bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Beyond RuBisCO: Convergent molecular evolution of multiple chloroplast genes in C_4 plants

Claudio Casola[1,2], Jingjia Li[1]

[1] Department of Ecology and Conservation Biology, Texas A&M University, College Station, TX, USA

[2] Interdisciplinary Graduate Program in Ecology and Evolutionary Biology, Texas A&M University, College Station, TX, USA

Corresponding author:

claudio.casola@ag.tamu.edu

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

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 introduced a robust empirical approach to test molecular convergence along a phylogeny including multiple C_3 and C_4 taxa. By analyzing proteins encoded by chloroplast genes, we tested if convergent replacements occurred more frequently than expected in C_4 lineages compared to C_3 lineages. Furthermore, we sought 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 ancestral branches, corresponding to the most recent common ancestor of each lineage. Pairwise comparisons between ancestral branches allowed us to identify both convergent and divergent 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 divergent replacements in 45/67 chloroplast proteins in both C_4 and C_3 ancestral branches. Pairs of C_4-C_4 ancestral branches showed higher levels of convergent replacements than C_3-C_3 and C_3-C_4 pairs. 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/divergent replacements ratio in C_4 branches. Notably, significantly more C_4-C_4 pairs of ancestral branches showed higher numbers of convergent vs. divergent replacements than C_3-C_3 and C_3-C_4 pairs. Our results demonstrated 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 both our understanding of the evolution of photosynthesis and the goal of engineering improved crop varieties that integrates components of the C_4 pathway.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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).

Several traits are also known to have convergently evolved in land plants (e.g., Li et al., 2018; Lü 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 originated 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 II, 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 et al., 2013a; 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 that often result in the convergent emergence of the same amino acid replacements across C_4 lineages (Christin et al., 2013b; Sage et al., 2012; Williams et al., 2013).

Evidence of convergent changes in proteins associated with photosynthetic processes has steadily accumulated since genomic data from multiple C_4 lineages have become available in the past couple of decades. Most of these studies have focused on the ribulose-1,5-bisphosphate

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

carboxylase/oxygenase (RuBisCO), a large multimeric enzyme that catalyzes the carboxylation of ribulose-1,5-bisphosphate (RuBP), allowing plants to fix atmospheric carbon (Andersson and Backlund, 2008). RuBisCO also initiates oxygenation of RuBP, which leads to a more limited production of energy and to loss of carbon in the process of photorespiration (Andersson and Backlund, 2008; Maurino and Peterhansel, 2010). RuBisCO's limited ability to discriminate between CO_2 and O_2 has been attributed to the much higher CO_2 to O_2 atmospheric partial pressure until ~400 million years ago (Sage, 1999, 2004; Sage et al., 2012).

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

Given the number of biochemical, physiological and anatomical traits that were affected in each evolutionary transition from C_3 to C_4 photosynthesis (Heyduk et al. 2019), it is likely that many genes experienced analogous selective pressures across taxa that include C_4 plants. This could have led to the widespread occurrence of convergent amino acid replacements among a significant fraction of proteins encoded by genes involved in photosynthesis processes. A recent, important work has produced the first analysis of convergent replacements across multiple proteins involved in the metabolism of C_4 and crassulacean acid metabolism (CAM) among species belonging to the portullugo clade (Caryophyllales). Goolsby and colleagues (2018) compared evolutionary patterns in 19 gene families with critical roles in metabolic pathways of both C_4 and CAM plants, also known as carbon-concentration mechanisms (CCMs) genes, and in 64 non-CCM gene families. They found convergent replacements in proteins from C_4 and CAM lineages, as well as higher levels of convergent replacements in CCM vs. non-CCM gene families (Goolsby et al., 2018). Additionally, several amino acid replacements that are prevalent among C_4 and CAM taxa compared to C_3 lineages were identified in this study (Goolsby et al., 2018).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Altogether, the results of this and other studies demonstrated that convergent molecular evolution occurred across multiple genes in both C_4 and CAM groups. However, a rigorous framework to assess the full extent of molecular convergence in C_3 to CCMs transitions has yet to be presented. For example, analyses of convergent evolution should include null hypotheses that assume no differences between taxa with and without convergence. In the case of CCMs evolution, a plausible null hypothesis consists in statistically equivalent numbers of convergent replacements between C_4 (or CAM) lineages and C_3 lineages.

Additionally, nonadaptive replacements should be used to normalize convergent replacements, in order to account for variation in the rates of nonsynonymous substitutions across lineages. This approach has been successfully applied in studies of molecular convergent evolution in vertebrates by assessing both convergent replacements and protein sequence changes that result in different amino acids, or *divergent replacements* (Castoe et al., 2009; Thomas and Hahn, 2015; Zou and Zhang, 2015). Furthermore, testing hypotheses about the extent of convergent molecular evolution remains particularly challenging for many nuclear genes, because of the prevalence of duplicated copies, particularly in plants (Christin et al., 2007; Goolsby et al., 2018). Single-copy nuclear or organelle genes allow to more easily recognize convergent changes and overcome possible confounding compensatory effects due to the presence of paralogous copies.

Given these premises, we sought to test if convergent amino acid changes occur more frequently in proteins encoded by chloroplast genes in a taxon that includes multiple well-characterized lineages of C_4 and C_3 grasses. Chloroplast proteins represent an ideal set of targets to study the role of convergent evolution in C_3 to C_4 transitions for a variety of reasons. First, most chloroplast proteins are involved in biochemical and biophysical processes that are critical to photosynthesis. For instance, out of ~75 functionally annotated protein-coding genes in the maize chloroplast genome, 45 genes are implicated in photosynthesis-related processes, including *rbcL*, 17 genes coding for subunits of the photosystems I and II (PS I and PS II), 12 genes coding for subunits of the NADH dehydrogenase complex, 6 genes coding for chloroplast ATPase subunits, 4 genes coding for cytochrome b6f complex subunits, and a few more genes implicated in the assembly of other protein complexes (Maier et al., 1995). Second, nonannotated orthologous copies of chloroplast genes can be readily identified across plants through sequence homology searches, taking advantage of the thousands of complete chloroplast genome sequences currently available for green plants. Third, comparative studies of convergent evolution in C_4 photosynthesis are facilitated by detailed reconstruction of phylogenetic relationships within groups with both C_4 and C_3 lineages. Fourth, signatures of positive selection have been found in multiple chloroplast genes in taxa that contain both C_3 and C_4 plants, although only the genes *rbcL* and *psaJ*, which encodes a small subunit of the Photosystem I complex, showed evidence of adaptive changes exclusively in C_4 lineages (Christin et al., 2008;

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Goolsby et al., 2018; Piot et al., 2018). Finally, most chloroplast genes occur as single copy loci, as opposed to the multiple paralogs typically present for plant genes encoded in the nucleus.

In this study, we analyzed 67 chloroplast genes from 64 grass species, including 43 C_4 and 19 C_3 species belonging to the PACMAD clade, named after six of its most representative subfamilies: Panicoideae, Arundinoideae, Chloridoideae, Micrairoideae, Aristidoideae and Danthonioideae. Using published phylogenetic information, we identified thirteen independent C_3 to C_4 transitions in this group of species. We applied a series of tests based on convergent vs. divergent amino acid replacements and determined that convergent molecular evolution occurred at a higher rate in chloroplast genes of C_4 lineages compared to C_3 lineages, a pattern that remained largely unchanged after excluding the RbcL protein from the convergence analyses. Our findings suggest that the evolutionary trajectories of multiple chloroplast genes have been remarkably affected during the emergence of the C_4 adaptation in the PACMAD clade, a result that has significant implications for our understanding of C_4 photosynthesis evolution and organelle-nucleus interactions, and for the identification of molecular changes that might be critical to the successful development of engineered C_3 crops that incorporate carbon-concentration mechanisms.

Methods

Data source and filtering

We queried NCBI GenBank (Sayers et al., 2019) for complete chloroplast genome sequences of grass species that were included in phylogenetic analyses by the Grass Phylogeny Working Group II (2012) and downloaded the corresponding coding sequences. Each species was assigned to either C_3 or C_4 type following the results of the Grass Phylogeny Working Group II (2012). Additionally, we downloaded the coding chloroplast sequences for *Dichanthelium acuminatum*, *Thyridolepis xerophila*, *Sartidia dewinteri* and *Sartidia perrieri* (C_3 species) (Brown and Smith, 1972; Smith and Brown, 1973; Hattersley and Stone, 1986; Hattersley et al., 1986; Besnard et al., 2014). We used the standalone blastn ver. 2.2.29+(Camacho et al., 2009) with the Expect value (E) cutoff of 1e-10 to determine putative sequence orthology with coding sequences of the *Zea mays* chloroplast genes (Maier et al., 1995). Single copy putative orthologs that were present in more than 95% of the species were retained for further analysis (Table S1).

Multiple sequence alignment

We aligned the individual sequences using TranslatorX ver. 1.1 (Abascal et al., 2010), and further adjusted the alignments manually using BioEdit ver. 7.0.9.0 (Hall, 1999). Stop codons and sites that could not be aligned unambiguously were removed.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Phylogeny reconstruction

We concatenated the individual sequence alignments and extracted third codon position sites for phylogeny reconstruction. We ran PartitionFinder ver. 1.1.1 (Lanfear et al., 2012) to identify the best partitioning scheme (partitioning by gene) for the downstream analysis using both Akaike information criterion (AIC) (Akaike, 1973) and Bayesian information criterion (BIC) (Schwarz, 1978). We then used maximum likelihood framework as implemented in RAxML ver. 8.2.10 (Stamatakis, 2014) to reconstruct the phylogeny. Branch support was estimated using 1,000 bootstrap replicates. *Oryza sativa* and *Brachypodium distachyon* from the BOP (Bambusoideae, Oryzoideae and Pooideae) clade were used as outgroup, whereas all ingroup species belonged to the PACMAD clade. We used FigTree ver. 1.4.0 (Rambaut, 2012) to rearrange and visualize the phylogeny, and the figures were edited further to improve readability and to indicate C_4/C_3 classification.

Ancestral state reconstruction

We reconstructed ancestral states at each phylogenetic node for each individual gene using the program codeml from the software package PAML ver. 4.9a (Yang, 2007) and the basic codon substitution model (model = 0, NSsites = 0).

Inference of convergent and divergent replacements

We extracted the reconstructed ancestral states from the codeml output. The corresponding amino acid sequences were then compared to investigate individual site changes along selected branches in the reconstructed phylogenetic tree in the context of emergence of the C_4 trait. For each group of species descendant from a single C_4 ancestor, we chose the branch between the most recent C_3 ancestor and the most ancestral C_4 node, i.e., the branch along which the C_4 adaptation presumably emerged (referred to as "C_4 ancestral branch" throughout this article, see Figs. 1 and 2). For C_3 species, we chose the most ancestral branch that did not share ancestry with any C_4 lineage ("C_3 ancestral branch", see Figs. 1 and 2). In either case, if only a single species was available in a given lineage, that terminal branch was used. The outgroup species (*O. sativa* and *B. distachyon*) were not included in this analysis (Fig. 1).

We searched for amino acid changes that occurred along pairs of ancestral branches. Replacements in both branches that resulted in the same state at a given site in the two descendants were considered convergent, regardless of whether the corresponding ancestral states of ancestral were the same or different (Castoe et al., 2009). Likewise, two replacements were considered divergent if states at the descendant orthologous sites were different, regardless of the corresponding ancestral states (Castoe et al., 2009). Although two orthologous sites, by definition, descend from one ancestral site, the actual state transitions, as well as their number, between ancestral and descendant states along a given branch are not known because the states are reconstructed only at discrete time steps (i.e., at selected nodes) and represent only those

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

specific evolutionary time stamps. Therefore, potential intermediate stages, including a transient convergent phase, would remain undetected.

We identified putative convergent and divergent amino acid changes in each gene product individually. We summarized those data within each of the three categories: (1) two C_4 ancestral branches (C_4-C_4), (2) C_3 ancestral branch and C_4 ancestral branch (C_3-C_4), and (3) two C_3 ancestral branches (C_3-C_3). the Boschloo's statistical exact unconditional test (Boschloo, 1970) was performed was performed to test the significance of the convergent replacement excess when comparing two of the three photosynthesis type pairs using the SciPy library ver. 1.7.1 in python3 (Virtanen et al. 2020).

Data availability

Raw data, including alignments, fasta sequences, and phylogenetic analyses data, are available through the following Figshare repository: https://figshare.com/articles/dataset/Convergence-chlorplast-genes-C4-Casola-Li-2021/15180690.

Results

Phylogeny reconstructions

We examined 63 grass chloroplast genomes to identify gene orthologs for *Zea mays* chloroplast genes and extracted the corresponding coding and protein sequences. The resulting dataset included up to 67 DNA/protein sequences in 64 grass species that were retained for further analysis (Table S1). One to four sequences were absent in thirteen species. Out of 64 species, 43 were classified as C_4 and 21 (including two outroup species) as C_3. The reconstructed phylogeny is well supported, except for three branches with low to moderate bootstrap values, and it is consistent for both AIC and BIC (Fig. 1 and Figs. S1-S3). We identified thirteen C_4 ancestral branches that represent putative C_3 to C_4 transitions, and nine C_3 ancestral branches (Fig. 1).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

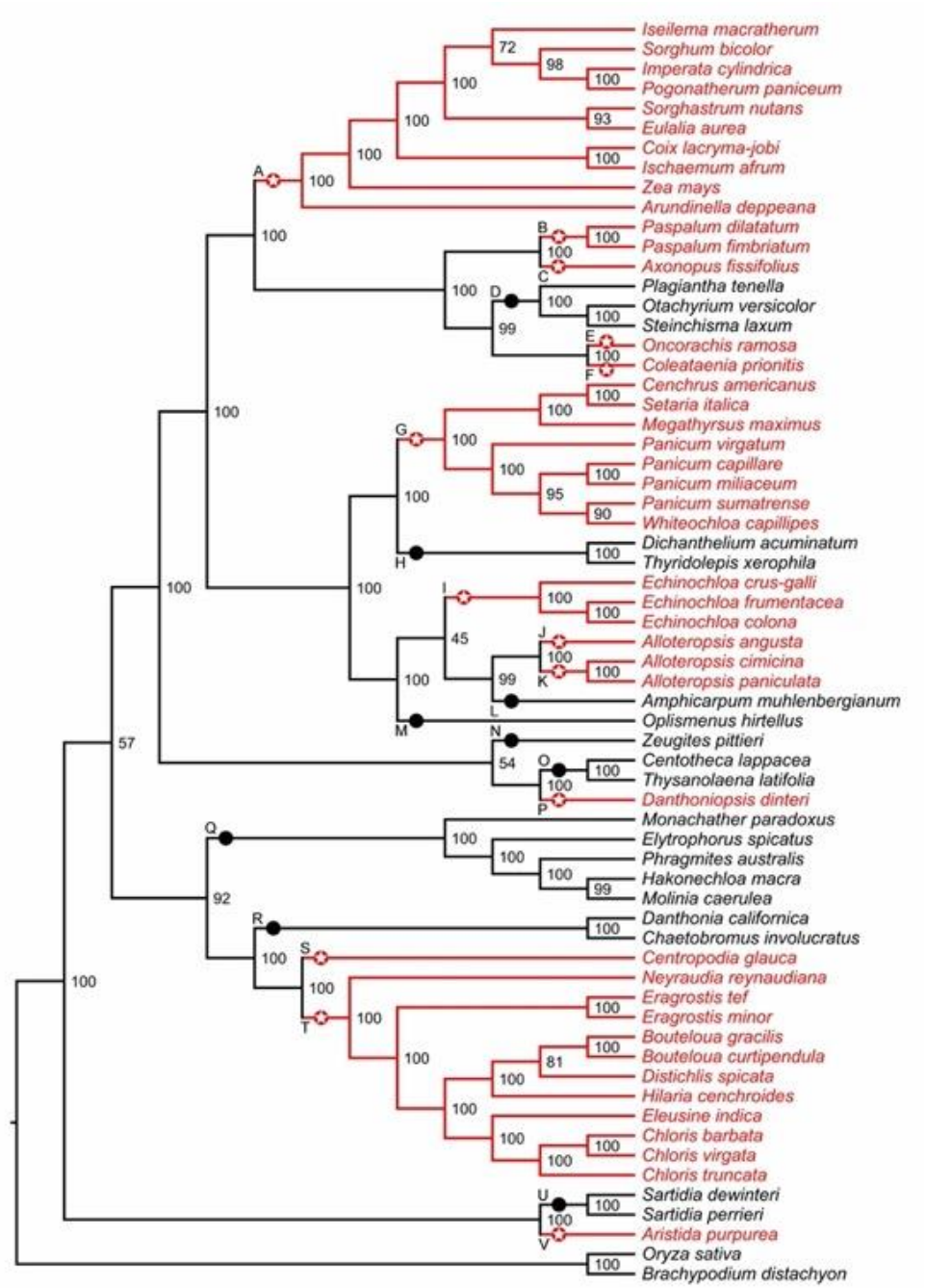

Figure 1. Phylogenetic relationships among 64 C_4 and C_3 grass species.

The phylogeny tree was obtained using RAxML (GTR+Γ 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 ancestral branches are shown in red and black, respectively. Red stars and black circles (labels A-V) indicate C_4 and C_3 ancestral branches, respectively. Numbers represent bootstrap support.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Convergent and divergent amino acid replacements across chloroplast proteins

We assessed the level of molecular convergence in C_3 to C_4 transitions by quantifying convergent and divergent amino acid replacements across the PACMAD phylogeny by performing pairwise comparisons of reconstructed ancestral sequences (Figs. 2 and 3, Table S2; see Methods).

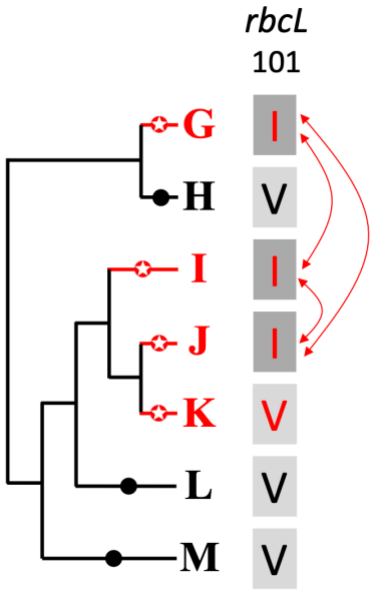

Figure 2. Example of ancestral C_4 and C_3 ancestral branches and convergent changes in C_4 ancestral branches.
Pairwise comparisons were carried out between the ancestral branches (in this example, G to M; see also Fig. 1). The C_4 branches G, I and J shared a convergent V→I amino acid replacement (red arrows) at position 101 of the large RuBisCO subunit encoded by the gene *rbcL*. The ancestral amino acid and the convergently changed amino acidic are highlighted in light and dark gray, respectively. C_4 and C_3 ancestral branches are shown in red and black, respectively.

A total of 217 sites showed at least one convergent replacement: 104 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 divergent replacements: 96 in C_4-C_4, 121 in C_3-C_4, and 39 in C_3-C_3 pairs (Table 1). The difference in convergent-divergent site distributions between the three photosynthesis types was not statistically significant ($P \geq 0.05$, Boschloo's test; Table 1).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1. Numbers of amino acid sites and genes with convergent and divergent replacements in ancestral branch comparisons.

	C_4-C_4			C_3-C_4			C_3-C_3		
	Con	**Div**	**Ratio**	**Con**	**Div**	**Ratio**	**Con**	**Div**	**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

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*:*Div* (*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. Div: divergent.

Among the C_4 ancestral branches, several individual sites showed high contrast in the number of branches involved in convergent and divergent replacements (Fig. 3, Tables S2 and S3). For example, seven C_4 branches (54%) shared the H18Q replacement in the product of *ndhH*, with no divergent replacements. Six, five, and four C_4 branches (46%, 38%, and 31%) Showed convergent replacements at three sites in the RbcL protein (V101I, M309I, and A328S, respectively). Furthermore, six C_4 branches shared the S25G replacement in the product of *ndhI* and four L204F changes in the protein encoded by *matK*. In all these cases, there were no other convergent or divergent replacements in C_3-C_3 or C_3-C_4 branch comparisons, except for one H18Q change in NdhH in a C_3-C_3 branch. Two sites with convergent replacements in the proteins encoded by *ndhF* (L557F) and *rpoC2* (H875Y) were found uniquely in C_3-C_3 pairs, and only one site in the protein Rps3 showed convergence independently in C_4-C_4 and C_3-C_3 pairs (Fig. 3).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

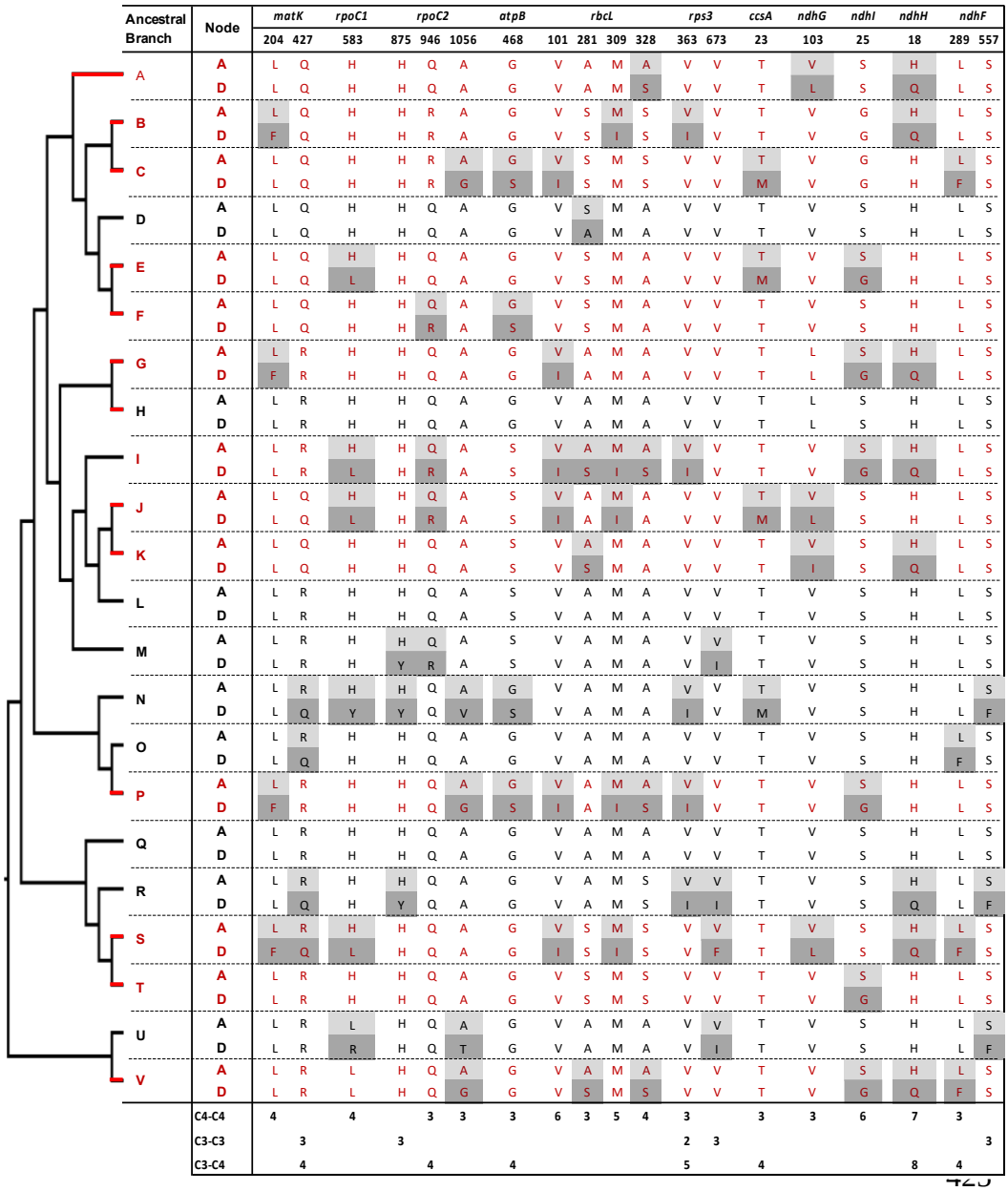

Figure 3. Amino acid replacements shared by at least three C_4 or C_3 reference branches.
Ancestral (A) and derived (D) amino acids at replacement sites are shown. Site numbers correspond to the *Zea mays* orthologous sequence annotation. Red and black letters and branches represent C_4 and C_3 ancestral branches, respectively (see also Figs. 1 and 2).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We then searched for convergent replacements that occurred along more than two C_4 branches at sites that remained otherwise conserved in C_3 and C_4 lineages, arguing that such changes could result from selective pressure rather than drift. We identified fourteen C_4-specific convergent sites in proteins from 8 genes: *atpB*, *ccsA*, *matK*, *ndhF*, *ndhH*, *ndhI*, *rbcL* and *rpoC2* (Table S3). Five of these sites were found in RbcL, whereas two sites were identified in both NdhF and NdhI.

Molecular convergence in individual chloroplast proteins

Convergent and divergent amino acid replacements were detected in the products of 45 chloroplast genes, thirteen of which had at least one site with four or more replacements (Fig. 4, Table 1 and Table S2). Twenty-four genes had convergent changes in C_4-C_4, 26 in C_3-C_4, and 13 in C_3-C_3 types of pairs (Table 1). Although the convergent/divergent replacement 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 photosynthesis types was not statistically significant ($P \geq 0.05$, Boschloo's test; Table 1). The lack of replacements was the single most common state for chloroplast proteins across photosynthesis types; however, in C_4-C_4 there were more genes with a higher number convergent vs. divergent replacements (Fig. 4 and Table S4).

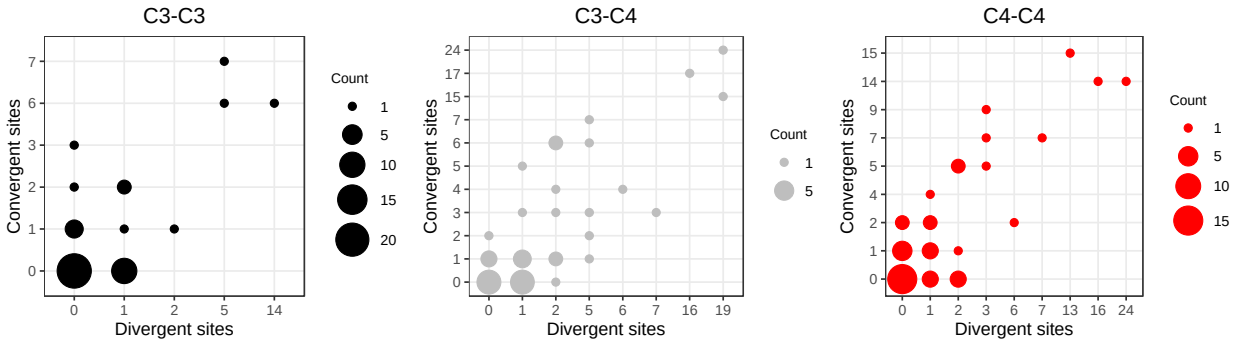

Figure 4. Distribution of convergent and divergent amino acid replacements in pairs of ancestral branches.

Overall, 26 proteins showed a higher number of convergent vs. replacement sites, of 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 S4). We found statistically significant differences in the number of replacements between C_4-C_4 and C_3-C_4 pairs, but not C_3-C_3 pairs, in the products of the genes *rbcL*, *rpoC1* and *rpoC2* ($P < 0.05$, Boschloo's test; Table S4). In RbcL and RpoC1, C_4-C_4 pairs shared much higher number of convergent replacements, whereas the opposite was true in RpoC2. RpoC1 was also the only protein showing more convergent than divergent replacements in C_4-C_4 pairs compared to C_3-C_3 and C_3-C_4 pairs. In C_4-C_4 pairs, RpoC1 shared 4 convergent and 1 divergent replacement,

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

compared to 1 and 2 in C_3-C_3 pairs and 1 and 5 in C_3-C_4 pairs, respectively. Additionally, the proteins NdhG, NdhI, PsaI, RpoA, Rps4 and Rps11 exhibited convergent replacements only in C_4-C_4 pairs (Table S4). When considering the number of affected sites rather than the number of replacements, no genes showed a significantly different pattern between photosynthesis types ($P \geq 0.05$, Boschloo's test; Table S4).

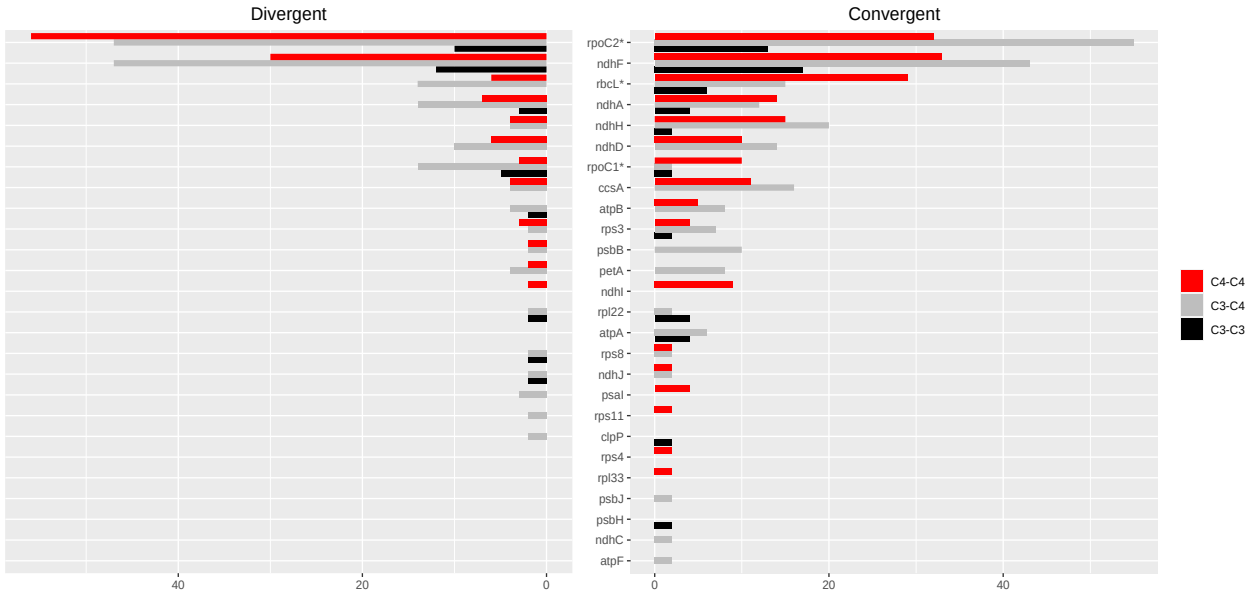

Figure 5. Amino acid replacements in chloroplast proteins with more convergent than divergent changes in at least one photosynthesis type.
Twenty-six chloroplast proteins with more convergent than divergent changes in C_4-C_4, C_3-C_4 and/or C_3-C_3 pairs. Asterisks indicate proteins with significantly different replacements between C_4-C_4 and C_3-C_4 pairs.

The proteins encoded by *matK*, *rpoC2* and *ndhF* shared much higher numbers of both convergent and divergent replacements than other chloroplast proteins across all photosynthesis type comparisons (Table S4). Both *matK* and *ndhF* are known to be rapidly evolving and have been consistently used in low taxonomic level phylogenetic studies in flowering plants (Barthet and Hilu, 2008; Patterson and Givnish, 2002). The gene *rpoC2* has also been recently described as a useful phylogenetic marker in angiosperms (Walker et al., 2019).

Molecular convergence across ancestral branches

The comparison of ancestral branch pairs with convergent and divergent replacements revealed remarkable differences between photosynthesis types. Overall, C_4-C_4 pairs of ancestral branches showed a distribution skewed toward more convergent and divergent replacements than the two other categories (Fig. 6). There were significantly fewer pairs of C_4-C_4 ancestral branches with no replacements and with no convergent replacements than C_3-C_4 and C_3-C_3 pairs

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

($P < 0.05$, Boschloo's test; Table 2). Conversely, significantly more C_4-C_4 pairs shared more convergent than divergent replacements, and at least two convergent changes compared to C_3-C_4 and C_3-C_3 pairs ($P < 0.05$, Boschloo's test; Table 2).

No significant difference was observed between pairs of C_3-C_4 and pairs of C_3-C_3. We found identical patterns when the same analyses were performed after excluding all replacements in the RbcL protein, except for the lack of a significant difference between C_4-C_4 and C_3-C_3 in the proportion of pairs with divergent replacements and pairs with more convergent than divergent changes (Table S6).

Table 2. Number of ancestral branches with convergent and divergent replacements.

	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/Div	63 (.81)	67 (.57)	18 (.50)
Con>Div	40 (.51)	36 (.31)	10 (.28)
Con>1	49 (.63)	39 (.33)	8 (.22)

Comparisons were made between pairs of C_4-C_4, C_3-C_3 and C_3-C_4 branches. Proportions of pairs of ancestral branches over all branches by category are shown in parenthesis. The total number of pairs of ancestral 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. Div: divergent. Con>Div: pairs of branches with more convergent than divergent replacements. Con>1: pairs of branches with more than one convergent replacement.

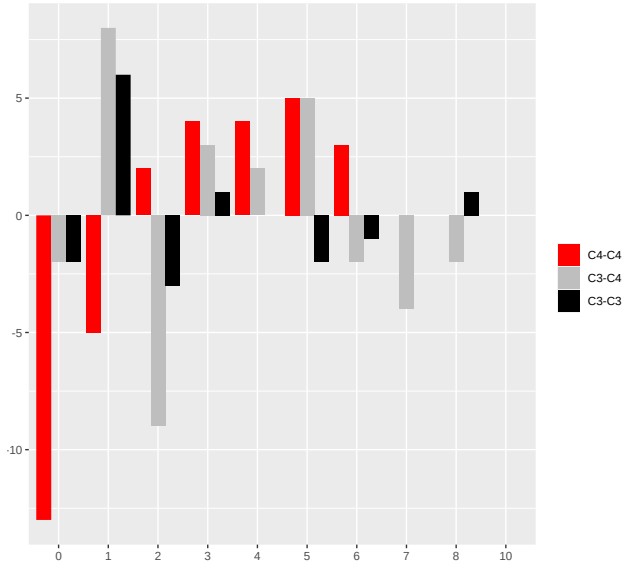

Figure 6. Pairs of ancestral branches by convergent and divergent replacements.
Difference in the number of pairs of ancestral branches for convergent and divergent categories (0-8 and 10 replacements).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Distribution of amino acid replacements across PACMAD lineages

Convergent and divergent replacements were preferentially found in specific pairs of ancestral branches. In C_4 pairs, convergent sites were most abundant between *Danthoniopsis dinteri* and *Aristida purpurea* (ten sites, branches P and V in Fig. 1), whereas divergent sites were most common between *Centropodia glauca* and *Aristida purpurea* (ten sites, branches S and V in Fig. 1). In pairwise C_3 branch comparisons, most convergent sites were identified between both *Zeugites pittieri* and Danthonieae (branches N and R in Fig. 1) and Danthonieae and *Sartidia* spp. (branches R and U in Fig. 1), whereas the most divergent site-rich pair was formed by *Zeugites pittieri* and *Sartidia* spp. (eight sites, branches N and U in Fig. 1; Table S5).

Molecular convergence in the RuBisCO large subunit

We further inspected the evolution of the RuBisCO large subunit across the PACMAD clade. A total of 4 out of 9 RbcL amino acids with convergent changes in C_4 ancestral branches—V101I, A281S, M309I and A328S—have been identified in previous studies on PACMAD grasses (Christin et al., 2008; Piot et al., 2018) as sites that experienced adaptive evolution in C_4 species (Table 3).

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

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

Ancestral AA: ancestral amino acid. Convergent change/p.s.s.: derived amino acid in multiple C_4 ancestral branches and positively selected sites from previous studies. #Convergent a.b.: number of ancestral 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).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A further site, T143A, was found to evolve under positive selection in C_3 to C_4 transitions in monocots (Studer et al., 2014). Interestingly, an adaptive S143A replacement has also been detected in the gymnosperm *Podocarpus* (Sen et al., 2011). Three more sites with convergent replacements—at positions 93, 94 and 461—correspond to amino acids that were reported to evolve under positive selection in different groups of seed plants by Kapralov and Filatov (2007). Thus, all of the *rbcL* codons that appear to have evolved convergently among the PACMAD C_4 lineages we have examined are also known to have experienced adaptive evolution in seed plants, but not all of them have been shown to evolve adaptively in C_4 grasses.

Discussion

The recurrent emergence of carbon-concentration mechanisms (CCMs) across multiple angiosperm clades in the past 35 million years represents one of the most striking examples of convergent evolution of a complex phenotypic trait. Several investigations have shown that the phenotypic parallelism across C_4 lineages is to some extent mirrored by convergent changes in the sequence of proteins with key metabolic roles in the biochemistry of C_4 photosynthesis, both in monocots and eudicots (Christin et al., 2007; Besnard et al., 2009, Christin et al., 2009a, Christin et al., 2009b, Kapralov et al., 2011, Goolsby et al., 2018). Furthermore, biochemical analyses have determined that some of these changes reflect adaptive shifts, as in the case of the increased availability of CO_2 at the RuBisCO site (Studer et al., 2014). Further evidence of changes in the selective pressure associated to the C_3 to C_4 transitions have emerged from the detection of several positively selected sites in multiple genes associated with photosynthetic processes (Christin et al., 2008; Studer et al., 2014; Goolsby et al., 2018; Piot et al., 2018). These and other discoveries have paved the way to a more nuanced understanding of the molecular basis of phenotypic convergence in CCM plants and may accelerate the development of crop varieties with augmented resistance to high temperature and low water availability.

For these aims to be fully realized, a robust framework to assess the extent and phenotypic impact of convergent molecular changes is necessary. Along the lines of strategies applied in vertebrates research (Castoe et al., 2009, Thomas and Hahn, 2015), we presented here the results of a novel methodological approach to the study of molecular convergence in C_4 grasses. We investigated patterns of convergent and divergent amino acid changes in nearly 70 chloroplast proteins across multiple C_4 and C_3 lineages in the PACMAD clade, with the goal of testing a specific hypothesis: is the evolution of chloroplast proteins showing stronger signatures of convergent amino acid replacements in C_4 lineages compared to C_3 lineages? This analysis also allowed us to establish if proteins other than enzymes involved in the CCM biochemistry underwent parallel amino acid changes in C_4 lineages. Our reasoning is that many proteins

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

expressed in the chloroplast could have experienced similar selective pressure across multiple C_3 to C_4 transitions and might have accumulated convergence replacements as a result.

We based our analysis on the identification of amino acid replacements shared by pairs of ancestral C_4 branches, defined here as branches corresponding to C_3 to C_4 transitions in the PACMAD phylogeny. We compared these changes to those identified in ancestral C_3 branches, namely all C_3 lineages that include only C_3 species (Figs. 1 and 2), and to changes found between ancestral C_3 and C_4 branches. For each of the three possible pairs of photosynthesis types C_4-C_4, C_3-C_4 and C_3-C_3, we determined the number of amino acid sites, genes and pairs of ancestral branches with convergent replacements. We detected signatures of convergent evolution in all types of datasets. First, we identified many individual replacements that emerged repeatedly and uniquely in C_4 ancestral branches, particularly in the proteins RbcL, NdhH, NdhI and MatK. We also observed C_3-specific convergent replacements in NdhF and RpoC2, and a case of multiple C_4 and C_3 convergent changes in Rps3. Additionally, we identified 8 chloroplast genes with one or more C_4-specific convergent sites. Second, we found evidence of significantly higher rates of convergent replacements in C_4 lineages in both RbcL and RpoC1, and several convergent replacements that occurred exclusively in C_4-C_4 pairs in proteins encoded by *ndhG*, *ndhI*, *psaI*, *rpoA*, *rps4* and *rps11*. These genes are involved in a variety of biological processes in the chloroplast, from the cyclic electron transport in (*ndhG* and *ndhI*) and the stabilization of (*psaI*) the photosystem I, to transcription (*rpoA* and *rpoC1*), translation (*rps4* and *rps11*) and CO_2 fixation (*rbcL*). Third, we identified statistically significant differences in pairs of C_4 branches with convergent replacements (Table 2). Crucially, we observed more pairs with higher convergent than divergent replacements in C_4-C_4 compared to both C_3-C_3 and C_3-C_4, even after removing replacements identified in the RuBisCO large subunit, RbcL.

Altogether, these findings suggest that multiple biochemical processes occurring in the chloroplast might have experienced recurrent adaptive changes associated with the emergence of C_4 photosynthesis. Notably, some of these proteins are not directly involved in the light-dependent or light-independent reactions of the photosynthesis, implying that processes such as the regulation of gene expression and protein synthesis in the chloroplast are also experiencing significant selective pressures during the transition from C_3 to C_4 plants. These results should motivate further studies to determine the prevalence of convergent amino acid replacements due transitions to CCMs among the thousands of proteins encoded by nuclear genes but expressed in the chloroplast (Jarvis and López-Juez, 2013). Although such analyses are currently hindered by the limited number of sequenced nuclear genomes in taxa with multiple C_3 and C_4 lineages, including the PACMAD clade, genome-wide investigations of convergent replacements will be possible in the near future given the current pace of DNA sequencing in plants.

A further important conclusion drawn from these results is that convergent replacements are not uncommon between C_3-C_3 and C_3-C_4 lineages. This is possibly due to some

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

environmental factors affecting the evolution of chloroplast genes that are shared across grass lineages regardless of their photosynthesis type.

The analysis of individual convergent replacements in the RuBisCO large subunit both confirmed previous findings and highlighted novel potentially adaptive changes among PACMAD species. Importantly, these novel convergent replacements are known to evolve under positive selection in non-PACMAD seed plants. This underscores the potential of our approach to identify novel changes with functional significance in the transition to CCMs in grasses, as opposed to standard statistical tests of positive selection. Alternatively, some RbcL sites could experience convergence across a variety of seed plants because of selective pressure other than those associated with C_3 to C_4 transitions.

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

We recognize some potential caveats in our approach. By relying on a relatively small sample of PACMAD species, our statistical power to detect signatures of convergent evolution was limited. Increasing the number of ancestral C_4 and C_3 lineages should provide a broader representation of convergent replacements in C_4 clades. Furthermore, we applied a strict definition of convergence that ignores changes to amino acids with similar chemical properties. We think that a conservative approach was necessary given that amino acids with similar chemical properties might have a very different functional effect on protein activity given their size and tridimensional interactions with nearby residues. Third, we assumed that all the observed convergent replacements were the result of convergent phenotypic changes, which fall under the general category of homoplasy (Avise and Robinson, 2008). However, some of these replacements could instead represent hemiplasy, or character state changes due to introgression between different C_4 lineages, incomplete lineage sorting (ILS) of ancestral alleles or horizontal gene transfer (Avise and Robinson, 2008). Recombination between chloroplast genomes, which is required for introgression to occur, has been documented but appears to be rare (Carbonell-Caballero et al., 2015, Greiner and Sobanski, 2015, Sancho et al., 2018). Introgression or horizontal gene transfer between congeneric species has been associated to the acquisition of part

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of the C_4 biochemical pathway in the PACMAD genus *Alloteropsis* (Christin et al., 2012; Olofsson et al., 2016). However, these transfers were limited to a few nuclear genes. Moreover, only a very few cases of horizontal transfer between chloroplast genomes have been reported in plants (Stegemann et al., 2012). Therefore, the contribution of hemiplasy to the observed pattern of convergent replacements in C_4 lineages is likely to be minimal. Finally, we treated C_4 species regardless of their photosynthesis subtype (NAPD-ME, NAD-ME and PEPCK), which is known to vary among PACMAD subfamilies (Taylor et al., 2010). We argue that our results are conservative with regard to this aspect because convergent replacements should be expected to occur more often between C_4 groups sharing the same photosynthesis subtype.

Conclusions

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

Acknowledgements

The project was supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, under award number TEX0-1-9599, the Texas A&M AgriLife Research, and the Texas A&M Forest Service.

References

Abascal, F., R. Zardoya, and M. J. Telford. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res* 38: W7–W13.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Proceedings of the Second International Symposium on Information Theory: 267–281.

Andersson, I., and A. Backlund. 2008. Structure and function of Rubisco. *Plant Physiol Biochem* 46: 275-291.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

720 Avise, J. C., and T. J. Robinson. 2008. Hemiplasy: a new term in the lexicon of phylogenetics.
721 *SystBiol* 57: 503–507.
722 Barthet, K M.M., W. Hilu. 2008. Evaluating evolutionary constraint on the rapidly evolving gene
723 matK using protein composition. *J Mol Evol* 66: 85-97.
724 Besnard, G., M. Muasya, F. Russier, E. H. Roalson, N. Salamin, P.-A. Christin. 2009.
725 Phylogenomics of C4 Photosynthesis in Sedges (Cyperaceae): Multiple Appearances and
726 Genetic Convergence. *Mol Biol Evol* 26: 1909–1919.
727 Besnard, G., P.-A. Christin, P.-J. G. Malé, E. Lhuillier, C. Lauzeral, E. Coissac, and M. S.
728 Vorontsova. 2014. From museums to genomics: old herbarium specimens shed light on a
729 C_3 to C_4 transition. *J Exp Bot* 65: 6711–6721.
730 Boschloo, R. D. 1970. Raised conditional level of significance for the 2 × 2-table when testing
731 the equality of two probabilities. *Stat Neerl* 24: 1-9.
732 Brown, W. V., and B. N. Smith. 1972. Grass evolution, the Kranz syndrome, $^{13}C/^{12}C$ ratios, and
733 continental drift. *Nature* 239: 345–346.
734 Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden.
735 2009. BLAST+: architecture and applications. *BMC Bioinfor* 10: 421.
736 Carbonell-Caballero, J., R. Alonso, V. Ibañez, J. Terol, M. Talon, J. Dopazo. 2015. A
737 phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between
738 wild and domestic species within the genus Citrus. *Mol Biol Evol* 32: 2015–2035.
739 Castoe, T. A., A. P. J. d. Koning, H.-M. Kim, W. Gu, B. P. Noonan, G. Naylor, Z. J. Jiang, et al.
740 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc*
741 *Natl Acad Sci USA* 106: 8986–8991.
742 Christin, P.-A., M. Arakaki, C. P. Osborne, and E. J. Edwards. 2015. Genetic enablers
743 underlying the clustered evolutionary origins of C_4 photosynthesis in angiosperms. *Mol*
744 *Biol Evol* 32: 846–858.
745 Christin, P.-A., S. F. Boxall, R. Gregory, E. J. Edwards, J. Hartwell, and C. P. Osborne. 2013a.
746 Parallel recruitment of multiple genes into C_4 photosynthesis. *Genome Biol Evol* 5: 2174–
747 2187.
748 Christin, P.-A., E. J. Edwards, G. Besnard, S. F. Boxall, R. Gregory, E. A. Kellogg, J. Hartwell,
749 C.P. Osborne. 2012. Adaptive evolution of C(4) photosynthesis through recurrent lateral
750 gene transfer. *Curr Biol* 22: 445-9.
751 Christin, P.-A., C. P. Osborne, D. S. Chatelet, J. T. Columbus, G. Besnard, T. R. Hodkinson, L.
752 M. Garrison, et al. 2013b. Anatomical enablers and the evolu- tion of C4 photosynthesis
753 in grasses. *Proc Natl Acad Sci USA* 110: 1381–1386.
754 Christin P.-A., B. Petitpierre, N. Salamin, L. Büchi, G. Besnard. 2009b. Evolution of C(4)
755 phosphoenolpyruvate carboxykinase in grasses, from genotype to phenotype. *Mol*
756 *Biol Evol* 26: 357-65.
757 Christin P.-A., E. Samaritani, B. Petitpierre B, N. Salamin, G. Besnard. 2009a. Evolutionary
758 insights on C4 photosynthetic subtypes in grasses from genomics and
759 phylogenetics. *Genome Biol Evol* 1:221-30.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

760 Christin, P.-A., N. Salamin, A. M. Muasya, E. H. Roalson, F. Russier, and G. Besnard. 2008.
761 Evolutionary switch and genetic convergence on *rbcL* following the evolution of C_4
762 photosynthesis. *Mol Biol Evol* 25: 2361–2368.
763 Christin, P.-A., N. Salamin, V. Savolainen, M. R. Duvall, and G. Besnard. 2007. C_4
764 photosynthesis evolved in grasses via parallel adaptive genetic changes. *Curr Biol* 17:
765 1241-1247.
766 Foote, A. D., Y. Liu, G. W. C. Thomas, T. Vinař, J. Alföldi, J. Deng, S. Dugan, et al. 2015.
767 Convergent evolution of the genomes of marine mammals. *Nat Genet* 47: 272–275.
768 Goolsby, E. W., A. J. Moore, L. P. Hancock, J. M. De Vos, and E. J. Edwards. 2018. Molecular
769 evolution of key metabolic genes during transitions to C_4 and CAM photosynthesis. *Am J*
770 *Bot* 105: 602–613.
771 Greiner S., J. Sobanski, R. Bock. 2015. Why are most organelle genomes transmitted
772 maternally? *Bioessays* 37: 80-94.
773 Grass Phylogeny Working Group II. 2012. New grass phylogeny resolves deep evolutionary
774 relationships and discovers C_4 origins. *New Phytol* 193: 304–312.
775 Hall, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis
776 program for Windows 95/98/NT. *Nucl Acids Symp Series* 41: 95-98.
777 Hattersley, P. W., and N. E. Stone. 1986. Photosynthetic enzyme activities in the C_3-C_4
778 intermediate *Neurachne minor* S. T . Blake (Poaceae). *Aust J Plant Physiol* 13: 399-408.
779 Hattersley, P. W., S.-C. Wong, S. Perry, and Z. Roksandic. 1986. Comparative ultrastructure and
780 gas exchange characteristics of the C_3–C_4 intermediate *Neurachne minor* S. T. Blake
781 (Poaceae). *Plant Cell Environ* 9: 217-233.
782 Heyduk, K., J. J. Moreno-Villena, I. S. Gilman, P.-A. Christin, and E. J. Edwards. 2019. The
783 genetics of convergent evolution: insights from plant photosynthesis. *Nat Rev Genet* 20:
784 485–493.
785 Jarvis, P., E. López-Juez. 2013. Biogenesis and homeostasis of chloroplasts and other
786 plastids. *Nat Rev Mol Cell Biol* 14: 787-802.
787 Kapralov, M. V., and D. A. Filatov. 2007. Widespread positive selection in the photosynthetic
788 Rubisco enzyme. *BMC Evol Biol* 7: 73.
789 Kapralov, M. V., J. A. C. Smith, and D. A. Filatov. 2012. Rubisco evolution in C_4 eudicots: An
790 analysis of Amaranthaceae *sensu lato*. *PLoS One* 7: e52974.
791 Kapralov, M. V., D. S. Kubien, I. Andersson, and D. A. Filatov. 2011. Changes in rubisco
792 kinetics during the evolution of C_4 photosynthesis in *Flaveria* (Asteraceae) are associated
793 with positive selection on genes encoding the enzyme. *Mol Biol Evol* 28: 1491–1503.
794 Kellogg, E. A. 1999. Phylogenetic Aspects of the Evolution of C_4 Photosynthesis. *In* R. F. Sage
795 and R. K. Monson [eds.], C_4 Plant Biology, 411-444. Academic Press, San Diego,
796 California, USA.
797 Knapp, A. K., and E. Medina. 1999. Success of C_4 Photosynthesis in the Field: Lessons from
798 Communities Dominated by C_4 Plants. *In* R. F. Sage and R. K. Monson [eds.], C_4 Plant
799 Biology, 251-283. Academic Press, San Diego, California, USA.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

800 Lanfear, R., B. Calcott, S. Y. W. Ho, and S. Guindon. 2012. PartitionFinder: Combined selection
801 of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol*
802 29: 1695–1701.
803 Li, B., C. Förster, C. A. M. Robert, T. Züst, L. Hu, R. A. R. Machado, J.-D. Berset, et al. 2018.
804 Convergent evolution of a metabolic switch between aphid and caterpillar resistance in
805 cereals. *Sci Adv* 4: eaat6797.
806 Long, S. P. 1999. Environmental Responses. *In* R. F. Sage and R. K. Monson [eds.], C_4 Plant
807 Biology, 215-249. Academic Press, San Diego, California, USA.
808 Lü, P., S. Yu, N. Zhu, Y.-R. Chen, B. Zhou, Y. Pan, D. Tzeng, et al. 2018. Genome encode
809 analyses reveal the basis of convergent evolution of fleshy fruit ripening. *Nat Plants* 4:
810 784–791.
811 Maier R. M., K. Neckermann, G. L. Igloi, H. Kössel. 1995. Complete sequence of the maize
812 chloroplast genome: gene content, hotspots of divergence and fine tuning of
813 genetic information by transcript editing. *J Mol Biol* 251: 614-28.
814 Maurino, V. G., and C. Peterhansel. 2010. Photorespiration: current status and approaches for
815 metabolic engineering. *Curr Opin Plant Biol* 13: 248-255.
816 Natarajan, C., F. G. Hoffmann, R. E. Weber, A. Fago, C. C. Witt, and J. F. Storz. 2016.
817 Predictable convergence in hemoglobin function has unpredictable molecular
818 underpinnings. *Science* 354: 336-339.
819 Olofsson, J.K., M. Bianconi, G. Besnard, L. T. Dunning, M. R. Lundgren, H. Holota, M. S.
820 Vorontsova, O. Hidalgo, I. L. Leitch, P. Nosil, C. P. Osborne, P.-A. Christin. 2016.
821 Genome biogeography reveals the intraspecific spread of adaptive mutations for a
822 complex trait. *Mol Ecol* 25: 6107-6123.
823 Patterson, T. B., T. J. Givnish. 2002. Phylogeny, concerted convergence, and phylogenetic niche
824 conservatism in the core Liliales: insights from rbcL and ndhF sequence data. *Evolution*
825 56: 233–252.
826 Parker, J., G. Tsagkogeorga, J. A. Cotton, Y. Liu, P. Provero, E. Stupka, and S. J. Rossiter. 2013.
827 Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* 502:
828 228–231.
829 Piot, A., J. Hackel, P.-A. Christin, and G. Besnard. 2018. One-third of the plastid genes evolved
830 under positive selection in PACMAD grasses. *Planta* 247: 255–266.
831 Preite, V., C. Sailer, L. Syllwasschy, S. Bray, H. Ahmadi, U. Krämer, and L. Yant. 2019.
832 Convergent evolution in *Arabidopsis halleri* and *Arabidopsis arenosa* on calamine
833 metalliferous soils. *Philos Trans R Soc Lond B Biol Sci* 374: 20180243.
834 Rambaut, A. 2012. FigTree. Tree Figure Drawing Tool, version 1.4.0. website:
835 http://tree.bio.ed.ac.uk/software/figtree/.
836 Rosenblum, E. B., C. E. Parent, and E. E. Brandt. 2014. The molecular basis of phenotypic
837 convergence. *Annu Rev Ecol Evol Syst* 45: 203-226.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

838 Saarela, J. M., S. V. Burke, W. P. Wysocki, M. D. Barrett, L. G. Clark, J. M. Craine, P. M.
839 Peterson, et al. 2018. A 250 plastome phylogeny of the grass family (Poaceae):
840 topological support under different data partitions. *PeerJ* 6: e4299.
841 Sage, R. F. 1999. Why C_4 Photosynthesis? *In* R. F. Sage and R. K. Monson [eds.], C_4 Plant
842 Biology, 3-16. Academic Press, San Diego, California, USA.
843 Sage, R. F. 2004. The evolution of C_4 photosynthesis. *New Phytol* 161: 341–370.
844 Sage, R. F., P.-A. Christin, and E. J. Edwards. 2011. The C_4 plant lineages of planet Earth. *J Exp*
845 *Bot* 62: 3155–3169.
846 Sage, R. F., T. L. Sage, and F. Kocacinar. 2012. Photorespiration and the evolution of C_4
847 photosynthesis. *Annu Rev Plant Biol* 63: 19-47.
848 Sancho, R., C.P. Cantalapiedra, D. López-Alvarez, S. P. Gordon, J. P. Vogel, P. Catalán,
849 B. Contreras-Moreira. 2018. Comparative plastome genomics and phylogenomics of
850 Brachypodium: flowering time signatures, introgression and recombination in
851 recently diverged ecotypes. *New Phytol* 218:1631-1644.
852 Sayers, E. W., M. Cavanaugh, K. Clark, J. Ostell, K. D. Pruitt, and I. Karsch-Mizrachi. 2019.
853 GenBank. *Nucleic Acids Res* 47: D94–D99.
854 Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6: 461-464.
855 Sen L, Fares MA, Liang B, Gao L, Wang B, Wang T, Su YJ. 2011. Molecular evolution
856 of *rbcL* in three gymnosperm families: identifying adaptive and coevolutionary
857 patterns. *Biol Direct* 6:29.
858 Smith, B. N., and W. V. Brown. 1973. The Kranz syndrome in the Gramineae as indicated by
859 carbon isotopic ratios. *American Journal of Botany* 60: 505-513.
860 Stegemann, S., M. Keuthe, S. Greiner, R. Bock. 2012. Horizontal transfer of chloroplast
861 genomes between plant species. *Proc Natl Acad Sci U S A* 109:2434-8.
862 Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
863 large phylogenies. *Bioinformatics* 30: 1312–1313.
864 Storz, J. F. 2016. Causes of molecular convergence and parallelism in protein evolution. *Nature*
865 *Reviews Genetics* 17: 239-250.
866 Studer, R. A., P.-A. Christin, M. A. Williams, and C. A. Orengo. 2014. Stability-activity
867 tradeoffs constrain the adaptive evolution of RubisCO. *Proc Natl Acad Sci U S A* 111:
868 2223-2228.
869 Taylor, S. H., S. P. Hulme, M. Rees, B. S. Ripley, F. I. Woodward, and C. P. Osborne. 2010.
870 Ecophysiological traits in C_3 and C_4 grasses: a phylogenetically controlled screening
871 experiment. *New Phytol* 185: 780–791.
872 Thomas, G. W. C., and M. W. Hahn. 2015. Determining the null model for detecting adaptive
873 convergence from genomic data: A case study using echolocating mammals. *Mol Biol*
874 *Evol* 32: 1232–1236.
875 Virtanen, P., R. Gommers, T. E. Oliphant. M. Haberland, T. Reddy, D. Cournapeau, E. Burovski,
876 P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett M, J. Wilson, K. J.
877 Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat,

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.30.457919; this version posted August 31, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

878 Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen,
879 E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van
880 Mulbregt; SciPy 1.0 Contributors. 2020. SciPy 1.0: fundamental algorithms for scientific
881 computing in Python. Nat Methods 17: 261-272.
882 Walker J. F., N. Walker-Hale, O. M. Vargas, D. A. Larson, G. W. Stull. 2019. Characterizing
883 gene tree conflict in plastome-inferred phylogenies. *PeerJ* 7:e7747.
884 Williams, B. P., I. G. Johnston, S. Covshoff, and J. M. Hibberd. 2013. Phenotypic landscape
885 inference reveals multiple evolutionary paths to C4 photosynthesis. *eLife* 2: e00961.
886 Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:
887 1586–1591.
888 Zhou, X., I. Seim, and V. N. Gladyshev. 2015. Convergent evolution of marine mammals is
889 associated with distinct substitutions in common genes. *Sci Rep* 5: 16550.
890 Zou, Z., and J. Zhang. 2015. No genome-wide protein sequence convergence for echolocation.
891 *Mol Biol Evol* 32: 1237–1241.