Title

Rubisco is evolving for improved catalytic efficiency and CO_2 assimilation in plants

Authors

Jacques W. Bouvier[1], David M. Emms[1], Steven Kelly[1]

Affiliations

[1] Department of Biology, University of Oxford, South Parks Road, Oxford, OX1 3RB, United Kingdom

Corresponding Author

Name: Steven Kelly

Email: steven.kelly@biology.ox.ac.uk

Address: Department of Biology, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK

Keywords

Rubisco; Evolution; Adaptation; Kinetics; Phylogenetics; Photosynthesis; CO_2 assimilation

Abstract

Rubisco is the primary entry point for carbon into the biosphere. However, rubisco is widely regarded as inefficient leading many to question whether the enzyme can adapt to become a better catalyst. Through a phylogenetic investigation of the molecular and kinetic evolution of Form I rubisco we demonstrate that rubisco is not stagnant. Instead, we demonstrate *rbcL* is among the 1% of slowest evolving genes and enzymes on Earth, accumulating one nucleotide substitution every 0.9 million years and one amino acid mutation every 7.2 million years. Despite this, we demonstrate that rubisco catalysis is continuing to evolve toward improved CO_2/O_2 specificity, carboxylase turnover, and carboxylation efficiency. Consistent with this kinetic adaptation, we reveal that increased rubisco evolution leads to a concomitant improvement in leaf-level CO_2 assimilation. Thus, rubisco is continually evolving toward improved catalytic efficiency and CO_2 assimilation in plants.

Introduction

Ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) converts atmospheric CO_2 into sugars that fuel the majority of life on Earth. The enzyme evolved ~3 billion years ago when the atmosphere contained high levels of CO_2 (≥10,000% present atmospheric levels) and comparatively little O_2 (≤0.1% present atmospheric levels) (Figure 1)[1–7]. Since emergence, the enzyme has helped guide the atmosphere on a trajectory of increasing O_2 and declining CO_2 (Figure 1)[1,8] such that current concentrations of CO_2 (0.04%) and O_2 (20.95%) are inverted compared to when the enzyme first evolved (Figure 1).

Although all extant rubisco are descended from a single ancestral rubisco-like protein[9–11], the enzyme is found in a variety of compositional forms across the tree of life (Figure 1)[12,13]. The simplest manifestations are the Form II and Form III variants found in protists, archaea, and some bacteria which are composed of a dimer, or dimers, of the ~50 kDa rubisco large (RbcL) subunit[13–16]. In contrast, Form I rubisco is a hexadecamer comprised of four RbcL dimers organised in an antiparallel core capped at either end by the ~15 kDa rubisco small subunit (RbcS)[13,17]. Of these three Forms, only Form I and II have been recruited for oxygenic photosynthesis[16], with Form I being responsible for the vast majority of global CO_2 assimilation[16,18].

Within Form I rubisco the active site is located in RbcL[15,19,20]. As a result, interspecific differences in Form I kinetics are primarily attributable to sequence variation in RbcL[21–32]. Despite not playing a direct role in catalysis RbcS influences the function of rubisco[33], and its incorporation in the holoenzyme enables its higher kinetic efficiency[34]. Specifically, RbcS enhances the stability and assembly of the holoenzyme complex[20,35–40], improves the efficiency CO_2 binding[41], and is thought to act as a reservoir for CO_2 accumulation[42]. Accordingly, rubisco function is altered when RbcS is mutated[43–45], or when chimeric holoenzymes are created *in vivo*[46–50] and *in vitro*[51–56]. Moreover, there is increasing recognition of the importance of both environment[57] and organ-specific[58,59] differences in plant RbcS isoform expression on holoenzyme catalysis. However, even though RbcS influences holoenzyme function, sequence variation in RbcL remains the primary determinant of variation in kinetics[21–32].

Although there is kinetic variability between rubisco orthologs, the enzyme is considered to be an inefficient catalyst. For example, the maximum substrate-saturated turnover rate of Form I rubisco (<12 s^{-1})[60] is slower than average[61]. In addition, rubisco catalyses a reaction with O_2[62–64] that is competitive with CO_2 and results in the loss of fixed carbon via photorespiration[65–67]. As a consequence, rubisco appears poorly suited to the current O_2-rich/CO_2-poor atmosphere (Figure 1). Moreover, it appears that instead of improving enzyme function, multiple lineages have evolved alternative strategies to overcome rubisco's shortcomings. For example, higher rates of CO_2 assimilation are often achieved either by synthesising large quantities of rubisco (~50% of soluble protein in plants[68] and some microbes[69,70]), or by operating CO_2-concentrating mechanisms[71–73]. As a result, many have questioned whether the enzyme is already perfectly adapted, and whether further kinetic improvements are possible[15,63,67,74–78]. Obtaining answers to these questions would shed light on the "rubisco paradox" – helping to explain why this enzyme of such paramount importance appears poorly adapted for its role.

The initial hypothesis that attempted to explain the above rubisco paradox proposed that rubisco is constrained by catalytic trade-offs that limit the enzyme's adaptation. This theory was pioneered by two studies[79,80] which found antagonistic correlations between rubisco kinetic traits and proposed that these trade-offs were caused by constraints on its catalytic mechanism. However, recent evidence has questioned this hypothesis as the sole mechanism to explain the rubisco paradox. Specifically, analysis of larger species sets have revealed that kinetic trait correlations are not strong[81–83]. In addition, phylogenetic signal in rubisco kinetics causes kinetic trait correlations to be overestimated unless phylogenetic comparative methods are employed[21,22]. Thus, when larger datasets are analysed with phylogenetic methods, the strength of catalytic trade-offs are substantially reduced[21,22]. Instead, phylogenetic constraints have had a larger impact on limiting enzyme adaptation compared to catalytic trade-offs[21,22]. These recent findings motivate a revaluation of the rubisco paradox, and an interrogation of whether rubisco is evolving for improved catalysis and CO_2 assimilation in plants.

Here, we address these issues through a phylogenetic interrogation of the molecular and kinetic evolution of the Form I holoenzyme. We reveal that RbcL has evolved at a slower rate than >98% of

all other gene/protein sequences across the tree of life. Through simultaneous evaluation of molecular and kinetic evolution of rubisco during the radiation of C_3 angiosperms, we reveal that the enzyme has been continually evolving toward improved CO_2/O_2 specificity, carboxylase turnover rate, and carboxylation efficiency. Furthermore, we demonstrate that enhanced rubisco evolution is associated with enhanced rates of CO_2 assimilation and higher photosynthetic nitrogen-use efficiencies. Thus, rubisco is not perfectly adapted, but is slowly evolving towards improved catalytic efficiency and CO_2 assimilation.

Results

RbcL has evolved slower than RbcS and has experienced stronger purifying selection

Sequences encoding Form I rubisco were obtained from the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/). This dataset was filtered to retain sequences for a given species only if a full-length sequence for both *rbcL* and *rbcS* were present. Although *rbcL* exists as a single copy gene in all species, many species harbour multiple *rbcS* genes in their genomes. Thus, for each species a single *rbcL* sequence and all available *rbcS* sequences were taken forward. In total, this resulted in a set of 488 *rbcL*/RbcL and 1140 *rbcS*/RbcS sequences across 488 species (Supplemental File 1, Figure S1 and table S1).

In order to compare the rate at which the two rubisco subunits have evolved, species were partitioned into distinct taxonomic groups comprising the red algae (*Rhodophyta*; *n* = 201), the SAR supergroup (*Stramenopiles*, *Alveolates*, and *Rhizaria*; *n* = 129), the bacteria (*Bacteria*; *n* = 78), land plants (*Streptophyta*; *n* = 68) and green algae (*Chlorophyta*; *n* = 12) (Supplemental File 1, Figure S1 and table S1). Hereinafter, the total amount of molecular evolution of the nucleotide sequences (*rbcL* and *rbcS*) and the total amount of molecular evolution of the protein sequences (RbcL and RbcS) in a taxonomic group is referred to as “the extent of nucleotide evolution” and “the extent of protein evolution”, respectively. The term “the extent of molecular evolution” jointly refers to both.

Comparison of the two rubisco subunits revealed that the extent of molecular evolution in *rbcL*/RbcL is lower than that experienced by *rbcS*/RbcS (Figure 2A). This was not an artefact of the higher gene copy number of *rbcS,* as a 1,000 bootstrapped stratified sampling recovered the same result when

only a single *rbcS*/RbcS sequence was randomly sampled per species (see Methods; Figure 2B). Therefore, *rbcL*/RbcL has explored less nucleotide and protein sequence space than *rbcS*/RbcS in the same sets of species over the same period of time (Figure 2C). Furthermore, *rbcL* also experienced fewer amino acid changes per change in nucleotide sequence compared to *rbcS* (Figure 2D), indicating a higher degree of purifying selection. Thus, *rbcL*/RbcL has evolved more slowly and has been subject to a higher degree of functional constraint on the encoded protein sequence than *rbcS*/RbcS.

RbcL is one of the slowest evolving genes in the tree of life

To evaluate the rate of molecular evolution in context of all other genes in the species under consideration, the percentile rank of *rbcL*/RbcL and *rbcs*/Rbcs was evaluated for all genes in all species (see Methods). This revealed that 99.3% of all gene nucleotide sequences and 98.1% of all gene protein sequences evolved faster than *rbcL*/RbcL in the same sets of species over the same period of time (Figure 3A; Supplemental File 1, table S2). This held true even if *rbcL*/RbcL was only compared only to the subset of genes that encode enzymes, with 99.2% of enzyme nucleotide sequences and 98.3% of enzyme protein sequences evolving faster than *rbcL*/RbcL (Figure 3B; Supplemental File 1, table S3). Furthermore, in land plants *rbcL*/RbcL was also the slowest evolving component of the Calvin-Benson-Bassham cycle (Figure 3C; Supplemental File 1, table S4 and S5). This slow pace of evolution is not simply an artefact of being encoded in the plastid genome, as *rbcL*/RbcL was also one of the slowest evolving genes/proteins in bacteria which encode all of their genes in a single cytoplasmic genome. Thus, *rbcL*/RbcL is one of the slowest evolving genes/enzymes in all species in which it is found, irrespective of the taxonomic group or genome in which it is encoded.

In contrast to *rbcL*/RbcL, considerable variability in the extent of molecular evolution in the small subunit was observed both within and between taxonomic groups (Figure 3A; Supplemental File 1, table S2). Analogous results in each taxonomic group were recovered when this analysis was restricted to the subset of genes that encode enzymes (Figure 3B; Supplemental File 1, table S3). Moreover, in land plants *rbcS*/RbcS was the fastest evolving component of the Calvin-Benson-Bassham cycle (Figure 3C; Supplemental File 1, table S4 and S5). Thus, while the pace of molecular

evolution in *rbcL*/RbcL is ubiquitously slow, the extent of molecular evolution of *rbcS*/RbcS is highly variable explaining the disparity in the rate of both subunits across the tree of life (Figure 2C; Supplemental File 1, table S6). A similar variable rate was also observed for rubisco's ancillary chaperones (Supplemental File 1). Thus, the rate of molecular evolution of *rbcL*/RbcL is ubiquitously low, and lower than *rbcS*/RbcS or any associated chaperone.

Rubisco is evolving for improved kinetic efficiency in plants

Although *rbcL* is among the slowest evolving genes on Earth, the analysis above demonstrates that it is not stagnant. This raises the question as to whether the sequence evolution is adaptive and is improving the catalysis of the enzyme. We hypothesised that if rubisco was undergoing directional selection for improved catalysis, then orthologs that have experienced the largest extent of molecular evolution would be the most efficient catalysts. To test this hypothesis, a dataset of kinetic measurements from C_3 angiosperms[21,22,81] was evaluated in the context of the molecular evolution of RbcL (Figure 4A,B). This analysis focused on RbcL as it is the primary determinant of kinetics[21–32], and because sufficient sequence data for RbcS are unavailable. This revealed that the more RbcL has evolved from the most recent common ancestral sequence, the better its CO_2/O_2 specificity ($S_{C/O}$; 10.1% variance explained), CO_2 turnover rate (k_{catC}; 4.6% variance explained) and carboxylation efficiency (k_{catC}/K_C; 3.8% variance explained) (Figure 4B). This result is not an artefact caused by potential systematic methodological biases associated with species sampling or potential uncertainties or errors in the underlying phylogenetic tree (See Methods, Supplemental File 1). Thus, rubisco has been adaptively evolving for improved $S_{C/O}$, k_{catC}, and k_{catC}/K_C during the radiation of the angiosperms.

Given that the origin of the angiosperms is estimated to have occurred 160 million years ago[84] (Figure 4A), it is possible to put the above kinetic change in context of both molecular sequence changes and evolutionary time (table 1). As the large subunit acquired one nucleotide substitution every 0.9 million years and one amino acid substitution every 7.2 million years (Supplemental File 1, figure S2), each amino acid substitution resulted in an average increase in $S_{C/O}$ by 2.7×10^{-1} mol.mol^{-1}, in k_{catC} by 3.6×10^{-2} s^{-1}, and in k_{catC}/K_C by 1.8×10^{-3} s^{-1} μM^{-1}. This is equivalent to a relative improvement of 0.3% ($S_{C/O}$), 1.4% (k_{catC}), and 1.1% (k_{catC}/K_C) per amino acid substitution, and a

relative improvement of 0.05% ($S_{C/O}$), 0.2% (k_{catC}), and 0.2% (k_{catC}/K_C) per million years. Thus, there has been continual improvement in rubisco kinetics during the radiation of the angiosperms at a rate that is dependent on the extent of its molecular sequence change.

Rubisco is evolving for improved leaf-level CO_2 assimilation

Given that rubisco is evolving to become a better catalyst, we hypothesised that this adaptation would also drive adaptation in the rate of leaf-level CO_2 assimilation. To test this hypothesis we analysed a large dataset of photosynthetic measurements from C_3 angiosperms[85] in the context of the extent of their RbcL evolution (Figure 5A-C). This revealed that the rate of leaf-level CO_2 assimilation was also dependent on the extent of molecular sequence change in rubisco such that that C_3 angiosperms with more evolved rubisco also higher rates of CO_2 assimilation (A_{mass}; 19.2% variance explained, Figure 5B). This is not a consequence of increased nitrogen investment in the leaf, as the association between rubisco evolution and increased CO_2 assimilation is strengthened when measurements are controlled for leaf nitrogen content ($PNUE_{mass}$, 22.1% variance explained, Figure 5B). Analogous results were obtained when measurements of CO_2 assimilation were evaluated on a leaf area basis (Figure 5C). This result is most parsimoniously explained by directional selection towards enhanced leaf-level CO_2 assimilation driven by the kinetic adaptation described above. Thus, the adaptive evolution of rubisco during the radiation of the angiosperms has resulted in the improvement in leaf-level CO_2 assimilation.

Discussion

Rubisco is the primary entry point for carbon into the biosphere, responsible for fixing 250 billion tons of CO_2 annually[18]. Despite this immense throughput, the enzyme is a surprisingly inefficient catalyst with a modest carboxylase turnover rate of <12 s^{-1} [60] and a competing oxygenase activity that results in the loss of fixed carbon[63,64,86]. This discord presents an evolutionary paradox that has attracted significant attention[15,21,22,63,67,74–78], with the prevailing assumption being that rubisco is evolutionarily stagnant. Here we demonstrate that the enzyme is not stagnant, but that it is encoded by one of the slowest evolving genes on Earth. Despite this, we demonstrate that rubisco has been evolving for higher CO_2/O_2 specificity ($S_{C/O}$), faster carboxylase turnover rates (k_{catC}), and improved carboxylation efficiencies (k_{catC}/K_C) in angiosperms. Moreover, we demonstrate that plants with more evolved

rubisco exhibit higher leaf-level CO_2 assimilation and enhanced photosynthetic nitrogen-use efficiencies. Thus, rubisco has been continually evolving towards improved catalytic efficiency and CO_2 assimilation during the radiation of the angiosperms.

A slow rate of molecular evolution in *rbcL* has long been assumed and has underpinned the use of this gene for systematics and phylogenetics[87–89]. However, to our knowledge there has been no contextualised measurement of the rate of *rbcL* evolution across the tree of life. The analysis presented here addresses this gap by revealing that across the tree of life, *rbcL*/RbcL has experienced a lower extent of molecular evolution than 99% of all gene nucleotide sequences and 98% of all gene protein sequences. It is interesting to note here that this is not due to the presence of *rbcL* in the chloroplast genome, as *rbcL* is also one of the slowest evolving sequences in bacteria which lack organellar genomes. Thus, RbcL is universally one of the slowest evolving sequences on Earth, irrespective of the taxon or genome in which it resides.

Although dissecting the factors which constrain the rate of *rbcL* evolution is beyond the scope of the current study, the slow pace of *rbcL* molecular evolution is most likely a consequence of several synergistic factors[90] including constraints imposed by expression[91–95], selection to preserve protein function[96–100], and the requirements for protein-protein interactions *in vivo*[101–104]. These factors would be particularly pertinent for rubisco given that it is the most abundant protein in organisms in which it is found[68,70], it is subject to catalytic trade-offs[21,22,80,81] and molecular activity-stability trade-offs[105–108], and given that it relies on multiple interacting partners and chaperones for folding, assembly and metabolic regulation[68,109]. Thus, a perfect storm of features exist which could limit the molecular evolution of *rbcL* and thereby cause it to be one of the slowest evolving genes on Earth. Further work to elucidate the exact contribution of each of these biological determinants on rubisco's rate of molecular evolution is warranted, building upon the work here and previous investigations[21,22,110].

Our integrated analysis of rubisco evolution revealed a continual improvement in $S_{C/O}$, k_{catC} and k_{catC}/K_C during the radiation of C_3 angiosperms. Thus, although rubisco is slowly-evolving, sequence changes have enhanced the catalytic properties of the enzyme. In the context of the C_3 leaf, such directional selection towards improved $S_{C/O}$ is consistent with adaptation to maintain adequate carbon assimilation in response to declining CO_2 and increasing O_2 (Figure 1). This evolutionary

strategy has been proposed previously[80], and is suggested to apply broadly across photoautotrophs lacking a CO_2 concentrating mechanism[82,111]. In addition to adaptation for higher $S_{C/O}$, we also discover simultaneous improvement in k_{catC} and k_{catC}/K_C without antagonism in any other kinetic trait. These results are also consistent with the inferior $S_{C/O}$ and k_{catC}/K_C reported for extinct rubisco resurrected at the dawn of the Form IB[6] and Form 1[34] lineages. It is noteworthy that on first appearances, all of these studies seem to contradict an analysis within the *Solanaceae* in which resurrected ancestral rubisco variants exhibited superior k_{catC} and k_{catC}/K_C values. However, in this instance the kinetic differences were proposed to be driven by sequence changes in RbcS[112], and therefore do not contradict the analysis of RbcL presented here or in other studies[6,34]. Thus, sequence change in RbcL during the radiation of angiosperms has driven the continual improvement of the enzyme in the presence of a declining atmospheric CO_2:O_2 concentration.

The 'FvCB model' of photosynthesis[113], as well as a suite of other experimental[114–121] and computational[122,123] studies all demonstrate that rubisco is a major rate-limiting factor for CO_2 assimilation under ambient steady-state conditions. The findings presented here link these mechanistic studies with evolutionary biology, and reveal that rubisco has experienced directional selection to improve kinetic efficiency and CO_2 assimilation. Ultimately, this changes our understanding of the rubisco paradox. Rubisco is not locked in evolutionary stasis, but is instead slowly evolving towards improved CO_2 assimilation. This discovery has significant implications for our understanding of the past, present, and future potential of rubisco in natural and engineered contexts.

Materials and Methods

Rubisco sequence data

All publicly available coding sequences of the rubisco large (*rbcL*) and small (*rbcS*) subunit genes in the National Center for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/) as of July 2020 were downloaded. Manual inspection of nucleotide and translated protein sequences was performed to remove any partial, chimeric, or erroneously annotated sequences. In addition, this dataset was further restricted to include only those species which possess a Form I rubisco and for which both a full-length *rbcL* and *rbcS* gene sequence could be obtained. Given that *rbcL* exists

as a single copy gene in all species, only one *rbcL* sequence per species was retained for downstream analysis. In contrast, all possible full-length *rbcS* sequences were taken forward to account for the fact that *rbcS* is multicopy in some genomes.

Translated *rbcL* and *rbcS* protein sequences were aligned using the MAFFT L-INS-i algorithm[124]. The corresponding codon alignments of the nucleotide sequences were generated by threading the nucleotide sequences through the aligned protein sequences that they encode using PAL2NAL software[125]. Multiple sequence alignments were trimmed to remove non-aligned codon or residue positions such that only ungapped columns remained. During this process, the putative transit peptide of the *rbcS*/RbcS sequences in taxa in which this gene is encoded by the nuclear genome was computationally cleaved. Following these data processing steps, alignments were partitioned depending on species membership to either the bacteria (*Bacteria*; *n* = 78), land plants (*Streptophyta*; *n* = 68), green algae (*Chlorophyta*; *n* = 12), red algae (*Rhodophyta*; *n* = 201) or the SAR supergroup (*Stramenopiles*, *Alveolates*, and *Rhizaria; n* = 129) by use of the NCBI taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi). Any sequences belonging to species in either the *Haptophyta*, *Cryptophyta*, *Glaucocystophyta* or *Excavata* taxonomic groups were excluded from the dataset at this point due to insufficient data availability. In total, this resulted in a combined set of 488 *rbcL*/RbcL and 1140 *rbcS*/RbcS gene and protein sequences across 488 species spanning 5 taxonomic groups (Supplemental File 1, Figure S1 and table S1). The complete set of raw *rbcL*/RbcL and *rbcS*/RbcS sequences, as well as the complete set of aligned and trimmed *rbcL*/RbcL and *rbcS*/RbcS sequences can be found in Supplemental File 2.

Rubisco phylogenetic tree inference

Maximum-likelihood *rbcL*/RbcL and *rbcS*/RbcS phylogenetic gene trees were inferred across all sequences within a taxonomic group by IQ-TREE[126] using the ultrafast bootstrapping method with 1,000 replicates and the Shimodaira–Hasegawa approximate–likelihood ratio branch test. The best fitting models of nucleotide (SYM+R8) and amino acid (LG+R5) sequence evolution were respectively determined as those which exhibit the lowest combined Bayesian information criterion rank score across the complete sets of both rubisco large and small subunit sequences[126]. Across

all taxonomic groups, the models of nucleotide and amino acid sequence evolution were held constant between the gene trees for *rbcL* and *rbcS*, and RbcL and RbcS, respectively, such that branch lengths are comparable across both subunits. The complete set of these *rbcL*/RbcL and *rbcS*/RbcS phylogenetic gene trees used as the basis of the analysis herein can be found in Supplemental File 3.

Stratified sampling of rbcS sequences and phylogenetic tree inference

To account for potential biases in our analysis caused by some species exhibiting multiple copies of *rbcS*, random stratified sampling of the non-gapped *rbcS*/RbcS sequence alignments was conducted by species using 1,000 replicates with replacement. This process resulted in the generation of 1,000 unique *rbcS*/RbcS alignments for each taxonomic group, whereby each of these respective alignments contain only a single randomly selected copy of the *rbcS*/RbcS per species. In turn, each of these alignments were subject to data processing and phylogenetic tree inference using IQ-TREE[126] following the method described above.

Quantification of the total extent of nucleotide and protein molecular evolution in rbcL/RbcL and rbcS/RbcS

The extent of molecular evolution in both rubisco subunits was assessed across all species in a given taxonomic group as the total length (sequence substitutions per aligned sequence site) of the phylogenetic tree describing the evolutionary history of each respective gene. For this purpose, tree length was calculated as the combined sum of branch lengths leading from the root at the last common ancestor of the tree to the set of sequences at the terminal nodes. In this way, using the trees inferred across the complete cohort of *rbcL*/RbcL and *rbcS*/RbcS sequences in each taxonomic group, it was possible to capture all nucleotide and amino acid evolution which has arisen in each subunit since the most recent common ancestor of all sampled species in the group. An identical analysis was also performed for each *rbcS*/RbcS tree generated by stratified sampling, with mean and standard errors of estimates being calculated in this case across the 1,000 unique bootstrap replicate trees.

Genomes and gene models

Complete sets of representative gene models for as many species in the rubisco sequence dataset as possible were acquired from either NCBI (https://www.ncbi.nlm.nih.gov/) or Phytozome V13[127]. Where more than one such gene model resource was available for a given species, the most recent assembly version was chosen. In this way, complete sets of representative gene models were acquired for a total of 32 of the bacteria species, 27 of the land plant species, 8 of the SAR species, 6 of the red algae species and 4 of the green algae species analysed in the present study, respectively (Supplemental File 1, table S1 and S7).

Predicted gene model sets were filtered to remove sequences with internal in-frame stop codons. Gene model sets were also filtered to keep only the longest gene model variant per gene. Moreover, owing to a lack of data availability of publicly available chloroplast or mitochondrial genomes for the eukaryotic species in the present analysis, and as organellar genomes contain fewer than 1% of genes encoded in the corresponding nuclear genome, only gene sequences encoded by the nuclear genomes of species in the land plant, green algae, red algae and SAR taxonomic groups were taken forward for analysis. Finally, after the above quality control checks were completed, a corresponding proteome was generated from each species gene model set by *in silico* translation of the respective coding sequences.

Orthogroup classification and phylogenetic tree inference

The complete set of translated proteomes for species in each respective taxonomic group were subject to orthogroup inference using OrthoFinder V2.5.2[128,129] software run with default settings and with the DIAMOND ultra-sensitive mode[130]. Protein sequences within each orthogroup were aligned using the MAFFT L-INS-I algorithm with 1,000 cycles of iterative refinement[124]. The corresponding codon alignments of the nucleotide sequences were generated by threading the nucleotide sequences through the aligned protein sequences that they encode using PAL2NAL software[125]. Alignments were trimmed to remove positions which contain gap characters. Sequences that were <50% of the median length of the cohort of all other sequences in the given orthogroup were excluded to avoid analysis of partial or truncated genes that could influence downstream analysis. All nucleotide and protein multiple sequence alignments which satisfied the above criteria and which

possessed >50 ubiquitously aligned codon or amino acid positions were subject to bootstrapped maximum likelihood phylogenetic tree inference using IQ-TREE[126] following the exact method and evolutionary substitution models described above. In total, this resulted in a combined set of 16,631 orthogroup phylogenies comprising 5,126,017 ortholog pairwise comparisons across 351 species pairwise comparisons for the land plant clade, 6,953 orthogroup phylogenies comprising 153,288 ortholog pairwise comparisons across 28 species pairwise comparisons for the SAR clade, 5,422 orthogroup phylogenies comprising 642,057 ortholog pairwise comparisons across 496 species pairwise comparisons for the bacteria clade, 4,269 orthogroup phylogenies comprising 31,133 ortholog pairwise comparisons across 6 species pairwise comparisons for the green algae clade and 3,966 orthogroup phylogenies comprising 54,091 ortholog pairwise comparisons across 15 species pairwise comparisons for the red algae clade, from which to base the analyses herein. A further breakdown of these metrics for each species comparison can be found in Supplemental File 4.

Characterization of the set of enzymatic gene and protein sequences within orthogroups

The set of all genes within each species proteome that encode enzymes was determined using the DeepEC[131] deep learning-based classifier algorithm. For this purpose, enzymes were defined as those protein sequences that could be assigned at least a partial enzyme commission (EC) number (i.e., at minimum, a single digit EC top-level code). On average 42.2% of all genes in the analysis encoded enzymes. A detailed breakdown of the metrics of enzyme ortholog pairwise comparisons for each species comparison can be found in Supplemental File 4.

Quantification of percentiles of the rate of molecular evolution

To evaluate the extent of molecular evolution in rubisco in the context of all other genes, only species in the rubisco sequence dataset possessing a publicly available whole-genome gene assembly were considered. Across each pairwise combination of species in a given taxonomic group which satisfied this criteria, the extent of *rbcL*/RbcL and *rbcS*/RbcS molecular evolution since the time point of species divergence was measured by computing the sum of branch lengths (sequence substitutions per aligned sequence site) separating these respective sequences in the rubisco phylogenetic trees previously inferred. Following this, the extent of molecular evolution separating all other pairs of

orthologous (but not paralogous) gene and protein sequences for that given species pair was measured across all inferred orthogroup phylogenies, and the percentile rank rate of rubisco nucleotide or protein evolution was computed relative to the cohort of these measurements. To assess the extent of rubisco molecular evolution in context all other enzymes, the exact same steps were followed but only the subset of genes and proteins predicted to encode enzymes were included. In both of these analyses, a minimum threshold of 100 measurements for orthologous genes and protein sequences was ensured per species pair. In cases where multiple percentiles are calculated for a rubisco subunit in a given species pair (due to gene duplications in the *rbcS* of some species, or due to a single species gene assembly matching multiple sub-species in the rubisco sequence dataset) the mean percentile was taken. The full set of data generated from these analyses quantifying the relative percentile extent of rubisco molecular evolution to all other genes and proteins, and to all other enzyme-encoding genes and proteins can be found in raw and processed forms in Supplemental File 5.

Identification and classification Calvin-Benson-Bassham cycle enzyme isoforms in land plants

The set of genes which encode Calvin-Benson-Bassham cycle enzymes was first resolved in the model plant species *Arabidopsis thaliana*. To achieve this, the complete gene families to which each Calvin-Benson-Bassham cycle enzyme in *A. thaliana* belongs was determined based on available data in The Arabidopsis Information Resource (TAIR) database (http://arabidopsis.org)[132,133]. Following this, the photosynthetic isoforms in these gene families which are active in the Calvin-Benson-Bassham cycle in the chloroplast stroma were then identified based on several lines of evidence. 1) A high protein abundance based on whole-organism integrated protein abundance data obtained from the Protein Abundance Database (https://pax-db.org/) dataset 3702/323. 2) Leaf mRNA expression based on tissue-specific RNA sequencing data obtained from both the Arabidopsis eFP Browser V2.0 (http://bar.utoronto.ca/efp2/Arabidopsis/Arabidopsis_eFPBrowser2.html) and the EMBL-EBI (https://www.ebi.ac.uk/) dataset E-GEOD-53197. 3) Chloroplast-targeted protein subcellular localisation as predicted using both TargetP V2.0[134,135] and Predotar V1.04[136]. 4) Gene orthology as

inferred from trees generated for each Calvin-Benson-Bassham cycle gene family using IQ-TREE[126]. The resulting set of photosynthetic isoforms encoding each Calvin-Benson-Bassham cycle enzyme in *A. thaliana* inferred from this multi-faceted analytical pipeline can be found in Supplemental File 1, table S8.

The set of genes which encode the photosynthetic isoforms of Calvin Bensen Bassham cycle enzymes in all other 26 land plant species (apart from *A. thaliana*) for which genome sequence data was available in this study were then determined by orthology using data from the orthogroup inference analysis performed above. Each group of orthologous protein sequences determined to encode a given Calvin-Benson-Bassham cycle enzyme across all species were aligned using the MAFFT L-INS-i algorithm[124], and corresponding nucleotide coding sequence alignments were generated using PAL2NAL[125]. Multiple sequence alignments were subject to the same data filtering and quality control criteria previously described to remove partial or incomplete sequences and subsequently delete any column positions which contain gaps. Finally, bootstrapped maximum likelihood phylogenetic trees were inferred by IQ-TREE[126] following the method outlined above. A similar analysis of Calvin-Bensen-Bassham cycle enzymes in other taxonomic groups was not able to be performed owing to a lack of the required data described here to determine the photosynthetic gene isoforms in these species.

Quantification of the relative extent of rbcL/RbcL and rbcS/RbcS molecular evolution relative to all Calvin-Benson-Bassham cycle enzymes in land plants

To determine the extent of molecular evolution between orthologous Calvin-Benson-Bassham cycle isoforms compared to Form I rubisco, the extent of molecular evolution measured between all other Calvin-Benson-Bassham cycle gene and protein orthologous sequences for each species pair were expressed as a percentage ratio of that measured in the corresponding *rbcL*/RbcL sequence. In cases where multiple percentage ratios are calculated for a given Calvin-Bensen-Bassham component in a given species pair (due to gene duplications, or due to a single species gene assembly matching multiple sub-species in the rubisco sequence dataset) the mean value was taken. The full set of data generated from this analysis can be found in raw and processed forms in Supplemental File 5.

Quantification of the percentile extent of rubisco chaperone molecular evolution within each taxonomic group

To evaluate the percentile rate of molecular evolution in the known chaperones of Form I rubisco in the context of all other genes in each taxonomic group, the exact same method was followed as above for *rbcL*/RbcL and *rbcS*/RbcS though the subject of the analysis was respectively altered. Here, for this investigation, the putative set of genes which encode each Form I ancillary chaperone involved in holoenzyme metabolic regulation (RUBISCO ACTIVASE (Rca)) and in holoenzyme folding and assembly (BUNDLE SHEATH DEFECTIVE 2 (BSD2), CHAPERONIN 10 (Cpn10), CHAPERONIN 20 (Cpn20), CHAPERONIN-60 (Cpn60), RBCX (RbcX), RUBISCO ACCUMULATION FACTOR 1 (Raf1), RUBISCO ASSEMBLY FACTOR 2 (Raf2)) were first resolved in the model plant species *A. thaliana*. This was achieved using a previously published dataset[137] supplemented by information available in the TAIR database (http://arabidopsis.org)[132,133]. The resulting set of *Arabidopsis* chaperone genes thus identified can be found in Supplemental File 1, table S9. Following this step, the corresponding set of genes encoding rubisco chaperones in all other species for which a complete gene assembly could be obtained were inferred using data from a separate OrthoFinder run performed with identical settings as previously described, but based on the translated proteomes of all organisms across all taxonomic groups.

After the cohort of Form I rubisco chaperone genes were identified in all species, the percentile rates of nucleotide and protein evolution in these genes were calculated between each pairwise combination of species relative to all other pairs of orthologous sequences using the identical measurements previously generated from the analysis of *rbcL*/RbcL and *rbcS*/RbcS above. In this way, analysis of some chaperones were omitted in certain taxonomic groups owing to the data quality and filtering steps that were previously performed as described above. In cases where multiple percentiles are calculated for a chaperone in a given species pair (due to gene duplications, or due to a single species gene assembly matching multiple sub-species in the rubisco sequence dataset) the mean percentile was taken as above. The full set of data generated from this analysis can be found in raw and processed forms in Supplemental File 5. A combined dataset including the

relative percentile extent of evolution in both rubisco subunits and all rubisco chaperones for each unique pairwise species comparison can be found in Supplemental File 6.

Integrated analysis of rubisco molecular and kinetic evolution

To interrogate the relationship between the molecular and kinetic evolution of extant Form I rubisco, a dataset of rubisco kinetic traits was downloaded from[21,22], as modified from that originally compiled by Flamholz and colleagues[81]. For the purpose of this study, only species in this dataset with a complete set of experimentally determined measurements of rubisco specificity ($S_{C/O}$) for CO_2 relative to O_2 (i.e., the overall carboxylation/oxygenation ratio of rubisco under defined concentrations of CO_2 and O_2 gases), maximum carboxylase turnover rate per active site (k_{catC}), and the respective Michaelis constant (i.e., the substrate concentration at half-saturated catalysed rate) for both CO_2 (K_C) and O_2 (K_O) substrates were selected. For each of the 137 species which satisfied this criteria (all of which were angiosperm land plants), an estimate of the Michaelis constant for CO_2 in 20.95% O_2 air (K_{Cair}) was also available[21,22]. In addition, the ratio of the Michaelis constant for CO_2 relative to O_2 (K_C/K_O), as well as carboxylation efficiency, defined as the ratio of the maximum carboxylase turnover to the Michaelis constant for CO_2 (k_{catC}/K_C), were also inferred. Measurements of the Michaelis constant for RuBP (K_{RuBP}) were not considered owing to a limited sample size (n = 19). All *Limonium* species in the dataset were also ignored on the basis that trait values obtained across different studies have been deemed to not be consistent[24,138]. In total, this left a dataset of rubisco kinetic trait measurements for 123 angiosperms. Of these, only the subset of 93 species which perform C_3 photosynthesis were considered for the purpose of the integrated molecular and kinetic evolution analysis herein. This is because of both a limited sample size of C_3-C_4 (n = 6), C_4-like species (n = 3) and C_4 species (n = 21) in the kinetic dataset, and given that transition toward C_4 photosynthesis is associated with a change in rubisco kinetic evolution[21,22] that would confound the directional selection analysis being conducted.

Coding sequences of the *rbcL* gene were obtained from[22] for each species in the kinetic dataset. In order to facilitate more accurate downstream phylogenetic tree inference across these sequences and to minimize the impact of long-branch effects[139], the complete set of publicly available *rbcL* coding sequences in land plants were also acquired in parallel from NCBI

(https://www.ncbi.nlm.nih.gov/) using the query term "rbcL[Gene Name] AND "plants"[porgn:_txid3193]". These sequences thus obtained were subject to the exact same data processing steps to remove ambiguous, partial or chimeric sequences as performed previously for the *rbcL* sequences of species in the rubisco kinetic dataset[22]. In total, this step resulted in an additional set of 29,218 full-length *rbcL* coding sequences to aid downstream phylogenetic inference. Protein sequences were inferred from each *rbcL* coding sequences via *in silico* translation. Next, the complete set of translated RbcL sequences (including the set of sequences from angiosperms in the rubisco kinetic dataset, as well as the set of all publicly available sequences for land plants) were respectively aligned using MAFFT L-INS-I[124], and a corresponding *rbcL* coding sequence alignment was generated using PAL2NAL software[125]. The resulting multiple sequence alignments were trimmed to remove non-aligned residue positions and bootstrapped phylogenetic trees were inferred using IQ-TREE[126] following the exact method described above and using the best-fit models of nucleotide and protein sequence evolution previously inferred. To facilitate downstream analysis, the *rbcL* and RbcL gene trees were subsequently modified to keep only internal and terminal branches leading to the set of species in the rubisco kinetic dataset, with pruned trees manually rooted in Dendroscope[140].

To compute the relative extent of protein evolution which has occurred in each angiosperm in the kinetic dataset, the summed branch length (sequence substitutions per aligned sequence site) leading from the last common ancestor at the root of this clade to each respective terminal node in the RbcL phylogeny generated above were measured. The kinetic trait values and extent of molecular evolution for all C_3 angiosperm rubisco can be found in Supplemental File 7. The predicted kinetic trait values at the last common ancestor at the base of the angiosperm clade were inferred from the estimated y intercept values from these regression models and can be found in table 1. The *rbcL*/RbcL phylogenetic gene trees used as the basis of this analysis, including the trees inferred across the full set of sequences, as well as the pruned versions of these trees containing only the subset of C_3 species in the kinetic dataset, can be found in Supplemental File 8.

To assess the robustness of the above integrated molecular and kinetic investigation of rubisco evolution, the same analysis was repeated but including only the minimal subset of species in the

kinetic dataset which captured the majority of phylogenetic diversity across all sampled species (see below), so as to control for biases associated with species sampling and overrepresentation of certain groups (see Supplemental File 1). An identical analysis was also performed using the complete set of species in the kinetic dataset but based on analogous trees generated following the exact same method as above but based on alternate best-fitting models of sequence evolution inferred for the specific alignment, so as to control for potential artefacts associated with errors or uncertainties in phylogenetic tree inference (Supplemental File 1). As the results of these supplementary analyses were identical to that generated from our original analysis, our conclusions were demonstrated to be valid and robust and not an artefact caused by either systematic biases in species sampling or by errors in phylogenetic reconstruction.

Accounting for potential species sampling error

To identify a minimal subset of species which capture all of the phylogenetic diversity (PD) contributed to by the complete set of 93 C_3 species in the rubisco kinetic dataset, the Phylogenetic Diversity Analyzer V1.0.3 software[142] was employed using the 'greedy' algorithm. Specifically, the unrooted RbcL phylogenetic tree of the 93 C_3 species was subject to systematic interrogation to identify the optimal combination of species at each iterative tree size (n = 2 – 93 species) which maximizes the PD score. For the purpose of this method, PD is defined as the total tree length (i.e., the combined sum of all internal and terminal branch lengths) of the pruned phylogeny comprising the selected subset of sampled species. Based on the results of this analysis, it was observed that phylogenetic diversity saturated at a threshold of 50 species (accounting for 54.8% of the total 93 C_3 species in the kinetic dataset) (Supplemental File 1). The optimal composition of species at this respective threshold included 31 dicotyledonous individuals and 19 monocotyledonous individuals, and are listed in Supplemental File 1, table S10. This set of 50 species were used to assess the robustness of the molecular and kinetic analysis of rubisco to potential artefacts associated with biases in species sampling.

Integrated analysis of rubisco molecular evolution and CO_2 assimilation

To investigate the relationship between rubisco molecular evolution and whole-plant photosynthetic performance, a comprehensive meta-dataset of photosynthetic measurements from species

spanning the whole land plant phylogeny was provided by Gago and colleagues[85]. This dataset contained measurements of light-saturated net photosynthetic rates expressed both per unit leaf mass (A_{mass}) and per unit leaf area (A_{area}), as well as measurements of total nitrogen content expressed both per unit leaf mass (N_{mass}) and per unit leaf area (N_{area}). In addition, for each unique species observation in this dataset with a corresponding measurement for both A_{mass} and N_{mass} or for both A_{area} and N_{area}, the mass-based and area-based photosynthetic nitrogen-use efficiencies were also derived using the calculations A_{mass}/N_{mass} ($PNUE_{mass}$) and A_{area}/N_{area} ($PNUE_{area}$), respectively. In cases where duplicate entries for a parameter were present across species, the mean value was taken so as to collapse the dataset to contain only a single row per species. Finally, although photosynthetic measurements were available from individuals belonging to all major land plant lineages (including the mosses, liverworts, fern allies, ferns, gymnosperms, and angiosperms), only the subset of angiosperms in the dataset for which a publicly available *rbcL* sequence could be obtained were taken forward. This is because various diffusional and biochemical factors other than rubisco are known to cause reduced photosynthetic capacities in non-angiosperm plants[85] that would bias the results of the current study. For the same reasons, only the subset of C_3 angiosperms in this dataset were taken forward in the present analysis to avoid picking up photosynthetic effects which result from CO_2 concentrating mechanisms that act upstream of rubisco. In total, this left a photosynthetic dataset of 366 C_3 angiosperms from which to base the analyses herein. Combined, this resulting dataset included 272 unique species measurements for N_{mass}, 137 unique species measurements for A_{mass} and 118 unique species measurements for $PNUE_{mass}$, as well as 270 unique species measurements for N_{area}, 151 unique species measurements for A_{area}, and 120 unique species measurements for $PNUE_{area}$, respectively.

To compute the relative extents of RbcL molecular evolution which has occurred in each angiosperm in the photosynthetic dataset, the exact same method was followed as described above. First, the full RbcL phylogenetic gene tree in Supplemental File 8 that was previously inferred from the complete set of publicly available RbcL sequences in NCBI was pruned so as to contain only terminal and internal branches corresponding to angiosperms in the photosynthetic dataset. Here, in situations where duplicate sequences in the alignment resulted in multiple terminal nodes for a given

species, only a single node was retained based on the sequence which is first in the alphabetical order of the gene accession numbers. As above, this reduced RbcL tree was then manually rooted in Dendroscope[140], and the relative extent of RbcL protein evolution in each angiosperm was computed as the summed branch length (sequence substitutions per aligned sequence site) leading from the last common ancestor at the root of this clade to each respective terminal node. Finally, linear regression models were employed to assess the pairwise relationships between the variation in rubisco molecular evolution and each respective photosynthetic parameter. The resulting full integrated dataset containing photosynthetic measurements and comparable extents of RbcL molecular evolution for all 366 C_3 angiosperms can be found in Supplemental File 9. The RbcL phylogenetic gene tree which has been pruned from that in Supplemental File 8 to contain the subset of C_3 angiosperms in the photosynthetic dataset used for the basis of this analysis can be found in Supplemental File 10.

References

1. Canfield, D. E. The early history of atmospheric oxygen: Homage to Robert M. Garrels. *Annu. Rev. Earth Planet. Sci.* **33**, 1–36 (2005).

2. Kasting, J. F. Earth's early atmosphere. *Science (80-.).* **259**, 920–926 (1993).

3. Kaufman, A. J. *et al.* Late archean biospheric oxygenation and atmospheric evolution. *Science (80-.).* **317**, 1900–1903 (2007).

4. Nisbet, E. G. *et al.* The age of Rubisco: The evolution of oxygenic photosynthesis. *Geobiology* **5**, 311–335 (2007).

5. Nisbet, E. G. & Nisbet, R. E. R. Methane, oxygen, photosynthesis, rubisco and the regulation of the air through time. *Philos. Trans. R. Soc. B Biol. Sci.* **363**, 2745–2754 (2008).

6. Shih, P. M. *et al.* Biochemical characterization of predicted Precambrian RuBisCO. *Nat. Commun.* **7**, (2016).

7. Ślesak, I., Ślesak, H. & Kruk, J. RubisCO Early Oxygenase Activity: A Kinetic and Evolutionary Perspective. *BioEssays* **39**, (2017).

8. Holland, H. D. Volcanic gases, black smokers, and the great oxidation event. *Geochim. Cosmochim. Acta* **66**, 3811–3826 (2002).

9. Ashida, H. *et al.* RuBisCO-like proteins as the enolase enzyme in the methionine salvage pathway: Functional and evolutionary relationships between RuBisCO-like proteins and photosynthetic RuBisCO. *J. Exp. Bot.* **59**, 1543–1554 (2008).

10. Ashida, H., Danchin, A. & Yokota, A. Was photosynthetic RuBisCO recruited by acquisitive evolution from RuBisCO-like proteins involved in sulfur metabolism? *Res. Microbiol.* **156**, 611–618 (2005).

11. Erb, T. J. & Zarzycki, J. A short history of RubisCO: the rise and fall (?) of Nature's predominant CO2 fixing enzyme. *Current Opinion in Biotechnology* vol. 49 100–107 https://reader.elsevier.com/reader/sd/pii/S095816691730099X?token=3ECF17E8852AFD765C4445B954C57E5AB9ED5265BF433EACA574481D29827EF2C02377F044E4135B5CB4

B897D3E0478A (2018).

12. Banda, D. M. *et al.* Novel bacterial clade reveals origin of form I Rubisco. *Nat. Plants* **6**, 1158–1166 (2020).

13. Tabita, F. R., Satagopan, S., Hanson, T. E., Kreel, N. E. & Scott, S. S. Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships. *J. Exp. Bot.* **59**, 1515–1524 (2008).

14. Andersson, I. & Backlund, A. Structure and function of Rubisco. *Plant Physiol. Biochem.* **46**, 275–291 (2008).

15. Whitney, S. M., Houtz, R. L. & Alonso, H. Advancing our understanding and capacity to engineer nature's CO2-sequestering enzyme, Rubisco. *Plant Physiol.* **155**, 27–35 (2011).

16. Tabita, F. R., Hanson, T. E., Satagopan, S., Witte, B. H. & Kreel, N. E. Phylogenetic and evolutionary relationships of RubisCO and the RubisCO-like proteins and the functional lessons provided by diverse molecular forms. *Philos. Trans. R. Soc. B Biol. Sci.* **363**, 2629–2640 (2008).

17. Schneider, G., Lindqvist, Y. & Brändén, C. I. RUBISCO: Structure and mechanism. *Annual Review of Biophysics and Biomolecular Structure* vol. 21 119–143 at https://doi.org/10.1146/annurev.bb.21.060192.001003 (1992).

18. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science (80-.).* **281**, 237–240 (1998).

19. Lee, B. & Tabita, F. R. Purification of Recombinant Ribulose-l,5-bisphosphate Carboxylase/Oxygenase Large Subunits Suitable for Reconstitution and Assembly of Active L8S8 Enzyme. *Biochemistry* **29**, 9352–9357 (1990).

20. Andrews, T. J. Catalysis by cyanobacterial ribulose-bisphosphate carboxylase large subunits in the complete absence of small subunits. *J. Biol. Chem.* **263**, 12213–12219 (1988).

21. Bouvier, J. W. & Kelly, S. Response to Tcherkez and Farquhar: Rubisco adaptation is more

limited by phylogenetic constraint than by catalytic trade-off. *J. Plant Physiol.* **287**, 154021 (2023).

22. Bouvier, J. W. *et al.* Rubisco Adaptation Is More Limited by Phylogenetic Constraint Than by Catalytic Trade-off. *Mol. Biol. Evol.* **38**, 2880–2896 (2021).

23. Galmés, J. *et al.* Expanding knowledge of the Rubisco kinetics variability in plant species: Environmental and evolutionary trends. *Plant, Cell Environ.* **37**, 1989–2001 (2014).

24. Galmés, J. *et al.* Environmentally driven evolution of Rubisco and improved photosynthesis and growth within the C3 genus Limonium (Plumbaginaceae). *New Phytol.* **203**, 989–999 (2014).

25. Chatterjee, A. & Anindya, B. Rubisco: Limitations and Re-engineering for a Better Enzyme. *Int. Res. J. Plant Sci.* **2**, 022–024 (2011).

26. Christin, P. A. *et al.* Evolutionary switch and genetic convergence on rbcL following the evolution of C4 photosynthesis. *Mol. Biol. Evol.* **25**, 2361–2368 (2008).

27. Kapralov, M. V., Kubien, D. S., Andersson, I. & Filatov, D. A. Changes in Rubisco kinetics during the evolution of C4 Photosynthesis in Flaveria (Asteraceae) are associated with positive selection on genes encoding the enzyme. *Mol. Biol. Evol.* **28**, 1491–1503 (2011).

28. Parry, M. A. J., Andralojc, P. J., Mitchell, R. A. C., Madgwick, P. J. & Keys, A. J. Manipulation of Rubisco: The amount, activity, function and regulation. *J. Exp. Bot.* **54**, 1321–1333 (2003).

29. Whitney, S. M. *et al.* Isoleucine 309 acts as a C 4 catalytic switch that increases ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco) carboxylation rate in flaveria. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 14688–14693 (2011).

30. Zhou, Y., Gunn, L. H., Birch, R., Andersson, I. & Whitney, S. M. Grafting Rhodobacter sphaeroides with red algae Rubisco to accelerate catalysis and plant growth. *Nat. Plants* **9**, 978–986 (2023).

31. Kapralov, M. V. & Filatov, D. A. Widespread positive selection in the photosynthetic Rubisco enzyme. *BMC Evol. Biol.* **7**, (2007).

32. Lin, M. T. *et al.* A procedure to introduce point mutations into the Rubisco large subunit gene in wild-type plants. *Plant J.* **106**, 876–887 (2021).

33. Spreitzer, R. J. Role of the small subunit in ribulose-1,5-bisphosphate carboxylase/oxygenase. *Arch. Biochem. Biophys.* **414**, 141–149 (2003).

34. Schulz, L. *et al.* Evolution of increased complexity and specificity at the dawn of form I Rubiscos. *Science (80-.).* **378**, (2022).

35. Fitchen, J. H., Knight, S., Andersson, I., Branden, C. I. & Mcintosh, L. Residues in three conserved regions of the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase are required for quaternary structure. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 5768–5772 (1990).

36. Genkov, T. & Spreitzer, R. J. Highly conserved small subunit residues influence Rubisco large subunit catalysis. *J. Biol. Chem.* **284**, 30105–30112 (2009).

37. Grabsztunowicz, M., Górski, Z., Luciński, R. & Jackowski, G. A reversible decrease in ribulose 1,5-bisphosphate carboxylase/oxygenase carboxylation activity caused by the aggregation of the enzyme's large subunit is triggered in response to the exposure of moderate irradiance-grown plants to low irradiance. *Physiol. Plant.* **154**, 591–608 (2015).

38. Joshi, J., Mueller-Cajar, O., Tsai, Y. C. C., Hartl, F. U. & Hayer-Hartl, M. Role of small subunit in mediating assembly of red-type Form I Rubisco. *J. Biol. Chem.* **290**, 1066–1074 (2015).

39. Lee, B., Berka, R. M. & Tabita, F. R. Mutations in the small subunit of cyanobacterial ribulose-bisphosphate carboxylase/oxygenase that modulate interactions with large subunits. *J. Biol. Chem.* **266**, 7417–7422 (1991).

40. McFadden, B. A. & Small, C. L. Cloning, expression and directed mutagenesis of the genes for ribulose bisphosphate carboxylase/oxygenase. *Photosynth. Res.* **18**, 245–260 (1988).

41. Schneider, G. *et al.* Comparison of the crystal structures of L2 and L8S8 Rubisco suggests a functional role for the small subunit. *EMBO J.* **9**, 2045–2050 (1990).

42. Van Lun, M., Hub, J. S., Van Der Spoel, D. & Andersson, I. CO2 and O2 distribution in rubisco suggests the small subunit functions as a CO2 reservoir. *J. Am. Chem. Soc.* **136**, 3165–3171

(2014).

43. Kostov, R. V., Small, C. L. & McFadden, B. A. Mutations in a sequence near the N-terminus of the small subunit alter the CO2/O2 specificity factor for ribulose bisphosphate carboxylase/oxygenase. *Photosynth. Res.* **54**, 127–134 (1997).

44. Voordouw, G., De Vries, P. A., Van Den Berg, W. A. M. & De Clerck, E. P. J. Site-directed mutagenesis of the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase from Anacystis nidulans. *Eur. J. Biochem.* **163**, 591–598 (1987).

45. Read, B. A. & Tabita, F. R. Amino Acid Substitutions in the Small Subunit of Ribulose-1, 5-bisphosphate Carboxylase/Oxygenase That Influence Catalytic Activity of the Holoenzyme. *Biochemistry* **31**, 519–525 (1992).

46. Getzoff, T. P., Zhu, G., Bohnert, H. J. & Jensen, R. G. Chimeric Arabidopsis thaliana Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase Containing a Pea Small Subunit Protein Is Compromised in Carbamylation. *Plant Physiol.* **116**, 695–702 (1998).

47. Ishikawa, C., Hatanaka, T., Misoo, S., Miyake, C. & Fukayama, H. Functional incorporation of sorghum small subunit increases the catalytic turnover rate of rubisco in transgenic rice. *Plant Physiol.* **156**, 1603–1611 (2011).

48. Kanevski, I., Maliga, P., Rhoades, D. F. & Gutteridge, S. Plastome engineering of ribulose-1,5-bisphosphate carboxylase/oxygenase in tobacco to form a sunflower large subunit and tobacco small subunit hybrid. *Plant Physiol.* **119**, 133–141 (1999).

49. Matsumura, H. *et al.* Hybrid Rubisco with Complete Replacement of Rice Rubisco Small Subunits by Sorghum Counterparts Confers C4 Plant-like High Catalytic Activity. *Mol. Plant* **13**, 1570–1581 (2020).

50. Sakoda, K. *et al.* Effects of introduction of sorghum RbcS with rice RbcS knockdown by RNAi on photosynthetic activity and dry weight in rice. *Plant Prod. Sci.* **24**, 346–353 (2021).

51. Andrews, T. J. & Lorimer, G. H. Catalytic properties of a hybrid between cyanobacterial large subunits and higher plant small subunits of ribulosebisphosphate carboxylase-oxygenase. *J.*

Biol. Chem. **260**, 4632–4636 (1985).

52. Horken, K. M. & Tabita, F. R. The ‘green’ form I ribulose 1,5-bisphosphate carboxylase/oxygenase from the nonsulfur purple bacterium Rhodobacter capsulatus. *J. Bacteriol.* **181**, 3935–3941 (1999).

53. Lee, B., Read, B. A. & Tabita, F. R. Catalytic properties of recombinant octameric, hexadecameric, and heterologous cyanobacterial/ bacterial ribulose-1,5-bisphosphate carboxylase/oxygenase. *Arch. Biochem. Biophys.* **291**, 263–269 (1991).

54. van der Vies, S. M., Bradley, D. & Gatenby, A. A. Assembly of cyanobacterial and higher plant ribulose bisphosphate carboxylase subunits into functional homologous and heterologous enzyme molecules in Escherichia coli . *EMBO J.* **5**, 2439–2444 (1986).

55. Wang, Y. L., Zhou, J. H., Wang, Y. F., Bao, J. S. & Chen, H. B. Properties of hybrid enzymes between Synechococcus large subunits and higher plant small subunits of ribulose-1,5-bisphosphate carboxylase/oxygenase in Escherichia coli. *Arch. Biochem. Biophys.* **396**, 35–42 (2001).

56. Read, B. A. & Tabita, F. R. A Hybrid Ribulosebisphosphate Carboxylase/Oxygenase Enzyme Exhibiting a Substantial Increase in Substrate Specificity Factor. *Biochemistry* **31**, 5553–5560 (1992).

57. Cavanagh, A. P., Slattery, R. & Kubien, D. S. Temperature-induced changes in Arabidopsis Rubisco activity and isoform expression. *J. Exp. Bot.* (2022) doi:10.1093/jxb/erac379.

58. Martin-Avila, E. *et al.* Modifying plant photosynthesis and growth via simultaneous chloroplast transformation of rubisco large and small subunits. *Plant Cell* **32**, 2898–2916 (2020).

59. Morita, K., Hatanaka, T., Misoo, S. & Fukayama, H. Unusual small subunit that is not expressed in photosynthetic cells alters the catalytic properties of Rubisco in rice. *Plant Physiol.* **164**, 69–79 (2014).

60. Badger, M. R. *et al.* The diversity and coevolution of Rubisco, plastids, pyrenoids, and chloroplast-based CO2-concentrating mechanisms in algae. *Can. J. Bot.* **76**, 1052–1071 (1998).

61. Bar-Even, A. *et al.* The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* **50**, 4402–4410 (2011).

62. Chollet, R. *The biochemistry of photorespiration*. *Trends in Biochemical Sciences* vol. 2 (1977).

63. Sharkey, T. D. Emerging research in plant photosynthesis. *Emerg. Top. Life Sci.* **4**, 137–150 (2020).

64. Bowes, G., Ogren, W. L. & Hageman, R. H. *Phosphoglycolate production catalyzed by ribulose diphosphate carboxylase*. *Biochemical and Biophysical Research Communications* vol. 45 (1971).

65. Busch, F. A. Photorespiration in the context of Rubisco biochemistry, CO2 diffusion and metabolism. *Plant J.* **101**, 919–939 (2020).

66. Eckardt, N. A. Photorespiration revisited. *Plant Cell* vol. 17 2139–2141 Article at https://doi.org/10.1105/tpc.105.035873 (2005).

67. Sharwood, R. E. Engineering chloroplasts to improve Rubisco catalysis: prospects for translating improvements into food and fiber crops. *New Phytologist* vol. 213 494–510 at https://doi.org/10.1111/nph.14351 (2017).

68. Carmo-Silva, E., Scales, J. C., Madgwick, P. J. & Parry, M. A. J. Optimizing Rubisco and its regulation for greater resource use efficiency. *Plant, Cell and Environment* vol. 38 1817–1832 at https://doi.org/10.1111/pce.12425 (2015).

69. Sarles, L. S. & Tabita, F. R. Derepression of the synthesis of D-ribulose 1,5-bisphosphate carboxylase/oxygenase from Rhodospirillum rubrum. *J. Bacteriol.* **153**, 458–464 (1983).

70. Tabita, F. R. Microbial ribulose 1,5-bisphosphate carboxylase/oxygenase: A different perspective. *Photosynth. Res.* **60**, 1–28 (1999).

71. Flamholz, A. & Shih, P. M. Cell biology of photosynthesis over geologic time. *Curr. Biol.* **30**, R490–R494 (2020).

72. Meyer, M. & Griffiths, H. Origins and diversity of eukaryotic CO2-concentrating mechanisms:

Lessons for the future. *Journal of Experimental Botany* vol. 64 769–786 Article at https://doi.org/10.1093/jxb/ers390 (2013).

73. Schlüter, U. *et al.* Brassicaceae display variation in efficiency of photorespiratory carbon recapturing mechanisms. *J. Exp. Bot.* (2023) doi:10.1093/jxb/erad250.

74. Ogren, W. L. Photorespiration: Pathways, Regulation, and Modification. *Annu. Rev. Plant Physiol.* **35**, 415–442 (1984).

75. Orr, D. J. *et al.* Engineering photosynthesis: Progress and perspectives. *F1000Research* vol. 6 at https://doi.org/10.12688/f1000research.12181.1 (2017).

76. Parry, M. A. J. *et al.* Rubisco activity and regulation as targets for crop improvement. *J. Exp. Bot.* **64**, 717–730 (2013).

77. Parry, M. A. J., Madgwick, P. J., Carvalho, J. F. C. & Andralojc, P. J. Prospects for increasing photosynthesis by overcoming the limitations of Rubisco. in *Journal of Agricultural Science* vol. 145 31–43 (2007).

78. Sharwood, R. E., Ghannoum, O. & Whitney, S. M. Prospects for improving CO2 fixation in C3-crops through understanding C4-Rubisco biogenesis and catalytic diversity. *Current Opinion in Plant Biology* vol. 31 135–142 at https://doi.org/10.1016/j.pbi.2016.04.002 (2016).

79. Savir, Y., Noor, E., Milo, R. & Tlusty, T. Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 3475–3480 (2010).

80. Tcherkez, G. G. B., Farquhar, G. D. & Andrews, T. J. Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 7246–7251 (2006).

81. Flamholz, A. I. *et al.* Revisiting Trade-offs between Rubisco Kinetic Parameters. *Biochemistry* vol. 58 3365–3376 Article at https://doi.org/10.1021/acs.biochem.9b00237 (2019).

82. Iñiguez, C. *et al.* Evolutionary trends in RuBisCO kinetics and their co-evolution with CO2 concentrating mechanisms. *Plant J.* **101**, 897–918 (2020).

83. Cummins, P. L. The Coevolution of RuBisCO, Photorespiration, and Carbon Concentrating Mechanisms in Higher Plants. *Front. Plant Sci.* **12**, (2021).

84. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).

85. Gago, J. *et al.* Photosynthesis Optimized across Land Plant Phylogeny. *Trends in Plant Science* vol. 24 947–958 at https://doi.org/10.1016/j.tplants.2019.07.002 (2019).

86. Ogren, W. L. & Bowes, G. Ribulose diphosphate carboxylase regulates soybean photorespiration. *Nat. New Biol.* **230**, 159–160 (1971).

87. APG. An Ordinal Classification for the Families of Flowering Plants. *Annals of the Missouri Botanical Garden* vol. 85 531–553 https://www.researchgate.net/publication/279592674_An_ordinal_classification_for_the_families_of_flowering_plants?enrichId=rgreq-264f96f50187de487ddbc0180b68f4ac-XXX&enrichSource=Y292ZXJQYWdlOzI3OTU5MjY3NDtBUzo0MTYwNDQ5NzI4MjI1MjhAMTQ3NjIwNDI1OTUxNg%3D%3D (1998).

88. APG. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* **181**, 1–20 (2016).

89. Gielly, L. & Taberlet, P. *The use of chloroplast DNA to resolve plant phylogenies: Noncoding versus rbcL sequences*. *Molecular Biology and Evolution* vol. 11 https://academic.oup.com/mbe/article-abstract/11/5/769/1008715 (1994).

90. Pál, C., Papp, B. & Lercher, M. J. An integrated view of protein evolution. *Nat. Rev. Genet.* **7**, 337–348 (2006).

91. Robbins, E. H. J. & Kelly, S. The Evolutionary Constraints on Angiosperm Chloroplast Adaptation. *Genome Biol. Evol.* **15**, (2023).

92. Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 14338–14343 (2005).

93. Gout, J. F., Kahn, D. & Duret, L. The relationship among gene expression, the evolution of

gene dosage, and the rate of protein evolution. *PLoS Genet.* **6**, 20 (2010).

94. Seward, E. A. & Kelly, S. Selection-driven cost-efficiency optimization of transcripts modulates gene evolutionary rate in bacteria. *Genome Biol.* **19**, (2018).

95. Wall, D. P. *et al.* Functional genomic analysis of the rates of protein evolution. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 5483–5488 (2005).

96. Bloom, J. D., Drummond, D. A., Arnold, F. H. & Wilke, C. O. Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.* **23**, 1751–1761 (2006).

97. Echave, J. Beyond Stability Constraints: A Biophysical Model of Enzyme Evolution with Selection on Stability and Activity. *Mol. Biol. Evol.* **36**, 613–620 (2019).

98. Lobkovsky, A. E., Wolf, Y. I. & Koonin, E. V. Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 2983–2988 (2010).

99. Soskine, M. & Tawfik, D. S. Mutational effects and the evolution of new protein functions. *Nat. Rev. Genet.* **11**, 572–582 (2010).

100. Tourasse, N. J. & Li, W. H. Selective constraints, amino acid composition, and the rate of protein evolution. *Mol. Biol. Evol.* **17**, 656–664 (2000).

101. Fraser, H. B., Wall, D. P. & Hirsh, A. E. A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC Evol. Biol.* **3**, 1–6 (2003).

102. Fraser, H. B. & Hirsh, A. E. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. *BMC Evol. Biol.* **4**, 1–5 (2004).

103. Saeed, R. & Deane, C. M. Protein protein interactions, evolutionary rate, abundance and age. *BMC Bioinformatics* **7**, (2006).

104. Tóth-Petróczy, Á. & Tawfik, D. S. Slow protein evolutionary rates are dictated by surface - core association. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 11151–11156 (2011).

105. Cummins, P. L., Kannappan, B. & Gready, J. E. Directions for optimization of photosynthetic carbon fixation: Rubisco's efficiency may not be so constrained after all. *Front. Plant Sci.* **9**,

(2018).

106. Duraõ, P. *et al.* Opposing effects of folding and assembly chaperones on evolvability of Rubisco. *Nat. Chem. Biol.* **11**, 148–155 (2015).

107. Studer, R. A., Christin, P. A., Williams, M. A. & Orengo, C. A. Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2223–2228 (2014).

108. Cummins, P. L., Kannappan, B. & Gready, J. E. Response: Commentary: Directions for Optimization of Photosynthetic Carbon Fixation: RuBisCO's Efficiency May Not Be So Constrained After All. *Front. Plant Sci.* **10**, 1426 (2019).

109. Aigner, H. *et al.* Plant RuBisCo assembly in E. coli with five chloroplast chaperones including BSD2. *Science (80-.).* **358**, 1272–1278 (2017).

110. Whitney, S. M., Birch, R., Kelso, C., Beck, J. L. & Kapralov, M. V. Improving recombinant Rubisco biogenesis, plant photosynthesis and growth by coexpressing its ancillary RAF1 chaperone. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 3564–3569 (2015).

111. Rickaby, R. E. M. & Eason Hubbard, M. R. Upper ocean oxygenation, evolution of RuBisCO and the Phanerozoic succession of phytoplankton. *Free Radic. Biol. Med.* **140**, 295–304 (2019).

112. Lin, M. T., Salihovic, H., Clark, F. K. & Hanson, M. R. Improving the efficiency of Rubisco by resurrecting its ancestors in the family Solanaceae. *Sci. Adv.* **8**, (2022).

113. Farquhar, G. D., von Caemmerer, S. & Berry, J. A. A biochemical model of photosynthetic CO2 assimilation in leaves of C3 species. *Planta* **149**, 78–90 (1980).

114. Bassham, J. A. & Krause, G. H. Free energy changes and metabolic regulation in steady-state photosynthetic carbon reduction. *BBA - Bioenerg.* **189**, 207–221 (1969).

115. Dietz, K. J. & Heber, U. Rate-limiting factors in leaf photosynthesis. I. Carbon fluxes in the calvin cycle. *BBA - Bioenerg.* **767**, 432–443 (1984).

116. Fichtner, K. *et al.* Decreased ribulose-1,5-bisphosphate carboxylase-oxygenase in transgenic

tobacco transformed with ‘antisense’ rbcS - V. Relationship between photosynthetic rate, storage strategy, biomass allocation and vegetative plant growth at three different nitrogen s. *Planta* **190**, 1–9 (1993).

117. Furbank, R. T., Chitty, J. A., Von Caemmerer, S. & Jenkins, C. L. D. Antisense RNA Inhibition of RbcS Gene Expression Reduces Rubisco Level and Photosynthesis in the C4 Plant Flaveria bidentis. *Plant Physiol.* **111**, 725–734 (1996).

118. Hudson, G. S., Evans, J. R., Von Caemmerer, S., Arvidsson, Y. B. C. & Andrews, T. J. Reduction of ribulose-1,5-bisphosphate carboxylase/oxygenase content by antisense RNA reduces photosynthesis in transgenic tobacco plants. *Plant Physiol.* **98**, 294–302 (1992).

119. Maheshwari, C. *et al.* Targeted knockdown of ribulose-1, 5-bisphosphate carboxylase-oxygenase in rice mesophyll cells. *J. Plant Physiol.* **260**, (2021).

120. Quick, W. P. *et al.* Decreased ribulose-1,5-bisphosphate carboxylase-oxygenase in transgenic tobacco transformed with ‘antisense’ rbcS - I. Impact on photosynthesis in ambient growth conditions. *Planta* **183**, 542–554 (1991).

121. Stitt, M. *et al.* Decreased ribulose-1,5-bisphosphate carboxylase-oxygenase in transgenic tobacco transformed with ‘antisense’ rbcS - II. Flux-control coefficients for photosynthesis in varying light, CO2, and air humidity. *Planta* **183**, 555–566 (1991).

122. Poolman, M. G., Fell, D. A. & Thomas, S. Modelling photosynthesis and its control. *J. Exp. Bot.* **51**, 319–328 (2000).

123. Zhu, X. G., De Sturler, E. & Long, S. P. Optimizing the distribution of resources between enzymes of carbon metabolism can dramatically increase photosynthetic rate: A numerical simulation using an evolutionary algorithm. *Plant Physiol.* **145**, 513–526 (2007).

124. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* vol. 30 772–780 Article at https://doi.org/10.1093/molbev/mst010 (2013).

125. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609 (2006).

126. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* vol. 32 268–274 Article at https://doi.org/10.1093/molbev/msu300 (2015).

127. Goodstein, D. M. *et al.* Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, (2012).

128. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).

129. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 1–14 (2015).

130. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2014).

131. Ryu, J. Y., Kim, H. U. & Lee, S. Y. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 13996–14001 (2019).

132. Swarbreck, D. *et al.* The Arabidopsis Information Resource (TAIR): Gene structure and function annotation. *Nucleic Acids Res.* **36**, 1009–1014 (2008).

133. Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res.* **40**, D1202–D1210 (2012).

134. Armenteros, J. J. A. *et al.* Detecting sequence signals in targeting peptides using deep learning. *Life Sci. Alliance* **2**, (2019).

135. Emanuelsson, O., Nielsen, H., Brunak, S. & Von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300**, 1005–1016 (2000).

136. Small, I., Peeters, N., Legeai, F. & Lurin, C. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* **4**, 1581–1590 (2004).

137. Gruber, A. V. & Feiz, L. Rubisco assembly in the chloroplast. *Front. Mol. Biosci.* **5**, (2018).

138. Galmés, J. *et al.* Rubisco specificity factor tends to be larger in plant species from drier habitats and in species with persistent leaves. *Plant, Cell Environ.* **28**, 571–579 (2005).

139. Su, Z. & Townsend, J. P. Utility of characters evolving at diverse rates of evolution to resolve quartet trees with unequal branch lengths: Analytical predictions of long-branch effects. *BMC Evolutionary Biology* vol. 15 86 Article at https://doi.org/10.1186/s12862-015-0364-7 (2015).

140. Huson, D. H. & Scornavacca, C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Systematic Biology* vol. 61 1061–1067 Article at https://doi.org/10.1093/sysbio/sys062 (2012).

141. Orr, D. J. *et al.* Surveying rubisco diversity and temperature response to improve crop photosynthetic efficiency. *Plant Physiol.* **172**, 707–717 (2016).

142. Chernomor, O. *et al.* Split diversity in constrained conservation prioritization using integer linear programming. *Methods Ecol. Evol.* **6**, 83–91 (2015).

Funding

This work was funded by the Royal Society and the European Union's Horizon 2020 research and innovation program under grant agreement number 637765. JWB was funded by the BBSRC through BB/J014427/1. This research was funded in whole, or in part, by the BBSRC number BB/J014427/1. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

Data Availability

All data used in this study is provided in the supplemental material.

Author Contributions

JWB and SK conceived the study and wrote and edited the manuscript. JWB conducted the analysis. DME provided advice on design and implementation of the study.

Figures

Figure 1

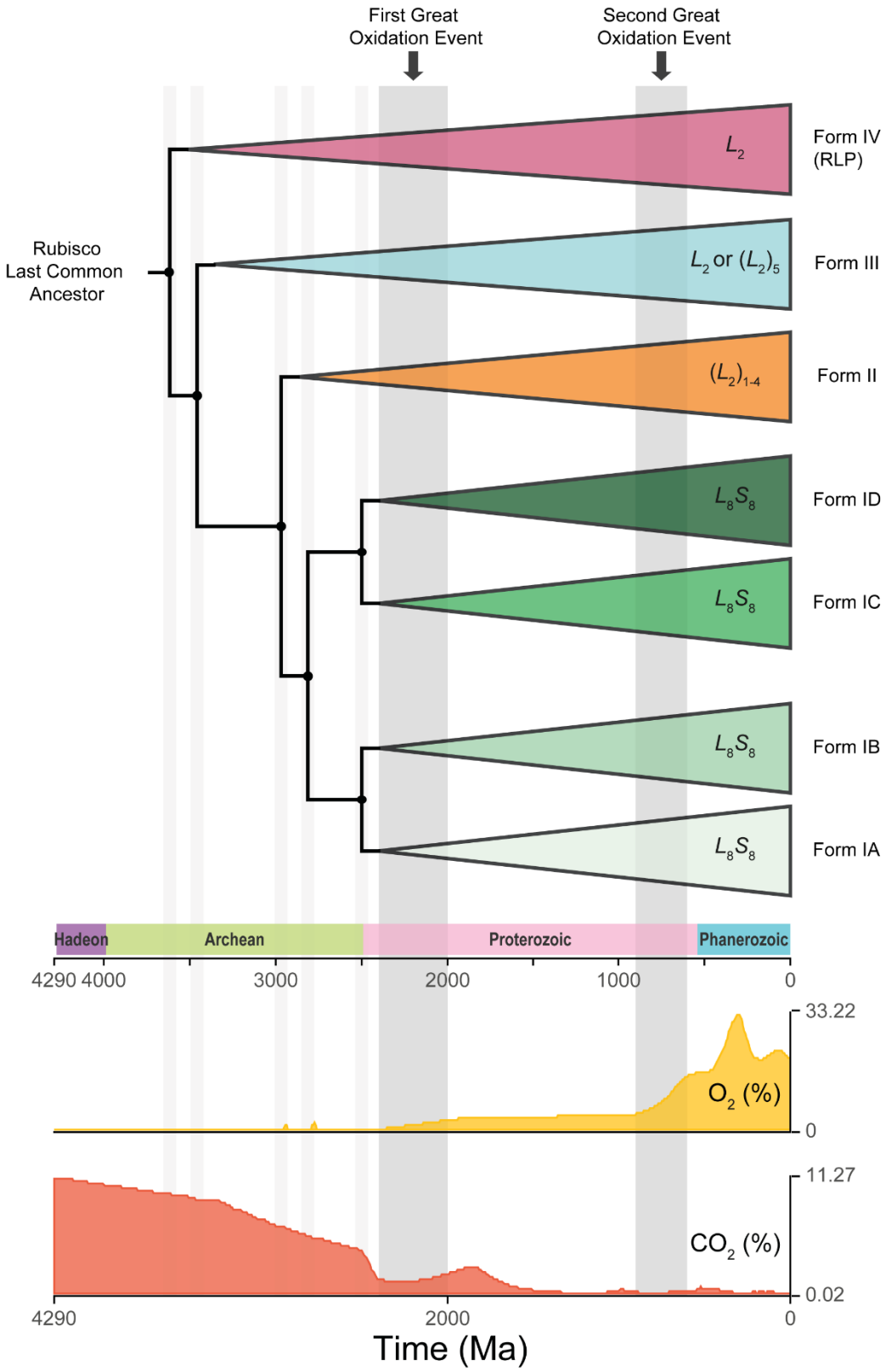
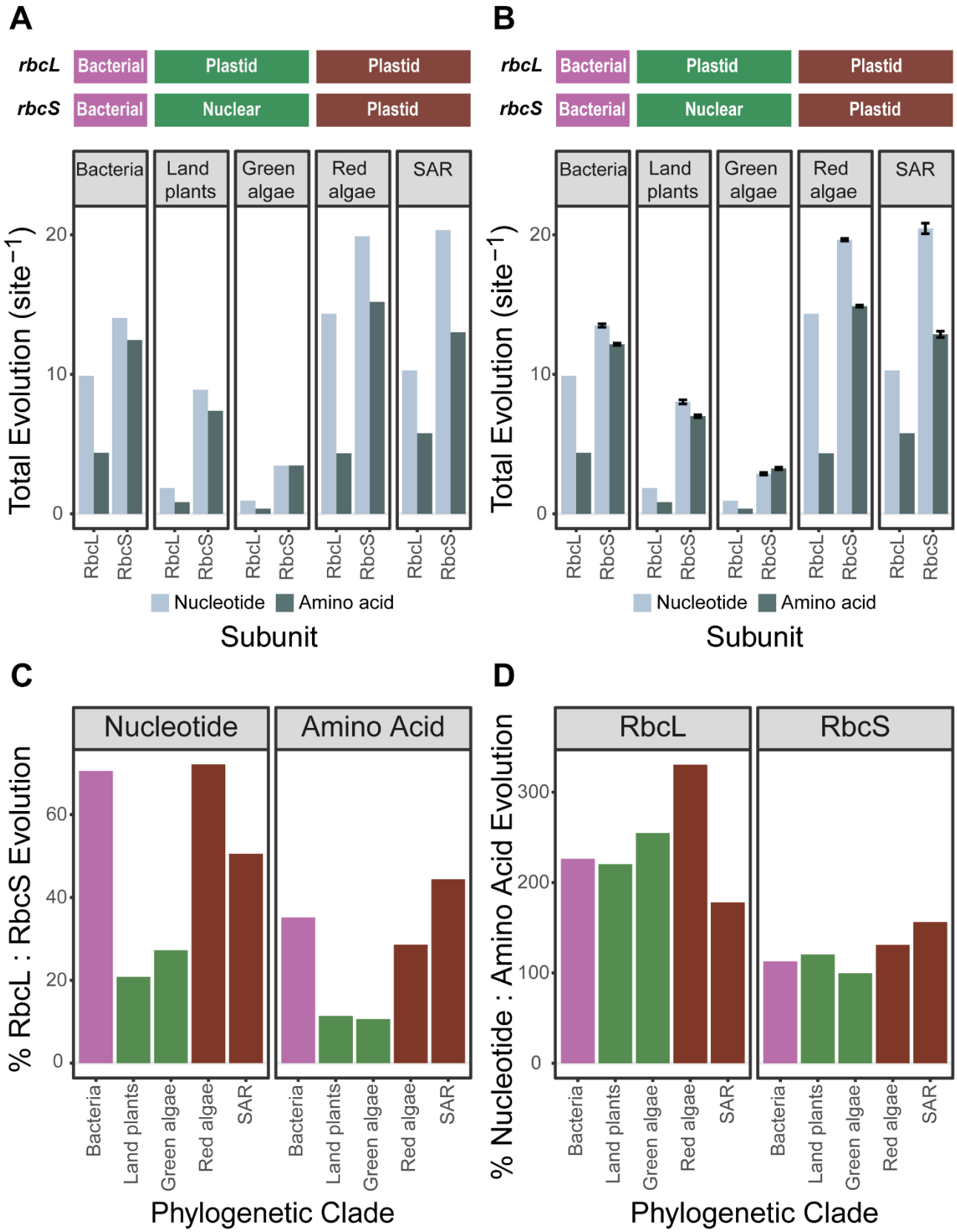


Figure 1. The evolutionary history of rubisco in the context of atmospheric CO_2 (%) and O_2 (%) following divergence from the ancestral rubisco-like protein (RLP). For ease of visualisation, branch points in the phylogeny are indicated by grey vertical bars. The First and Second Great Oxidation events are also indicated by grey vertical bars and have been labelled. Graphics of atmospheric CO_2 and O_2 levels were adapted from the *TimeTree* resource (http://www.timetree.org; [84]).

936 ***Figure 2***



937

Figure 2. The extent of molecular evolution in rubisco during the radiation of each taxonomic group. **A)** Bar plot depicting the total amount of molecular evolution (substitutions per sequence site) in the nucleotide and protein sequences of the rubisco large (*rbcL*/RbcL) and small (*rbcS*/RbcS) subunit across taxonomic groups. The genome in which *rbcL* and *rbcS* genes reside within each group is indicated above the plot (bacterial, plastid, nuclear). **B)** As in (A) but using 1,000 bootstrapped stratified sampling of *rbcS*/RbcS per species to account for the higher copy number of this gene as compared to *rbcL*/RbcL in the dataset used for analysis (see Methods). Error bars represent ± 1 S.E of the mean. **C)** Bar plot depicting the percentage ratio (%) of nucleotide and amino acid evolution between rubisco subunits (*rbcL* to *rbcS* and RbcL to RbcS, respectively) in each taxonomic group. The colour of each bar is determined by the genome in which the *rbcL* and *rbcS* gene resides, following the colour scale in (A) and (B). **D)** Bar plot depicting the percentage ratio (%) of nucleotide to amino acid evolution in each rubisco subunit (*rbcL* to RbcL and *rbcS* to RbcS, respectively) in each taxonomic group. The colour of each bar is the same as described in (C).

Figure 3

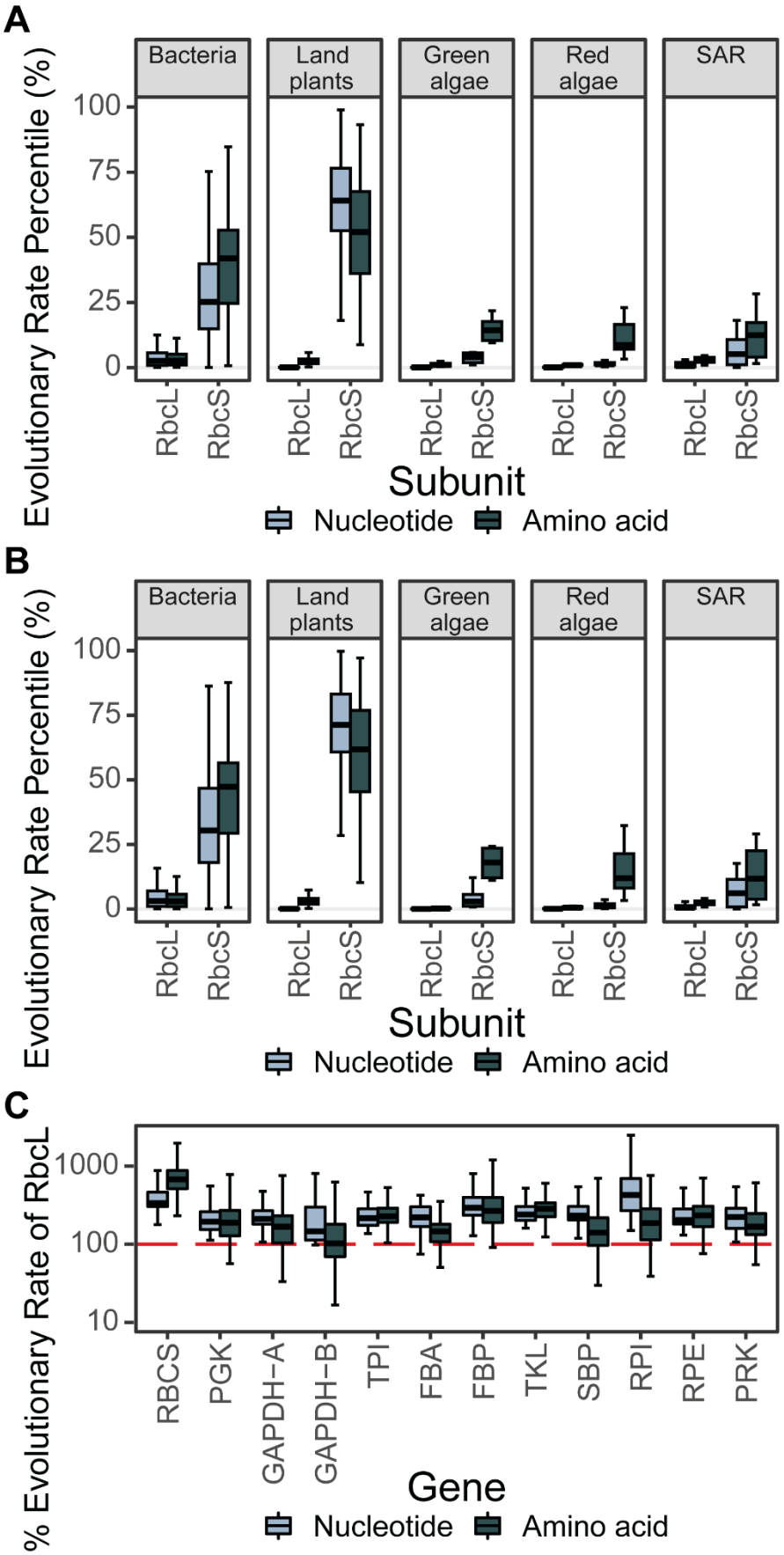
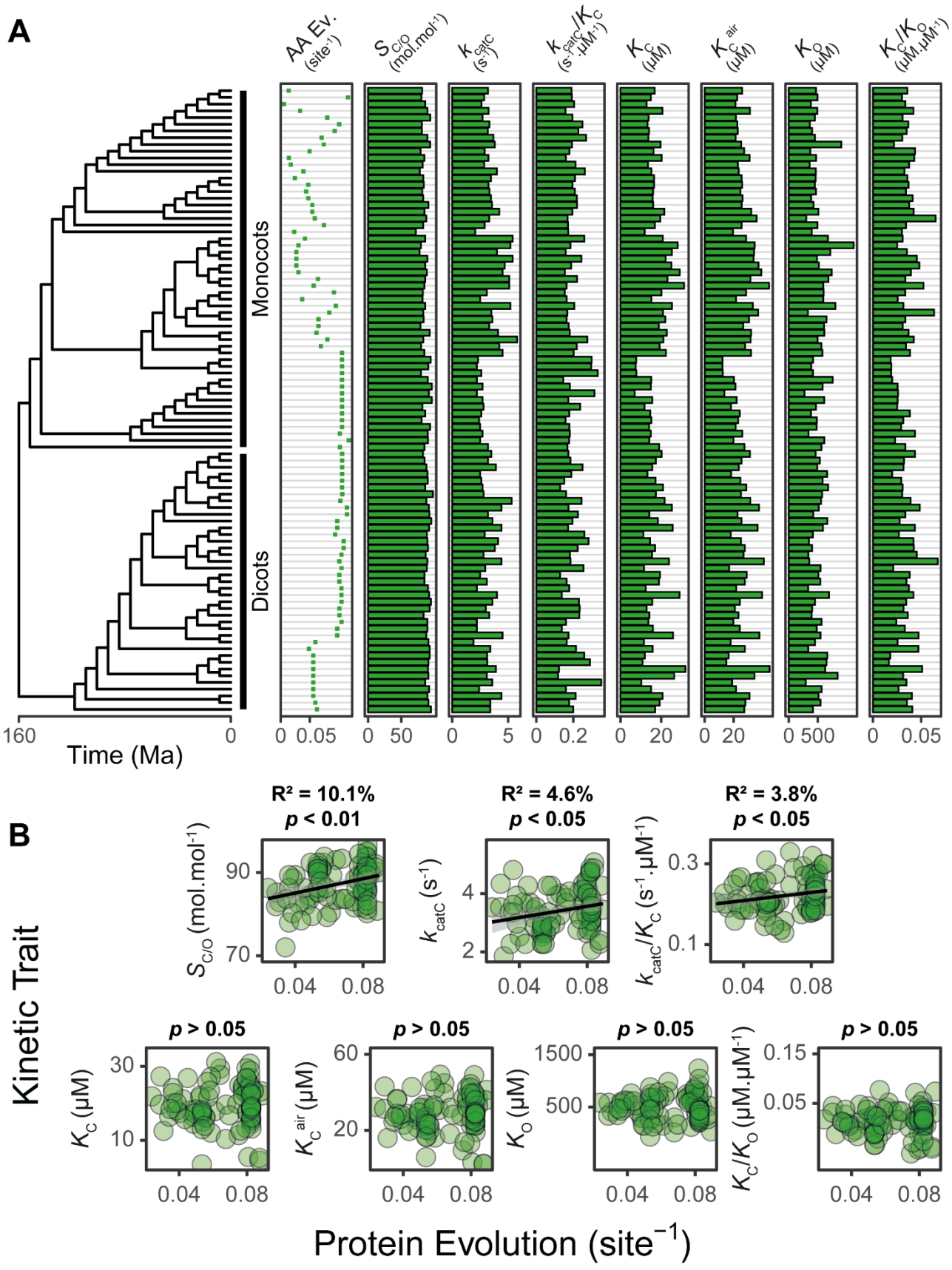

A
Bacteria
Land plants
Green algae
Red algae
SAR
Evolutionary Rate Percentile (%)
100
75
50
25
0
RbcL
RbcS
Subunit
Nucleotide
Amino acid
B
Bacteria
Land plants
Green algae
Red algae
SAR
Evolutionary Rate Percentile (%)
100
75
50
25
0
RbcL
RbcS
Subunit
Nucleotide
Amino acid
C
% Evolutionary Rate of RbcL
1000
100
10
RBCS
PGK
GAPDH−A
GAPDH−B
TPI
FBA
FBP
TKL
SBP
RPI
RPE
PRK
Gene
Nucleotide
Amino acid

Figure 3. The extent of molecular evolution in rubisco in context other genes. **A)** Boxplot of the extent of molecular evolution (substitutions per sequence site) in the nucleotide and protein sequences of the rubisco large (*rbcL*/RbcL) and small (*rbcS*/RbcS) subunit expressed as a percentile (%) of that measured across all other genes and proteins, respectively. See also Supplemental File 1, table S2. **B)** As in (A) but calculating the percentile (%) extent of rubisco molecular evolution (substitutions per sequence site) relative to only the subset of genes and proteins in each species which encode enzymes. See also Supplemental File 1, table S3. **C)** Boxplot of the total amount of molecular evolution (substitutions per sequence site) in the nucleotide and protein sequences of each Calvin-Bensen-Bassham cycle enzyme expressed as a percentage (%) of that measured in *rbcL*/RbcL (100%; red horizontal line) across land plants. Phosphoglycerate kinase: PGK. Glyceraldehyde-3-phosphate dehydrogenase A/B subunit: GAPDH-A/GAPDH-B. Triose phosphate isomerase: TPI. Fructose-bisphosphate aldolase: FBA. Fructose-1,6-bisphosphatase: FBP. Transketolase: TKL. Sedoheptulose-bisphosphatase: SBP. Ribose 5-phosphate isomerase: RPI. Ribulose-p-3-epimerase: RPE. Phosphoribulokinase: PRK. See also Supplemental File 1, table S4 and table S5. The raw data for this figure can be found in Supplemental File 5.

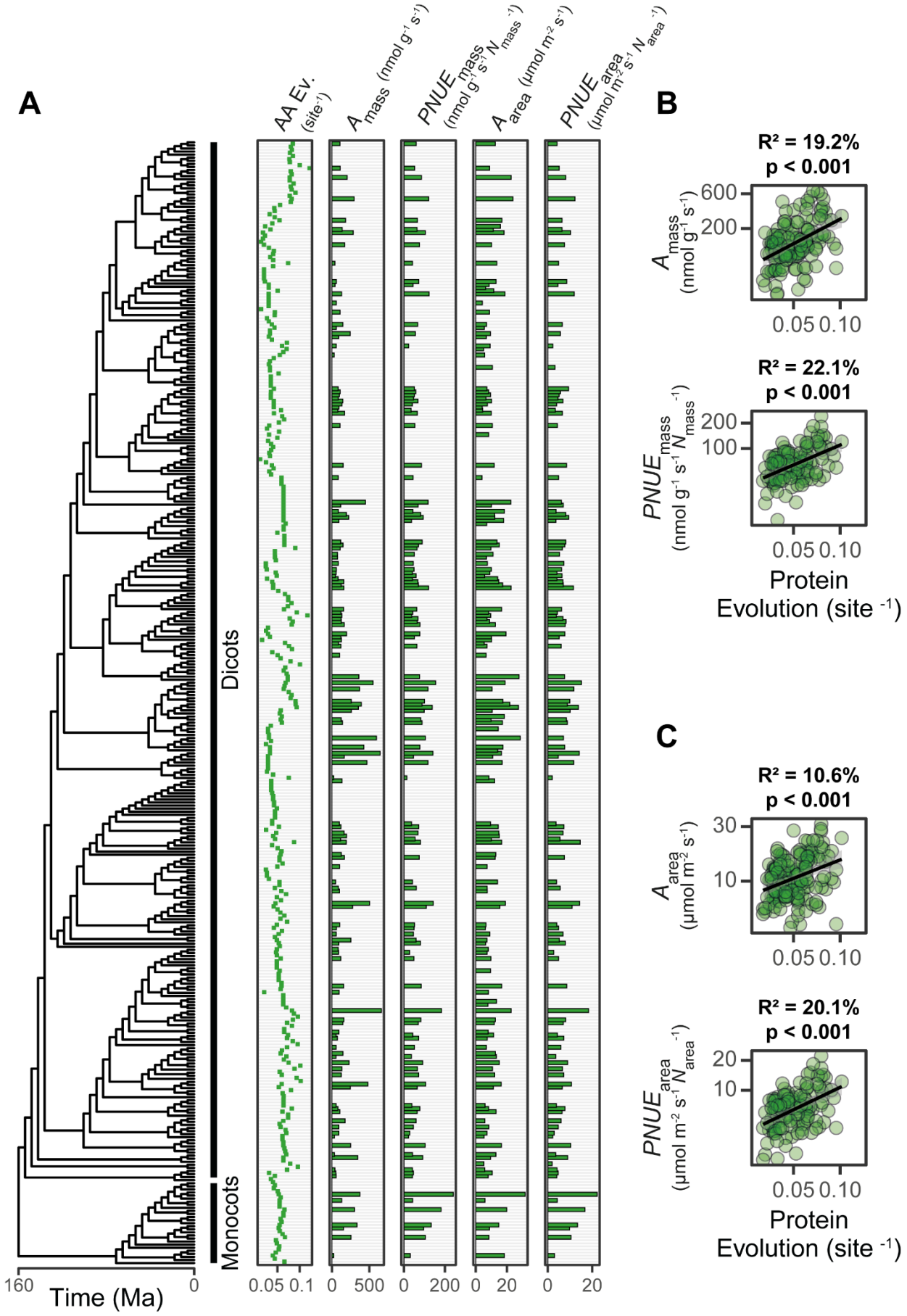
969 ***Figure 4***



970

Figure 4. The relationship between rubisco molecular and kinetic evolution in C_3 angiosperms. **A)** The relationship between RbcL evolution and its corresponding kinetic trait values. AA Ev.: The extent of RbcL amino acid evolution that has occurred since the last common ancestor at the root of the angiosperm phylogeny. $S_{C/O}$: specificity. k_{catC}: carboxylase turnover per site. k_{catC}/K_C: carboxylation efficiency. K_C: the Michaelis constant for CO_2. K_C^{air}: the inferred Michaelis constant for CO_2 in 20.95% O_2. K_O: the Michaelis constant for O_2. K_C/K_O: the ratio of the Michaelis constant for CO_2 compared to O_2. **B)** The relationship between the extent of RbcL protein evolution (substitutions per sequence site) and each rubisco kinetic trait in (A) as assessed using least squares regression models. The raw data can be found in Supplemental File 7.

980 ***Figure 5***



981

Figure 5. The relationship between rubisco molecular evolution and CO_2 assimilation in C_3 angiosperms. **A)** The relationship between the extent of RbcL evolution and leaf level CO_2 assimilation. AA Ev.: The extent of RbcL amino acid evolution that has occurred since the most recent common ancestor at the root of the angiosperm phylogeny. A_{mass}: Photosynthetic rate per unit leaf mass. $PNUE_{mass}$: Photosynthetic nitrogen use efficiency rate per unit leaf mass per unit leaf nitrogen content (N_{mass}; % N). A_{area}: Photosynthetic rate per unit leaf area. $PNUE_{area}$: Photosynthetic nitrogen use efficiency rate per unit leaf area expressed per unit leaf area nitrogen content (N_{area}; g m^{-2} N). **B)** The relationship between the extent of RbcL protein evolution (substitutions per sequence site) and each photosynthetic trait in (A) evaluated on a mass-basis (A_{mass}, $PNUE_{mass}$) as assessed using least squares regression models. **C)** As in (B) but for each photosynthetic trait evaluated on an area-basis (A_{area}, $PNUE_{area}$). The raw data can be found in Supplemental File 9.

Tables

Table 1

Table 1. Rubisco kinetics in extinct and extant angiosperms. Kinetic trait values for the last common ancestor of the angiosperms were computed based on the estimated y intercept (mean ± 1 S.E.) of the linear regression analysis performed between the extent of RbcL protein evolution and each rubisco kinetic trait in Figure 4B. Mean values of rubisco kinetic traits and associated variation (± 1 S.E.) in extant C_3 species are shown for comparison. The raw data set used can be found in Supplemental File 7.

Rubisco	$S_{C/O}$ **(mol mol^{-1})**	k_{catC} **(s^{-1})**	k_{catC}/K_C **(s^{-1} μM^{-1})**	K_C **(μM)**	K_C^{air} **(μM)**	K_O **(μM)**	K_C/K_O **(μM μM^{-1})**
Last Common Angiosperm Ancestor	81.1 ± 1.9	2.6 ± 0.3	0.16 ± 0.02	16.3 ± 2.1	24.8 ± 2.8	484.1 ± 56.4	0.034 ± 0.004
Extant Angiosperms	87.1 ± 0.5	3.4 ± 0.1	0.20 ± 0.01	17.6 ± 0.5	26.4 ± 0.7	517.2 ± 14.7	0.035 ± 0.001