251 **Running HyPhy for *ACC***
252
253 In addition to running a PAML branch-site test on the *ACC* alignment (Yang, 2007; Yang and
254 Nielsen, 2002), tests for positive and relaxed selection were implemented in HyPhy v2.5.32
255 (Kosakovsky Pond et al., 2020). Positive selection was tested for using the aBSREL and
256 BUSTED methods (Murrell et al., 2015; Smith et al., 2015). The RELAX method was used to
257 test for relaxed vs. intensified selection (Wertheim et al., 2015). As with the PAML runs, the
258 constraint tree used for HyPhy methods had the branches separated into two categories (*ACC1*
259 and *ACC2*).
260
261 **Comparisons between *ACC1* and *ACC2* genes**
262
263 To compare d_N and d_S between cytosolic-targeted and plastid-targeted *ACC* genes (*ACC1* and
264 *ACC2*, respectively), a mean root-to-tip distance was calculated for each family in the tree. The
265 base of each duplication event was used as the root for each family. For both d_N and d_S, the four
266 mean distances for *ACC1* were compared to those of *ACC2* using a paired t-test in R. Because of
267 the *a priori* prediction that retargeting to the plastid would be associated with accelerated protein
268 sequence evolution, a one-sided test (*ACC2 > ACC1*) was used for d_N, while a two-sided test was
269 used for d_S.

270

271 **Fisher's exact test on Clp subunit paralogs**

272

273 Using the output from the free ratios (model = 1) PAML runs, Fisher's exact test was used to test
274 for asymmetry in the ratio of the estimated numbers of nonsynonymous and synonymous
275 substitutions (Pegueroles et al., 2013). Nonsynonymous and synonymous substitution estimates
276 were entered into the *fisher.test()* function in R with default parameters. For each pair of
277 duplicates, a test between paralog 1 and paralog 2 was performed (**Figure 1**). If the paralogs
278 were found to be evolving symmetrically, their combined numbers of substitutions were
279 compared to those of the ancestral branch (**Figure 1**). If the paralogs were found to be evolving
280 asymmetrically, each one was compared individually against the ancestral branch (**Figure 1**).
281 The four cases in which there were more than two species-specific paralogs (*Soja max* and
282 *Gossypium raimondii CLPP5*; *Musa acuminata* and *Vitis vinifera CLPR4*) were excluded from
283 this analysis.

284

285 **Data availability**

286

287 Scripts, untrimmed and trimmed alignments, PAML output, and HyPhy output are provided for
288 both *ACC* and Clp subunits at https://github.com/alissawilliams/Gene_duplication_ACCase_Clp.

289

290 **<u>Results:</u>**

291

292 **Plastid-targeted ACCases evolve more rapidly than cytosol-targeted ACCases across**
293 **angiosperms**

294

295 Across the sampled clades (Geraniaceae, Caryophyllaceae, Brassicaceae, and Poaceae), nearly
296 all plastid-targeted *ACC2* genes have higher d_N/d_S values than their cytosol-targeted *ACC1*
297 counterparts (**Figure 2, Figure S1**). The single-partition model assigned all branches a d_N/d_S
298 value of 0.1266, while the two-partition model assigned *ACC1* branches a value of 0.0883 and
299 *ACC2* branches a value of 0.1936 ($\chi^2 = 466.84$, $p << 0.0001$). This pattern is true for both
300 terminal and internal branches. The increase in in d_N/d_S ratios in *ACC2* branches is generally

301 driven by increases in d_N rather than reductions in d_S (t = 4.48, p = 0.01 for d_N; t = 0.72, p =
302 0.5249 for d_S; **Figure 2, Figure S2**), suggesting changes in selective pressure.
303
304 Using a branch-sites test in PAML (Yang, 2007), we did not find a significant signature of
305 positive selection spanning the alignment ($\chi^2 = 0$, p = 1), although there were multiple individual
306 sites found to be under positive selection (**Table S1**). Two HyPhy methods found limited, though
307 significant, evidence for positive selection—the aBSREL run (Smith et al., 2015) detected one
308 branch under positive selection (p = 0.04) and the BUSTED run (Murrell et al., 2015) assigned
309 0.12% of sites in foreground (*ACC2*) branches to the positive selection class relative to 0.05% of
310 sites in background (*ACC1*) branches (p = 0.0026). The HyPhy RELAX method (Wertheim et
311 al., 2015) found significant evidence for relaxed selection in the *ACC2* branches relative to the
312 rest of the tree (K = 0.09, p <<0.001).
313
314 **Characterizing ongoing duplication of nuclear-encoded Clp core subunit genes in**
315 **angiosperms**
316
317 Of the 23 angiosperm species in our dataset, 11 had one or more duplications of nuclear genes
318 encoding Clp core subunits, and all eight of these genes were duplicated in at least one species
319 (**Figure 3**). Most of these duplications were represented by two paralogs, but in four cases, we
320 identified more than two paralogs for a particular subunit in a particular species. For *CLPP5*,
321 *Soja max* and *Gossypium raimondii* have five and seven copies, respectively, and for *CLPR4*,
322 both *Musa acuminata* and *Vitis vinifera* have four copies.
323
324 *Soja max* had duplications of the largest number of subunits (six of eight), followed by *Plantago*
325 *maritima* and *Populus trichocarpa* with duplications of five subunits. Of the 11 species with
326 duplications, *Eucalyptus grandis* and *Oenothera biennis* were the only species that had
327 duplications of just one subunit. Across subunits, *CLPP5* had the highest number of paralogs (37
328 in 23 species) and *CLPR2* had the lowest (24 in 23 species).
329
330 In total, we identified 72 gene copies of Clp core subunits resulting from duplication events,
331 including 40 catalytic subunits (*CLPP3-CLPP6*) (**Figure 3**). Of the 40 catalytic paralogs, we

332 found evidence of loss of one or more catalytic sites in multiple genes (**Table S2**). Across all 72
333 paralogs, we also found evidence of truncation of multiple different gene copies (including some
334 with catalytic site loss) (**Table S2, Table S3**).
335
336 **Recent paralogs of Clp core subunits tend to have higher rates of protein sequence**
337 **evolution than their single-copy counterparts**
338
339 Out of the eight nuclear-encoded Clp core subunit trees (**Figures S3-S10**), seven showed
340 statistically significant differences between a model that allowed for different d_N/d_S rates in gene
341 duplicates vs. single-copy genes (the three-partition model) and one that forces the same d_N/d_S
342 rate on these two types of branches (the two-partition model) based on an uncorrected
343 significance threshold of $p = 0.05$. (**Figure 4, Table 1**). In six of those cases, duplicated terminal
344 branches had a higher d_N/d_S rate than non-duplicated terminal branches, while in the remaining
345 case, the reverse was true. We separated internal branches from terminal branches to account for
346 the fact that terminal branches will, on average, have higher d_N/d_S estimates than internal
347 branches because selection has had more time to act on older deleterious mutations (Hasegawa et
348 al., 1998; Ho et al., 2005). Further, terminal branches represent both interspecific divergence and
349 intraspecific polymorphism, which is important because the latter inflates evolutionary rate
350 calculations (Ho et al., 2005; Moilanen and Majamaa, 2003; Nielsen and Weinreich, 1999).
351
352 We also compared the evolutionary rates of paralogs to one another as well as to their common
353 ancestor, again using an uncorrected significance threshold of p = 0.05 (**Table 2**). Of the 26 pairs
354 of paralogs, 13 (50%) showed statistically significant rate asymmetry relative to each another. In
355 10 (77%) of those cases, only one paralog had a significantly different evolutionary rate than the
356 common ancestor (and in all 10 of those cases, that paralog was evolving at a faster rate than the
357 common ancestor). Of the 13 pairs with symmetric evolutionary rates, five (38%) were
358 asymmetric relative to the common ancestor. In three of those cases, the combined paralog
359 evolutionary rate was significantly faster than that of the ancestor.
360
361 **<u>Discussion:</u>**
362

363 **Neofunctionalization and accelerated evolution of duplicated *ACC* genes in multiple clades**
364 **of flowering plants**
365
366 Independent duplications of *ACC* and subsequent retargeting events have been previously
367 reported in multiple angiosperm clades (Babiychuk et al., 2011; Konishi and Sasaki, 1994; Park
368 et al., 2017; Parker et al., 2014; Rockenbach et al., 2016; Schulte et al., 1997). The process of
369 retargeting of a paralog is inherently a form of neofunctionalization because the newly retargeted
370 protein functions in a different cellular compartment than it did ancestrally. A hallmark of
371 neofunctionalization is evolutionary rate asymmetry between paralogs due to selection associated
372 with gaining a new function (Hahn, 2009; Pegueroles et al., 2013; Zhang, 2003). We found that
373 branches of our *ACC* tree representing paralogs targeted to the plastid had statistically
374 significantly higher d_N/d_S values than branches representing paralogs targeted to the cytosol
375 (**Figure 2, Figure S1**), consistent with the predictions under neofunctionalization. These results
376 were based on a trimmed alignment lacking the target peptide, which we excluded because target
377 peptides exhibit fast rates of evolution and reduced constraints on primary amino acid sequence
378 (Bruce, 2001, 2000; Jarvis, 2008). Thus, our results show that *ACC* genes retargeted to the
379 plastid are undergoing evolutionary rate increases unrelated to the target peptide, suggesting that
380 other functional domains are also evolving rapidly.
381
382 Retargeting of the cytosolic, homomeric ACCase protein to the plastid is somewhat unexpected
383 given that a heteromeric ACCase complex already exists in plastids. Whether the retargeted
384 homomeric ACCases functionally replaces or coexists with the heteromeric version appears to
385 vary across clades. In some angiosperm groups, the two complexes coexist, including in
386 *Arabidopsis thaliana* and likely in other members of the Brassicaceae (Babiychuk et al., 2011;
387 Rousseau-Gueutin et al., 2013). In other clades, the homomeric ACCase has replaced the
388 heteromeric version, as was reported in the Poaceae (Konishi and Sasaki, 1994). The duplication
389 found in *Silene noctiflora* and *Silene paradoxa* may also represent a replacement event given that
390 both species lack at least one heteromeric ACCase gene each, where *S. noctiflora* lacks all of
391 them (Rockenbach et al., 2016). In some cases, including *Monsonia emarginata* in the
392 Geraniaceae, the plastid-encoded *accD* gene of the heteromeric complex has been transferred to
393 the nuclear genome, again suggesting that the heteromeric version is still functional (Park et al.,

394 2017; Rousseau-Gueutin et al., 2013). These contrasting histories of replacement vs. coexistence
395 may mean that duplicates in different clades are evolving under different selection regimes.
396
397 Variation in post-duplication fates could confound tests of selection conducted across the entire
398 *ACC* tree. Using PAML and HyPhy (Murrell et al., 2015; Smith et al., 2015; Wertheim et al.,
399 2015; Yang, 2007), we tested for positive selection and relaxed selection in *ACC2* genes relative
400 to *ACC1* genes, both of which can contribute to increased rates of protein sequence evolution.
401 The results were mixed; there is some evidence for relaxed selection across all *ACC2* branches as
402 well as for positive selection in a small number of branches and sites (**Table S1**). Across the four
403 families in our sample, the smallest ratio between mean *ACC2* d_N and mean *ACC1* d_N was found
404 in the Poaceae (1.5 vs. 2.2-2.6 for the other three families). Since the heteromeric ACCase is
405 completely absent in the Poaceae (Konishi and Sasaki, 1994), we would expect stronger
406 purifying selection on the plastid homomeric ACCase in this clade compared to clades in which
407 the two versions coexist. Thus, these results are consistent with the hypothesis that relaxed
408 selection is contributing to rate accelerations and that there is greater relaxation of selection
409 when homomeric and heteromeric ACCases functions redundantly in the plastid, though the
410 evidence is still limited. The potential for positive selection on retargeted ACCases is intriguing
411 given that these proteins are thought to perform the same function as the ancestral protein; it is
412 possible that retargeted proteins are adapting to specific biochemical and/or osmotic conditions
413 within the new destination. Increased evolutionary rates after subcellular retargeting have been
414 previously noted, though we do not fully understand their underlying causes (Byun-McKay and
415 Geeta, 2007; Marques et al., 2008).
416
417 **Ongoing duplication of nuclear-encoded Clp core subunit genes is common in angiosperms**
418
419 Across green plants, duplication of the plastid-encoded Clp core subunit gene *clpP1* has only
420 been found in a handful of lineages (Williams et al., 2019). While other studies have identified
421 recent duplications of nuclear-encoded Clp core subunit genes (Rockenbach et al., 2016;
422 Williams et al., 2021), our work shows that duplications of these nuclear-encoded subunits are
423 pervasive across angiosperms (**Figure 3**). Because we used a mix of transcriptomic and genomic
424 data, we took into consideration the possibility of misidentifying transcript variants as paralogs

425 but our use of primary transcripts only and manual curation to remove hits that appeared to be
426 splice variants (see Materials and Methods) minimizes the risk of this type of error.
427
428 The prevalence of whole genome duplication in plants may partially explain the prevalence of
429 Clp core subunit duplication (Clark and Donoghue, 2018; De Bodt et al., 2005; del Pozo and
430 Ramirez-Parra, 2015; Flagel and Wendel, 2009; Panchy et al., 2016; Wendel et al., 2018). For
431 instance, *Soja max* is a partially diploidized tetraploid, meaning that this lineage underwent a
432 polyploidization event very recently and has only just started the subsequent process of genome
433 reduction (Shultz et al., 2006). *Soja max* had the largest number of duplicated subunits across our
434 sample, which is consistent with this history of whole genome duplication. Similarly, *Populus*
435 *trichopoda*, which tied for the second largest number of duplicated subunits, only recently
436 underwent genome reduction after whole genome duplication (Tuskan et al., 2006). In these
437 cases, we may simply be observing the short-term effects of polyploidization prior to returning to
438 a single copy of each of these genes.
439
440 **Subunit stoichiometry and subfunctionalization in the evolution of the plastid Clp complex**
441
442 Clp core subunit ratios have been studied in *Arabidopsis thaliana* (Olinares et al., 2011a). The
443 core consists of two rings—a ClpP1/ClpR1-4 ring with a 3:1:1:1:1 subunit ratio, respectively,
444 and a ClpP3-6 ring with a 1:2:3:1 subunit ratio, respectively (Olinares et al., 2011a). Despite the
445 high degree of structural similarity amongst the plastid Clp core subunits, core composition (i.e.
446 the number of each type of core subunit) does not appear to vary in *A. thaliana* (Olinares et al.,
447 2011a; Peltier et al., 2004). Due to the stability of subunit interactions in *A. thaliana*, Clp
448 complexes in other angiosperms are typically assumed to have the same ratios of core subunits,
449 but our results suggest that varied numbers of core subunit paralogs may lead to varied
450 stoichiometry across species. Additional work has shown that loss of catalytic activity in ClpP5
451 (present in three copies in *A. thaliana*) is lethal while loss of catalytic activity in ClpP3 (present
452 in one copy in *A. thaliana*) is tolerated, suggesting that core subunit composition may be flexible
453 given a threshold number of catalytic subunits (Liao et al., 2018).
454

455 In fact, core subunit composition has been dynamic throughout the evolutionary history of the
456 green lineage. The Clp complex is widely conserved across bacteria; in most bacteria, including
457 *E. coli*, the Clp core consists of 14 identical subunits (Nishimura and van Wijk, 2015; Yu and
458 Houry, 2007). However, in cyanobacteria, several duplications have produced four core
459 subunits—three catalytic ClpP subunits and one catalytically inactive ClpR subunit (Andersson
460 et al., 2009; Stanne et al., 2007). In green lineage (Viridiplantae) plastids, which are descended
461 from ancient cyanobacteria, gene duplication has continued to expand the number of genes
462 incorporated into the Clp core to yield nine genes (*CLPP1,3-6*, *CLPR1-4*) (Nishimura and van
463 Wijk, 2015). Interestingly, ClpR subunits are incorporated into the core despite their lack of
464 catalytic activity; they are thought to play a structural role in the complex, including chaperone
465 docking (Nishimura and van Wijk, 2015; Olinares et al., 2011b, 2011a; Sjögren and Clarke,
466 2011). In *A. thaliana*, the Clp chaperone is believed to bind only to the ClpP1/ClpR1-4 ring,
467 whereas chaperone proteins bind to both rings of the Clp core in bacteria (Peltier et al., 2004; Yu
468 and Houry, 2007). This ClpP/ClpR division of function (catalytic activity vs chaperone binding)
469 is indicative of subfunctionalization. Further, though the plastid Clp core subunit genes share
470 common ancestry and are structurally similar, knockouts of individual subunits tend to produce
471 severe phenotypes, including lethality in several cases (Kim et al., 2009; Koussevitzky et al.,
472 2007; Rudella et al., 2006).
473
474 **Possible subfunctionalization in recent paralogs of Clp core subunits**
475
476 Given that subfunctionalization has likely played a major role in plastid Clp complex evolution,
477 we were particularly interested in whether we could identify subfunctionalization after more
478 recent duplication events. Taken to an extreme, subfunctionalization would involve having one
479 gene for each of the 14 core subunits, which would lead to further expansion of the typical nine
480 core subunit genes. The total number of core subunits after including recent paralogs and the
481 plastid-encoded ClpP1 was less than 14 in most species. *Musa acuminata*, *Plantago maritima*,
482 and *Populus trichocarpa* had 14 each, *Soja max* had 17, and *Gossypium raimondii* had 15
483 (**Figure 3)**. The numbers larger than 14 were driven in both cases by multiple paralogs of
484 *CLPP5*, with five and seven copies, respectively. ClpP5 has the largest number of subunits
485 stoichiometrically among the eight nuclear-encoded subunits, so the fact that the two largest

numbers of paralogs were both found in *CLPP5* could potentially suggest that some species are moving toward a 1:1 relationship between genes and core subunits. However, this explanation is not supported by other evidence. For example, the other cases of >2 paralogs were found for *CLPR4*, which encodes a protein that is present in just a single copy in the core in *Arabidopsis*. Further, it is not clear that all of these paralogs are capable of producing functional proteins given truncations and loss of catalytic sites (**Table S2, Table S3**).

We tested for signatures of subfunctionalization by looking at evolutionary rate asymmetry. Under subfunctionalization, we would expect paralogs to evolve at symmetric rates relative to one another but asymmetrically relative to their common ancestor (Pegueroles et al., 2013). We found five of these cases in our dataset: *Plantago maritima CLPP3*, *Geranium maderense CLPP4, Medicago truncatula CLPP5, Populus trichocarpa CLPP5,* and *Soja max CLPR1* (**Table 2**). In cases of subfunctionalization, we would expect the paralogs to evolve more quickly than the common ancestor because of relaxed selection due to their more limited functional roles, which was only the case for the former three. In those three cases, the evidence is consistent with subfunctionalization, particularly given that all six involved paralogs are full length. Further, the *P. maritima CLPP3* paralogs share the same substitutions in all three catalytic sites, which indicates duplication after the loss of catalytic activity, and the *G. maderense CLPP4* and *M. truncatula CLPP5* paralogs all have fully retained catalytic triads (**Table S2**).

Possible pseudogenization or neofunctionalization in recent paralogs of Clp core subunits

Predictions about evolutionary rates under neofunctionalization are similar to predictions under the degeneration/gene loss model—one paralog will maintain the ancestral evolutionary rate while the other undergoes evolutionary rate acceleration (Hahn, 2009; Pegueroles et al., 2013; Zhang, 2003). Previous work in this complex has shown that even ClpP1 subunits demonstrating massive accelerations in evolutionary rate can still be functional, meaning that high evolutionary rates alone do not necessarily indicate pseudogenization (Barnard-Kubow et al., 2014; Williams et al., 2019, 2015). Other sequence features can help us differentiate between pseudogenization and neofunctionalization. For example, truncation of a sequence can be evidence that it is no longer producing a functional protein; additionally, for ClpP subunits, loss of catalytic sites may

517 also be an indication of degeneration/pseudogenization (though there may be exceptions,
518 including the *P. maritima CLPP3* paralogs mentioned above). In our dataset, the paralogs of
519 *Musa acuminata CLPP4* and *P. trichocarpa CLPP4* follow these patterns (**Table 2**). In each of
520 these pairs, the paralogs are evolving asymmetrically, and the paralog with a faster rate of
521 evolution is truncated and lacking all three catalytic sites, suggesting loss of function (**Table S2**).
522 Another example of probable pseudogenization is found for the second copy of *M. acuminata*
523 *CLPR1*. This paralog was annotated as two separate genes due to an internal stop codon, which
524 would lead to a truncation in the resultant protein.
525
526 As for neofunctionalization, there are other cases in our dataset where paralogs evolving
527 asymmetrically both have retained catalytic sites and are full length (for instance, *Geranium*
528 *maderense CLPP5* and *Oenothera biennis CLPP5*). There are no known instances of retargeting
529 of plastid Clp core subunits; thus, evolutionary drivers of neofunctionalization of duplicated
530 subunits are unknown. It is possible that neofunctionalization in this complex could involve
531 recruiting additional interacting proteins—the ClpT proteins, for instance, are involved in
532 assembly of the core and are a recently evolutionary innovation specific to green plants
533 (Colombo et al., 2014; Kim et al., 2015; Nishimura and van Wijk, 2015; Sjögren and Clarke,
534 2011). Additionally, ongoing work has identified potential new adapter proteins in the plastid
535 Clp complex (Montandon et al., 2019; Nishimura et al., 2015). Another possibility is tissue-
536 specific expression of paralogs, which has not been documented in the Clp complex but has been
537 identified in mitochondrial complexes (Boss et al., 1997; Guerrero-Castillo et al., 2017; Sinkler
538 et al., 2017).
539
540 **Possible retention of Clp core paralogs under the gene dosage advantage hypothesis**
541
542 We also have eight cases of symmetrically evolving paralogs that are also evolving
543 symmetrically relative to the common ancestor (**Table 2**). Under our initial predictions, these
544 would represent paralogs retained under the gene dosage advantage hypothesis (Hahn, 2009;
545 Ohno, 1970; Pegueroles et al., 2013; Zhang, 2003). Of these eight paralog pairs, four are from
546 *Soja max* (which had six total pairs of paralogs), and three are from *Populus trichocarpa* (which
547 had five total pairs of paralogs). As described above, both of these species are in the process of

548 rediploidization after a recent whole genome duplication (Shultz et al., 2006; Tuskan et al.,
549 2006). It is possible that these results reflect the fact that the gene duplications happened so
550 recently that the paralogs have not had time to diverge. This possibility is further supported given
551 that the estimates of numbers of substitutions for many of these paralogs were so low that there
552 was virtually no power to detect significant asymmetry.
553
554 **Alternative hypotheses and future directions**
555
556 While we based our analyses on established expectations for evolutionary rates under different
557 post-duplication fates (gene dosage advantage, neofunctionalization, and subfunctionalization),
558 other work has challenged the universality of these predictions. He and Zhang (2005) outline the
559 subneofunctionalization model, in which gene duplicates undergo rapid subfunctionalization
560 followed by prolonged neofunctionalization. Asymmetric evolutionary rates are often assumed to
561 be the result of either neofunctionalization or degeneration, but subfunctionalization can also
562 occur in an asymmetric fashion (He and Zhang, 2005). This hypothesis could relate to some of
563 our results; cases of asymmetric evolutionary rates could be due to subfunctionalization rather
564 than neofunctionalization. Additionally, functional constraint can also exist under
565 neofunctionalization, leading to lower substitution rates and possibly symmetric rates of
566 evolution, meaning that symmetrically evolving paralogs could represent cases of
567 neofunctionalization rather than subfunctionalization or gene dosage advantage (He and Zhang,
568 2005).
569
570 Regardless, our results demonstrate that post-duplication evolutionary fates of paralogs vary
571 widely across clades, even when the same genes are involved. Duplications of the homomeric
572 ACCase complex gene (*ACC*) and subsequent retargeting of one copy to the plastid have been
573 previously reported (Babiychuk et al., 2011; Konishi and Sasaki, 1994; Park et al., 2017; Parker
574 et al., 2014; Rockenbach et al., 2016; Schulte et al., 1997). Our results show that the retargeted
575 duplicates almost universally have increased d_N/d_S rates (**Figure 2, Figure S1**). As for plastid
576 Clp core subunit duplications, duplication has clearly shaped this complex over the course of
577 Viridiplantae evolution. We provide evidence of all possible post-duplication routes of recent
578 paralogs amongst the different subunits and different species in our dataset. Overall, our results

are compelling evidence that subunit ratios and stoichiometry may be dynamic across angiosperm lineages. Isolation of plastid Clp complexes and analyses of subunit composition have been performed in a handful of species (Moreno et al., 2017; Olinares et al., 2011a; Williams et al., 2019); future work could determine these compositions in other angiosperms, including those that have undergone recent gene duplications. Our work demonstrates that gene duplication has been and continues to be an important force in plastid evolution.

Acknowledgements

This work was supported by a National Science Foundation (NSF) grant (MCB-1733227), graduate fellowships from NSF (DGE-1321845) and the National Institutes of Health (T32-GM132057), and the Wolves to Rams undergraduate research program (NSF Grant Numbers 1930150 and 19300092, NIH Grant Number 1T34GM137861-01).

References

Andersson, F.I., Tryggvesson, A., Sharon, M., Diemand, A.V., Classen, M., Best, C., Schmidt, R., Schelin, J., Stanne, T.M., Bukau, B., Robinson, C.V., Witt, S., Mogk, A., Clarke, A.K., 2009. Structure and function of a novel type of ATP-dependent Clp protease. J. Biol. Chem. 284, 13519–13532. https://doi.org/10.1074/jbc.M809588200

Apitz, J., Nishimura, K., Schmied, J., Wolf, A., Hedtke, B., Wijk, K.J. van, Grimm, B., 2016. Posttranslational Control of ALA Synthesis Includes GluTR Degradation by Clp Protease and Stabilization by GluTR-Binding Protein. Plant Physiology 170, 2040–2051. https://doi.org/10.1104/pp.15.01945

Babiychuk, E., Vandepoele, K., Wissing, J., Garcia-Diaz, M., Rycke, R.D., Akbari, H., Joubès, J., Beeckman, T., Jänsch, L., Frentzen, M., Montagu, M.C.E.V., Kushnir, S., 2011. Plastid gene expression and plant development require a plastidic protein of the mitochondrial transcription termination factor family. PNAS 108, 6674–6679. https://doi.org/10.1073/pnas.1103442108

Barnard-Kubow, K.B., Sloan, D.B., Galloway, L.F., 2014. Correlation between sequence divergence and polymorphism reveals similar evolutionary mechanisms acting across multiple timescales in a rapidly evolving plastid genome. BMC Evol Biol 14. https://doi.org/10.1186/s12862-014-0268-y

Boss, O., Samec, S., Paoloni-Giacobino, A., Rossier, C., Dulloo, A., Seydoux, J., Muzzin, P., Giacobino, J.P., 1997. Uncoupling protein-3: a new member of the mitochondrial carrier family with tissue-specific expression. FEBS Lett 408, 39–42. https://doi.org/10.1016/s0014-5793(97)00384-0

616 Bouchnak, I., van Wijk, K.J., 2021. Structure, Function and Substrates of Clp AAA+ protease
617 systems in cyanobacteria, plastids and apicoplasts; a comparative analysis. Journal of
618 Biological Chemistry 100338. https://doi.org/10.1016/j.jbc.2021.100338
619 Bruce, B.D., 2001. The paradox of plastid transit peptides: conservation of function despite
620 divergence in primary structure. Biochimica et Biophysica Acta (BBA) - Molecular Cell
621 Research 1541, 2–21. https://doi.org/10.1016/S0167-4889(01)00149-5
622 Bruce, B.D., 2000. Chloroplast transit peptides: structure, function and evolution. Trends in Cell
623 Biology 10, 440–447. https://doi.org/10.1016/S0962-8924(00)01833-X
624 Byun-McKay, S.A., Geeta, R., 2007. Protein subcellular relocalization: a new perspective on the
625 origin of novel genes. Trends in Ecology & Evolution 22, 338–344.
626 https://doi.org/10.1016/j.tree.2007.05.002
627 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L.,
628 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421.
629 https://doi.org/10.1186/1471-2105-10-421
630 Caroca, R., Howell, K.A., Malinova, I., Burgos, A., Tiller, N., Pellizzer, T., Annunziata, M.G.,
631 Hasse, C., Ruf, S., Karcher, D., Bock, R., 2021. Knock-down of the plastid-encoded
632 acetyl-CoA carboxylase gene uncovers functions in metabolism and development. Plant
633 Physiology. https://doi.org/10.1093/plphys/kiaa106
634 Castresana, J., 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in
635 Phylogenetic Analysis. Molecular Biology and Evolution 17, 540–552.
636 https://doi.org/10.1093/oxfordjournals.molbev.a026334
637 Clark, J.W., Donoghue, P.C.J., 2018. Whole-Genome Duplication and Plant Macroevolution.
638 Trends in Plant Science 23, 933–945. https://doi.org/10.1016/j.tplants.2018.07.006
639 Colombo, C.V., Ceccarelli, E.A., Rosano, G.L., 2014. Characterization of the accessory protein
640 ClpT1 from Arabidopsis thaliana: oligomerization status and interaction with Hsp100
641 chaperones. BMC Plant Biology 14, 228. https://doi.org/10.1186/s12870-014-0228-0
642 Conant, G.C., Wagner, A., 2003. Asymmetric Sequence Divergence of Duplicate Genes.
643 Genome Res. 13, 2052–2058. https://doi.org/10.1101/gr.1252603
644 Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2011. ProtTest 3: fast selection of best-fit
645 models of protein evolution. Bioinformatics 27, 1164–1165.
646 https://doi.org/10.1093/bioinformatics/btr088
647 De Bodt, S., Maere, S., Van de Peer, Y., 2005. Genome duplication and the origin of
648 angiosperms. Trends in Ecology & Evolution 20, 591–597.
649 https://doi.org/10.1016/j.tree.2005.07.008
650 del Pozo, J.C., Ramirez-Parra, E., 2015. Whole genome duplications in plants: an overview from
651 Arabidopsis. Journal of Experimental Botany 66, 6991–7003.
652 https://doi.org/10.1093/jxb/erv432
653 Dermitzakis, E.T., Clark, A.G., 2001. Differential selection after duplication in mammalian
654 developmental genes. Mol Biol Evol 18, 557–562.
655 https://doi.org/10.1093/oxfordjournals.molbev.a003835
656 Des Marais, D.L., Rausher, M.D., 2008. Escape from adaptive conflict after duplication in an
657 anthocyanin pathway gene. Nature 454, 762–765. https://doi.org/10.1038/nature07092
658 Erixon, P., Oxelman, B., 2008. Whole-Gene Positive Selection, Elevated Synonymous
659 Substitution Rates, Duplication, and Indel Evolution of the Chloroplast clpP1 Gene.
660 PLOS ONE 3, e1386. https://doi.org/10.1371/journal.pone.0001386

661 Flagel, L.E., Wendel, J.F., 2009. Gene duplication and evolutionary novelty in plants. New
662 Phytologist 183, 557–564. https://doi.org/10.1111/j.1469-8137.2009.02923.x
663 Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y., Postlethwait, J., 1999. Preservation of
664 Duplicate Genes by Complementary, Degenerative Mutations. Genetics 151, 1531–1545.
665 Freeling, M., Lyons, E., Pedersen, B., Alam, M., Ming, R., Lisch, D., 2008. Many or most genes
666 in Arabidopsis transposed after the origin of the order Brassicales. Genome Res 18,
667 1924–1937. https://doi.org/10.1101/gr.081026.108
668 Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W.,
669 Hellsten, U., Putnam, N., Rokhsar, D.S., 2012. Phytozome: a comparative platform for
670 green plant genomics. Nucleic Acids Research 40, D1178–D1186.
671 https://doi.org/10.1093/nar/gkr944
672 Green, B.R., 2011. Chloroplast genomes of photosynthetic eukaryotes. The Plant Journal 66, 34–
673 44. https://doi.org/10.1111/j.1365-313X.2011.04541.x
674 Guerrero-Castillo, S., Cabrera-Orefice, A., Huynen, M.A., Arnold, S., 2017. Identification and
675 evolutionary analysis of tissue-specific isoforms of mitochondrial complex I subunit
676 NDUFV3. Biochim Biophys Acta Bioenerg 1858, 208–217.
677 https://doi.org/10.1016/j.bbabio.2016.12.004
678 Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New
679 Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the
680 Performance of PhyML 3.0. Syst Biol 59, 307–321.
681 https://doi.org/10.1093/sysbio/syq010
682 Hahn, M.W., 2009. Distinguishing among evolutionary models for the maintenance of gene
683 duplicates. J Hered 100, 605–617. https://doi.org/10.1093/jhered/esp047
684 Hasegawa, M., Cao, Y., Yang, Z., 1998. Preponderance of slightly deleterious polymorphism in
685 mitochondrial DNA: nonsynonymous/synonymous rate ratio is much higher within
686 species than between species. Mol Biol Evol 15, 1499–1505.
687 https://doi.org/10.1093/oxfordjournals.molbev.a025877
688 He, X., Zhang, J., 2005. Rapid Subfunctionalization Accompanied by Prolonged and Substantial
689 Neofunctionalization in Duplicate Gene Evolution. Genetics 169, 1157–1164.
690 https://doi.org/10.1534/genetics.104.037051
691 Ho, S.Y.W., Phillips, M.J., Cooper, A., Drummond, A.J., 2005. Time Dependency of Molecular
692 Rate Estimates and Systematic Overestimation of Recent Divergence Times. Molecular
693 Biology and Evolution 22, 1561–1568. https://doi.org/10.1093/molbev/msi145
694 Huang, Y., Kendall, T., Forsythe, E.S., Dorantes-Acosta, A., Li, S., Caballero-Pérez, J., Chen,
695 X., Arteaga-Vázquez, M., Beilstein, M.A., Mosher, R.A., 2015. Ancient Origin and
696 Recent Innovations of RNA Polymerase IV and V. Mol Biol Evol 32, 1788–1799.
697 https://doi.org/10.1093/molbev/msv060
698 Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., dePamphilis, C.W., Leebens-Mack, J.,
699 Müller, K.F., Guisinger-Bellian, M., Haberle, R.C., Hansen, A.K., Chumley, T.W., Lee,
700 S.-B., Peery, R., McNeal, J.R., Kuehl, J.V., Boore, J.L., 2007. Analysis of 81 genes from
701 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale
702 evolutionary patterns. PNAS 104, 19369–19374.
703 https://doi.org/10.1073/pnas.0709121104
704 Jarvis, P., 2008. Targeting of nucleus-encoded proteins to chloroplasts in plants. New
705 Phytologist 179, 257–285. https://doi.org/10.1111/j.1469-8137.2008.02452.x

706 Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7:
707 Improvements in Performance and Usability. Mol Biol Evol 30, 772–780.
708 https://doi.org/10.1093/molbev/mst010
709 Kim, J., Kimber, M.S., Nishimura, K., Friso, G., Schultz, L., Ponnala, L., Wijk, K.J. van, 2015.
710 Structures, Functions, and Interactions of ClpT1 and ClpT2 in the Clp Protease System of
711 Arabidopsis Chloroplasts. The Plant Cell 27, 1477–1496.
712 https://doi.org/10.1105/tpc.15.00106
713 Kim, J., Rudella, A., Rodriguez, V.R., Zybailov, B., Olinares, P.D.B., Wijk, K.J. van, 2009.
714 Subunits of the Plastid ClpPR Protease Complex Have Differential Contributions to
715 Embryogenesis, Plastid Biogenesis, and Plant Development in Arabidopsis. The Plant
716 Cell 21, 1669–1692. https://doi.org/10.1105/tpc.108.063784
717 Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., Koonin, E.V., 2002. Selection in the evolution of
718 gene duplications. Genome Biol 3, RESEARCH0008. https://doi.org/10.1186/gb-2002-3-
719 2-research0008
720 Konishi, T., Sasaki, Y., 1994. Compartmentalization of two forms of acetyl-CoA carboxylase in
721 plants and the origin of their tolerance toward herbicides. PNAS 91, 3598–3601.
722 https://doi.org/10.1073/pnas.91.9.3598
723 Konishi, T., Shinohara, K., Yamada, K., Sasaki, Y., 1996. Acetyl-CoA Carboxylase in Higher
724 Plants: Most Plants Other Than Gramineae Have Both the Prokaryotic and the Eukaryotic
725 Forms of This Enzyme. Plant and Cell Physiology 37, 117–122.
726 https://doi.org/10.1093/oxfordjournals.pcp.a028920
727 Kosakovsky Pond, S.L., Poon, A.F.Y., Velazquez, R., Weaver, S., Hepler, N.L., Murrell, B.,
728 Shank, S.D., Magalis, B.R., Bouvier, D., Nekrutenko, A., Wisotsky, S., Spielman, S.J.,
729 Frost, S.D.W., Muse, S.V., 2020. HyPhy 2.5—A Customizable Platform for Evolutionary
730 Hypothesis Testing Using Phylogenies. Molecular Biology and Evolution 37, 295–299.
731 https://doi.org/10.1093/molbev/msz197
732 Koussevitzky, S., Stanne, T.M., Peto, C.A., Giap, T., Sjögren, L.L.E., Zhao, Y., Clarke, A.K.,
733 Chory, J., 2007. An <Emphasis Type="Italic">Arabidopsis thaliana</Emphasis>
734 virescent mutant reveals a role for ClpR1 in plastid development. Plant Mol Biol 63, 85–
735 96. https://doi.org/10.1007/s11103-006-9074-2
736 Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A.,
737 Graham, S.W., Grosse, I., Li, Z., Melkonian, M., Mirarab, S., Porsch, M., Quint, M.,
738 Rensing, S.A., Soltis, D.E., Soltis, P.S., Stevenson, D.W., Ullrich, K.K., Wickett, N.J.,
739 DeGironimo, L., Edger, P.P., Jordon-Thaden, I.E., Joya, S., Liu, T., Melkonian, B.,
740 Miles, N.W., Pokorny, L., Quigley, C., Thomas, P., Villarreal, J.C., Augustin, M.M.,
741 Barrett, M.D., Baucom, R.S., Beerling, D.J., Benstein, R.M., Biffin, E., Brockington,
742 S.F., Burge, D.O., Burris, J.N., Burris, K.P., Burtet-Sarramegna, V., Caicedo, A.L.,
743 Cannon, S.B., Çebi, Z., Chang, Y., Chater, C., Cheeseman, J.M., Chen, T., Clarke, N.D.,
744 Clayton, H., Covshoff, S., Crandall-Stotler, B.J., Cross, H., dePamphilis, C.W., Der, J.P.,
745 Determann, R., Dickson, R.C., Di Stilio, V.S., Ellis, S., Fast, E., Feja, N., Field, K.J.,
746 Filatov, D.A., Finnegan, P.M., Floyd, S.K., Fogliani, B., García, N., Gâteblé, G., Godden,
747 G.T., Goh, F. (Qi Y., Greiner, S., Harkess, A., Heaney, J.M., Helliwell, K.E., Heyduk, K.,
748 Hibberd, J.M., Hodel, R.G.J., Hollingsworth, P.M., Johnson, M.T.J., Jost, R., Joyce, B.,
749 Kapralov, M.V., Kazamia, E., Kellogg, E.A., Koch, M.A., Von Konrat, M., Könyves, K.,
750 Kutchan, T.M., Lam, V., Larsson, A., Leitch, A.R., Lentz, R., Li, F.-W., Lowe, A.J.,
751 Ludwig, M., Manos, P.S., Mavrodiev, E., McCormick, M.K., McKain, M., McLellan, T.,

752 McNeal, J.R., Miller, R.E., Nelson, M.N., Peng, Y., Ralph, P., Real, D., Riggins, C.W.,
753 Ruhsam, M., Sage, R.F., Sakai, A.K., Scascitella, M., Schilling, E.E., Schlösser, E.-M.,
754 Sederoff, H., Servick, S., Sessa, E.B., Shaw, A.J., Shaw, S.W., Sigel, E.M., Skema, C.,
755 Smith, A.G., Smithson, A., Stewart, C.N., Stinchcombe, J.R., Szövényi, P., Tate, J.A.,
756 Tiebel, H., Trapnell, D., Villegente, M., Wang, C.-N., Weller, S.G., Wenzel, M.,
757 Weststrand, S., Westwood, J.H., Whigham, D.F., Wu, S., Wulff, A.S., Yang, Y., Zhu, D.,
758 Zhuang, C., Zuidof, J., Chase, M.W., Pires, J.C., Rothfels, C.J., Yu, J., Chen, C., Chen,
759 L., Cheng, S., Li, J., Li, R., Li, X., Lu, H., Ou, Y., Sun, X., Tan, X., Tang, J., Tian, Z.,
760 Wang, F., Wang, J., Wei, X., Xu, X., Yan, Z., Yang, F., Zhong, X., Zhou, F., Zhu, Y.,
761 Zhang, Y., Ayyampalayam, S., Barkman, T.J., Nguyen, N., Matasci, N., Nelson, D.R.,
762 Sayyari, E., Wafula, E.K., Walls, R.L., Warnow, T., An, H., Arrigo, N., Baniaga, A.E.,
763 Galuska, S., Jorgensen, S.A., Kidder, T.I., Kong, H., Lu-Irving, P., Marx, H.E., Qi, X.,
764 Reardon, C.R., Sutherland, B.L., Tiley, G.P., Welles, S.R., Yu, R., Zhan, S., Gramzow,
765 L., Theißen, G., Wong, G.K.-S., One Thousand Plant Transcriptomes Initiative, 2019.
766 One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574,
767 679–685. https://doi.org/10.1038/s41586-019-1693-2
768 Liao, J.-Y.R., Friso, G., Kim, J., Wijk, K.J. van, 2018. Consequences of the loss of catalytic
769 triads in chloroplast CLPPR protease core complexes in vivo. Plant Direct 2, e00086.
770 https://doi.org/10.1002/pld3.86
771 Lynch, M., Conery, J.S., 2000. The Evolutionary Fate and Consequences of Duplicate Genes.
772 Science 290, 1151–1155. https://doi.org/10.1126/science.290.5494.1151
773 Lynch, M., Force, A., 2000. The Probability of Duplicate Gene Preservation by
774 Subfunctionalization. Genetics 154, 459–473.
775 Maere, S., Bodt, S.D., Raes, J., Casneuf, T., Montagu, M.V., Kuiper, M., Peer, Y.V. de, 2005.
776 Modeling gene and genome duplications in eukaryotes. PNAS 102, 5454–5459.
777 https://doi.org/10.1073/pnas.0501102102
778 Majeran, W., Wollman, F.-A., Vallon, O., 2000. Evidence for a Role of ClpP in the Degradation
779 of the Chloroplast Cytochrome b6f Complex. The Plant Cell 12, 137–149.
780 https://doi.org/10.1105/tpc.12.1.137
781 Marques, A.C., Vinckenbosch, N., Brawand, D., Kaessmann, H., 2008. Functional
782 diversification of duplicate genes through subcellular adaptation of encoded proteins.
783 Genome Biol 9, R54. https://doi.org/10.1186/gb-2008-9-3-r54
784 Moilanen, J.S., Majamaa, K., 2003. Phylogenetic network and physicochemical properties of
785 nonsynonymous mutations in the protein-coding genes of human mitochondrial DNA.
786 Mol Biol Evol 20, 1195–1210. https://doi.org/10.1093/molbev/msg121
787 Montandon, C., Friso, G., Liao, J.-Y.R., Choi, J., van Wijk, K.J., 2019. In Vivo Trapping of
788 Proteins Interacting with the Chloroplast CLPC1 Chaperone: Potential Substrates and
789 Adaptors. J. Proteome Res. 18, 2585–2600.
790 https://doi.org/10.1021/acs.jproteome.9b00112
791 Moreno, J.C., Tiller, N., Diez, M., Karcher, D., Tillich, M., Schöttler, M.A., Bock, R., 2017.
792 Generation and characterization of a collection of knock-down lines for the chloroplast
793 Clp protease complex in tobacco. J Exp Bot 68, 2199–2218.
794 https://doi.org/10.1093/jxb/erx066
795 Murrell, B., Weaver, S., Smith, M.D., Wertheim, J.O., Murrell, S., Aylward, A., Eren, K.,
796 Pollner, T., Martin, D.P., Smith, D.M., Scheffler, K., Kosakovsky Pond, S.L., 2015.

797 Gene-Wide Identification of Episodic Selection. Molecular Biology and Evolution 32,
798 1365–1371. https://doi.org/10.1093/molbev/msv035
799 Nielsen, R., Weinreich, D.M., 1999. The age of nonsynonymous and synonymous mutations in
800 animal mtDNA and implications for the mildly deleterious theory. Genetics 153, 497–
801 506.
802 Nishimura, K., Apitz, J., Friso, G., Kim, J., Ponnala, L., Grimm, B., van Wijk, K.J., 2015.
803 Discovery of a Unique Clp Component, ClpF, in Chloroplasts: A Proposed Binary ClpF-
804 ClpS1 Adaptor Complex Functions in Substrate Recognition and Delivery. Plant Cell 27,
805 2677–2691. https://doi.org/10.1105/tpc.15.00574
806 Nishimura, K., Kato, Y., Sakamoto, W., 2017. Essentials of Proteolytic Machineries in
807 Chloroplasts. Molecular Plant 10, 4–19. https://doi.org/10.1016/j.molp.2016.08.005
808 Nishimura, K., van Wijk, K.J., 2015. Organization, function and substrates of the essential Clp
809 protease system in plastids. Biochimica et Biophysica Acta (BBA) - Bioenergetics, SI:
810 Chloroplast Biogenesis 1847, 915–930. https://doi.org/10.1016/j.bbabio.2014.11.012
811 Ohno, S., 1970. Evolution by Gene Duplication. Springer Science & Business Media.
812 Olinares, P.D.B., Kim, J., Davis, J.I., van Wijk, K.J., 2011a. Subunit stoichiometry, evolution,
813 and functional implications of an asymmetric plant plastid ClpP/R protease complex in
814 Arabidopsis. Plant Cell 23, 2348–2361. https://doi.org/10.1105/tpc.111.086454
815 Olinares, P.D.B., Kim, J., van Wijk, K.J., 2011b. The Clp protease system; a central component
816 of the chloroplast protease network. Biochimica et Biophysica Acta (BBA) -
817 Bioenergetics, Regulation of Electron Transport in Chloroplasts 1807, 999–1011.
818 https://doi.org/10.1016/j.bbabio.2010.12.003
819 Panchin, A.Y., Gelfand, M.S., Ramensky, V.E., Artamonova, I.I., 2010. Asymmetric and non-
820 uniform evolution of recently duplicated human genes. Biol Direct 5, 54.
821 https://doi.org/10.1186/1745-6150-5-54
822 Panchy, N., Lehti-Shiu, M., Shiu, S.-H., 2016. Evolution of Gene Duplication in Plants. Plant
823 Physiology 171, 2294–2316. https://doi.org/10.1104/pp.16.00523
824 Park, S., Ruhlman, T.A., Weng, M.-L., Hajrah, N.H., Sabir, J.S.M., Jansen, R.K., 2017.
825 Contrasting Patterns of Nucleotide Substitution Rates Provide Insight into Dynamic
826 Evolution of Plastid and Mitochondrial Genomes of Geranium. Genome Biol Evol 9,
827 1766–1780. https://doi.org/10.1093/gbe/evx124
828 Parker, N., Wang, Y., Meinke, D., 2014. Natural Variation in Sensitivity to a Loss of Chloroplast
829 Translation in Arabidopsis. Plant Physiology 166, 2013–2027.
830 https://doi.org/10.1104/pp.114.249052
831 Paterson, A.H., Chapman, B.A., Kissinger, J.C., Bowers, J.E., Feltus, F.A., Estill, J.C., 2006.
832 Many gene and domain families have convergent fates following independent whole-
833 genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. Trends
834 Genet 22, 597–602. https://doi.org/10.1016/j.tig.2006.09.003
835 Pegueroles, C., Laurie, S., Albà, M.M., 2013. Accelerated Evolution after Gene Duplication: A
836 Time-Dependent Process Affecting Just One Copy. Molecular Biology and Evolution 30,
837 1830–1842. https://doi.org/10.1093/molbev/mst083
838 Peltier, J.-B., Ripoll, D.R., Friso, G., Rudella, A., Cai, Y., Ytterberg, J., Giacomelli, L., Pillardy,
839 J., Wijk, K.J. van, 2004. Clp Protease Complexes from Photosynthetic and Non-
840 photosynthetic Plastids and Mitochondria of Plants, Their Predicted Three-dimensional
841 Structures, and Functional Implications. J. Biol. Chem. 279, 4768–4781.
842 https://doi.org/10.1074/jbc.M309212200

843 Porankiewicz, J., Wang, J., Clarke, A.K., 1999. New insights into the ATP-dependent Clp
844 protease: Escherichia coli and beyond. Molecular Microbiology 32, 449–458.
845 https://doi.org/10.1046/j.1365-2958.1999.01357.x
846 Renny-Byfield, S., Wendel, J.F., 2014. Doubling down on genomes: polyploidy and crop plants.
847 Am J Bot 101, 1711–1725. https://doi.org/10.3732/ajb.1400119
848 Rizzon, C., Ponger, L., Gaut, B.S., 2006. Striking Similarities in the Genomic Distribution of
849 Tandemly Arrayed Genes in Arabidopsis and Rice. PLOS Computational Biology 2,
850 e115. https://doi.org/10.1371/journal.pcbi.0020115
851 Rockenbach, K., Havird, J.C., Monroe, J.G., Triant, D.A., Taylor, D.R., Sloan, D.B., 2016.
852 Positive Selection in Rapidly Evolving Plastid–Nuclear Enzyme Complexes. Genetics
853 204, 1507–1522. https://doi.org/10.1534/genetics.116.188268
854 Rousseau-Gueutin, M., Huang, X., Higginson, E., Ayliffe, M., Day, A., Timmis, J.N., 2013.
855 Potential Functional Replacement of the Plastidic Acetyl-CoA Carboxylase Subunit
856 (accD) Gene by Recent Transfers to the Nucleus in Some Angiosperm Lineages. Plant
857 Physiology 161, 1918–1929. https://doi.org/10.1104/pp.113.214528
858 Rudella, A., Friso, G., Alonso, J.M., Ecker, J.R., Wijk, K.J. van, 2006. Downregulation of ClpR2
859 Leads to Reduced Accumulation of the ClpPRS Protease Complex and Defects in
860 Chloroplast Biogenesis in Arabidopsis. The Plant Cell 18, 1704–1721.
861 https://doi.org/10.1105/tpc.106.042861
862 Salie, M.J., Thelen, J.J., 2016. Regulation and structure of the heteromeric acetyl-CoA
863 carboxylase. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of
864 Lipids, Plant Lipid Biology 1861, 1207–1213.
865 https://doi.org/10.1016/j.bbalip.2016.04.004
866 Sasaki, Y., Nagano, Y., 2004. Plant Acetyl-CoA Carboxylase: Structure, Biosynthesis,
867 Regulation, and Gene Manipulation for Plant Breeding. Bioscience, Biotechnology, and
868 Biochemistry 68, 1175–1184. https://doi.org/10.1271/bbb.68.1175
869 Schulte, W., Töpfer, R., Stracke, R., Schell, J., Martini, N., 1997. Multi-functional acetyl-CoA
870 carboxylase from Brassica napus is encoded by a multi-gene family: Indication for
871 plastidic localization of at least one isoform. PNAS 94, 3465–3470.
872 https://doi.org/10.1073/pnas.94.7.3465
873 Shultz, J.L., Kurunam, D., Shopinski, K., Iqbal, M.J., Kazi, S., Zobrist, K., Bashir, R., Yaegashi,
874 S., Lavu, N., Afzal, A.J., Yesudas, C.R., Kassem, M.A., Wu, C., Zhang, H.B., Town,
875 C.D., Meksem, K., Lightfoot, D.A., 2006. The Soybean Genome Database (SoyGD): a
876 browser for display of duplicated, polyploid, regions and sequence tagged sites on the
877 integrated physical and genetic maps of Glycine max. Nucleic Acids Res 34, D758-765.
878 https://doi.org/10.1093/nar/gkj050
879 Sikosek, T., Chan, H.S., Bornberg-Bauer, E., 2012. Escape from Adaptive Conflict follows from
880 weak functional trade-offs and mutational robustness. Proc Natl Acad Sci U S A 109,
881 14888–14893. https://doi.org/10.1073/pnas.1115620109
882 Sinkler, C.A., Kalpage, H., Shay, J., Lee, I., Malek, M.H., Grossman, L.I., Hüttemann, M., 2017.
883 Tissue- and Condition-Specific Isoforms of Mammalian Cytochrome c Oxidase Subunits:
884 From Function to Human Disease [WWW Document]. Oxidative Medicine and Cellular
885 Longevity. https://doi.org/10.1155/2017/1534056
886 Sjögren, L.L.E., Clarke, A.K., 2011. Assembly of the chloroplast ATP-dependent Clp protease in
887 Arabidopsis is regulated by the ClpT accessory proteins. Plant Cell 23, 322–332.
888 https://doi.org/10.1105/tpc.110.082321

889 Sjögren, L.L.E., Stanne, T.M., Zheng, B., Sutinen, S., Clarke, A.K., 2006. Structural and
890 Functional Insights into the Chloroplast ATP-Dependent Clp Protease in Arabidopsis.
891 The Plant Cell 18, 2635–2649. https://doi.org/10.1105/tpc.106.044594
892 Sloan, D.B., Triant, D.A., Forrester, N.J., Bergner, L.M., Wu, M., Taylor, D.R., 2014. A
893 recurring syndrome of accelerated plastid genome evolution in the angiosperm tribe
894 Sileneae (Caryophyllaceae). Molecular Phylogenetics and Evolution 72, 82–89.
895 https://doi.org/10.1016/j.ympev.2013.12.004
896 Smith, M.D., Wertheim, J.O., Weaver, S., Murrell, B., Scheffler, K., Kosakovsky Pond, S.L.,
897 2015. Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient
898 Detection of Episodic Diversifying Selection. Molecular Biology and Evolution 32,
899 1342–1353. https://doi.org/10.1093/molbev/msv022
900 Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D.,
901 Depamphilis, C.W., Wall, P.K., Soltis, P.S., 2009. Polyploidy and angiosperm
902 diversification. Am J Bot 96, 336–348. https://doi.org/10.3732/ajb.0800079
903 Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
904 large phylogenies. Bioinformatics 30, 1312–1313.
905 https://doi.org/10.1093/bioinformatics/btu033
906 Stanne, T.M., Pojidaeva, E., Andersson, F.I., Clarke, A.K., 2007. Distinctive Types of ATP-
907 dependent Clp Proteases in Cyanobacteria. J. Biol. Chem. 282, 14394–14402.
908 https://doi.org/10.1074/jbc.M700275200
909 Taylor, J.S., Raes, J., 2004. Duplication and Divergence: The Evolution of New Genes and Old
910 Ideas. Annual Review of Genetics 38, 615–643.
911 https://doi.org/10.1146/annurev.genet.38.072902.092831
912 The Angiosperm Phylogeny Group, Chase, M.W., Christenhusz, M.J.M., Fay, M.F., Byng, J.W.,
913 Judd, W.S., Soltis, D.E., Mabberley, D.J., Sennikov, A.N., Soltis, P.S., Stevens, P.F.,
914 2016. An update of the Angiosperm Phylogeny Group classification for the orders and
915 families of flowering plants: APG IV. Botanical Journal of the Linnean Society 181, 1–
916 20. https://doi.org/10.1111/boj.12385
917 Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N.,
918 Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R.R.,
919 Bhalerao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell,
920 M., Carlson, J., Chalot, M., Chapman, J., Chen, G.-L., Cooper, D., Coutinho, P.M.,
921 Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Déjardin,
922 A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B.,
923 Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L.,
924 Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang,
925 W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjärvi,
926 J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer,
927 F., Leebens-Mack, J., Leplé, J.-C., Locascio, P., Lou, Y., Lucas, S., Martin, F.,
928 Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda,
929 V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P.,
930 Rinaldi, C., Ritland, K., Rouzé, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B.,
931 Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.-J., Uberbacher, E., Unneberg, P.,
932 Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg,
933 G., Van de Peer, Y., Rokhsar, D., 2006. The genome of black cottonwood, Populus

trichocarpa (Torr. & Gray). Science 313, 1596–1604. https://doi.org/10.1126/science.1128691

Van de Peer, Y., Taylor, J.S., Braasch, I., Meyer, A., 2001. The Ghost of Selection Past: Rates of Evolution and Functional Divergence of Anciently Duplicated Genes. J Mol Evol 53, 436–446. https://doi.org/10.1007/s002390010233

Wang, W., Zheng, H., Fan, C., Li, J., Shi, J., Cai, Z., Zhang, G., Liu, D., Zhang, J., Vang, S., Lu, Z., Wong, G.K.-S., Long, M., Wang, J., 2006. High Rate of Chimeric Gene Origination by Retroposition in Plant Genomes. The Plant Cell 18, 1791–1802. https://doi.org/10.1105/tpc.106.041905

Welsch, R., Zhou, X., Yuan, H., Álvarez, D., Sun, T., Schlossarek, D., Yang, Y., Shen, G., Zhang, H., Rodriguez-Concepcion, M., Thannhauser, T.W., Li, L., 2018. Clp Protease and OR Directly Control the Proteostasis of Phytoene Synthase, the Crucial Enzyme for Carotenoid Biosynthesis in Arabidopsis. Mol Plant 11, 149–162. https://doi.org/10.1016/j.molp.2017.11.003

Wendel, J.F., Lisch, D., Hu, G., Mason, A.S., 2018. The long and short of doubling down: polyploidy, epigenetics, and the temporal dynamics of genome fractionation. Curr Opin Genet Dev 49, 1–7. https://doi.org/10.1016/j.gde.2018.01.004

Wertheim, J.O., Murrell, B., Smith, M.D., Kosakovsky Pond, S.L., Scheffler, K., 2015. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. Molecular Biology and Evolution 32, 820–832. https://doi.org/10.1093/molbev/msu400

Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A., Wagner, L., 2003. Database resources of the National Center for Biotechnology. Nucleic Acids Res 31, 28–33.

Wicke, S., Schneeweiss, G.M., dePamphilis, C.W., Müller, K.F., Quandt, D., 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol 76, 273–297. https://doi.org/10.1007/s11103-011-9762-4

Williams, A.M., Friso, G., Wijk, K.J. van, Sloan, D.B., 2019. Extreme variation in rates of evolution in the plastid Clp protease complex. The Plant Journal 98, 243–259. https://doi.org/10.1111/tpj.14208

Williams, A.M., Itgen, M.W., Broz, A.K., Carter, O.G., Sloan, D.B., 2021. Long-read transcriptome and other genomic resources for the angiosperm Silene noctiflora. G3 Genes|Genomes|Genetics. https://doi.org/10.1093/g3journal/jkab189

Williams, A.V., Boykin, L.M., Howell, K.A., Nevill, P.G., Small, I., 2015. The Complete Sequence of the Acacia ligulata Chloroplast Genome Reveals a Highly Divergent clpP1 Gene. PLOS ONE 10, e0125768. https://doi.org/10.1371/journal.pone.0125768

Yang, Z., 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24, 1586–1591. https://doi.org/10.1093/molbev/msm088

Yang, Z., Nielsen, R., 2002. Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. Molecular Biology and Evolution 19, 908–917. https://doi.org/10.1093/oxfordjournals.molbev.a004148

Yu, A.Y.H., Houry, W.A., 2007. ClpP: A distinctive family of cylindrical energy-dependent serine proteases. FEBS Letters 581, 3749–3757. https://doi.org/10.1016/j.febslet.2007.04.076

Zhang, J., 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution 18, 292–298. https://doi.org/10.1016/S0169-5347(03)00033-8

979 Zhang, P., Gu, Z., Li, W.-H., 2003. Different evolutionary patterns between young duplicate
980 genes in the human genome. Genome Biology 4, R56. https://doi.org/10.1186/gb-2003-4-
981 9-r56
982 Zhang, Y., Ma, J., Yang, B., Li, R., Zhu, W., Sun, L., Tian, J., Zhang, L., 2014. The complete
983 chloroplast genome sequence of Taxus chinensis var. mairei (Taxaceae): loss of an
984 inverted repeat region and comparative analysis with related species. Gene 540, 201–209.
985 https://doi.org/10.1016/j.gene.2014.02.037
986

Table 1: Differences in evolutionary rates between duplicated and non-duplicated plastid Clp core subunits. Reported *p*-values are based on likelihood ratio tests for 2-partition vs. 3-partition PAML models (see Materials and Methods). Log-likelihood (lnL) values are reported for each model.

Subunit	**lnL 2-partition model**	**lnL 3-partition model**	***p*-value**	**Class with higher d_N/d_S**
CLPP3	-10058.01	-10047.93	7.12e-06	Duplicated
CLPP4	-9606.41	-9552.26	<1.00e-10	Duplicated
CLPP5	-8121.80	-8070.90	<1.00e-10	Duplicated
CLPP6	-7697.62	-7691.31	3.82e-04	Non-duplicated
CLPR1	-14534.75	-14517.85	6.07e-09	Duplicated
CLPR2	-12018.69	-12002.74	1.63e-08	Duplicated
CLPR3	-10556.37	-10556.05	0.42	n/a
CLPR4	-10395.60	-10337.35	<1.00e-10	Duplicated

Table 2: *p*-values for asymmetries between paralogs and between paralogs and their common ancestor.

Species	Gene	Paralog 1 vs. paralog 2	Paralog 1 vs. ancestor	Paralog 2 vs. ancestor	Paralogs 1+2 vs. ancestor
P. maritima	*CLPP3*	0.83			1.25e-04*
S. max	*CLPP3*	1			0.20
P. maritima	*CLPP4*	1.28e-07	1.07e-09*	0.71	
M. acuminata	*CLPP4*	0.04	0.15	2.12e-04*	
S. max	*CLPP4*	0.57			1
P. trichocarpa	*CLPP4*	3.31e-04	3.44e-12*	1.76e-13*	
G. maderense	*CLPP4*	0.08			1.72e-03*
M. truncatula	*CLPP5*	1			3.61e-05*
P. trichocarpa	*CLPP5*	1			0.04^
O. biennis	*CLPP5*	9.12e-3	0.10	5.37e-4*	
G. maderense	*CLPP5*	1.71e-04	0.42	6.23e-4*	
S. max	*CLPP6*	0.27			1
P. trichocarpa	*CLPP6*	0.46			1
G. raimondii	*CLPP6*	0.71			1
M. acuminata	*CLPR1*	2.20e-16	0.82	2.20e-16*	
S. max	*CLPR1*	0.47			0.05^
P. trichocarpa	*CLPR1*	0.77			0.42
V. vinifera	*CLPR1*	4.45e-3	0.17	1.38e-09*	
M. guttatus	*CLPR1*	0.05	0.59	4.70e-04*	
P. maritima	*CLPR1*	4.70e-05	0.53	2.54e-05*	
P. maritima	*CLPR2*	0.01	0.18	1.63e-06*	
S. max	*CLPR3*	1			0.55
P. trichocarpa	*CLPR3*	0.34			0.35
M. truncatula	*CLPR4*	2.29e-03	0.21	0.11	
E. grandis	*CLPR4*	2.51e-05	0.68	5.42e-11*	
P. maritima	*CLPR4*	9.73e-05	9.53e-04	4.17e-14*	

* denotes that paralog(s) has/have significantly higher evolutionary rate than ancestor branch
^ denotes that paralog(s) has/have significantly lower evolutionary rate than ancestor branch

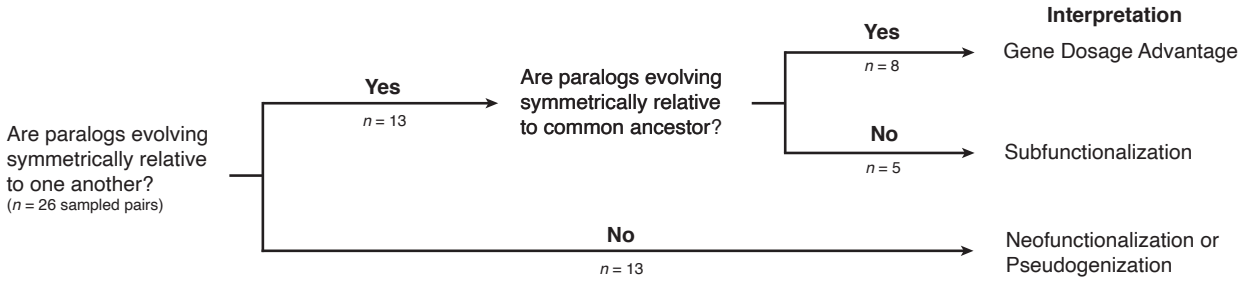

Figure 1: Expectations under different post-duplication models. For each pair of paralogs (n = 26), we first determined whether they were evolving symmetrically relative to one another using N and S estimates from PAML output. Paralogs evolving asymmetrically are predicted to represent neofunctionalization or pseudogenization events. For paralogs evolving symmetrically (n = 8), combined N and S values were compared to those of the immediate ancestor branch. Pairs evolving symmetrically relative to the common ancestor (n = 8) are predicted to represent gene dosage advantage while those evolving asymmetrically relative to the common ancestor (n = 5) are predicted to represent subfunctionalization.

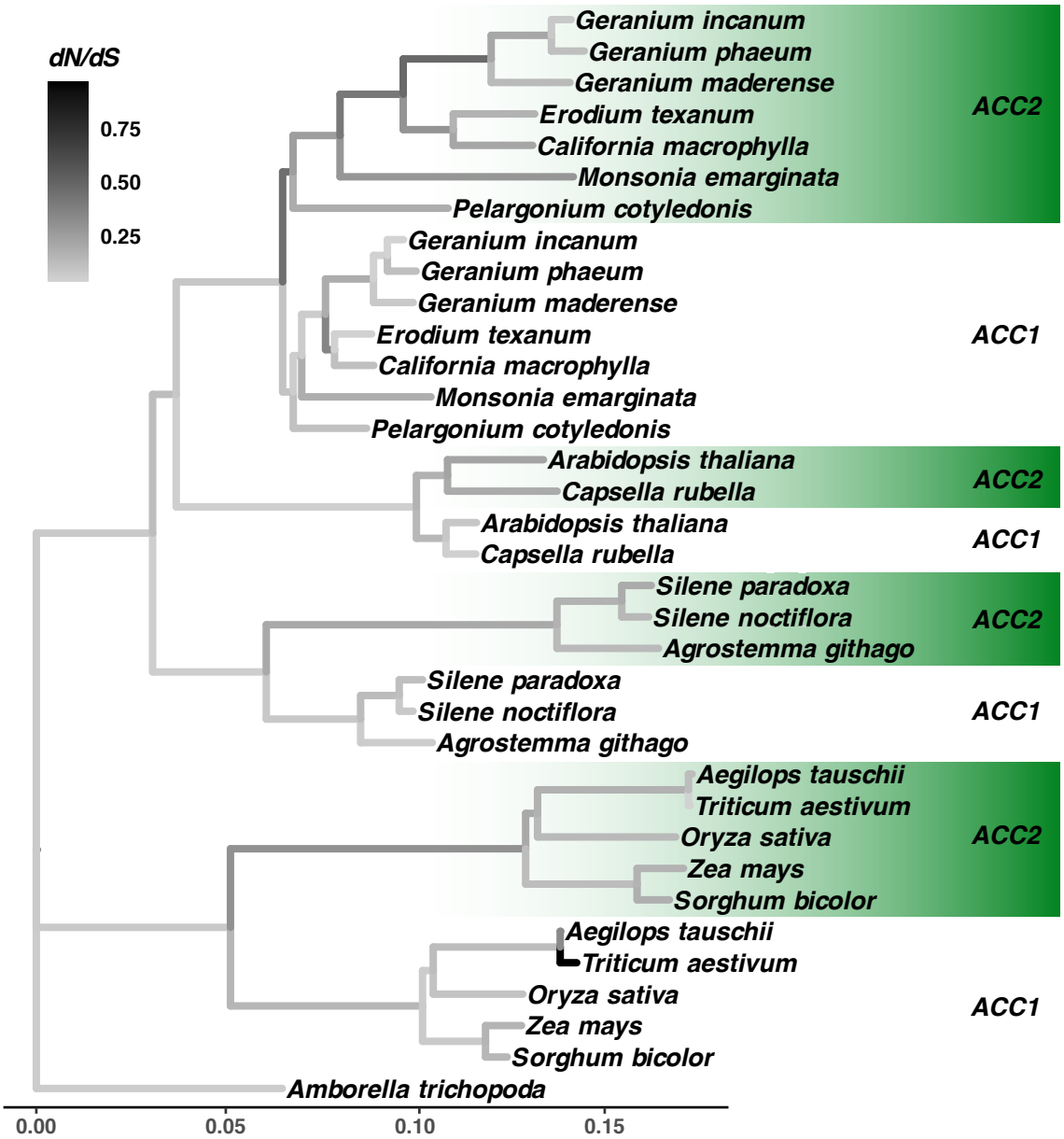

Figure 2: *ACC* genes across the Brassicaceae, Caryophyllaceae, Geraniaceae, and Poaceae, with the single copy of *ACC* in *Amborella trichopoda* as an outgroup. Branch lengths represent d_N values and branch colors represent d_N/d_S ratios.

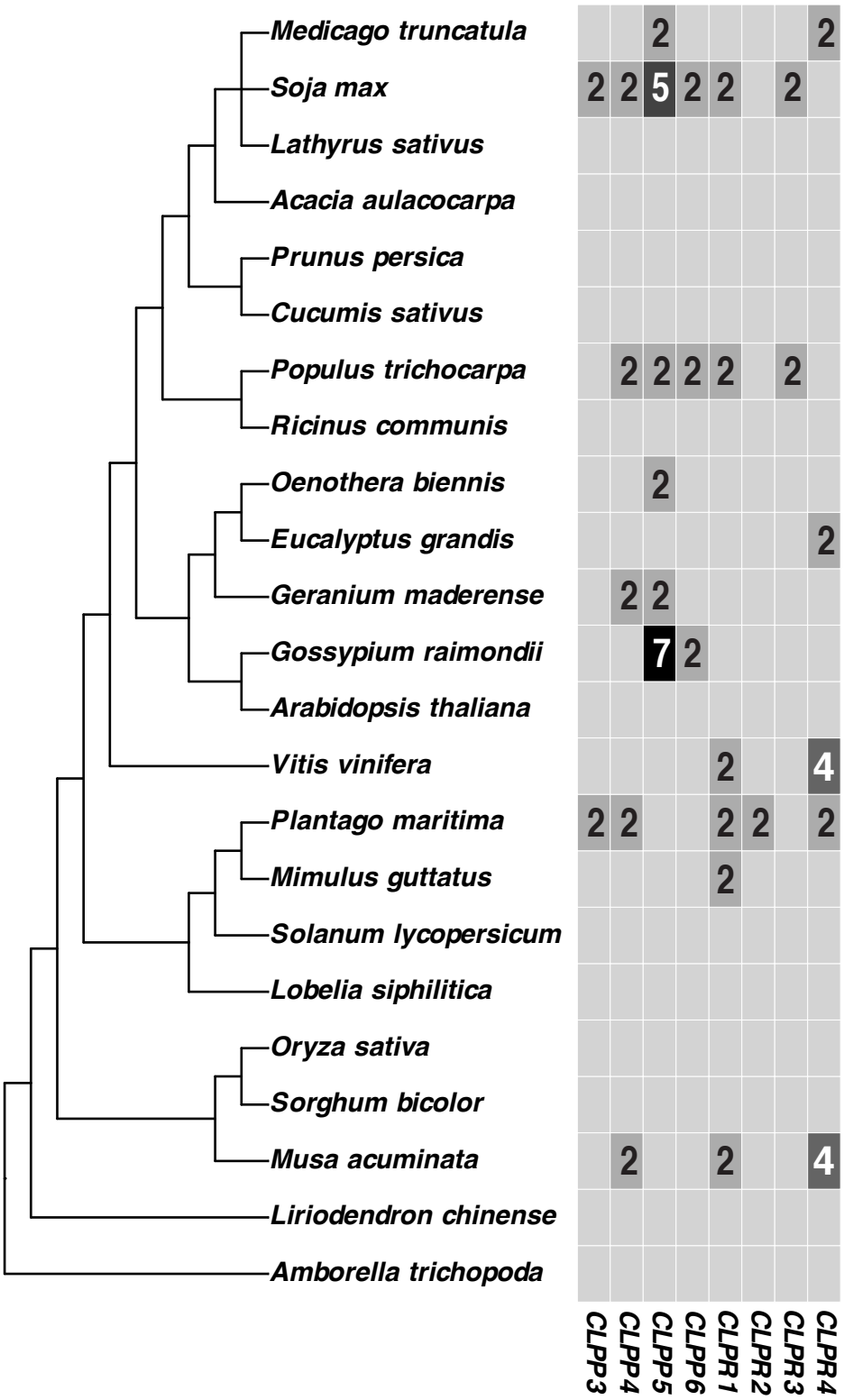

Figure 3: Copy numbers of the nuclear-encoded subunits of the plastid Clp core across angiosperms. Boxes without numbers indicate single-copy genes.

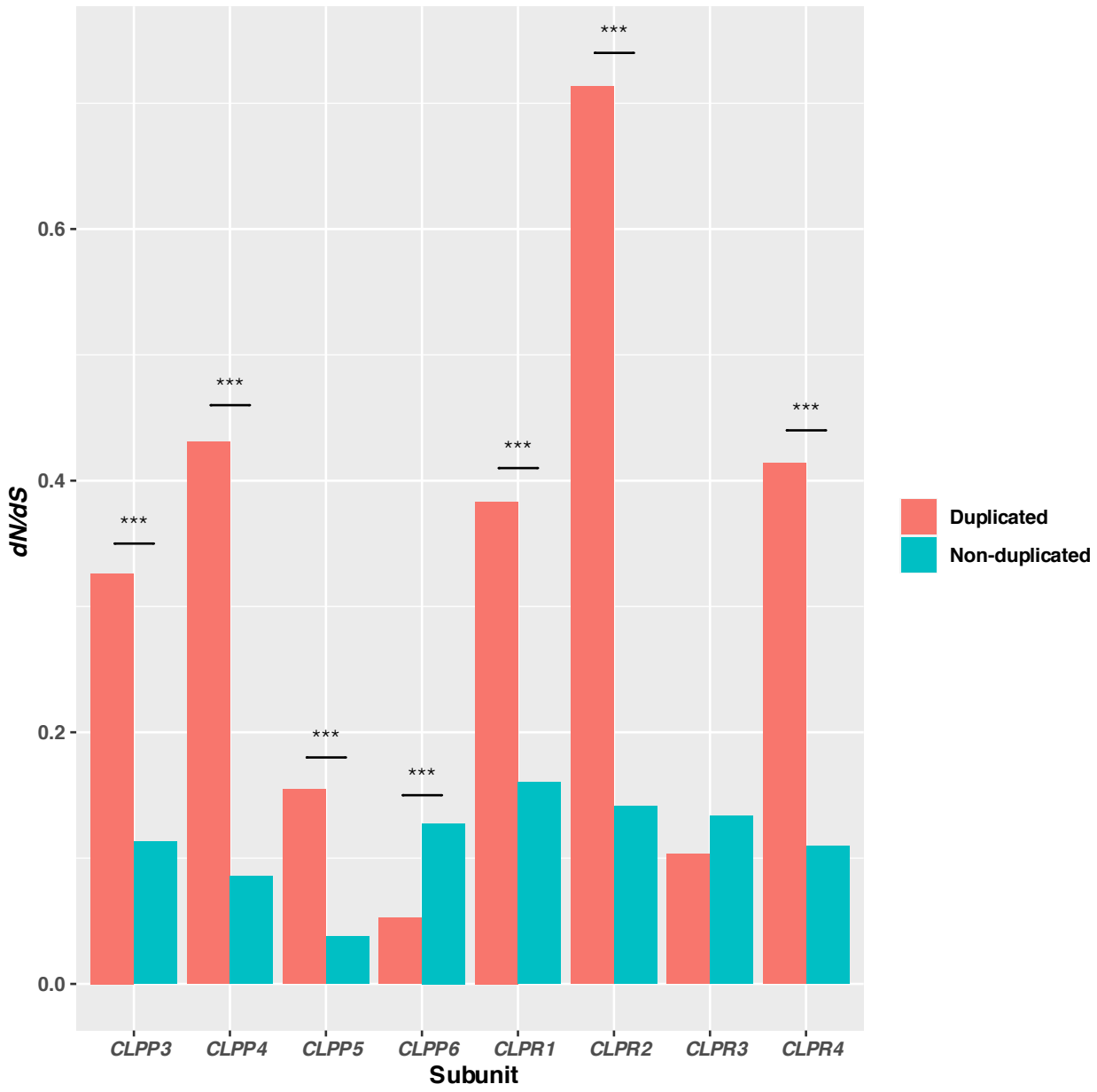

Figure 4: d_N/d_S ratios of duplicated and non-duplicated plastid Clp core subunits across angiosperms. The values were calculated using a PAML branch test with three groups, where each group was assigned its own d_N/d_S value: non-duplicated terminal branches, duplicated terminal branches (and in the cases of *CLPP5* and *CLPR4*, internal post-duplication branches), and internal branches. Significant differences ($p<<0.001$) are indicated with ***.

Table S1: Sites inferred to be under positive selection in *ACC2* branches based on a branch-sites test in PAML using the trimmed alignment for *ACC*.

Site	**Amino acid**	**Probability**
6	L	1.000
92	G	0.965
125	T	0.998
153	V	0.967
182	L	0.983
200	V	0.982
212	L	0.951
333	E	0.993
387	E	0.952
425	E	0.999
428	S	0.982
429	L	0.999
475	S	1.000
480	R	0.968
533	T	0.968
539	S	0.991
546	V	0.977
562	V	0.989
597	L	0.994
652	L	0.999
660	H	0.982
663	M	0.995
687	R	0.961
693	H	0.996
699	L	0.991
700	G	1.000
713	F	0.999
715	A	0.973
739	L	0.995
740	N	0.999
744	S	0.993
753	Q	0.980
766	D	0.967
769	N	0.990
776	K	0.986

794	L	0.967
797	G	0.996
837	R	0.997
849	S	0.996
859	Q	0.986
865	R	1.000
868	L	0.999
872	K	0.995
922	T	0.979
948	Q	0.985
973	T	0.984
981	T	0.999
982	P	0.999
985	K	0.998
989	N	0.999
991	R	0.997
1015	P	0.999
1027	R	0.999
1042	Q	0.991
1043	W	1.000
1044	H	1.000
1045	R	0.995
1048	L	0.978
1078	E	0.995
1085	W	0.980
1096	L	0.984
1108	T	0.999
1110	H	0.958
1147	M	0.990
1151	Q	0.994
1160	Q	0.998
1161	E	0.957
1168	K	1.000
1177	S	0.999
1197	R	0.998
1200	M	0.986
1212	Y	0.989
1243	A	0.995
1326	A	0.963

1342	I	0.981
1344	R	0.983
1507	S	0.952
1580	K	0.956
1602	R	0.971
1655	S	0.966
1735	L	1.000
1887	V	0.984
1898	A	0.999
1904	Q	0.985
1928	E	0.982
2021	P	0.958
2027	S	0.976
2063	E	1.000
2131	K	0.998
2135	E	0.962
2137	A	0.998
2164	G	0.965
2194	E	1.000
2230	P	1.000
2235	Q	0.992
2241	R	0.985
2245	G	0.976

Table S2: Loss of catalytic sites and truncation of nuclear-encoded plastid ClpP core subunits

Species	Protein	Serine	Histidine	Aspartate	Length
Eucalyptus grandis	ClpP3		Replaced with R		
Musa acuminata	ClpP3		Replaced with R		
Plantago maritima 1	ClpP3	Replaced with Y	Replaced with R	Replaced with N	
Plantago maritima 2	ClpP3	Replaced with Y	Replaced with R	Replaced with N	
Lathyrus sativus	ClpP4		Replaced with T		
Medicago truncatula	ClpP4		Replaced with A		
Musa acuminata 2	ClpP4	Gap	Replaced with R	Gap	Truncated
Plantago maritima 2	ClpP4			Gap	
Populus trichocarpa 2	ClpP4	Replaced with M	Gap	Gap	Truncated
Gossypium raimondii 3	ClpP5	Gap			
Gossypium raimondii 4	ClpP5				Truncated
Gossypium raimondii 5	ClpP5	Replaced with N			Truncated
Gossipium Raimondii 6	ClpP5	Replaced with N		Replaced with N	Truncated
Gossypium raimondii 7	ClpP5	Replaced with N		Replaced with F	
Plantago maritima	ClpP5	Gap	Gap	Gap	Truncated
Lathyrus sativus	ClpP6		Replaced with G		
Lobelia siphilitica	ClpP6	Replaced with N			
Medicago truncatula	ClpP6		Replaced with N		
Plantago maritima	ClpP6	Replaced with G	Replaced with E	Replaced with F	
Populus trichocarpa 2	ClpP6				Truncated

*Empty cell indicates presence of catalytic site or full length.

Table S3: Truncation of nuclear-encoded plastid ClpR core subunits

Truncated ClpR1 Subunits
Vitis vinifera 2
Populus trichocarpa 2
Musa acuminata 2 (internal stop codon)
Mimulus guttatus 2
Truncated ClpR2 Subunits
Plantago maritima 2
Truncated ClpR3 Subunits
Populus trichocarpa 1
Populus trichocarpa 2
Truncated ClpR4 Subunits
Eucalyptus grandis 2
Medicago truncatula 2
Vitis vinifera 2
Vitis vinifera 3
Vitis vinifera 4
Musa acuminata 3
Musa acuminata 4

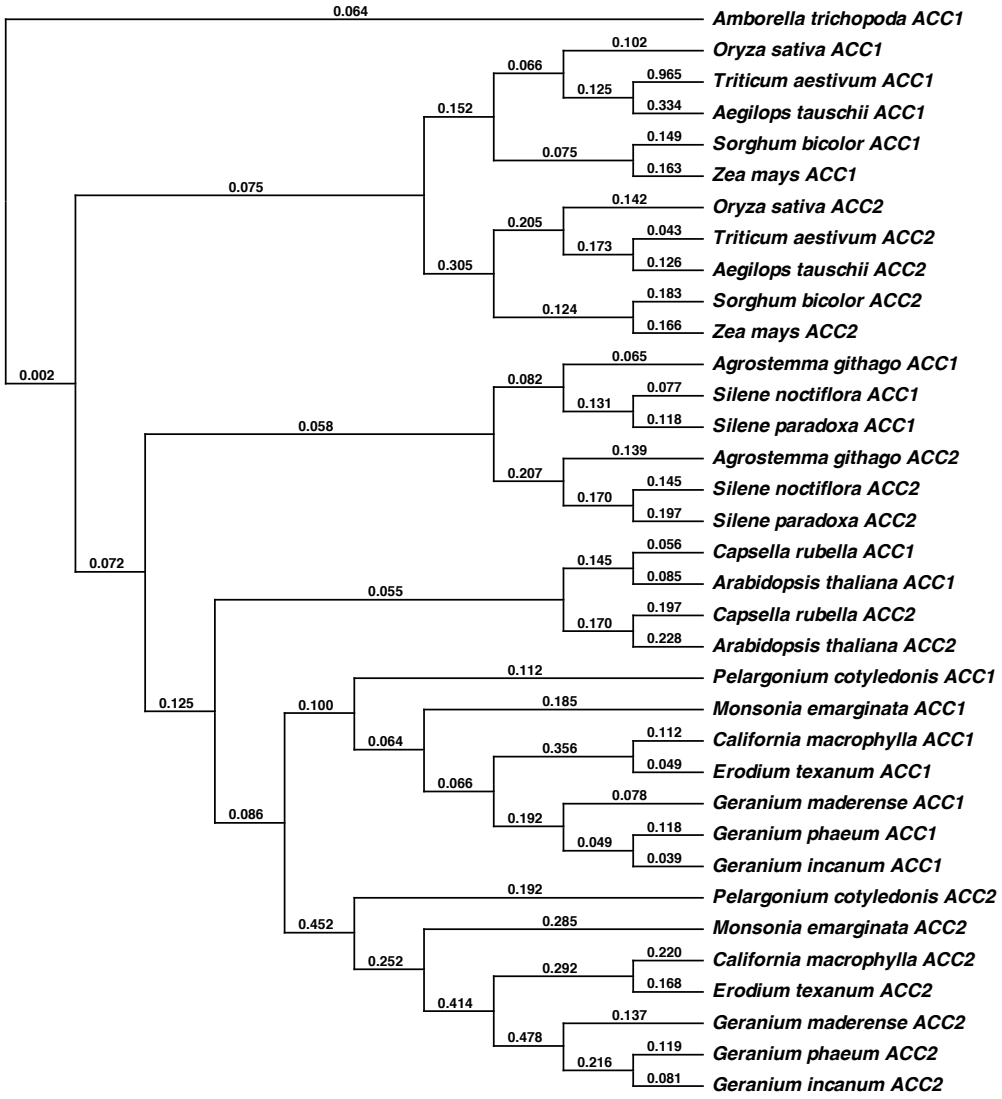

Figure S1. *ACC* tree. Branch labels are d_N/d_S values. *ACC1* represents cytosolic-targeted genes while *ACC2* represents plastid-targeted genes.

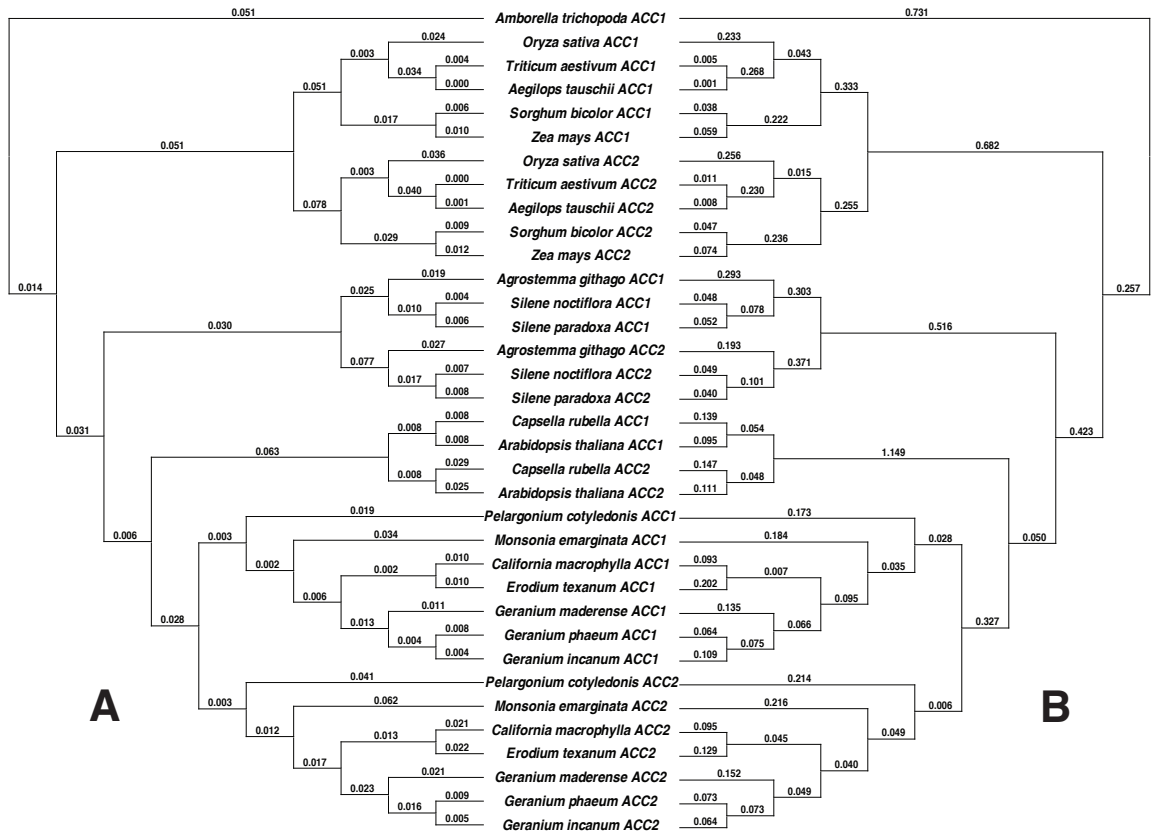

Figure S2. *ACC* tree. A) Branch labels are d_N values. B) Branch labels are d_S values.

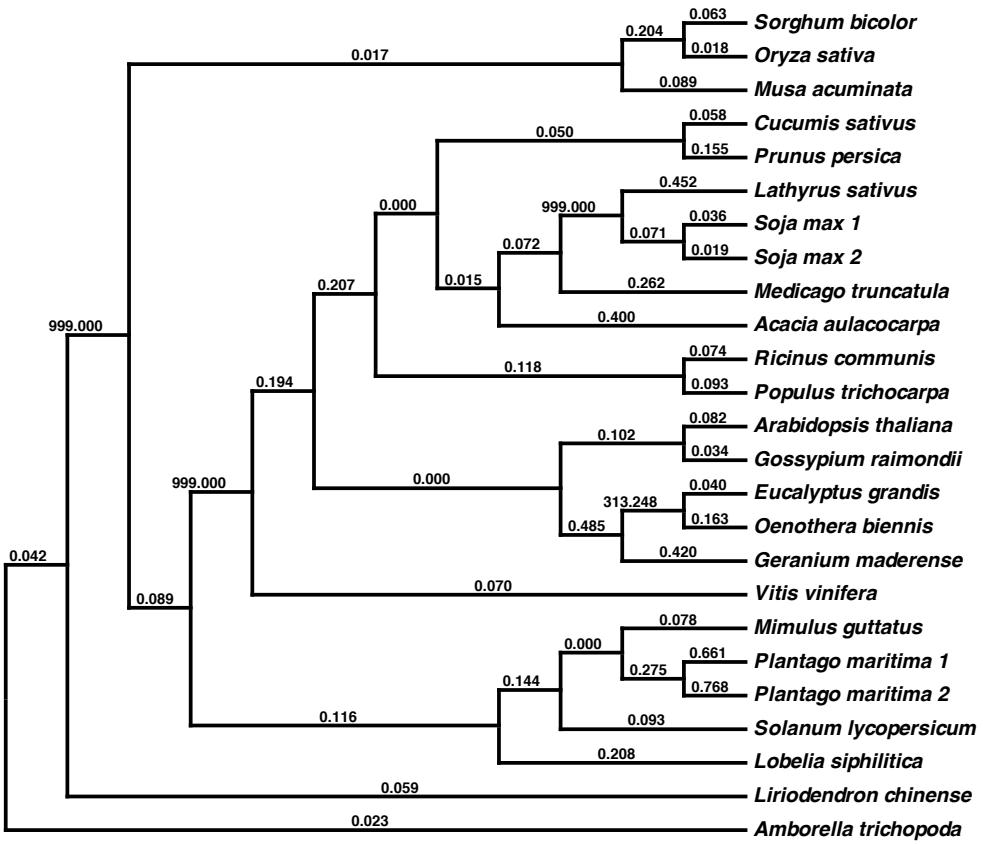

Figure S3. *CLPP3* tree. Branch labels are d_N/d_S values.

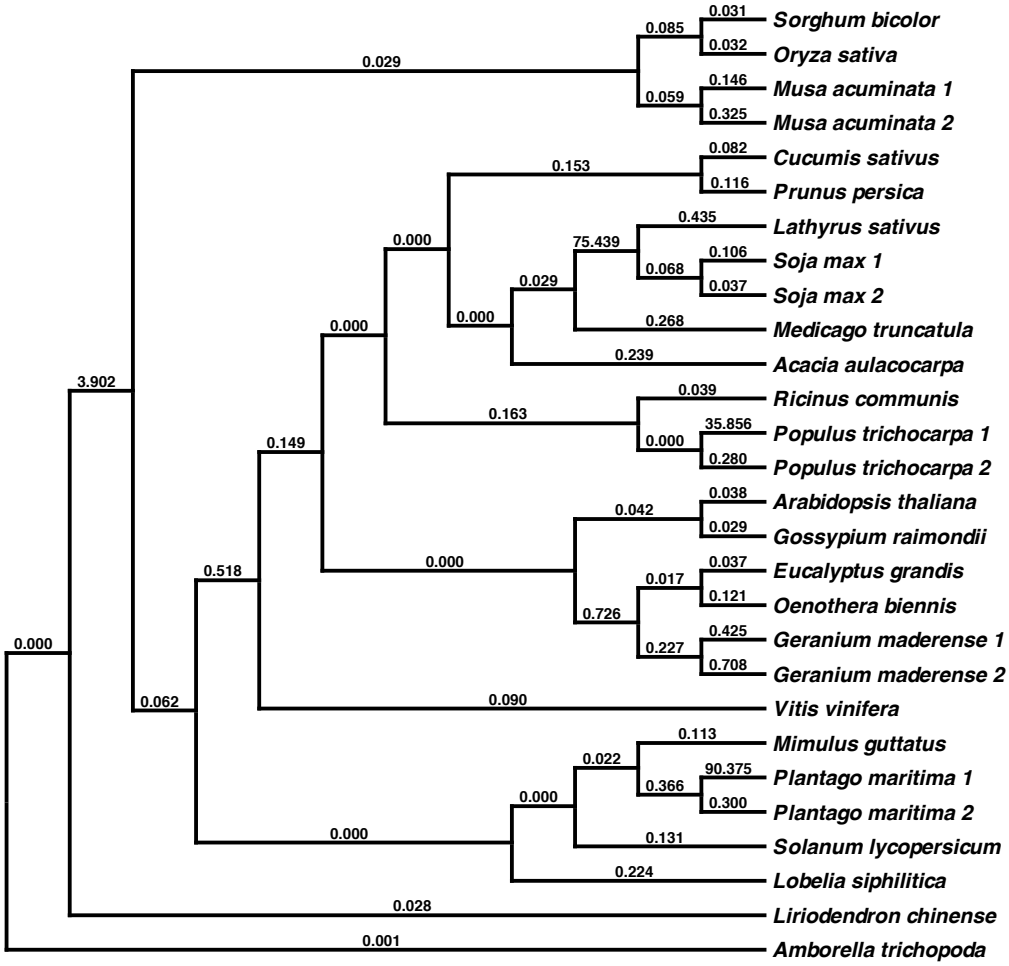

Figure S4. *CLPP4* tree. Branch labels are d_N/d_S values.

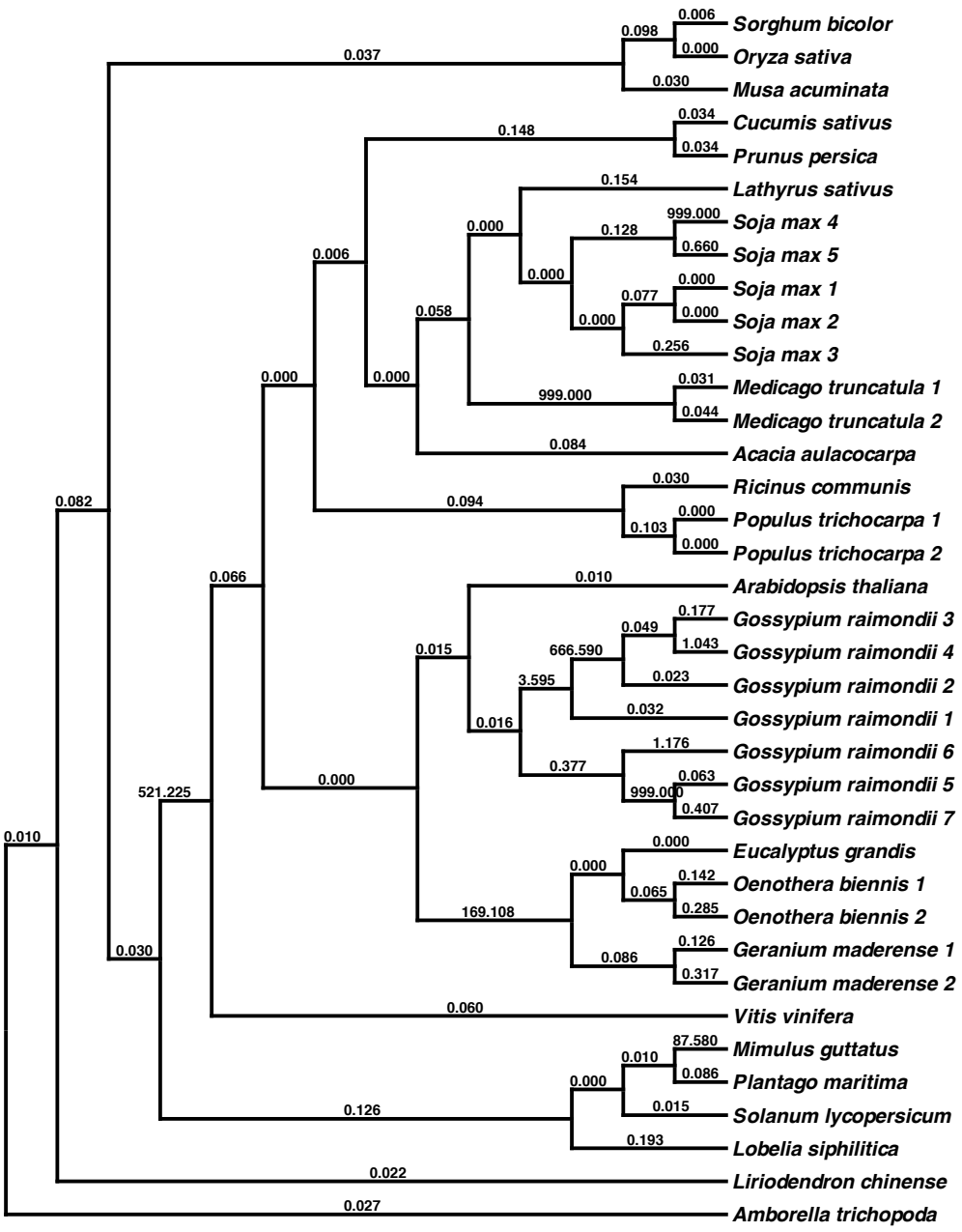

Figure S5. *CLPP5* tree. Branch labels are d_N/d_S values.

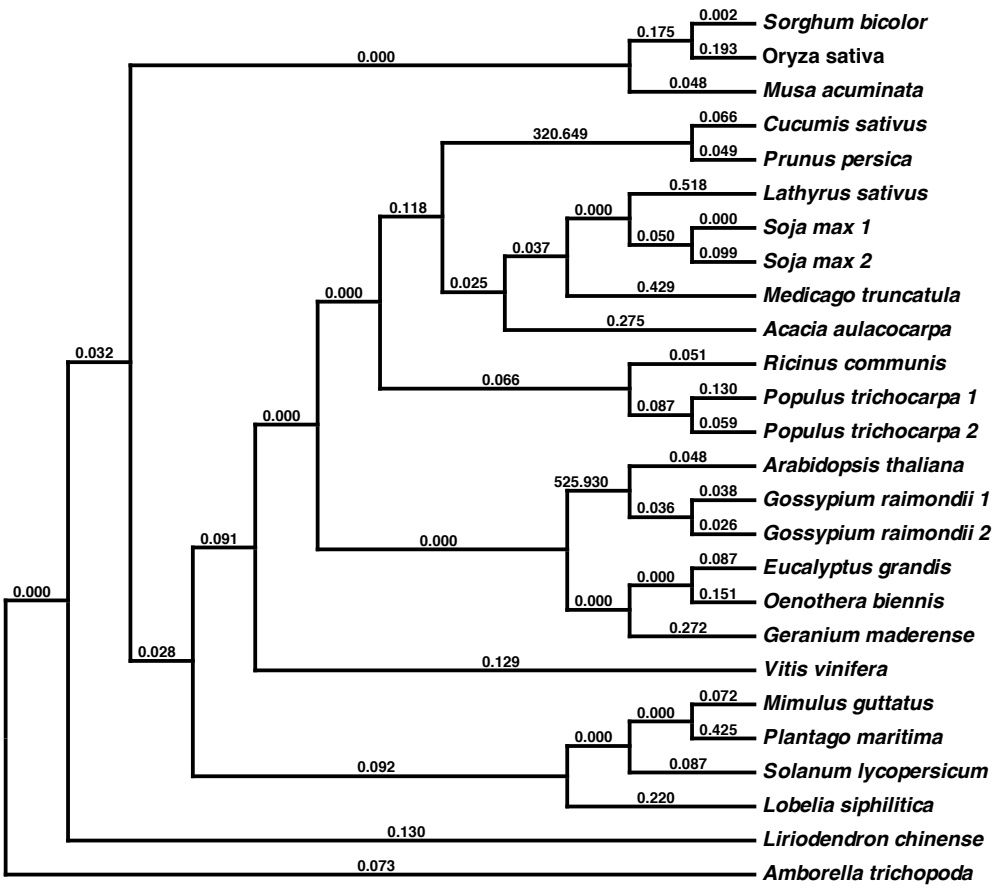

Figure S6. *CLPP6* tree. Branch labels are d_N/d_S values.

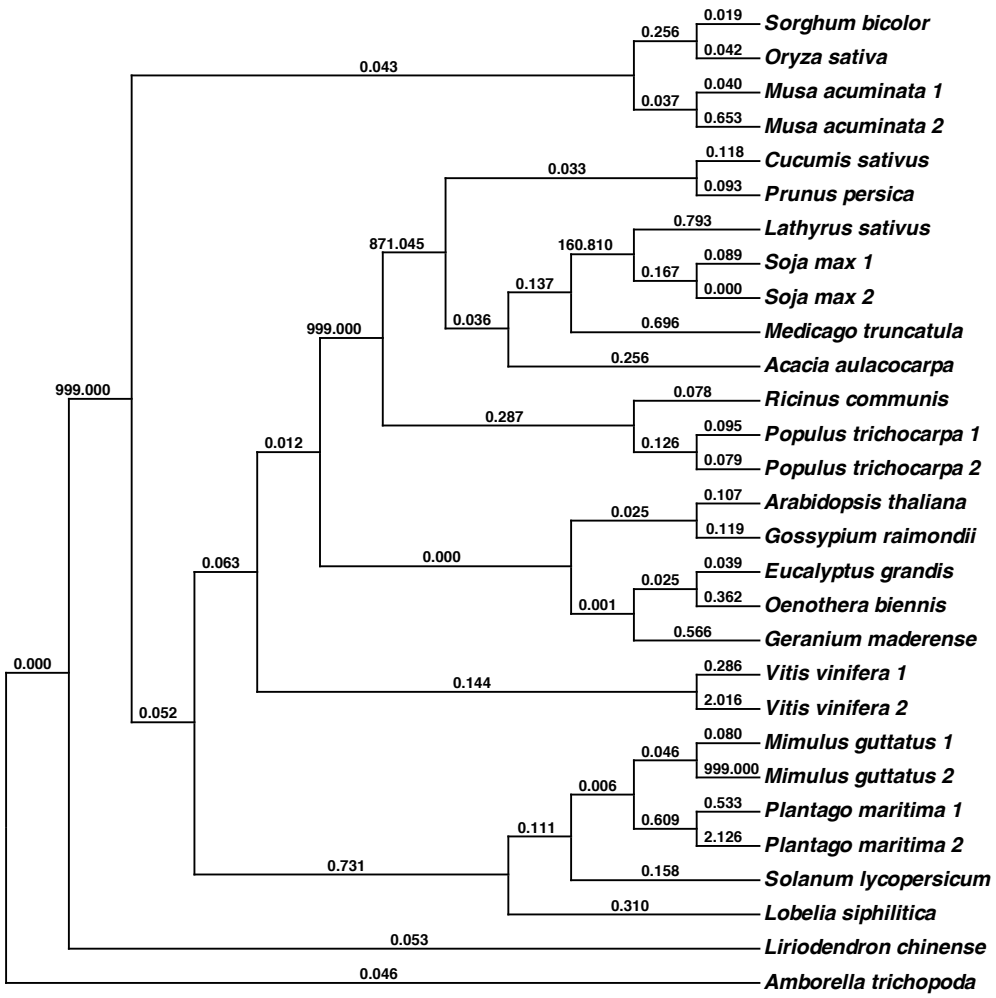

Figure S7. *CLPR1* tree. Branch labels are d_N/d_S values.

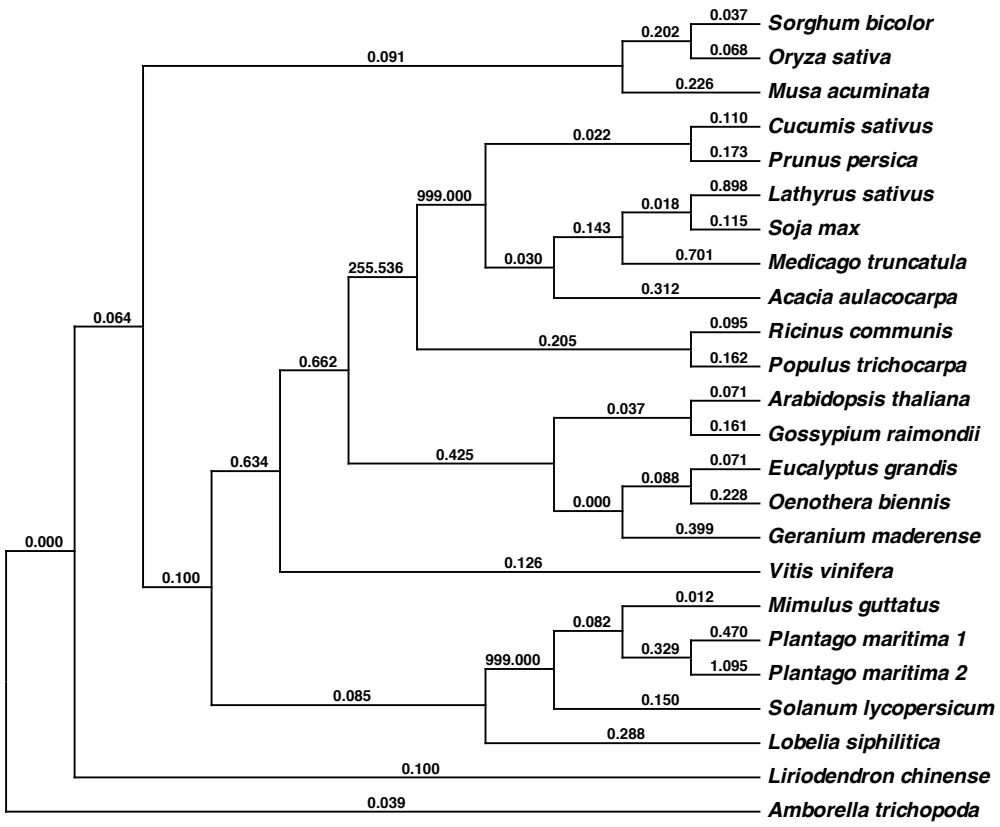

Figure S8. *CLPR2* tree. Branch labels are d_N/d_S values.

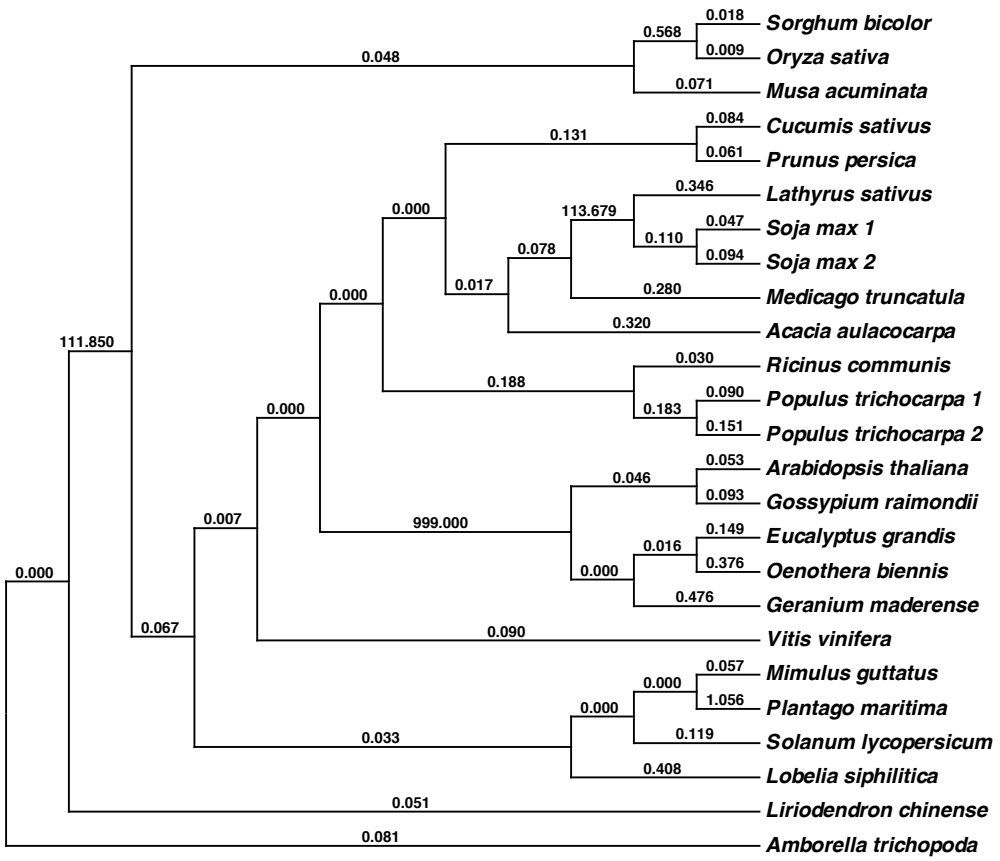

Figure S9. *CLPR3* tree. Branch labels are d_N/d_S values.

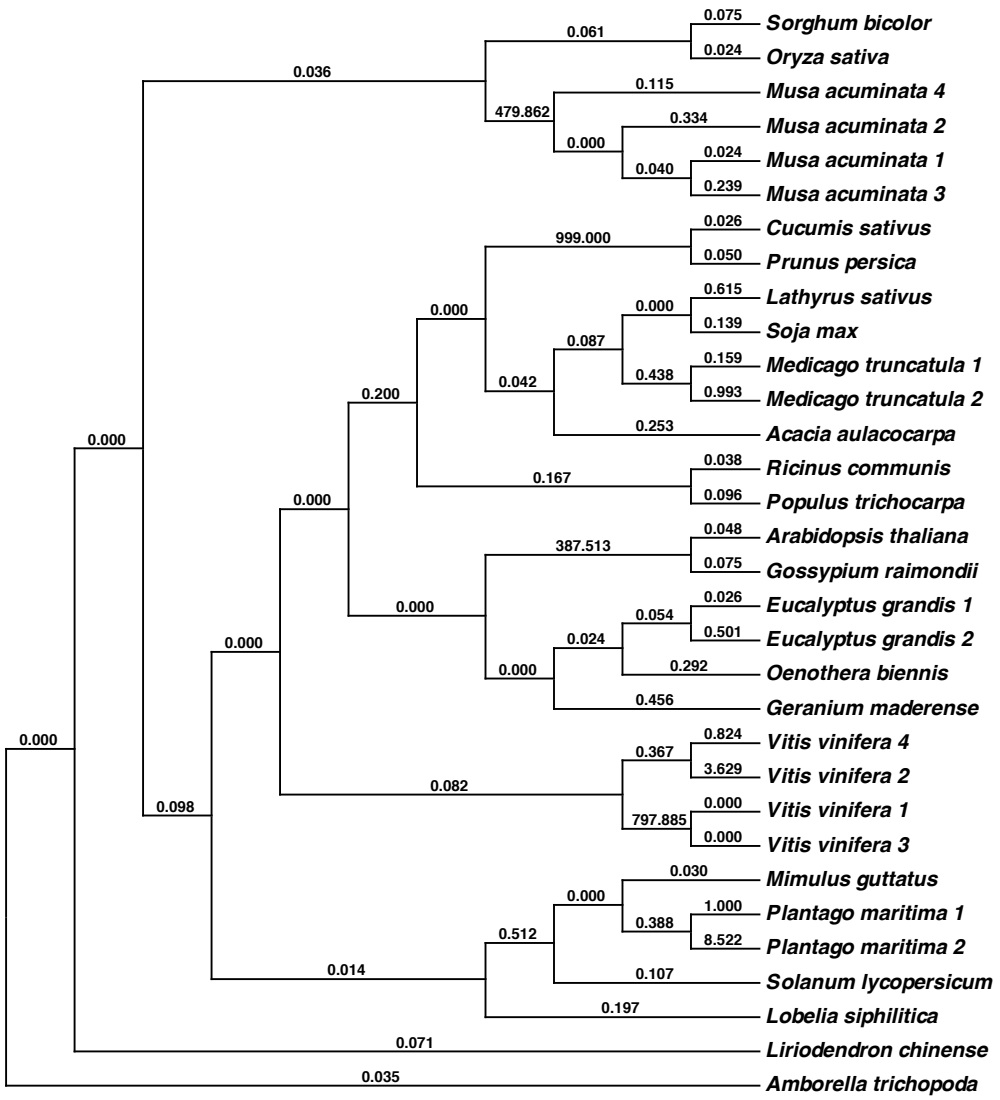

Figure S10. *CLPR4* tree. Branch labels are d_N/d_S values.