# Beyond RuBisCO: Convergent molecular evolution of multiple chloroplast genes in C$_4$ plants

**Claudio Casola** [Corresp., 1, 2] , **Jingjia Li** [1]

1  Department of Ecology and Conservation Biology, Texas A&M University, College Station, Texas, United States

2  Interdisciplinary Graduate Program in Ecology and Evolutionary Biology, Texas A&M University, College Station, Texas, United States

Corresponding Author: Claudio Casola
Email address: ccasola@tamu.edu

**Background**. The recurrent evolution of the C$_4$ photosynthetic pathway in angiosperms represents one of the most extraordinary examples of convergent evolution of a complex trait. Comparative genomic analyses have unveiled some of the molecular changes associated with the C$_4$ pathway. For instance, several key enzymes involved in the transition from C$_3$ to C$_4$ photosynthesis have been found to share convergent amino acid replacements along C$_4$ lineages. However, the extent of convergent replacements potentially associated with the emergence of C$_4$ plants remains to be fully assessed. Here, we conducted an organelle-wide analysis to determine if convergent evolution occurred in multiple chloroplast proteins beside the well-known case of the large RuBisCO subunit encoded by the chloroplast gene *rbcL*.

**Methods**. Our study was based on the comparative analysis of 43 C$_4$ and 21 C$_3$ grass species belonging to the PACMAD clade, a focal taxonomic group in many investigations of C$_4$ evolution. We first used protein sequences of 67 orthologous chloroplast genes to build an accurate phylogeny of these species. Then, we inferred amino acid replacements along 13 C$_4$lineages and 9 C$_3$ lineages using reconstructed protein sequences of their reference branches, corresponding to the branches containing the most recent common ancestors of C4-only clades and C3-only clades. Pairwise comparisons between reference branches allowed us to identify both convergent and non-convergent amino acid replacements between C$_4$:C$_4$, C$_3$:C$_3$ and C$_3$:C$_4$ lineages.

**Results**. The reconstructed phylogenetic tree of 64 PACMAD grasses was characterized by strong supports in all nodes used for analyses of convergence. We identified 217 convergent replacements and 201 non-convergent replacements in 45/67 chloroplast proteins in both C$_4$ and C$_3$ reference branches. C$_4$:C$_4$ branches showed higher levels of convergent replacements than C$_3$:C$_3$ and C$_3$:C$_4$ branches. Furthermore, we found that more proteins shared unique convergent replacements in C$_4$ lineages, with both RbcL and RpoC1 (the RNA polymerase beta' subunit 1) showing a significantly higher convergent/non-convergent replacements ratio in C$_4$ branches. Notably, more C$_4$:C$_4$ reference branches showed higher numbers of convergent vs. non-convergent replacements than C$_3$:C$_3$ and C$_3$:C$_4$ branches. Our results suggest that, in the PACMAD clade, C$_4$ grasses experienced higher levels of molecular convergence than C$_3$ species across multiple chloroplast genes. These findings have important implications for our understanding of the evolution of the C$_4$photosynthesis pathway.

1 **Beyond RuBisCO: Convergent molecular evolution of multiple chloroplast genes in C4**
2 **plants**

3 Claudio Casola[1,2], Jingjia Li[1]
4 [1] Department of Ecology and Conservation Biology, Texas A&M University, College Station,
5 TX, USA
6 [2] Interdisciplinary Graduate Program in Ecology and Evolutionary Biology, Texas A&M
7 University, College Station, TX, USA
8
9 Corresponding author:
10 Claudio Casola
11 534 John Kimbrough Blvd, College Station, TX 77843, USA
12 Email address: claudio.casola@ag.tamu.edu
13

15 **Abstract**

16 **Background**. The recurrent evolution of the $C_4$ photosynthetic pathway in angiosperms
17 represents one of the most extraordinary examples of convergent evolution of a complex trait.
18 Comparative genomic analyses have unveiled some of the molecular changes associated with the
19 $C_4$ pathway. For instance, several key enzymes involved in the transition from $C_3$ to $C_4$
20 photosynthesis have been found to share convergent amino acid replacements along $C_4$ lineages.
21 However, the extent of convergent replacements potentially associated with the emergence of $C_4$
22 plants remains to be fully assessed. Here, we conducted an organelle-wide analysis to determine
23 if convergent evolution occurred in multiple chloroplast proteins beside the well-known case of
24 the large RuBisCO subunit encoded by the chloroplast gene *rbcL*.
25 **Methods**. Our study was based on the comparative analysis of 43 $C_4$ and 21 $C_3$ grass species
26 belonging to the PACMAD clade, a focal taxonomic group in many investigations of $C_4$
27 evolution. We first used protein sequences of 67 orthologous chloroplast genes to build an
28 accurate phylogeny of these species. Then, we inferred amino acid replacements along 13 $C_4$
29 lineages and 9 $C_3$ lineages using reconstructed protein sequences of their reference branches,
30 corresponding to the branches containing the most recent common ancestors of C4-only clades
31 and C3-only clades. Pairwise comparisons between reference branches allowed us to identify
32 both convergent and non-convergent amino acid replacements between $C_4$:$C_4$, $C_3$:$C_3$ and $C_3$:$C_4$
33 lineages.
34 **Results**. The reconstructed phylogenetic tree of 64 PACMAD grasses was characterized by
35 strong supports in all nodes used for analyses of convergence. We identified 217 convergent
36 replacements and 201 non-convergent replacements in 45/67 chloroplast proteins in both $C_4$ and
37 $C_3$ reference branches. $C_4$:$C_4$ branches showed higher levels of convergent replacements than
38 $C_3$:$C_3$ and $C_3$:$C_4$ branches. Furthermore, we found that more proteins shared unique convergent
39 replacements in $C_4$ lineages, with both RbcL and RpoC1 (the RNA polymerase beta' subunit 1)
40 showing a significantly higher convergent/non-convergent replacements ratio in $C_4$ branches.
41 Notably, more $C_4$:$C_4$ reference branches showed higher numbers of convergent vs. non-
42 convergent replacements than $C_3$:$C_3$ and $C_3$:$C_4$ branches. Our results suggest that, in the
43 PACMAD clade, $C_4$ grasses experienced higher levels of molecular convergence than $C_3$ species
44 across multiple chloroplast genes. These findings have important implications for our
45 understanding of the evolution of the $C_4$ photosynthesis pathway.

46
47

**Introduction**

Convergent evolution represents the independent acquisition of similar phenotypic traits in phylogenetically distant organisms. Understanding the genomic changes underlying the recurrent emergence of phenotypes is a major goal of molecular evolution. The rapidly increasing taxonomic breadth of genomic resources combined with the development of rigorous frameworks to comparatively investigate molecular changes has accelerated the pace of discovery in this area. For instance, substitutions in coding regions of conserved genes have been implicated in phenotypic changes responsible for adaptation of marine mammals to an aquatic lifestyle (Foote, et al. 2015; Zhou, et al. 2015). Other examples of convergent phenotypes whose molecular underpinnings have been investigated include adaptations in snake and agamid lizard mitochondria (Castoe, et al. 2009), echolocation in mammals (Parker, et al. 2013; Thomas and Hahn 2015; Zou and Zhang 2015; Storz 2016), and hemoglobin function in birds (Natarajan, et al. 2016). Convergent traits can evolve via changes toward the same derived state (similar phenotype) from the same initial state, which is known as parallelism, or through changes of different initial states, referred to as convergence (Zhang and Kumar 1997; Storz 2016). For the sake of simplification, we will refer to these two processes using the general terminology 'convergence' and 'convergent replacements' throughout the manuscript, unless differently stated.

Several traits are also known to have convergently evolved in land plants (Li, et al. 2018; Lu, et al. 2018; Preite, et al. 2019). One of the most notable examples is represented by the repeated evolution of the $C_4$ photosynthetic pathway in flowering plants. The $C_4$ pathway is a complex functional adaptation that allows for better photosynthesis efficiency under certain environmental conditions, such as dry and warm climates, high light intensity, low $CO_2$ concentration, and limited availability of nutrients (Knapp and Medina 1999; Long 1999). The $C_4$ pathway involves cytological, anatomical and metabolic modifications thought to have evolved multiple times independently in various lineages from the $C_3$ type (Kellogg 1999; Sage 2004; Sage, et al. 2011). According to phylogenetic, anatomical and biochemical evidence, the few slightly different variants of the $C_4$ photosynthesis type evolved more than 60 times in angiosperms (Sage, et al. 2012; Heyduk, et al. 2019). In grasses (family Poaceae) alone, the $C_4$ pathway has evolved independently ~20 times (Grass Phylogeny Working Group 2012).

Transitions from $C_3$ to $C_4$ plants resulted from genetic changes that include nonsynonymous substitutions, gene duplications and gene expression alterations (Christin, et al. 2007; Christin, Boxall, et al. 2013; Christin, et al. 2015; Goolsby, et al. 2018; Heyduk, et al. 2019). It has been suggested that the evolution of the $C_4$ pathways proceeded throughout a series of evolutionary steps wherein the Kranz leaf anatomy typical of this pathway originated first, followed by changes in the expression patterns of key genes and finally by adaptive modifications of protein sequences (Sage, et al. 2012; Christin, Osborne, et al. 2013; Williams, et al. 2013). A model of the adaptive steps leading to $C_4$ photosynthesis showed that key

88    biochemical components of this pathway evolved modularly along a trajectory that was likely
89    very similar across lineages with $C_3$ to $C_4$ transitions (Heckmann, et al. 2013). Overall, these
90    scenarios suggest that enzymes involved in $C_3$ to $C_4$ transitions experienced similar selective
91    pressures that resulted in the convergent evolution of the same amino acid replacements across
92    $C_4$ lineages.

93

94          Evidence of convergent changes in proteins associated with photosynthetic processes has
95    steadily accumulated since genomic data from multiple $C_4$ lineages have become available in the
96    past couple of decades. Most of these studies have focused on the ribulose-1,5-bisphosphate
97    carboxylase/oxygenase (RuBisCO), a large multimeric enzyme that catalyzes the carboxylation
98    of ribulose-1,5-bisphosphate (RuBP), allowing plants to fix atmospheric carbon (Andersson and
99    Backlund 2008). RuBisCO also catalyzes oxygenation of RuBP, which leads to loss of carbon in
100   the process of photorespiration (Andersson and Backlund 2008; Maurino and Peterhansel 2010).
101   RuBisCO's limited ability to discriminate between $CO_2$ and $O_2$ has been attributed to the much
102   higher $CO_2$ to $O_2$ atmospheric partial pressure until ~400 million years ago (Sage 1999, 2004;
103   Sage, et al. 2012).

104

105          Previous studies have revealed multiple convergent amino acid replacements in the large
106   RuBisCO subunit in $C_4$ lineages, encoded by the chloroplast gene *rbcL* (Kapralov and Filatov
107   2007; Christin, et al. 2008; Kapralov, et al. 2011; Kapralov, et al. 2012; Piot, et al. 2018). Some
108   of these convergent replacements have been associated to positive selection of the corresponding
109   codons in $C_4$ monocot and eudicot lineages (Kapralov and Filatov 2007; Christin, et al. 2008;
110   Kapralov, et al. 2012; Piot, et al. 2018). Notably, biochemical analyses have demonstrated that
111   some recurrent amino acid changes in the large RuBisCO subunit of $C_4$ plants critically alter the
112   kinetics of RuBisCO, resulting in an accelerated rate of $CO_2$ fixation at the beginning of the
113   Calvin-Benson cycle (Studer, et al. 2014; Bouvier, et al. 2021). Convergent amino acid changes
114   have also been described in enzymes that are encoded by nuclear genes and play a primary role
115   in the $C_4$ pathway, including the phosphoenolpyruvate carboxylase PEPC (Christin, et al. 2007;
116   Besnard, et al. 2009), the NADP-malic enzymes NADP-me (Christin, Samaritani, et al. 2009),
117   the phosphoenolpyruvate carboxykinase PEPCK (Christin, Petitpierre, et al. 2009) and the small
118   RuBisCO subunit (Kapralov, et al. 2011).

119

120          Given the number of biochemical, physiological and anatomical traits that were affected
121   in each evolutionary transition from $C_3$ to $C_4$ photosynthesis (Heyduk, et al. 2019), it is likely
122   that many genes experienced analogous selective pressures across taxa that include $C_4$ plants.
123   This has been shown to be the case by Huang et al. (Huang, et al. 2017), who have developed an
124   approach to identify potential genes involved in the transition to $C_4$ photosynthesis using a
125   genome-wide scan for selection along a phylogeny of PACMAD grasses. Of the 88 genes
126   showing signatures of positive or relaxed selection in $C_4$ species, several were not previously
127   known to have a role in $C_4$ photosynthesis. Although this study did not focus on finding

128  convergent replacements, it provided a comprehensive strategy and statistical testing framework
129  to identify novel genes that have likely played a role in the evolution of $C_4$ grasses. It is possible
130  that a significant fraction of these genes accumulated convergent amino acid replacements during
131  $C_3$-to-$C_4$ transitions.
132      Another recent, important work has produced the first analysis of convergent
133  replacements across multiple proteins involved in the metabolism of $C_4$ and crassulacean acid
134  metabolism (CAM) among species belonging to the portullugo clade (Caryophyllales). Goolsby
135  and colleagues (2018) compared evolutionary patterns in 19 gene families with critical roles in
136  metabolic pathways of both $C_4$ and CAM plants, also known as carbon-concentration
137  mechanisms (CCMs) genes, and in 64 non-CCM gene families. They found convergent
138  replacements in proteins from $C_4$ and CAM lineages, as well as higher levels of convergent
139  replacements in CCM vs. non-CCM gene families (Goolsby, et al. 2018). Additionally, several
140  amino acid replacements that are prevalent among $C_4$ and CAM taxa compared to $C_3$ lineages
141  were identified in this study (Goolsby, et al. 2018).
142
143      Altogether, the results of this and other studies demonstrated that convergent molecular
144  evolution occurred across multiple genes in both $C_4$ and CAM groups. While significant progress
145  has been made towards the detection of signatures of selection associated to the evolution of
146  CCMs (Huang, et al. 2017; Piot, et al. 2018), a rigorous framework to assess the full extent of
147  molecular convergence in $C_3$ to CCMs transitions has yet to be presented. For example, analyses
148  of convergent evolution should include null hypotheses that assume no differences between taxa
149  with and without convergence. In the case of CCMs evolution, a plausible null hypothesis
150  consists in statistically equivalent numbers of convergent replacements between $C_4$ (or CAM)
151  lineages and $C_3$ lineages.
152
153      Additionally, nonadaptive replacements should be used to normalize convergent
154  replacements, in order to account for variation in the rates of nonsynonymous substitutions
155  across lineages. This approach has been successfully applied in studies of molecular convergent
156  evolution in vertebrates by assessing both convergent replacements and protein sequence
157  changes that result in different amino acids, or *divergent replacements* (Castoe, et al. 2009;
158  Thomas and Hahn 2015; Zou and Zhang 2015). A broader definition of the latter group
159  incorporates all replacements leading to different amino acids, regardless of their ancestral state.
160  We refer to such changes as *non-convergent replacements*.
161      Furthermore, testing hypotheses about the extent of convergent molecular evolution
162  remains particularly challenging for many nuclear genes, because of the prevalence of duplicated
163  copies, particularly in plants (Christin, et al. 2007; Goolsby, et al. 2018). Single-copy nuclear or
164  organelle genes allow to more easily recognize convergent changes and overcome possible
165  confounding compensatory effects due to the presence of paralogous copies.
166

167         Given these premises, we sought to test if convergent amino acid changes occur more
168 frequently in proteins encoded by chloroplast genes in a taxon that includes multiple well-
169 characterized lineages of $C_4$ and $C_3$ grasses. Chloroplast proteins represent an ideal set of targets
170 to study the role of convergent evolution in $C_3$ to $C_4$ transitions for a variety of reasons. First,
171 most chloroplast proteins are involved in biochemical and biophysical processes that are critical
172 to photosynthesis. For instance, out of ~75 functionally annotated protein-coding genes in the
173 maize chloroplast genome, 45 genes are implicated in photosynthesis-related processes,
174 including *rbcL*, 17 genes coding for subunits of the photosystems I and II (PS I and PS II), 12
175 genes coding for subunits of the NADH dehydrogenase complex, 6 genes coding for chloroplast
176 ATPase subunits, 4 genes coding for cytochrome b6f complex subunits, and a few more genes
177 implicated in the assembly of other protein complexes (Maier, et al. 1995). Second,
178 nonannotated orthologous copies of chloroplast genes can be readily identified across plants
179 through sequence homology searches, taking advantage of the thousands of complete chloroplast
180 genome sequences currently available for green plants. Third, comparative studies of convergent
181 evolution in $C_4$ photosynthesis are facilitated by detailed reconstruction of phylogenetic
182 relationships within groups with both $C_4$ and $C_3$ lineages. Fourth, signatures of positive selection
183 have been found in multiple chloroplast genes in taxa that contain both $C_3$ and $C_4$ plants,
184 although only the genes *rbcL* and *psaJ*, which encodes a small subunit of the Photosystem I
185 complex, showed evidence of adaptive changes exclusively in $C_4$ lineages (Christin, et al. 2008;
186 Goolsby, et al. 2018; Piot, et al. 2018). Finally, most chloroplast genes occur as single copy loci,
187 as opposed to the multiple paralogs typically present for plant genes encoded in the nucleus.
188
189         In this study, we analyzed 67 chloroplast genes from 64 grass species, including 43 $C_4$
190 and 19 $C_3$ species belonging to the PACMAD clade, named after six of its most representative
191 subfamilies: Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae and
192 Danthonioideae. Using published information, we placed thirteen known independent $C_3$ to $C_4$
193 transitions in the reconstructed phylogeny of these 64 species. We applied a series of tests based
194 on convergent vs. non-convergent amino acid replacements and determined that convergent
195 molecular evolution occurred at a higher rate in chloroplast genes of $C_4$ lineages compared to $C_3$
196 lineages, a pattern that remained largely unchanged after excluding the RbcL protein from the
197 convergence analyses. Our findings suggest that the evolutionary trajectories of multiple
198 chloroplast genes have been affected during the emergence of the $C_4$ adaptation in the PACMAD
199 clade, a result that has significant implications for our understanding of $C_4$ photosynthesis
200 evolution.
201
202
203 **Methods**
204
205 ***Data source and filtering***

206       We queried NCBI GenBank (Sayers, et al. 2019) for complete chloroplast genome
207 sequences of grass species that were included in phylogenetic analyses by the Grass Phylogeny
208 Working Group II (2012) and downloaded the corresponding coding sequences. Each species
209 was assigned to either $C_3$ or $C_4$ type following the results of the Grass Phylogeny Working
210 Group II (2012). Additionally, we downloaded the coding chloroplast sequences for
211 *Dichanthelium acuminatum*, *Thyridolepis xerophila*, *Sartidia dewinteri* and *Sartidia perrieri* ($C_3$
212 species) (Brown and Smith 1972; Smith and Brown 1973; Hattersley and Stone 1986; Hattersley,
213 et al. 1986; Besnard, et al. 2014). We used the standalone blastn ver. 2.2.29+ (Camacho, et al.
214 2009) with the Expect value (E) cutoff of $1e^{-10}$ to determine putative sequence orthology with
215 coding sequences of the *Zea mays* chloroplast genes (Maier, et al. 1995). Single copy putative
216 orthologs that were present in more than 95% of the species were retained for further analysis
217 (Table S1).
218
219 ***Multiple sequence alignment***
220       We aligned the individual sequences using TranslatorX ver. 1.1 (Abascal et al., 2010) and
221 the multiple sequence aligner MUSCLE with default parameters. Alignments were further
222 adjusted manually using BioEdit ver. 7.0.9.0 (Hall 1999). Stop codons and sites that could not be
223 aligned unambiguously were removed.
224
225 ***Phylogeny reconstruction***
226       We concatenated the individual sequence alignments and extracted third codon position
227 sites for phylogeny reconstruction. We ran PartitionFinder ver. 1.1.1 (Lanfear, et al. 2012) to
228 identify the best partitioning scheme (partitioning by gene) for the downstream analysis using
229 both Akaike information criterion (AIC) (Akaike 1973) and Bayesian information criterion
230 (BIC) (Schwarz, 1978). We then used maximum likelihood framework as implemented in
231 RAxML ver. 8.2.10 (Stamatakis 2014) to reconstruct the phylogeny. Branch support was
232 estimated using 1,000 bootstrap replicates. *Oryza sativa* and *Brachypodium distachyon* from the
233 BOP (Bambusoideae, Oryzoideae and Pooideae) clade were used as outgroup, whereas all
234 ingroup species belonged to the PACMAD clade. We used FigTree ver. 1.4.0 (Rambaut 2012) to
235 rearrange and visualize the phylogeny, and the figures were edited further to improve readability
236 and to indicate $C_4$/$C_3$ classification.
237
238 ***Ancestral state reconstruction***
239       We reconstructed ancestral states at each phylogenetic node for each individual gene
240 using the program codeml from the software package PAML ver. 4.9a (Yang 2007) and the basic
241 codon substitution model (model = 0, NSsites = 0). The guide tree consisted of the cladogram of
242 all species with available sequences for each individual gene. Sites with gaps in one or more
243 PACMAD species were excluded.
244
245 ***Definition and characteristics of "reference branches"***

246    In the reconstructed PACMAD phylogeny, we identified the branches including the most
247    recent common ancestors of $C_4$-only clades and $C_3$-only clades. We refer to these branches as
248    "$C_4$ reference branches" and "$C_3$ reference branches", respectively (see Figs. 1 and 2). We then
249    compared the inferred protein sequence of each reference branch with the inferred sequence in
250    their ancestral branch (next branch toward the root), in order to identify individual site changes
251    that occurred along reference branches.
252    To assess the number of convergent and non-convergent replacements, amino acid
253    changes were compared in all possible pairs of reference branches. Replacements in two
254    reference branches that resulted in the same state (amino acid) at a given site were considered
255    convergent, regardless of whether the corresponding ancestral states were the same or different
256    (Castoe, et al. 2009). After identifying convergent replacements, we separated them into parallel
257    and convergent changes (Zhang and Kumar 1997; Storz 2016). Likewise, two replacements were
258    considered non-convergent if states at the descendant orthologous sites were different, regardless
259    of the corresponding ancestral states (Castoe, et al. 2009).
260    The pairwise comparisons between reference branches are akin to the phylogenetically
261    independent contrast (PIC) method developed by Felsenstein (Felsenstein 1985). In the PIC
262    approach, the values to compare are represented by differences between branches. The
263    differences between two branches are independent of the differences between two other
264    branches. Therefore, pairwise comparisons of these values are independent and can be tested
265    using 2x2 contingency table tests (see also below). In our study, pairwise comparisons are
266    independent from each other, i.e., replacements in each pair of branches are independent of
267    replacements in each other pairs of branches. The difference from the PIC method is that we
268    compare both differences (non-convergent replacements) and similarities (convergent
269    replacements). A similar approach have been used in studies of convergent amino acid
270    replacements (Castoe, et al. 2009; Foote, et al. 2015; Thomas and Hahn 2015).
271    Reference branch lengths were extracted from the RAxML phylogeny obtained on the
272    AIC partitioning scheme (Fig. S2). Testing was performed on the sum of pairs of branch lengths
273    for each photosynthesis type using the R package exactRankTests (Table S2).
274
275    ***Inference of convergent and non-convergent replacements and statistical testing***
276    Using the approach described above, we identified putative convergent and non-
277    convergent amino acid changes in each gene product individually. We summarized those data
278    within each of the three categories: (1) two $C_4$ reference branches ($C_4$:$C_4$), (2) $C_3$ reference
279    branch and $C_4$ reference branch ($C_3$:$C_4$), and (3) two $C_3$ reference branches ($C_3$:$C_3$).
280    To test the significance of replacement differences between categories we used the
281    Boschloo's exact unconditional test (Boschloo 1970) implemented in the SciPy library ver. 1.7.1
282    in python3 (Virtanen, et al. 2020). In the Boschloo's test, the $p$-value from the Fisher's exact test
283    represents the test statistic of the exact unconditional test. It has been shown that Boschloo's test
284    is more powerful than Fisher's exact test (Mehrotra, et al. 2003). There is no restriction to using
285    contingency table tests, including Boschloo's test, on categories with different sample size, as

286    long as the categories are independent (Mehrotra, et al. 2003), as in the case of reference
287    branches in our phylogeny.
288
289    ***Data availability***
290        Raw data, including alignments, fasta sequences, and phylogenetic analyses data, are
291    available through the following Figshare repository:
292    https://figshare.com/articles/dataset/Convergence-chlorplast-genes-C4-Casola-Li-
293    2021/15180690.
294
295
296    **Results**
297
298    ***Phylogeny reconstructions***
299        We examined 63 grass chloroplast genomes to identify gene orthologs for *Zea mays*
300    chloroplast genes and extracted the corresponding coding and protein sequences. The resulting
301    dataset included up to 67 DNA/protein sequences in 64 grass species that were retained for
302    further analysis (Table S1). One to four sequences were absent in thirteen species. Out of 64
303    species, 43 were classified as $C_4$ and 21 (including two outroup species) as $C_3$. The reconstructed
304    phylogeny is well supported, except for three branches with low to moderate bootstrap values,
305    and it is consistent for both AIC and BIC partitioning schemes (Fig. 1 and Figs. S1-S3). We
306    identified thirteen $C_4$ reference branches that represent putative $C_3$ to $C_4$ transitions, and nine $C_3$
307    reference branches (Fig. 1). Four pairs of reference branches corresponding to $C_3$ to $C_4$
308    transitions—B-C, E-F, J-L and S-T—are sister to each other in Figure 1. Phylogenetic inferences
309    from deep-taxonomic sampling of the PACMAD clade has shown that each of the these four
310    pairs of reference branches is separated by at least one clade of $C_3$ species (Grass Phylogeny
311    Working Group 2012), supporting the independent origin of $C_4$ photosynthesis in all reference
312    branches shown in Figure 1. However, no high-quality chloroplast genomes are available for any
313    of the $C_3$ species between these pairs of reference $C_4$ branches, precluding their inclusion in our
314    study.
315        Overall, the reference branches A-V showed support values that were in close agreement
316    with those reported in the Grass Phylogeny Working Group II (2012), including the three
317    branches with low statistical support in our tree. Importantly, the species topology was identical
318    between the two phylogenies downstream these three branches. We also noticed three other
319    branches that shared higher statistical support in our phylogeny compared to the Grass
320    Phylogeny Working Group II tree. Two of these branches occurred in the subtribe Boivinellinae
321    and correspond to the split between the group J/K and the branch L, and the split between the
322    group I/J/K/L and the branch M (Fig. 1). The third node with higher support in our phylogeny
323    correspond to the reference branch Q (tribe Arundoideae).
324
325    ***Convergent and non-convergent amino acid replacements across chloroplast proteins***

326      We assessed the level of molecular convergence in $C_3$ to $C_4$ transitions by quantifying
327   convergent and non-convergent amino acid replacements across the PACMAD phylogeny by
328   performing pairwise comparisons of reconstructed sequences in reference branches (Figs. 2 and
329   3, Table S3; see Methods). A total of 217 sites showed at least one convergent replacement: 104
330   in $C_4$:$C_4$, 120 in $C_3$:$C_4$ and 34 in $C_3$:$C_3$ pairs. A further 201 sites exhibited one or more non-
331   convergent replacements: 96 in $C_4$:$C_4$, 121 in $C_3$:$C_4$, and 39 in $C_3$:$C_3$ pairs (Table 1). The
332   difference in convergent/non-convergent site distributions between the three photosynthesis
333   types was not statistically significant ($P \geq 0.05$, Boschloo's test; Table 1). The vast majority of
334   convergent replacements shared the same ancestral state and should thus be considered parallel
335   replacements according to widely accepted definitions of convergence (Zhang and Kumar 1997;
336   Storz 2016). Only two sites, one in MatK (T205S/K205S in two $C_4$ reference branches) and the
337   other in NdhF (L636I/K636I in one $C_4$ and three $C_3$ reference branches), shared replacements
338   with different ancestral states, representing true convergent sites (Table S3).
339      To control for possible biases in the counting of convergent replacements due to branch
340   length variation, we tested whether reference branch lengths in the three photosynthesis types
341   $C_4$:$C_4$, $C_3$:$C_4$ and $C_3$:$C_3$ were different (Table S2). We found no significant difference among
342   types ($P > 0.5$ for each of the three pairwise comparisons, Mann-Whitney $U$ test). We performed
343   the same test only on branches with convergent and non-convergent replacements and found no
344   significant difference between categories ($P > 0.5$, Mann-Whitney $U$ test; Table S2). Therefore,
345   branch length variation between the three types is not expected to affect our results.
346
347      Among the $C_4$ reference branches, several individual sites showed high contrast in the
348   number of branches involved in convergent and non-convergent replacements (Fig. 3, Tables S3
349   and S4). For example, seven $C_4$ branches (54%) shared the H18Q replacement in the product of
350   *ndhH*, with no non-convergent replacements. Six, five, and four $C_4$ branches (46%, 38%, and
351   31%) showed convergent replacements at three sites in the RbcL protein (V101I, M309I, and
352   A328S, respectively). Furthermore, six $C_4$ branches shared the S25G replacement in the product
353   of *ndhI* and four L204F changes in the protein encoded by *matK*. In all these cases, there were no
354   other convergent or non-convergent replacements in $C_3$:$C_3$ or $C_3$:$C_4$ branch comparisons, except
355   for one H18Q change in NdhH in a $C_3$:$C_3$ branch. Two sites with convergent replacements in the
356   proteins encoded by *ndhF* (L557F) and *rpoC2* (H875Y) were found uniquely in $C_3$:$C_3$ pairs, and
357   only one site in the protein Rps3 showed convergence independently in $C_4$:$C_4$ and $C_3$:$C_3$ pairs
358   (Fig. 3).
359      We then searched for convergent replacements that occurred along more than two $C_4$
360   branches at sites that remained otherwise conserved in $C_3$ and $C_4$ lineages, arguing that such
361   changes could result from selective pressure rather than drift. We identified twelve $C_4$-specific
362   convergent sites in proteins from 7 genes: *matK*, *ndhF*, *ndhG*, *ndhI*, *rbcL, rpoC1* and *rpoC2*
363   (Table S4). Five of these sites were found in RbcL, whereas two sites were identified in NdhI.
364   We also observed two convergent sites NdhF and one in RpoC2 that were uniquely found in
365   three $C_3$ branches.

366
367
368   ***Molecular convergence in individual chloroplast proteins***
369           Convergent and non-convergent amino acid replacements were detected in the products
370   of 45 chloroplast genes, thirteen of which had at least one site with four or more replacements
371   (Fig. 4, Table 1 and Table S3). Twenty-four genes had convergent changes in $C_4$:$C_4$, 26 in $C_3$:$C_4$,
372   and 13 in $C_3$:$C_3$ types of pairs (Table 1). Although the convergent/ non-convergent replacement
373   ratio was higher in $C_4$:$C_4$ pairs than $C_3$:$C_4$ and $C_3$:$C_3$ pairs, the differences between the three
374   photosynthesis types were not statistically significant ($P \geq 0.05$, Boschloo's test; Table 1). The
375   lack of replacements was the single most common state for chloroplast proteins across
376   photosynthesis types; however, in $C_4$:$C_4$ there were more genes with a higher number of
377   convergent vs. non-convergent replacements (Fig. 4 and Table S5).
378
379           Overall, 26 proteins showed a higher number of convergent vs. non-convergent sites, of
380   which 16, 13 and 10 were found in $C_4$:$C_4$, $C_3$:$C_4$ and $C_3$:$C_3$ pairs, respectively (Fig. 5 and Table
381   S5). We found statistically significant differences in the number of convergent vs. non-
382   convergent replacements between $C_4$:$C_4$ and $C_3$:$C_4$ pairs, but not $C_3$:$C_3$ pairs, in the products of
383   the genes *rbcL*, *rpoC1* and *rpoC2* ($P < 0.05$, Boschloo's test; Table S5). In RbcL and RpoC1,
384   $C_4$:$C_4$ pairs shared much higher proportion of convergent vs. non-convergent replacements,
385   whereas the opposite was true in RpoC2. RpoC1 was also the only protein showing more
386   convergent than non-convergent replacements in $C_4$:$C_4$ pairs compared to $C_3$:$C_3$ and $C_3$:$C_4$ pairs.
387   In $C_4$:$C_4$ pairs, RpoC1 shared 4 convergent and 1 non-convergent replacement, compared to 1
388   and 2 in $C_3$:$C_3$ pairs and 1 and 5 in $C_3$:$C_4$ pairs, respectively. Additionally, the proteins NdhG,
389   NdhI, PsaI, RpoA, Rps4 and Rps11 exhibited convergent replacements only in $C_4$:$C_4$ pairs
390   (Table S5). When considering the number of affected sites rather than the number of
391   replacements, no genes showed a significantly different pattern between photosynthesis types ($P$
392   $\geq 0.05$, Boschloo's test; Table S5).
393
394           The proteins encoded by *matK*, *rpoC2* and *ndhF* shared much higher numbers of both
395   convergent and non-convergent replacements than other chloroplast proteins across all
396   photosynthesis type comparisons (Table S5). Both *matK* and *ndhF* are known to be rapidly
397   evolving and have been consistently used in low taxonomic level phylogenetic studies in
398   flowering plants (Patterson and Givnish 2002; Barthet and Hilu 2008). The gene *rpoC2* has also
399   been recently described as a useful phylogenetic marker in angiosperms (Walker, et al. 2019).
400
401
402   ***Molecular convergence across reference branches***
403           The comparison of reference branch pairs with convergent and non-convergent
404   replacements revealed remarkable differences between photosynthesis types. Overall, $C_4$:$C_4$ pairs
405   of reference branches showed a distribution skewed toward more convergent and non-convergent

406  replacements than the two other categories (Fig. 6). There were significantly fewer pairs of C$_4$:C$_4$
407  reference branches with no replacements and with no convergent replacements than C$_3$:C$_4$ and
408  C$_3$:C$_3$ pairs ($P$ < 0.05, Boschloo's test; Table 2). Conversely, significantly more C$_4$:C$_4$ pairs
409  shared more convergent than non-convergent replacements, and at least two convergent changes
410  compared to C$_3$:C$_4$ and C$_3$:C$_3$ pairs ($P$ < 0.05, Boschloo's test; Table 2). No significant difference
411  was observed between pairs of C$_3$:C$_4$ and pairs of C$_3$:C$_3$. We found identical patterns when the
412  same analyses were performed after excluding all replacements in the RbcL protein, except for
413  the lack of a significant difference between C$_4$:C$_4$ and C$_3$:C$_3$ in the proportion of pairs with non-
414  convergent replacements and pairs with more convergent than non-convergent changes (Table
415  S6).

416

417  ***Distribution of amino acid replacements across PACMAD lineages***
418       Convergent and non-convergent replacements were preferentially found in specific pairs
419  of reference branches. In C$_4$ pairs, convergent sites were most abundant between *Danthoniopsis*
420  *dinteri* and *Aristida purpurea* (ten sites, branches P and V in Fig. 1), whereas non-convergent
421  sites were most common between *Centropodia glauca* and *Aristida purpurea* (ten sites, branches
422  S and V in Fig. 1). In pairwise C$_3$ branch comparisons, most convergent sites were identified
423  between both *Zeugites pittieri* and Danthonieae (branches N and R in Fig. 1) and Danthonieae
424  and *Sartidia* spp. (branches R and U in Fig. 1), whereas the most non-convergent site-rich pair
425  was formed by *Zeugites pittieri* and *Sartidia* spp. (eight sites, branches N and U in Fig. 1; Table
426  S7).

427

428  ***Molecular convergence in the RuBisCO large subunit***
429       We further inspected the evolution of the RuBisCO large subunit across the PACMAD
430  clade. A total of 4 out of 9 RbcL amino acids with convergent changes in C$_4$ reference
431  branches—V101I, A281S, M309I and A328S—have been identified in previous studies on
432  PACMAD grasses (Christin, et al. 2008; Piot, et al. 2018) as sites that experienced adaptive
433  evolution in C$_4$ species (Table 3). A further site, T143A, was found to evolve under positive
434  selection in C$_3$ to C$_4$ transitions in monocots (Studer, et al. 2014). Interestingly, an adaptive
435  S143A replacement has also been detected in the gymnosperm *Podocarpus* (Sen, et al. 2011).
436  Three more sites with convergent replacements—at positions 93, 94 and 461—correspond to
437  amino acids that were reported to evolve under positive selection in different groups of seed
438  plants by Kapralov and Filatov (2007). Thus, all of the *rbcL* codons that appear to have evolved
439  convergently among the PACMAD C$_4$ lineages we have examined are also known to have
440  experienced adaptive evolution in seed plants, but not all of them have been shown to evolve
441  adaptively in C$_4$ grasses.

442

443

444  **Discussion**

445   The recurrent emergence of carbon-concentration mechanisms (CCMs) across multiple
446   angiosperm clades in the past 35 million years represents one of the most striking examples of
447   convergent evolution of a complex phenotypic trait (Sage, et al. 2011; Heyduk, et al. 2019).
448   Several investigations have shown that the phenotypic parallelism across $C_4$ lineages is to some
449   extent mirrored by convergent changes in the sequence of proteins with key metabolic roles in
450   the biochemistry of $C_4$ photosynthesis, both in monocots and eudicots (Christin, et al. 2007;
451   Besnard, et al. 2009; Christin, Petitpierre, et al. 2009; Christin, Samaritani, et al. 2009; Kapralov,
452   et al. 2011; Goolsby, et al. 2018). Furthermore, biochemical analyses have determined that some
453   of these changes reflect adaptive shifts, as in the case of the increased availability of $CO_2$ at the
454   RuBisCO site (Studer, et al. 2014). Substantial changes in several RuBisCO kinetic traits
455   associated to $C_3$ to $C_4$ transitions have recently been described (Bouvier, et al. 2021). Further
456   evidence of changes in the selective pressure associated to the $C_3$ to $C_4$ transitions have emerged
457   from the detection of several positively selected sites in multiple genes associated with
458   photosynthetic processes (Christin, et al. 2008; Studer, et al. 2014; Goolsby, et al. 2018; Piot, et
459   al. 2018). These and other discoveries have paved the way to a more nuanced understanding of
460   the molecular basis of phenotypic convergence in CCM plants and may accelerate the
461   development of crop varieties with augmented resistance to high temperature and low water
462   availability.
463
464   For these aims to be fully realized, a robust framework to assess the extent and
465   phenotypic impact of convergent molecular changes is necessary. Along the lines of strategies
466   applied in vertebrates research (Castoe, et al. 2009; Foote, et al. 2015; Thomas and Hahn 2015;
467   Zou and Zhang 2015), we presented here the results of a novel methodological approach to the
468   study of molecular convergence in $C_4$ grasses. We investigated patterns of convergent and non-
469   convergent amino acid changes in nearly 70 chloroplast proteins across multiple $C_4$ and $C_3$
470   lineages in the PACMAD clade, with the goal of testing a specific hypothesis: is the evolution of
471   chloroplast proteins showing stronger signatures of convergent amino acid replacements in $C_4$
472   lineages compared to $C_3$ lineages? This analysis also allowed us to establish if proteins other
473   than enzymes involved in the CCM biochemistry underwent parallel amino acid changes in $C_4$
474   lineages. Our reasoning is that many proteins expressed in the chloroplast could have
475   experienced similar selective pressure across multiple $C_3$ to $C_4$ transitions and might have
476   accumulated convergence replacements as a result. In agreement with our expectation, dozens of
477   nuclear genes sharing signatures of positive or relaxed selection and likely associated with the
478   evolution of $C_4$ PACMAD grasses have been recently described, albeit these analyses relied on a
479   limited number of species (Huang, et al. 2017).
480
481   We based our analysis on the identification of amino acid replacements shared by pairs of
482   reference $C_4$ branches, defined here as branches corresponding to $C_3$ to $C_4$ transitions in the
483   PACMAD phylogeny. We compared these changes to those identified in reference $C_3$ branches,
484   namely all $C_3$ lineages that include only $C_3$ species (Figs. 1 and 2), and to changes found between

485    reference C₃ and C₄ branches. For each of the three possible pairs of photosynthesis types, C₄:C₄,
486    C₃:C₄ and C₃:C₃, we determined the number of amino acid sites, genes and pairs of reference
487    branches with convergent replacements.
488            We detected signatures of convergent evolution in all types of datasets. First, we
489    identified many individual replacements that emerged repeatedly and uniquely in C₄ reference
490    branches, particularly in the proteins RbcL, NdhH, NdhI and MatK. We also observed C₃-
491    specific convergent replacements in NdhF and RpoC2, and a case of multiple C₄ and C₃
492    convergent changes in Rps3. Additionally, we identified 7 chloroplast genes with one or more
493    C₄-specific convergent sites and 3 chloroplast genes with at least one C₃-specific convergent site.
494    Second, we found evidence of significantly higher rates of convergent replacements in C₄
495    lineages in both RbcL and RpoC1, and several convergent replacements that occurred
496    exclusively in C₄:C₄ pairs in proteins encoded by *ndhG, ndhI, psaI, rpoA, rps4* and *rps11*. These
497    genes are involved in a variety of biological processes in the chloroplast, from the cyclic electron
498    transport in (*ndhG* and *ndhI*) and the stabilization of (*psaI*) the photosystem I, to transcription
499    (*rpoA* and *rpoC1*), translation (*rps4* and *rps11*) and $CO_2$ fixation (*rbcL*). Third, we identified
500    statistically significant differences in pairs of C₄ branches with convergent replacements (Table
501    2). Crucially, we observed more pairs with higher convergent than non-convergent replacements
502    in C₄:C₄ compared to both C₃:C₃ and C₃:C₄, even after removing replacements identified in the
503    RuBisCO large subunit, RbcL.
504
505            Altogether, these findings suggest that multiple biochemical processes occurring in the
506    chloroplast might have experienced recurrent adaptive changes associated with the emergence of
507    C₄ photosynthesis. Notably, some of these proteins are not directly involved in the light-
508    dependent or light-independent reactions of the photosynthesis, implying that processes such as
509    regulation of gene expression and protein synthesis in the chloroplast are also experiencing
510    significant selective pressures during the transition from C₃ to C₄ plants. These results should
511    motivate further studies to determine the prevalence of convergent amino acid replacements in
512    transitions to CCMs among the thousands of proteins encoded by nuclear genes but expressed in
513    the chloroplast (Jarvis and Lopez-Juez 2013). Although such analyses are currently hindered by
514    the limited number of sequenced nuclear genomes in taxa with multiple C₃ and C₄ lineages,
515    including the PACMAD clade, genome-wide investigations of convergent replacements will be
516    possible in the near future given the current pace of DNA sequencing in plants.
517            A further important conclusion drawn from these results is that convergent replacements
518    are not uncommon between C₃:C₃ and C₃:C₄ lineages. This is possibly due to some
519    environmental factors affecting the evolution of chloroplast genes that are shared across grass
520    lineages regardless of their photosynthesis type.
521
522            The analysis of individual convergent replacements in the RuBisCO large subunit both
523    confirmed previous findings (Christin, et al. 2008; Studer, et al. 2014; Piot, et al. 2018) and
524    highlighted novel potentially adaptive changes among PACMAD species. Importantly, these

525 novel convergent replacements are known to evolve under positive selection in non-PACMAD
526 seed plants (Kapralov and Filatov 2007; Sen, et al. 2011). This underscores the potential of our
527 approach to identify novel changes with functional significance in the transition to CCMs in
528 grasses, as opposed to standard statistical tests of positive selection. Alternatively, some RbcL
529 sites could experience convergence across a variety of seed plants because of selective pressure
530 other than those associated with $C_3$ to $C_4$ transitions.

531

532        Overall, our results are robust to several possible confounding factors. First, we analyzed
533 branches that are strongly supported in our phylogeny reconstruction. The phylogenetic tree built
534 using the 67 chloroplast genes is well supported, with the exception of three branches with fairly
535 low bootstrap support. However, all three branches are short and have minimal impact upon our
536 conclusions regarding $C_4$ evolution (Fig. 1 and Figs. S1-S3). Moreover, the tree is largely
537 consistent with a comprehensive recent study of 250 grasses based on complete plastome data
538 (Saarela, et al. 2018). Second, by focusing only on reference branches and ignoring amino acid
539 replacements that may have occurred after the divergence of species within a given $C_4$ clade, our
540 strategy provided a conservative estimate of the number of convergent changes that could have
541 occurred during the evolution of PACMAD grasses. Third, we eliminated genes with possible
542 paralogous copies, which could have introduced false positive replacements.

543

544        We recognize some potential caveats in our approach. By relying on a relatively small
545 sample of PACMAD species, our statistical power to detect signatures of convergent evolution
546 was limited. Increasing the number of reference $C_4$ and $C_3$ lineages should provide a broader
547 representation of convergent replacements in $C_4$ clades. Furthermore, we applied a strict
548 definition of convergence that ignores changes to amino acids with similar chemical properties.
549 We think that a conservative approach was necessary given that amino acids with similar
550 chemical properties might have a very different functional effect on protein activity given their
551 size and tridimensional interactions with nearby residues. Third, we assumed that all the
552 observed convergent replacements were the result of convergent phenotypic changes, which fall
553 under the general category of homoplasy (Avise and Robinson 2008). However, some of these
554 replacements could instead represent hemiplasy, or character state changes due to introgression
555 between different $C_4$ lineages, incomplete lineage sorting (ILS) of reference alleles or horizontal
556 gene transfer (Avise and Robinson 2008). Recombination between chloroplast genomes, which
557 is required for introgression to occur, has been documented but appears to be rare (Carbonell-
558 Caballero, et al. 2015; Greiner, et al. 2015; Sancho, et al. 2018). Introgression or horizontal gene
559 transfer between congeneric species has been associated to the acquisition of part of the $C_4$
560 biochemical pathway in the PACMAD genus *Alloteropsis* (Christin, et al. 2012; Olofsson, et al.
561 2016). However, these transfers were limited to a few nuclear genes. Moreover, only a very few
562 cases of horizontal transfer between chloroplast genomes have been reported in plants
563 (Stegemann, et al. 2012). Therefore, the contribution of hemiplasy to the observed pattern of
564 convergent replacements in $C_4$ lineages is likely to be minimal. Finally, we treated $C_4$ species

565 regardless of their photosynthesis subtype (NAPD-ME, NAD-ME and PEPCK), which is known
566 to vary among PACMAD subfamilies (Taylor, et al. 2010). We argue that our results are
567 conservative with regard to this aspect because convergent replacements should be expected to
568 occur more often between $C_4$ groups sharing the same photosynthesis subtype.
569
570

571 **Conclusions**
572
573 In this study, we showed that molecular convergent evolution in the form of recurrent amino acid
574 replacements affected multiple chloroplast proteins in $C_4$ lineages of the PACMAD clade of
575 grasses. This finding significantly broadened the number of genes known to have evolved
576 convergently in $C_4$ species. We observed for the first time that genes not directly involved in
577 photosynthesis-related processes experienced convergent changes, suggesting that future efforts
578 should rely whenever possible on genome-wide analyses of amino acid changes rather than focus
579 primarily on candidate key metabolic genes, similarly to previous investigations on gene
580 expression patterns in $C_4$ and CAM plants. Our methodological approach based on the
581 comparison of convergent and non-convergent replacements among photosynthesis types
582 underscores the importance of a more rigorous hypothesis-based testing of convergent evolution
583 signatures in $C_4$ plant evolution. Our results should inform more nuanced approaches to
584 introduce CCM-like processes in $C_3$ crops.
585
586

587 **Acknowledgements**
588 We thank two reviewers for comments and suggestions that have led to a significant improved
589 version of this manuscript, and to A. Michelle Lawing for comments on the manuscript. This
590 project was supported by the National Institute of Food and Agriculture, U.S. Department of
591 Agriculture, under award number TEX0-1-9599, the Texas A&M AgriLife Research, and the
592 Texas A&M Forest Service.

593
594

595 **References**
596 Akaike H. 1973. Information theory and an extension of the maximum likelihood principle.
597 Proceedings of the Second International Symposium on Information Theory:267–281.
598 Andersson I, Backlund A. 2008. Structure and function of Rubisco. Plant Physiol Biochem
599 46:275-291.
600 Avise JC, Robinson TJ. 2008. Hemiplasy: a new term in the lexicon of phylogenetics. Syst Biol
601 57:503-507.
602 Barthet MM, Hilu KW. 2008. Evaluating evolutionary constraint on the rapidly evolving gene
603 matK using protein composition (vol 66, pg 85, 2008). Journal of Molecular Evolution 67:123-
604 123.

605 Besnard G, Christin PA, Male PJ, Lhuillier E, Lauzeral C, Coissac E, Vorontsova MS. 2014.
606 From museums to genomics: old herbarium specimens shed light on a C3 to C4 transition. J Exp
607 Bot 65:6711-6721.
608 Besnard G, Muasya AM, Russier F, Roalson EH, Salamin N, Christin PA. 2009. Phylogenomics
609 of C(4) photosynthesis in sedges (Cyperaceae): multiple appearances and genetic convergence.
610 Mol Biol Evol 26:1909-1919.
611 Boschloo RD. 1970. Raised conditional level of significance for the 2 × 2-table when testing the
612 equality of two probabilities. Stat Neerl 24:1-9.
613 Bouvier JW, Emms DM, Rhodes T, Bolton JS, Brasnett A, Eddershaw A, Nielsen JR, Unitt A,
614 Whitney SM, Kelly S. 2021. Rubisco Adaptation Is More Limited by Phylogenetic Constraint
615 Than by Catalytic Trade-off. Molecular Biology and Evolution 38:2880-2896.
616 Brown WV, Smith BN. 1972. Grass evolution, the Kranz syndrome, 13C/12C ratios, and
617 continental drift. Nature 239:345-346.
618 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
619 BLAST+: architecture and applications. BMC Bioinformatics 10:421.
620 Carbonell-Caballero J, Alonso R, Ibanez V, Terol J, Talon M, Dopazo J. 2015. A Phylogenetic
621 Analysis of 34 Chloroplast Genomes Elucidates the Relationships between Wild and Domestic
622 Species within the Genus Citrus. Molecular Biology and Evolution 32:2015-2035.
623 Castoe TA, de Koning AP, Kim HM, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL,
624 Pollock DD. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution.
625 Proc Natl Acad Sci U S A 106:8986-8991.
626 Christin PA, Arakaki M, Osborne CP, Edwards EJ. 2015. Genetic enablers underlying the
627 clustered evolutionary origins of C4 photosynthesis in angiosperms. Mol Biol Evol 32:846-858.
628 Christin PA, Boxall SF, Gregory R, Edwards EJ, Hartwell J, Osborne CP. 2013. Parallel
629 recruitment of multiple genes into c4 photosynthesis. Genome Biol Evol 5:2174-2187.
630 Christin PA, Edwards EJ, Besnard G, Boxall SF, Gregory R, Kellogg EA, Hartwell J, Osborne
631 CP. 2012. Adaptive evolution of C(4) photosynthesis through recurrent lateral gene transfer.
632 Curr Biol 22:445-449.
633 Christin PA, Osborne CP, Chatelet DS, Columbus JT, Besnard G, Hodkinson TR, Garrison LM,
634 Vorontsova MS, Edwards EJ. 2013. Anatomical enablers and the evolution of C4 photosynthesis
635 in grasses. Proc Natl Acad Sci U S A 110:1381-1386.
636 Christin PA, Petitpierre B, Salamin N, Buchi L, Besnard G. 2009. Evolution of C-4
637 Phosphoenolpyruvate Carboxykinase in Grasses, from Genotype to Phenotype. Molecular
638 Biology and Evolution 26:357-365.
639 Christin PA, Salamin N, Muasya AM, Roalson EH, Russier F, Besnard G. 2008. Evolutionary
640 switch and genetic convergence on rbcL following the evolution of C4 photosynthesis. Mol Biol
641 Evol 25:2361-2368.
642 Christin PA, Salamin N, Savolainen V, Duvall MR, Besnard G. 2007. C4 Photosynthesis
643 evolved in grasses via parallel adaptive genetic changes. Curr Biol 17:1241-1247.
644 Christin PA, Samaritani E, Petitpierre B, Salamin N, Besnard G. 2009. Evolutionary insights on
645 C4 photosynthetic subtypes in grasses from genomics and phylogenetics. Genome Biol Evol
646 1:221-230.
647 Felsenstein J. 1985. Phylogenies and the Comparative Method. American Naturalist 125:1-15.
648 Foote AD, Liu Y, Thomas GW, Vinar T, Alfoldi J, Deng J, Dugan S, van Elk CE, Hunter ME,
649 Joshi V, et al. 2015. Convergent evolution of the genomes of marine mammals. Nat Genet
650 47:272-275.

651 Goolsby EW, Moore AJ, Hancock LP, De Vos JM, Edwards EJ. 2018. Molecular evolution of
652 key metabolic genes during transitions to C4 and CAM photosynthesis. Am J Bot 105:602-613.
653 Grass Phylogeny Working Group I. 2012. New grass phylogeny resolves deep evolutionary
654 relationships and discovers C4 origins. New Phytol 193:304-312.
655 Greiner S, Sobanski J, Bock R. 2015. Why are most organelle genomes transmitted maternally?
656 Bioessays 37:80-94.
657 Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis
658 program for Windows 95/98/NT. Nucl Acids Symp Series 41:95-98.
659 Hattersley PW, Stone NE. 1986. Photosynthetic Enzyme-Activities in the C-3-C-4 Intermediate
660 Neurachne-Minor Blake,S.T. (Poaceae). Australian Journal of Plant Physiology 13:399-408.
661 Hattersley PW, Wong SC, Perry S, Roksandic Z. 1986. Comparative Ultrastructure and Gas-
662 Exchange Characteristics of the C3-C4 Intermediate Neurachne-Minor Blake,S.T. (Poaceae).
663 Plant Cell and Environment 9:217-233.
664 Heckmann D, Schulze S, Denton A, Gowik U, Westhoff P, Weber APM, Lercher MJ. 2013.
665 Predicting C-4 Photosynthesis Evolution: Modular, Individually Adaptive Steps on a Mount Fuji
666 Fitness Landscape. Cell 153:1579-1588.
667 Heyduk K, Moreno-Villena JJ, Gilman IS, Christin PA, Edwards EJ. 2019. The genetics of
668 convergent evolution: insights from plant photosynthesis. Nat Rev Genet.
669 Huang P, Studer AJ, Schnable JC, Kellogg EA, Brutnell TP. 2017. Cross species selection scans
670 identify components of C4 photosynthesis in the grasses. J Exp Bot 68:127-135.
671 Jarvis P, Lopez-Juez E. 2013. Biogenesis and homeostasis of chloroplasts and other plastids. Nat
672 Rev Mol Cell Biol 14:787-802.
673 Kapralov MV, Filatov DA. 2007. Widespread positive selection in the photosynthetic Rubisco
674 enzyme. BMC Evol Biol 7:73.
675 Kapralov MV, Kubien DS, Andersson I, Filatov DA. 2011. Changes in Rubisco kinetics during
676 the evolution of C4 photosynthesis in Flaveria (Asteraceae) are associated with positive selection
677 on genes encoding the enzyme. Mol Biol Evol 28:1491-1503.
678 Kapralov MV, Smith JA, Filatov DA. 2012. Rubisco evolution in C(4) eudicots: an analysis of
679 Amaranthaceae sensu lato. PLoS One 7:e52974.
680 Kellogg EA. 1999. Phylogenetic Aspects of the Evolution of C4 Photosynthesis. In:  Sage RF,
681 Monson RK, editors. C4 Plant Biology. San Diego, California, USA: Academic Press. p. 411-
682 444.
683 Knapp AK, Medina E. 1999. Lessons from Communities Dominated by C4 Plants. In:  Sage RF,
684 Monson RK, editors. C4 Plant Biology. San Diego, California, USA: Academic Press. p. 251-
685 283.
686 Lanfear R, Calcott B, Ho SY, Guindon S. 2012. Partitionfinder: combined selection of
687 partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol 29:1695-
688 1701.
689 Li B, Forster C, Robert CAM, Zust T, Hu L, Machado RAR, Berset JD, Handrick V, Knauer T,
690 Hensel G, et al. 2018. Convergent evolution of a metabolic switch between aphid and caterpillar
691 resistance in cereals. Science Advances 4.
692 Long SP. 1999. Environmental Responses. In:  Sage RF, K. MR, editors. C4 Plant Biology. San
693 Diego, California, USA: Academic Press. p. 215-249.
694 Lu PT, Yu S, Zhu N, Chen YR, Zhou BY, Pan Y, Tzeng D, Fabi JP, Argyris J, Garcia-Mas J, et
695 al. 2018. Genome encode analyses reveal the basis of convergent evolution of fleshy fruit
696 ripening. Nature Plants 4:784-791.

697    Maier RM, Neckermann K, Igloi GL, Kossel H. 1995. Complete Sequence of the Maize
698    Chloroplast Genome - Gene Content, Hotspots of Divergence and Fine-Tuning of Genetic
699    Information by Transcript Editing. Journal of Molecular Biology 251:614-628.
700    Maurino VG, Peterhansel C. 2010. Photorespiration: current status and approaches for metabolic
701    engineering. Curr Opin Plant Biol 13:249-256.
702    Mehrotra DV, Chan ISF, Berger RL. 2003. A cautionary note on exact unconditional inference
703    for a difference between two independent binomial proportions. Biometrics 59:441-450.
704    Natarajan C, Hoffmann FG, Weber RE, Fago A, Witt CC, Storz JF. 2016. Predictable
705    convergence in hemoglobin function has unpredictable molecular underpinnings. Science
706    354:336-339.
707    Olofsson JK, Bianconi M, Besnard G, Dunning LT, Lundgren MR, Holota H, Vorontsova MS,
708    Hidalgo O, Leitch IJ, Nosil P, et al. 2016. Genome biogeography reveals the intraspecific spread
709    of adaptive mutations for a complex trait. Mol Ecol 25:6107-6123.
710    Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ. 2013. Genome-
711    wide signatures of convergent evolution in echolocating mammals. Nature 502:228-231.
712    Patterson TB, Givnish TJ. 2002. Phylogeny, concerted convergence, and phylogenetic niche
713    conservatism in the core Liliales: insights from rbcL and ndhF sequence data. Evolution 56:233-
714    252.
715    Piot A, Hackel J, Christin PA, Besnard G. 2018. One-third of the plastid genes evolved under
716    positive selection in PACMAD grasses. Planta 247:255-266.
717    Preite V, Sailer C, Syllwasschy L, Bray S, Ahmadi H, Kramer U, Yant L. 2019. Convergent
718    evolution in Arabidopsis halleri and Arabidopsis arenosa on calamine metalliferous soils.
719    Philosophical Transactions of the Royal Society B-Biological Sciences 374.
720    FigTree. Tree Figure Drawing Tool, version 1.4.0 [Internet]. 2012. Available from:
721    http://tree.bio.ed.ac.uk/software/figtree/
722    Saarela JM, Burke SV, Wysocki WP, Barrett MD, Clark LG, Craine JM, Peterson PM, Soreng
723    RJ, Vorontsova MS, Duvall MR. 2018. A 250 plastome phylogeny of the grass family (Poaceae):
724    topological support under different data partitions. PeerJ 6:e4299.
725    Sage RF. 2004. The evolution of C4 photosynthesis. New Phytol 161:341-370.
726    Sage RF. 1999. Why C4 Photosynthesis? In: Sage RF, Monson RK, editors. C4 Plant Biology.
727    San Diego, California, USA: Academic Press. p. 3-16.
728    Sage RF, Christin PA, Edwards EJ. 2011. The C(4) plant lineages of planet Earth. J Exp Bot
729    62:3155-3169.
730    Sage RF, Sage TL, Kocacinar F. 2012. Photorespiration and the evolution of C4 photosynthesis.
731    Annu Rev Plant Biol 63:19-47.
732    Sancho R, Cantalapiedra CP, Lopez-Alvarez D, Gordon SP, Vogel JP, Catalan P, Contreras-
733    Moreira B. 2018. Comparative plastome genomics and phylogenomics of Brachypodium:
734    flowering time signatures, introgression and recombination in recently diverged ecotypes. New
735    Phytol 218:1631-1644.
736    Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. 2019. GenBank.
737    Nucleic Acids Res 47:D94-D99.
738    Sen L, Fares MA, Liang B, Gao L, Wang B, Wang T, Su YJ. 2011. Molecular evolution of rbcL
739    in three gymnosperm families: identifying adaptive and coevolutionary patterns. Biol Direct
740    6:29.
741    Smith BN, Brown WV. 1973. The Kranz syndrome in the Gramineae as indicated by carbon
742    isotopic ratios. Amer J Bot 60:505-513.

743    Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
744    large phylogenies. Bioinformatics 30:1312-1313.
745    Stegemann S, Keuthe M, Greiner S, Bock R. 2012. Horizontal transfer of chloroplast genomes
746    between plant species. Proceedings of the National Academy of Sciences of the United States of
747    America 109:2434-2438.
748    Storz JF. 2016. Causes of molecular convergence and parallelism in protein evolution. Nature
749    Reviews Genetics 17:239-250.
750    Studer RA, Christin PA, Williams MA, Orengo CA. 2014. Stability-activity tradeoffs constrain
751    the adaptive evolution of RubisCO. Proc Natl Acad Sci U S A 111:2223-2228.
752    Taylor SH, Hulme SP, Rees M, Ripley BS, Woodward FI, Osborne CP. 2010. Ecophysiological
753    traits in C3 and C4 grasses: a phylogenetically controlled screening experiment. New Phytol
754    185:780-791.
755    Thomas GW, Hahn MW. 2015. Determining the Null Model for Detecting Adaptive
756    Convergence from Genomic Data: A Case Study using Echolocating Mammals. Mol Biol Evol
757    32:1232-1236.
758    Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E,
759    Peterson P, Weckesser W, Bright J, et al. 2020. SciPy 1.0: fundamental algorithms for scientific
760    computing in Python. Nat Methods 17:261-272.
761    Walker JF, Walker-Hale N, Vargas OM, Larson DA, Stull GW. 2019. Characterizing gene tree
762    conflict in plastome-inferred phylogenies. PeerJ 7:e7747.
763    Williams BP, Johnston IG, Covshoff S, Hibberd JM. 2013. Phenotypic landscape inference
764    reveals multiple evolutionary paths to C4 photosynthesis. Elife 2:e00961.
765    Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-
766    1591.
767    Zhang J, Kumar S. 1997. Detection of convergent and parallel evolution at the amino acid
768    sequence level. Mol Biol Evol 14:527-536.
769    Zhou XM, Seim I, Gladyshev VN. 2015. Convergent evolution of marine mammals is associated
770    with distinct substitutions in common genes. Scientific Reports 5.
771    Zou Z, Zhang J. 2015. No genome-wide protein sequence convergence for echolocation. Mol
772    Biol Evol 32:1237-1241.
773

# Figure 1

Phylogenetic relationships among 64 $C_4$ and $C_3$ grass species.

The phylogeny tree was obtained using RAxML (GTR+$\Gamma$ model) based on the third codon position sites in 67 chloroplast genes. The partitioning scheme was selected according to Akaike information criterion (AIC). $C_4$ and $C_3$ reference branches are shown in red and black, respectively. Red stars and black circles (labels A-V) indicate $C_4$ and $C_3$ reference branches, respectively. Numbers represent bootstrap support.

*Iseilema macratherum*
*Sorghum bicolor*
*Imperata cylindrica*
*Pogonatherum paniceum*
*Sorghastrum nutans*
*Eulalia aurea*
*Coix lacryma-jobi*
*Ischaemum afrum*
*Zea mays*
*Arundinella deppeana*
*Paspalum dilatatum*
*Paspalum fimbriatum*
*Axonopus fissifolius*
*Plagiantha tenella*
*Otachyrium versicolor*
*Steinchisma laxum*
*Oncorachis ramosa*
*Coleataenia prionitis*
*Cenchrus americanus*
*Setaria italica*
*Megathyrsus maximus*
*Panicum virgatum*
*Panicum capillare*
*Panicum miliaceum*
*Panicum sumatrense*
*Whiteochloa capillipes*
*Dichanthelium acuminatum*
*Thyridolepis xerophila*
*Echinochloa crus-galli*
*Echinochloa frumentacea*
*Echinochloa colona*
*Alloteropsis angusta*
*Alloteropsis cimicina*
*Alloteropsis paniculata*
*Amphicarpum muhlenbergianum*
*Oplismenus hirtellus*
*Zeugites pittieri*
*Centotheca lappacea*
*Thysanolaena latifolia*
*Danthoniopsis dinteri*
*Monachather paradoxus*
*Elytrophorus spicatus*
*Phragmites australis*
*Hakonechloa macra*
*Molinia caerulea*
*Danthonia californica*
*Chaetobromus involucratus*
*Centropodia glauca*
*Neyraudia reynaudiana*
*Eragrostis tef*
*Eragrostis minor*
*Bouteloua gracilis*
*Bouteloua curtipendula*
*Distichlis spicata*
*Hilaria cenchroides*
*Eleusine indica*
*Chloris barbata*
*Chloris virgata*
*Chloris truncata*
*Sartidia dewinteri*
*Sartidia perrieri*
*Aristida purpurea*
*Oryza sativa*
*Brachypodium distachyon*

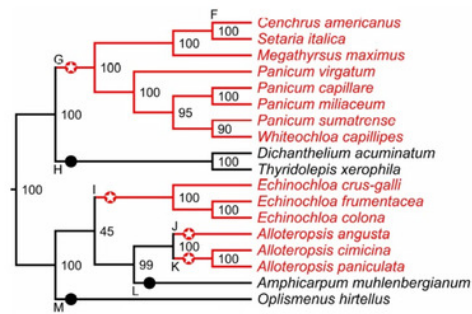
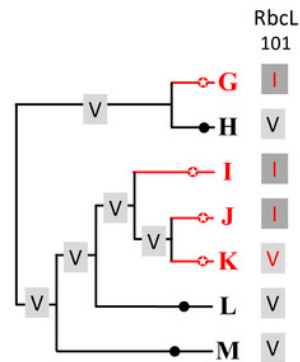
# Figure 2

Example of $C_4$ and $C_3$ reference branches and convergent changes in $C_4$ reference branches.

(A) PACMAD phylogeny and identification of reference branches. The $C_4$ reference branches (highlighted by red circles with stars) contain the common ancestor of a clade with only $C_4$ species (red lines). The $C_3$ reference branches (highlighted by black circles) contain the common ancestor of a clade with only $C_4$ species (black lines). $C_4$ reference branches that are next to each other represent lineages that independently acquired the $C_4$ pathway and are separated by species with the $C_3$ pathway that were no included in this study because of the lack of complete chloroplast genomes. For each species, the $C_4$ or $C_3$ photosynthesis type was obtained from the Supplementary figure 1 in the Grass Phylogeny Working Group II (2012).

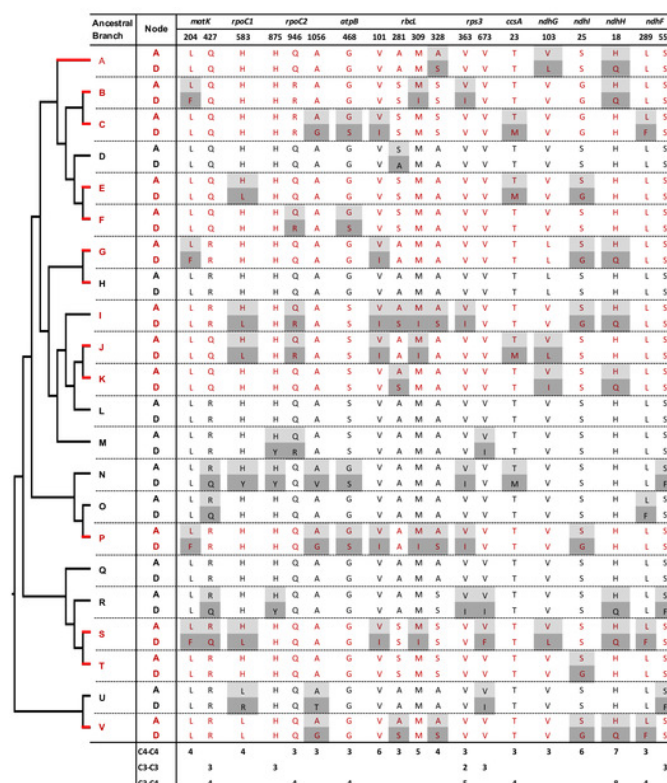
(B) Amino acid replacements in the reference branches. The sequence of chloroplast proteins was inferred in each reference branch and compared to the inferred sequence in the branch ancestral to the reference branch. In this example, the amino acid 101 in the protein RbcL is represented by a Valine (V) in branches ancestral to all reference branches, but a convergent V->I amino acid replacement occurred along the $C_4$ reference branches G, I and J.

(A)

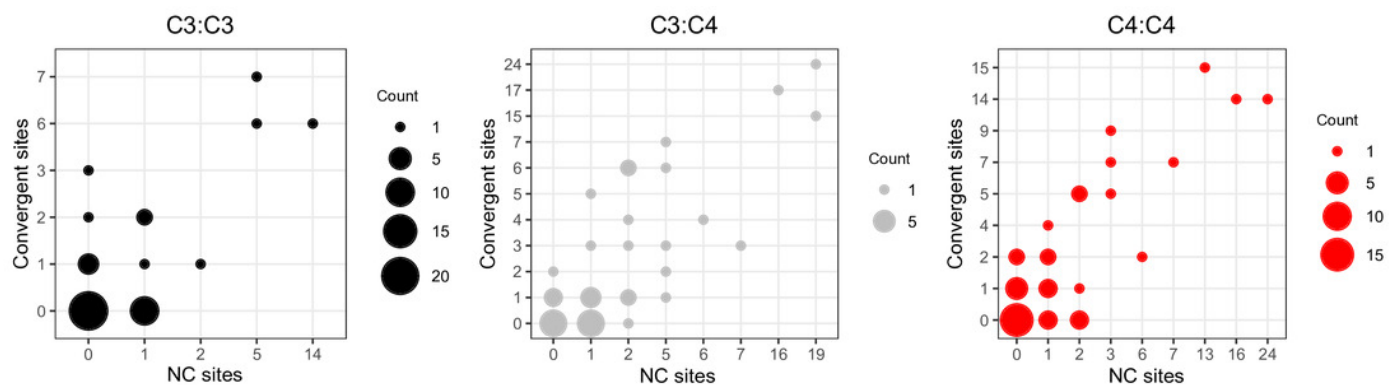
(B)

Phylogenetic reconstruction

Replacements in ancestral branches

# Figure 3

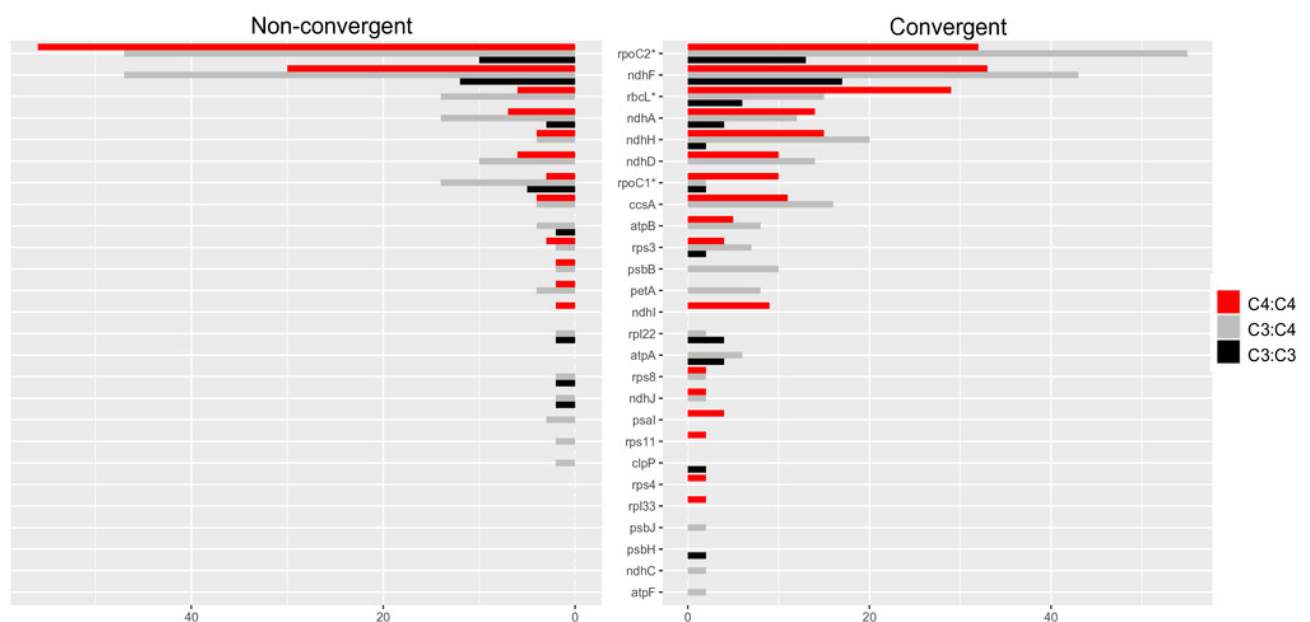
Amino acid replacements shared by at least three C$_4$ or C$_3$ reference branches.

Ancestral (A) and derived (D) amino acids at replacement sites are shown. Site numbers correspond to the *Zea mays* orthologous sequence annotation. Red and black letters and branches represent C$_4$ and C$_3$ reference branches, respectively (see also Figs. 1 and 2).

# Figure 4

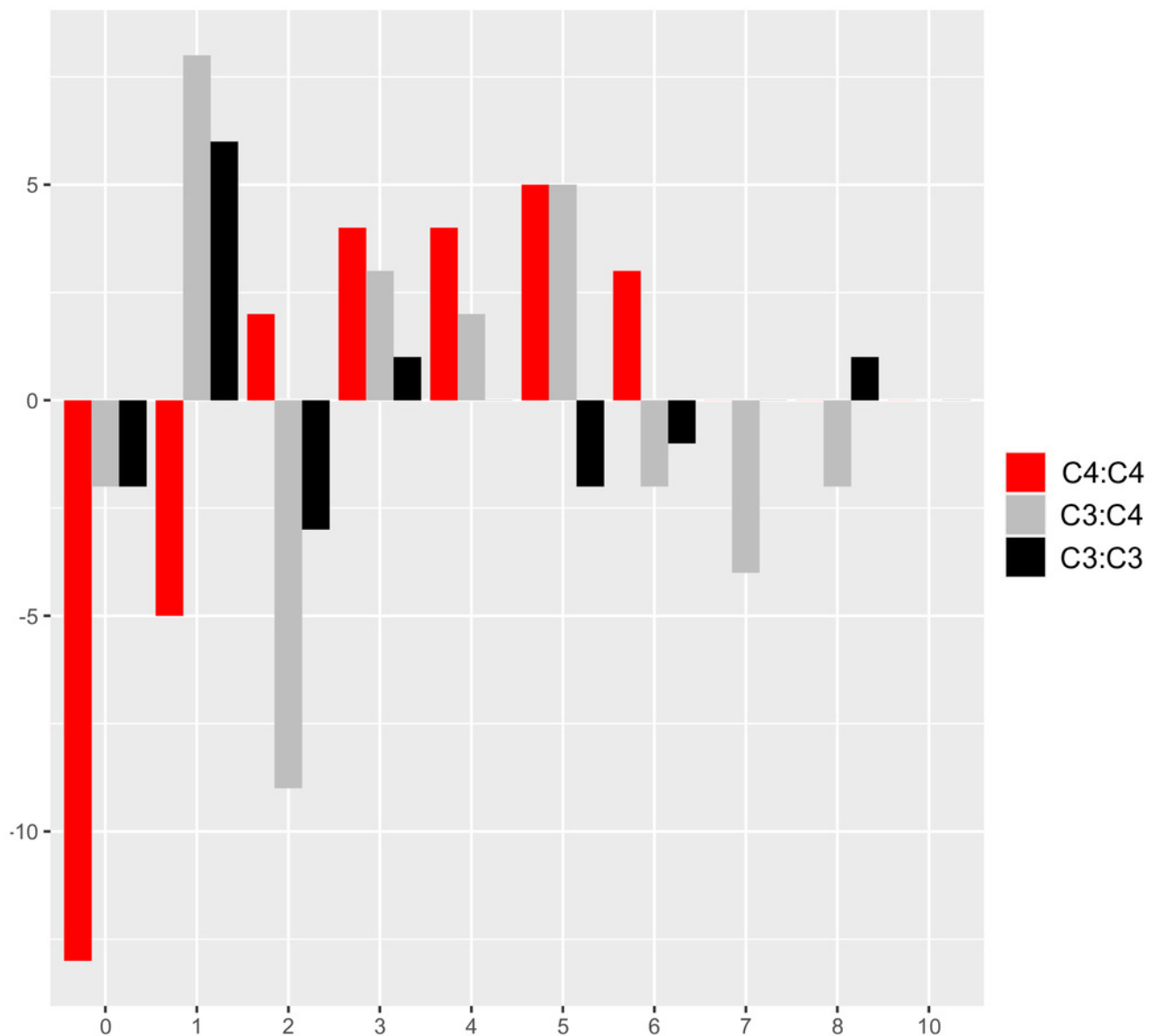
Distribution of convergent and non-convergent amino acid replacements in pairs of reference branches.

(A) $C_4$:$C_4$ pairs. (B) $C_3$:$C_3$ pairs. (C) $C_3$:$C_4$ pairs. NC: non-convergent.

# Figure 5

Amino acid replacements in chloroplast proteins with more convergent than non-convergent changes.

Twenty-six chloroplast proteins with more convergent than non- convergent changes in $C_4$:$C_4$, $C_3$:$C_4$ and $C_3$:$C_3$ pairs.

# Figure 6

Pairs of reference branches by convergent and non-convergent replacements.

Difference in the number of pairs of reference branches for convergent and non-convergent categories (0-8 and 10 replacements).

Numbers of amino acid sites and genes with convergent and non-convergent replacements in reference branch comparisons.

Comparisons were made between pairs of $C_4$:$C_4$, $C_3$:$C_3$ and $C_3$:$C_4$ branches. Numbers of replacements unique to a given category (*), and the corresponding ratios *Con*:*NC* (*Ratio*). Differences between the $C_3$:$C_3$ and $C_4$:$C_4$ categories are not statistically significant ($P \geq 0.05$, Boschloo's test). Con: convergent. NC: non-convergent.

| | C$_4$:C$_4$ | | | C$_3$:C$_4$ | | | C$_3$:C$_3$ | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Con** | **NC** | **Ratio** | **Con** | **NC** | **Ratio** | **Con** | **NC** | **Ratio** |
| Sites | 104 | 96 | 1.08 | 120 | 121 | 0.99 | 34 | 39 | 0.87 |
| Sites* | 80 | 64 | 1.25 | 82 | 69 | 1.19 | 17 | 16 | 1.06 |
| Genes | 24 | 23 | 1.04 | 26 | 32 | 0.81 | 13 | 17 | 0.76 |
| Genes* | 24 | 20 | 1.2 | 25 | 29 | 0.86 | 9 | 10 | 0.9 |

Number of reference branches with convergent and non-convergent replacements.

Comparisons were made between pairs of $C_4$:$C_4$, $C_3$:$C_3$ and $C_3$:$C_4$ branches. Proportions of pairs of reference branches over all branches by category are shown in parenthesis. The total number of pairs of reference branches are 78, 36 and 117 for $C_4$:$C_4$, $C_3$:$C_3$ and $C_3$:$C_4$ comparisons, respectively. All comparisons between $C_4$:$C_4$ pairs and both $C_3$:$C_3$ and $C_3$:$C_4$ pairs were statistically significantly different ($P < 0.05$, *Boschloo's test*). No comparison between $C_3$:$C_3$ and $C_3$:$C_4$ pairs was statistically significant ($P \geq 0.05$, *Boschloo's test*). Con: convergent. NC: non-convergent. Con>NC: pairs of branches with more convergent than non-convergent replacements. Con>1: pairs of branches with more than one convergent replacement.

|  | $C_4{:}C_4$ | $C_3{:}C_4$ | $C_3{:}C_3$ |
|---|---|---|---|
| No replacements | 6 (.08) | 30 (.26) | 12 (.33) |
| No Con | 12 (.15) | 48 (.41) | 16 (.44) |
| w/Con | 66 (.85) | 69 (.59) | 20 (.56) |
| w/NC | 63 (.81) | 67 (.57) | 18 (.50) |
| Con>NC | 40 (.51) | 36 (.31) | 10 (.28) |
| Con>1 | 49 (.63) | 39 (.33) | 8 (.22) |

Table 3. Summary of RbcL amino acid sites with signatures of convergent evolution or positive selection.

Ancestral AA: ancestral amino acid. Convergent change/p.s.s.: derived amino acid in multiple $C_4$ reference branches and positively selected sites from previous studies. #Convergent a.b.: number of reference branches with convergent changes. Boldface: sites with convergent changes identified in this study. Asterisk: positively selected sites in PACMAD $C_4$ lineages from Christin et al. (2008). Dagger: positively selected sites in PACMAD $C_4$ lineages from Piot et al. (2018).

| Codon | Ancestral AA | Convergent Change/p.s.s. | #Convergent a.b. |
|---|---|---|---|
| **10** | **S** | **G** | **2** |
| **93** | **E** | **D** | **2** |
| **94** | **A** | **P** | **2** |
| **101**[*†] | **V** | **I** | **6** |
| 142[*†] | P | Several | na |
| **143** | **T** | **A** | **3** |
| 145[*†] | S | A/V | na |
| 258[*] | R | K | na |
| 270[*] | L | I | na |
| **281**[*†] | **A** | **S** | **3** |
| 282[†] | H | Several | na |
| **309**[*†] | **M** | **I** | **5** |
| **328**[*†] | **A** | **S** | **4** |
| **461**[*] | **V** | **I** | **2** |
| 468[†] | E | D | na |
| 471[†] | E | Several | na |
| 476[†] | I | L/V | na |

2
3
4