Title

Slow molecular evolution of rubisco limits adaptive improvement of CO_2 assimilation

Authors

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

Affiliations

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

Corresponding Author

Name: Steven Kelly

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

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

Keywords

Rubisco; Evolution; Adaptation; Kinetics; Phylogenetics

Abstract

Rubisco assimilates CO_2 to form the sugars that fuel life on Earth. Although rubisco is the source of most carbon in the biosphere, it is a surprisingly inefficient catalyst with a modest carboxylase turnover rate and a competing oxygenase activity which results in the loss of fixed CO_2. These apparent shortcomings of rubisco present a puzzling evolutionary paradox: why does the enzyme appear well suited to the high CO_2 low O_2 conditions of its origin, rather than the low CO_2 high O_2 conditions of the present day? To help answer this question, we perform a phylogenetically resolved analysis of the molecular and kinetic evolution of Form I rubisco. We discover that rubisco is among the slowest evolving genes on Earth. Specifically, we find that the rubisco catalytic large subunit evolves substantially slower than its cognate small subunit, and is slower than >98% of all other genes and enzymes across the tree of life. Next, through simultaneous analysis of rubisco molecular and kinetic evolution in C_3 angiosperms, we demonstrate that despite its slow molecular evolution, rubisco kinetics have evolved, and are continuing to evolve, to improve CO_2/O_2 specificity, carboxylase turnover and overall carboxylation efficiency. Thus, slow molecular evolution has severely limited rubisco kinetic optimisation, resulting in the present enzyme that is poorly adapted for current conditions.

Introduction

Ribulose-1,5-bisphosphate carboxylase/oxygenase (rubisco, EC 4.1.1.39) is the enzyme that converts atmospheric CO_2 into the sugars that fuel the majority of life on Earth. This enzyme evolved approximately 3 billion years ago when the Earth's atmosphere contained high levels of CO_2 (≥10,000 % present atmospheric levels) and comparatively little O_2 (≤0.1% present atmospheric levels) (Figure 1) (Kasting, 1993; Canfield, 2005; Kaufman *et al.*, 2007; Nisbet *et al.*, 2007; Nisbet and Nisbet, 2008; Shih *et al.*, 2016; Ślesak, Ślesak and Kruk, 2017). However, since its emergence rubisco has guided the planet on a continual trajectory of increasing atmospheric O_2 and declining CO_2 (Holland, 2002; Canfield, 2005), such that today the concentrations of CO_2 (0.04%) and O_2 (20.95%) are maintained well-outside of their thermodynamic equilibria, and are inverted in relative abundance compared to the time when the enzyme first evolved (Figure 1).

Although all extant rubisco are descended from a single ancestral rubisco-like protein (Ashida, Danchin and Yokota, 2005; Ashida *et al.*, 2008; Erb and Zarzycki, 2018), the enzyme can be found in a variety of compositional forms across the tree of life (Figure 1) (Tabita, Satagopan, *et al.*, 2008; Banda *et al.*, 2020). The simplest manifestations of rubisco are that of the Form II and Form III variants found in protists, archaea, and some bacteria (Tabita, Hanson, *et al.*, 2008; Tabita, Satagopan, *et al.*, 2008) which are composed exclusively of a dimer of the ~50 kDa rubisco large (RbcL) subunit (Andersson and Backlund, 2008; Tabita, Satagopan, *et al.*, 2008; Whitney, Houtz and Alonso, 2011). In contrast, the more complex Form I rubisco is a hexadecamer comprised of four RbcL dimers organised in an antiparallel core capped by the ~15 kDa rubisco small subunit (RbcS) in an L_8S_8 arrangement (Schneider, Lindqvist and Brändén, 1992; Tabita, Satagopan, *et al.*, 2008). Of these three distinct rubisco lineages, only Form I and II have been recruited for use in oxygenic photosynthesis (Tabita, Hanson, *et al.*, 2008), with Form I being responsible for the vast majority of global CO_2 assimilation (Field *et al.*, 1998; Tabita, Hanson, *et al.*, 2008). In accordance with this role in manufacturing the precursors for nearly all carbon in the biosphere, Form I rubisco is also the most abundant enzyme on Earth (Ellis, 1979; Bar-On and Milo, 2019).

Although Form I rubisco is a hexadecamer of RbcL and RbcS subunits, the active site residues required for catalysis are exclusively located on RbcL (Andrews, 1988; Lee and Tabita, 1990;

Whitney, Houtz and Alonso, 2011). As a result, interspecific differences in Form I holoenzyme kinetics are primarily attributable to sequence variation in RbcL (Parry *et al.*, 2003; Kapralov and Filatov, 2007; Christin *et al.*, 2008; Chatterjee and Anindya, 2011; Kapralov *et al.*, 2011; Whitney *et al.*, 2011; Galmés, Andralojc, *et al.*, 2014; Galmés, Kapralov, *et al.*, 2014; Bouvier *et al.*, 2021; Lin *et al.*, 2021). However, despite not playing a direct role in catalysis, RbcS influences the function of Form I rubisco in a variety of ways (Spreitzer, 2003). For example, RbcS helps to enhance the stability and assembly of the holoenzyme complex (Andrews, 1988; McFadden and Small, 1988; Fitchen *et al.*, 1990; Lee, Berka and Tabita, 1991; Genkov and Spreitzer, 2009; Grabsztunowicz *et al.*, 2015; Joshi *et al.*, 2015). In addition, RbcS improves the efficiency of catalytic CO_2 binding by affecting both a conformational change in the active site (Schneider *et al.*, 1990) as well as acting as a reservoir for accumulating gaseous CO_2 (Van Lun *et al.*, 2014). Accordingly, rubisco holoenzyme kinetics and activity have been impacted during experiments in which the RbcS subunit has either been directly mutated (VOORDOUW *et al.*, 1987; Read and Tabita, 1992b; Kostov, Small and McFadden, 1997), or when chimeric holoenzymes have been created by substituting whole RbcS between species *in vivo* (Getzoff *et al.*, 1998; Kanevski *et al.*, 1999; Ishikawa *et al.*, 2011; Matsumura *et al.*, 2020; Sakoda *et al.*, 2021) and *in vitro* (Andrews and Lorimer, 1985; van der Vies, Bradley and Gatenby, 1986; Lee, Read and Tabita, 1991; Read and Tabita, 1992a; Horken and Tabita, 1999; Wang *et al.*, 2001). Thus, even though RbcL is the catalytic subunit, it is evident that sequence variation in both RbcL and RbcS are important determinants of the enzymatic properties of Form I rubisco.

While rubisco kinetic variability exists in nature, the enzyme is broadly considered to be an inefficient catalyst. For example, the maximum substrate saturated turnover rate of Form I rubisco at 1–12 s^{-1} (Badger *et al.*, 1998) is slower than average for an enzyme (Bar-Even *et al.*, 2011). In addition, all rubisco variants catalyse a secondary reaction with O_2 (Bowes, Ogren and Hageman, 1971; Chollet, 1977; Sharkey, 2020) that not only exacerbates the inherent slow carboxylase activity of rubisco, but also results in the loss of fixed carbon via photorespiration (Eckardt, 2005; Sharwood, 2017; Busch, 2020). As a consequence of these limitations, rubisco appears to be poorly suited to the O_2-rich, CO_2-poor atmosphere of the present day (Figure 1). Moreover, rather

than evolving to become better adapted to its current environment, it appears that organisms have instead adopted alternate strategies to overcome rubisco's shortcomings. For instance, many photoautotrophs achieve high rates of CO_2 assimilation by either synthesising enormous quantities of rubisco (up to 50% of the total soluble protein in microbes (Tabita, 1999) and plants (Carmo-Silva *et al.*, 2015)), or by operating complex CO_2-concentrating mechanisms to create supra-atmospheric CO_2 concentrations around the enzyme *in vivo* (Meyer and Griffiths, 2013; Flamholz and Shih, 2020). Taken together, the poor catalytic efficiency of rubisco coupled with the costly or complex strategies that have evolved to overcome its limitations have caused many to question whether the enzyme is already perfectly adapted, and indeed, whether further improvements in its kinetics are possible (Ogren, 1984; Parry *et al.*, 2007, 2013; Whitney, Houtz and Alonso, 2011; Sharwood, Ghannoum and Whitney, 2016; Orr *et al.*, 2017; Sharwood, 2017; Sharkey, 2020). Obtaining the answers to these questions would help clarify the mystery of the "rubisco paradox", providing an answer to why the enzyme that catalyses the entry point of carbon into the biosphere appears poorly adapted for this role.

The initial hypothesis that attempted to explain the "rubisco paradox" posited that rubisco is constrained by catalytic trade-offs that prevent the optimisation of its kinetics. This theory was pioneered by two comparative studies (Tcherkez, Farquhar and Andrews, 2006; Savir *et al.*, 2010) which reported strong antagonistic correlations between rubisco kinetic traits. From this, it was proposed that chemical constraints on the catalytic mechanism were the cause of these trade-offs (Tcherkez, Farquhar and Andrews, 2006; Savir *et al.*, 2010). However, new evidence has revealed that the catalytic constraints on rubisco evolution are not absolute, and that the extent of this constraint has been over-estimated (Flamholz *et al.*, 2019; Iñiguez *et al.*, 2020, 2021; Bouvier *et al.*, 2021). Specifically, when the number of rubisco variants being analysed was increased, the apparent strength of kinetic trait correlations were relaxed (Flamholz *et al.*, 2019; Iñiguez *et al.*, 2020). Moreover, it was also discovered that rubisco kinetic traits exhibit a high degree of phylogenetic signal that causes an artefactual inflation in kinetic trait correlations unless the correct phylogenetic comparative methods are applied (Bouvier *et al.*, 2021). Specifically, when kinetic trait correlations were re-evaluated after accounting for this phylogenetic signal, it was concluded that a

genuine antagonism only existed between the Michaelis constant for CO_2 and carboxylase turnover (variance explained = 21–37%), and between the Michaelis constants for CO_2 and O_2 (variance explained 9–19%), respectively (Bouvier *et al.*, 2021). Moreover, it was found that phylogenetic constraints explain more variation in rubisco kinetics (variance explained 30–61%), and have thus had a larger impact on limiting enzyme adaptation, than the combined action of catalytic trade-offs (variance explained 6–9%) (Bouvier *et al.*, 2021). However, although a mechanistic understanding of the source of the rubisco catalytic trade-offs exists (Tcherkez, Farquhar and Andrews, 2006; Savir *et al.*, 2010), a similar mechanistic understanding of this phylogenetic constraint is lacking. As such, investigation into the source of this phylogenetic constraint on rubisco would help shed light on the biological causes of the rubisco paradox.

In this study, we provide new insight into the phylogenetic constraint on rubisco kinetic adaptation by performing a comparative evaluation of the molecular and kinetic evolution of the Form I holoenzyme. We show that *rbcL* is evolving at a slower pace, and experiences a greater extent of purifying selection, than *rbcS*. We further demonstrate that 99% of all genes across the tree of life have evolved at a comparatively faster rate than *rbcL*/RbcL across equivalent periods of evolutionary time. Next, by simultaneously evaluating the extent of molecular and kinetic change in rubisco, we show that the enzyme has been continuously evolving toward increased CO_2/O_2 specificity, carboxylase turnover rate, and carboxylation efficiency during the past 160 million years of angiosperm diversification. On the basis of these results, we conclude that rubisco is not perfectly optimized. Instead, we argue that the slow molecular evolution of rubisco has limited its kinetic evolution, resulting in the present-day enzyme that is poorly optimized for current conditions. These findings have significant implications for our understanding of the past, present, and future potential of the enzyme in natural and engineered contexts.

Results

rbcL has evolved slower than rbcS and experiences stronger purifying selection

Sequences encoding Form I rubisco were obtained from the National Center for Biotechnology Information (NCBI) public sequence repository (https://www.ncbi.nlm.nih.gov/). This sequence dataset was filtered to retain data for a given species only if a full-length sequence for both *rbcL*

and *rbcS* was present. Although *rbcL* exists as a single copy gene in all species, many species harbour multiple *rbcS* genes in their genomes and thus all available *rbcS* sequences for a given species were retained for analysis. In total, this resulted in a set of 488 *rbcL*/RbcL and 1140 *rbcS*/RbcS sequences from 488 species (table 1; Supplemental File 1, Figure S1; Supplemental File 2).

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), the land plants (*Streptophyta*; *n* = 68) and the green algae (*Chlorophyta*; *n* = 12) (table 1; Supplemental File 2). The nucleotide and protein sequences encoding *rbcL*/RbcL and *rbcS*/RbcS were subject to multiple sequence alignment and maximum likelihood phylogenetic tree inference for each taxonomic group (see Methods). Here, the total branch length of each *rbcL*/RbcL and *rbcS*/RbcS gene tree captures the amount of molecular evolution of each gene during the radiation of the set of species under consideration. For ease of interpretation, the total amount of molecular evolution of the nucleotide sequences and the total amount of molecular evolution of the protein sequences across all sampled species 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” is used to jointly refer to both the extent of nucleotide evolution and the extent of protein evolution.

Comparison of the extent of molecular evolution of *rbcL*/RbcL and *rbcS*/RbcS in each taxonomic group revealed that the extent of *rbcL*/RbcL evolution is always lower than that experienced by *rbcS*/RbcS (Figure 2A). This finding was not an artefact of the higher gene copy number of the *rbcS,* as a 1000 bootstrapped stratified sampling recovered the same result when only a single *rbcS*/RbcS sequence (rather than all *rbcS*/RbcS sequences) was randomly sampled for each species (see Methods) (Figure 2B). Therefore, the *rbcL* gene has explored less nucleotide and protein sequence space than *rbcS* in the same sets of species over the same period of evolutionary time. Furthermore, in each sampled taxonomic group, *rbcL* also experienced fewer amino acid changes per change in nucleotide sequence compared to *rbcS* over the same period of

time (Figure 2C). As such, *rbcL* has experienced a higher degree of purifying selection than *rbcS*. Combined, *rbcL* has thus evolved more slowly and has been subject to a higher degree of functional constraint on the encoded protein sequence in comparison to *rbcS*.

Interestingly, the extent of molecular evolution of *rbcL*/RbcL and *rbcS*/RbcS varied between taxonomic groups (Figure 2D). The largest difference between the two genes was observed for the land plants and green algae (Figure 2D). Here, in these two taxonomic groups, *rbcL* resides in the plastid genome and *rbcS* in the nuclear genome, respectively. In contrast, for all other taxonomic groups in which both genes are encoded in the same genome (either both in the single bacterial genome, or both in the plastid genome), *rbcL*/RbcL and *rbcS*/RbcS exhibited more similar extents of molecular evolution (Figure 2D). Thus, the difference in the genomic context of *rbcL* and *rbcS* in land plants and green algae potentially contributes to the disparity in their extent of molecular evolution. This result is consistent with the lower rate of molecular evolution experienced by the plastid as compared to the nuclear genome (Wolfe, Li and Sharp, 1987; Smith, 2015).

rbcL is one of the slowest evolving genes in the tree of life whereas the rate of rbcS evolution is more variable

Using the measure of molecular evolution above, it is not possible to compare the extent of molecular evolution of *rbcL*/RbcL and *rbcS*/RbcS between taxonomic groups. This is because the total length of each gene tree is sensitive to the number of species contained in the tree and the time since the divergence of the species in the group. Moreover, it is not possible from the above analysis to assess whether the variation in the ratios of *rbcL* to *rbcS* evolution above is due to differences in the extent of *rbcL* evolution, differences in the extent of *rbcS* evolution, or both. To address this, the extent of molecular evolution of *rbcL*/RbcL and *rbcS*/RbcS was placed in context of all other genes present in the same set of species to compute a percentile rank for each gene (see Methods). Converting to percentiles in this way enables direct comparison of the rates of *rbcL*/RbcL and *rbcS*/RbcS within and between taxonomic groups. This analysis revealed that 99.2% of gene nucleotide sequences (average percentile = 0.76) and 98.0% of gene protein sequences (average percentile = 1.96) experienced a larger extent of molecular evolution than *rbcL*/RbcL in the same sets of species over the same periods of evolutionary time (Figure 3A;

Supplemental File 1, table S1). These findings were consistent when *rbcL*/RbcL was compared to just the subset of genes which encode enzymes, with 99.2% of enzyme nucleotide sequences (average percentile = 0.79) and 98.2% of enzyme protein sequences (average percentile = 1.84) having experienced a larger extent of molecular evolution than *rbcL*/RbcL (Figure 3B; Supplemental File 1, table S2). Furthermore, in land plants *rbcL*/RbcL was also the slowest evolving enzyme of the Calvin-Benson-Bassham cycle (Figure 3C; Supplemental File 1, table S3 and S4). Thus, *rbcL* is one of the slowest evolving genes in the species in which it is found, irrespective of taxonomic group, and the slowest evolving component of the photosynthetic carbon reduction cycle.

In contrast to *rbcL*, considerable variability in the extent of molecular evolution of *rbcS* was observed both within and between taxonomic groups (Figure 3A; Supplemental File 1, table S1). Here, in each of the green algae, red algae, and SAR clades, the extent of molecular evolution of *rbcS*/RbcS was lower than >90% of gene nucleotide sequences (average percentile = 4.085), and >80% of gene protein sequences (average percentile = 13.26) (Figure 3A; Supplemental File 1, table S1). Interestingly, the extent of molecular evolution of *rbcS*/RbcS was found to be higher in bacteria and land plants (Figure 3A; Supplemental File 1, table S1). For example, in bacteria, 74.9% of gene nucleotide sequences (percentile = 25.11) and 58.1% of gene protein sequences (percentile = 41.88) experienced a larger extent of molecular evolution than *rbcS*/RbcS, whilst in land plants, only 35.9% of gene nucleotide sequences (percentile = 64.06) and 48.0% of gene protein sequences (percentile = 52.02) experienced a larger extent of molecular evolution than *rbcS*/RbcS (Figure 3A; Supplemental File 1, table S1). Furthermore, substantial variation was observed in the extent of molecular evolution of *rbcS*/RbcS across species in both of these respective taxonomic groups (Figure 3A; Supplemental File 1, table S1). Analogous results in each taxonomic group were recovered when the analysis was restricted to the subset of genes that encode enzymes (Figure 3B; Supplemental File 1, table S2). Moreover, in land plants *rbcS*/RbcS was the fastest evolving enzyme in the Calvin-Benson-Bassham cycle (Figure 3C; Supplemental File 1, table S3 and S4). Therefore, the extent of molecular evolution of *rbcS* is highly variable across the tree of life. Moreover, the disparity in the ratio of *rbcL* to *rbcS* evolution observed

between taxonomic groups is predominantly attributable to variation in the extent of molecular evolution of *rbcS*.

The magnitude of RbcL molecular evolution correlates with kinetic adaptation in angiosperms

Although *rbcL* is among the slowest evolving genes in the tree of life, it is not stagnant. Thus, the sequence evolution that has occurred in rubisco is either adaptive for the catalytic properties of the enzyme or is simply a result of neutral variation that lacks significant consequences for organismal fitness. Here, if rubisco has undergone adaptive evolution, it would be expected that those enzymes that have experienced the largest extent of protein evolution would be the most efficient catalysts. To test this hypothesis, a dataset of kinetic measurements from angiosperms (Flamholz *et al.*, 2019; Bouvier *et al.*, 2021) was evaluated in the context of the extent of protein evolution of RbcL measured since the last common ancestor at the root of this clade (Figure 4A and 4B). For this purpose, only C_3 species in the dataset were considered. This is because C_4 photosynthesis is associated with a change in the direction of rubisco kinetic evolution (Bouvier *et al.*, 2021) that would confound the directional selection analysis being conducted. Here, simultaneous analysis of kinetic and molecular evolution in C_3 angiosperms revealed that the more an individual rubisco has evolved, the better its CO_2/O_2 specificity ($S_{C/O}$; 10.1% variance explained), and the higher its CO_2 turnover rate (k_{catC}; 4.6% variance explained) and carboxylation efficiency (k_{catC}/K_C; 3.8% variance explained) (Figure 4B; table 2), respectively. As such, these data are most parsimoniously explained by a directional selection gradient towards enhanced $S_{C/O}$, k_{catC}, and k_{catC}/K_C over the course of the radiation the angiosperms (table 3). In contrast, no correlation was observed among C_3 angiosperms between the extent of protein evolution in RbcL and any other kinetic trait (Figure 4B; table 2). Thus, all other kinetic traits except for $S_{C/O}$, k_{catC} and k_{catC}/K_C appear to display evolutionary profiles characteristic of undirected Brownian motion, and are thus indicative of neutral evolution. In summary, sequence changes in rubisco have been adaptive during the radiation of the angiosperms such that the greater the extent of molecular evolution which has occurred in a given *rbcL* gene since the last common ancestor of land plants, the more efficient the holoenzyme which exists in the present day.

Given that the time point of angiosperm divergence is estimated to have occurred 160 million years ago (Figure 4A) (Kumar *et al.*, 2017), it is possible to compute the rate at which rubisco kinetic traits have improved during the radiation of this clade as a function of both time and molecular change relative to the last common ancestor (table 3). This analysis revealed that on average, rubisco $S_{C/O}$ has increased by $3.8\times10^{-2} \pm 3.2\times10^{-3}$ mol.mol^{-1} (mean ± 1 S.E.) per million years during the course of angiosperm evolution. In addition, both k_{catC} and k_{catC}/K_C have improved over this equivalent time period at an average rate of $4.7\times10^{-3} \pm 5.7\times10^{-4}$ s^{-1} (mean ± 1 S.E.) per million years and $2.3\times10^{-4} \pm 3.1\times10^{-5}$ s^{-1} μM^{-1} (mean ± 1 S.E.) per million years, respectively. As such, for each million years of angiosperm evolution, rubisco $S_{C/O}$ in C_3 plants has been improving on average by 0.05 ± 0.004%, whilst k_{catC} and k_{catC}/K_C have been improving by 0.2 ± 0.02% and 0.1 ± 0.02%, respectively. Next, when the extent of molecular evolution was quantified over this same time period since the last common ancestor of the angiosperms, *rbcL* was revealed to have evolved by 9.7×10^{-4} changes per nucleotide site per million years, and by 4.1×10^{-4} changes per amino acid site per million years. To put this into context, this means that *rbcL* has evolved in angiosperms at a rate of one nucleotide mutation every 0.7 million years and one amino acid mutation every 5.1 million years. Therefore, an average *rbcL* in an extant angiosperm is predicted to have undergone a total of ~223 nucleotide substitutions and ~31 amino acid substitutions since the *rbcL* present in the last common ancestor at the root of this clade. Accordingly, when a comparison of the extent of molecular evolution and kinetic evolution was performed, it was revealed that each amino acid mutation that has occurred in RbcL has resulted in an average increase in $S_{C/O}$ by $2.0\times10^{-1} \pm 1.7\times10^{-2}$ mol.mol^{-1}, an average increase in k_{catC} by $2.4\times10^{-2} \pm 3.0\times10^{-3}$ s^{-1}, and an average increased in k_{catC}/K_C by $1.2\times10^{-3} \pm 1.6\times10^{-4}$ s^{-1} μM^{-1} (mean ± 1 S.E.), respectively. This rate of improvement in each kinetic trait per RbcL amino acid mutation is equivalent to 0.2 ± 0.02% ($S_{C/O}$), 1.0 ± 0.1% (k_{catC}), and 0.7 ± 0.1% (k_{catC}/K_C). Thus, rubisco has evolved, and is continuing to evolve, to be a better catalyst by optimising CO_2/O_2 specificity, CO_2 turnover rate and carboxylation efficiency, albeit at a rate that is limited by *rbcL* being one of the slowest evolving genes in the tree of life.

Discussion

Rubisco is the primary entry point for carbon into the biosphere and is responsible for fixing 250 billion tons of CO_2 globally each year (Field *et al.*, 1998). Despite this immense throughput, the enzyme is a surprisingly inefficient catalyst with a modest carboxylase turnover rate of 1–12 s^{1} (Badger *et al.*, 1998) and a competing oxygenase activity that culminates in the loss of fixed carbon (Bowes, Ogren and Hageman, 1971; Ogren and Bowes, 1971; Sharkey, 2020). This discord presents an evolutionary paradox that has attracted significant attention from disparate perspectives (Ogren, 1984; Parry *et al.*, 2007, 2013; Whitney, Houtz and Alonso, 2011; Sharwood, Ghannoum and Whitney, 2016; Orr *et al.*, 2017; Sharwood, 2017; Sharkey, 2020). In this study, we shed new light on the rubisco paradox by revealing that the large subunit (RbcL) of the enzyme, which contains the active site residues and is responsible for catalytic activity, is encoded by one of the slowest evolving genes on Earth. Despite this slow rate of molecular evolution, we find that more-evolved *rbcL* genes in C_3 angiosperms have higher CO_2/O_2 specificity ($S_{C/O}$), faster carboxylase turnover rates (k_{catC}), and improved carboxylation efficiency (k_{catC}/K_C) compared to less-evolved variants. Thus, rubisco is not perfectly optimised. Instead, its low rate of molecular evolution has constrained its kinetic evolution, resulting in the present-day enzyme that is poorly adapted to current atmospheric conditions.

The slow rate of *rbcL* molecular evolution has long been exploited in systematic and phylogenetic analyses to resolve evolutionary histories among taxa (Gielly and Taberlet, 1994; APG, 1998, 2016). However, to date there has been no direct assessment of just how slowly rubisco is evolving. The analysis presented here addresses this issue by resolving the relative rate of Form I rubisco molecular evolution compared to all other orthologous genes present in the same species genomes. We show that *rbcL*/RbcL has experienced a lower extent of molecular evolution than 99% of all gene nucleotide sequences, and 98% of all gene protein sequences over identical periods of time. Moreover, we show that rubisco is evolving slower than >98% of enzymes, and also represents the slowest evolving component of the Calvin Benson Bassham cycle. This extremely low rate of molecular evolution is likely due to several synergistic factors (Krylov *et al.*, 2003; Pál, Papp and Lercher, 2006) that include expression level (Drummond *et al.*, 2005; Wall *et*

al., 2005; Gout, Kahn and Duret, 2010; Seward and Kelly, 2018), selection to preserve protein structure and function (Tourasse and Li, 2000; Bloom *et al.*, 2006; Lobkovsky, Wolf and Koonin, 2010; Soskine and Tawfik, 2010; Echave, 2019), and the constraints imposed by the requirements for protein-protein interactions *in vivo* (Fraser, Wall and Hirsh, 2003; Fraser and Hirsh, 2004; Saeed and Deane, 2006; Tóth-Petróczy and Tawfik, 2011). These factors are particularly pertinent for rubisco given that it is the most abundant protein expressed by photoautotrophs (Tabita, 1999; Carmo-Silva *et al.*, 2015), has high protein instability (Studer *et al.*, 2014; Duraõ *et al.*, 2015; Cummins, Kannappan and Gready, 2018), and relies on multiple interacting partners and chaperones for folding, assembly and metabolic regulation (Carmo-Silva *et al.*, 2015; Aigner *et al.*, 2017). Furthermore, these constraints on rubisco molecular evolution are further compounded by *rbcL* being encoded in the organellar genome, which on average experiences a ten-fold lower rate of molecular evolution compared to that of the nuclear genome (Wolfe, Li and Sharp, 1987; Smith, 2015). Thus, a perfect storm of features acts to constrain the molecular evolution of *rbcL* causing it to be one of the slowest evolving genes on Earth.

Interestingly, although *rbcS* is located in the nuclear genome in green algae and in land plants, the relative extent of molecular evolution of *rbcS* is much lower in green algae than in land plants. This discrepancy between the two lineages is likely attributable to the role of RbcS in the structure and phase separation of the green algal pyrenoid (Meyer *et al.*, 2012; Mitchell *et al.*, 2017; Wunder *et al.*, 2018; Atkinson *et al.*, 2019; He *et al.*, 2020; Matsuzaki *et al.*, 2021). As this additional protein-protein interaction is not required of RbcS in land plants, this could provide an additional sequence constraint of *rbcS* in green algae that limits the relative extent of its molecular evolution.

The analysis of integrated molecular and kinetic evolution data presented here has revealed that there has been a continual improvement in rubisco specificity, carboxylase turnover rate and carboxylation efficiency over the past 160 million years of angiosperm evolution. In the context of the C_3 leaf, such selection for improved specificity makes sense as an adaptation to maintain a high rate of net CO_2 fixation in response to the declining atmospheric CO_2 and increasing O_2 (Holland, 2002; Canfield, 2005). This evolutionary strategy to maximize $S_{C/O}$ has previously been suggested to apply broadly across photoautotrophs lacking a CO_2 concentrating mechanism

(Rickaby and Eason Hubbard, 2019; Iñiguez *et al.*, 2020). This result is also in agreement with the lower specificity of the *in vitro* characterised extinct enzyme present at the last common ancestor of the Form IB lineage (Shih *et al.*, 2016). Moreover, the simultaneous improvement in rubisco carboxylase turnover rate and carboxylation efficiency that is concomitantly achieved alongside adaptation for higher CO_2/O_2 selectivity (without associated antagonism in any other kinetic trait), is further evidence in favour of rubisco being only weakly constrained by catalytic trade-offs (Bouvier *et al.*, 2021). Thus, rubisco has been adapted for enhanced CO_2 assimilation capacity among C_3 angiosperms by selection for CO_2 specificity, carboxylase turnover rate and carboxylation efficiency.

Although rubisco has adapted to the declining ratio of atmospheric CO_2 to O_2 by optimising $S_{C/O}$, an analogous change has not been observed in the Michaelis constant for CO_2 (K_C). This is because K_C values observed in angiosperms today appear well suited to the 8 µM chloroplastic pool of CO_2 (Andrews and Whitney, 2003), and thus adaptive change in this trait is not required. Our finding that K_C has evolved in a manner that is consistent with neutral Brownian motion is compatible with this hypothesis, helping to explain why little variation is observed in this trait. Additionally, neutral evolution is similarly observed in the rubisco Michaelis constant for O_2 (K_O). Thus, trait evolution in neither rubisco CO_2 affinity nor rubisco O_2 affinity has contributed to the catalytic optimisation of this enzyme since the radiation of the angiosperms.

In this work we show that the rate of *rbcL* molecular evolution is extremely slow, and further demonstrate that rubisco catalytic adaptation is dependent on the extent of RbcL molecular evolution. Combined, we therefore argue that the slow rate of *rbcL* evolution has presented a considerable constraint on rubisco optimization thereby causing the enzyme to be poorly optimized to current conditions. We term this the “rubisco molecular constraint hypothesis” to encapsulate all of the competing and synergistic factors (including catalytic trade-offs) that act in concert to limit the molecular evolution of the enzyme. This hypothesis provides a formal mechanism to explain the severe phylogenetic constraint on rubisco kinetic adaptation that has been recently reported (Bouvier *et al.*, 2021). It also takes a contrasting viewpoint to the previously proposed “evolutionary fine-tuning hypothesis” (Tcherkez, Farquhar and Andrews, 2006; Savir *et al.*, 2010) which invokes

the existence of severe biochemical constraints that act to provide a ceiling on catalytic efficiency and inhibit further adaptation once this ceiling is reached. We believe that three key pieces of evidence indicate that the "molecular constraint hypothesis" provides a better explanation for the rubisco paradox than the "fine-tuning hypothesis". First, in this work we show that rubisco has been continually optimizing $S_{C/O}$, k_{catC} and k_{catC}/K_C as a function of RbcL molecular evolution along a conserved trajectory. This finding that increased molecular evolution results in improved efficiency is also observed in natural contexts of Form I rubisco, in which the best variants are found in fast evolving species with short generation times and large population sizes (Bousquet *et al.*, 1992; Soria-Hernanz *et al.*, 2008; Weller and Wu, 2015; Rickaby and Eason Hubbard, 2019; Cagan *et al.*, 2021) such as in red algae and red algal-derived lineages (high $S_{C/O}$) (Uemura *et al.*, 1997; Whitney *et al.*, 2001; Savir *et al.*, 2010; Young *et al.*, 2016; Heureux *et al.*, 2017), and in cyanobacteria (high k_{catC}) (Yu *et al.*, 2015; Ungerer *et al.*, 2018; Davidi *et al.*, 2020). In addition, engineered variants of rubisco with improved kinetics have been generated in archaea (Wilson, Alonso and Whitney, 2016), photosynthetic bacteria (Zhou and Whitney, 2019), and cyanobacteria (Wilson *et al.*, 2018) when molecular constraints on enzyme optimisation have been circumvented using directed evolution. Together, these observations demonstrate that extant rubiscos are not perfectly adapted. Instead, they suggest that rubisco is constrained from evolving toward an improved state by slow rates of molecular evolution. Second, given that all rubisco are sub-optimal for the concentrations of CO_2 and O_2 at present, it is not plausible that the kinetic variability which exists among species represents micro-adaptations to their respective subcellular environment, as is a crucial piece of evidence in support of the idea that all rubisco are "nearly perfectly optimised" (Tcherkez, Farquhar and Andrews, 2006). Instead, this kinetic variability more likely reflects differing extents of optimization along the conserved trajectory towards improved specificity, carboxylase turnover and carboxylation efficiency, as argued above. Third, although genuine catalytic trade-offs exist, recent analyses have shown that these are considerably weaker than initial estimates (Cummins, Kannappan and Gready, 2018, 2019; Flamholz *et al.*, 2019; Iñiguez *et al.*, 2020; Bouvier *et al.*, 2021; Cummins, 2021). Therefore, rubisco kinetic traits have been evolving largely independently of one another in contradiction of the "fine-tuning hypothesis". In summary, compelling evidence exists to suggest rubisco is not "near perfectly optimized" within a

heavily constrained biochemical landscape. Instead, it is more likely that the limited kinetic variability in rubisco, and the poor efficiency of this enzyme, are a consequence of constraints imposed by a slow rate of molecular evolution.

Conclusion

Here, we find that rubisco (specifically the *rbcL*) is among the slowest evolving genes on Earth. We demonstrate that despite this slow rate of molecular evolution, the CO_2 assimilation capacity of the enzyme has been continually improved over the course of angiosperm evolution in a manner that is dependent on the extent of RbcL evolution. In summary, the improved understanding of rubisco kinetic evolution presented in this work motivates ongoing strategies to improve the enzyme in synthetic biology contexts by circumventing the evolutionary barriers on its optimisation. This could be achieved using a combination of techniques including laboratory-based directed evolution to generate better variants than found in nature (Wilson, Alonso and Whitney, 2016; Wilson *et al.*, 2018; Zhou and Whitney, 2019), as well as by substituting more efficient evolved variants between species along with their attendant chaperones (Gunn *et al.*, 2020; Iqbal *et al.*, 2021).

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. As *rbcS* can be multicopy in some genomes, all possible full-length *rbcS* sequences were taken forward for analysis.

Translated *rbcL* and *rbcS* protein sequences were aligned using the mafft L-INS-i algorithm (Katoh and Standley, 2013). 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 (Suyama, Torrents and Bork, 2006). 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 sequence in taxa in which this protein 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) (table 1; Supplemental File 2) by using the NCBI taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi). The *Haptophyta*, *Cryptophyta*, *Glaucocystophyta* and *Excavata* were not included due to insufficient data available. In total, this left a combined set of 488 *rbcL*/RbcL and 1140 *rbcS*/RbcS genes and protein sequences across 488 species, respectively (Supplemental File 2).

Rubisco phylogenetic tree inference

Maximum-likelihood phylogenetic trees were inferred by IQ-TREE (Nguyen *et al.*, 2015) using the ultrafast bootstrapping method with 1000 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 (Nguyen *et al.*, 2015). 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.

Stratified sampling of rbcS sequences

Random stratified sampling of the non-gapped *rbcS*/RbcS sequence alignment was conducted by species using 1000 replicates with replacement. Each alignment, containing only a single copy of the *rbcS* for each species, was subject to phylogenetic tree inference using IQ-TREE (Nguyen *et al.*, 2015) following the method described above.

Classification and phylogenetic tree inference of orthogroups

Complete sets of representative gene models for as many species as possible in this study were acquired from either NCBI (https://www.ncbi.nlm.nih.gov/) or Phytozome V13 (Goodstein *et al.*, 2012). Where more than one resource was available for a given species, the most recent assembly version was chosen. In total, complete sets of representative gene models were acquired for 33 bacterial species, 27 land plant species, 9 SAR species, 7 red algal species and 4 green algal species in this study (table 1). Assemblies were processed to keep only the longest gene model variant per gene and sequences with internal in-frame stop codons were removed. Finally, a corresponding proteome was generated from each transcriptome via *in silico* translation of the gene models.

The complete set of proteomes for the species in each taxonomic group were subject to orthogroup inference using OrthoFinder V2.5.2 (Emms and Kelly, 2015, 2019) software run with default settings and using the DIAMOND ultra-sensitive mode. Protein sequences within each orthogroup were aligned using the MAFFT L-INS-I algorithm with 1000 cycles of iterative refinement (Katoh and Standley, 2013). The corresponding set of nucleotide coding sequences within each orthogroup were codon aligned using PAL2NAL (Suyama, Torrents and Bork, 2006). Alignments were trimmed to remove positions which contain gap characters. Sequences that were <50% of the median length of sequences in the given orthogroup were excluded to avoid analysis of partial or truncated sequences that can influence phylogenetic tree branch length. 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 (Nguyen *et al.*, 2015) following the method described above. In total, this resulted in a set of 20351 orthogroup phylogenies and 5126493 ortholog pairwise comparisons for the land plant clade, 7667 orthogroup phylogenies and 157769 ortholog pairwise comparisons for the SAR clade, 5621 orthogroup phylogenies and 644673 ortholog pairwise comparisons for the bacteria clade, 4698 orthogroup phylogenies and 31275 ortholog pairwise comparisons for the green algae clade and 4408 orthogroup phylogenies and

54230 ortholog pairwise comparisons for the red algae clade from which to base the analyses herein.

Classification of genes enzyme encoding

To determine which genes encode enzymes in the sets of all genes used for orthogroup analysis, each species proteome was classified using the DeepEC (Ryu, Kim and Lee, 2019) deep learning-based algorithm. Enzymes were defined as those protein sequences that could be assigned at least a partial enzyme commission number.

Classification and phylogenetic tree inference of Calvin-Benson-Bassham cycle genes

The set of putative genes which encode enzymes in the Calvin-Benson-Bassham cycle were 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 were determined based on available data in The Arabidopsis Information Resource (TAIR) genome database (http://arabidopsis.org) (Swarbreck *et al.*, 2008; Lamesch *et al.*, 2012). 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 TargetP V2.0 (Emanuelsson *et al.*, 2000; Armenteros *et al.*, 2019) and Predotar V1.04 (Small *et al.*, 2004). 4) Evolutionary history as inferred from trees generated for each Calvin-Benson-Bassham cycle gene family using IQ-TREE (Nguyen *et al.*, 2015). Based on this combined data, the resulting putative photosynthetic isoforms encoding each Calvin-Benson-Bassham cycle enzyme in *A. thaliana* were thus inferred, and can be found in Supplemental File 1, table S5.

The set of genes which encode the photosynthetic isoforms of Calvin Bensen Bassham cycle enzymes in all other 27 land plant species for which genome sequence data was available in this study were then determined by orthology using data from the orthogroup inference analysis performed above. Here, each group of orthologous protein sequences that encode a given Calvin-Benson-Bassham cycle enzyme were aligned using the MAFFT L-INS-i algorithm (Katoh and Standley, 2013), and the corresponding set of nucleotide coding sequences were codon aligned using PAL2NAL (Suyama, Torrents and Bork, 2006). Gene and protein sequence alignments were then subject to the same data filtering and quality control criteria previously described to trim columns which contain gaps and remove partial or incomplete sequences. Finally, bootstrapped maximum likelihood phylogenetic trees were inferred by IQ-TREE (Nguyen *et al.*, 2015) following the method outlined above. An analogous determination of the gene isoforms used in the Calvin-Bensen-Bassham cycle in other taxonomic groups was not able to be performed owing to a lack of the required data described here that is needed to determine Calvin-Bensen-Bassham gene isoforms from those used in other pathways.

Quantification of the total extent of rbcL and rbcS molecular evolution

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 above 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. The same analysis was also performed for the stratified sampling of *rbcS*/RbcS trees, with mean and standard errors of estimates being calculated in this case across 1000 bootstrapped replicates.

Quantification of the relative extent of rbcL and rbcS 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 for which a publicly available whole-genome gene sets could be obtained 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 evolution since the 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 that were previously inferred. Following this method, the extent of evolution separating all other pairs of orthologous (but not paralogous) gene and protein sequences for a given species pair was measured from the orthogroup phylogenies, and the extent of rubisco molecular evolution was then assessed as a percentile of the distribution of these measurements. To next assess the extent of rubisco molecular evolution in context of all other enzymes, the exact same steps were followed as above however only the subset of genes and proteins for which an enzyme commission number could be generated were included. In both of these analyses, species pairs were only considered where a minimum threshold of 100 orthologous genes could be identified across all orthogroups. In cases where multiple percentiles are calculated for a given species pair (due to gene duplications in the *rbcS* of some species) the mean percentile was taken.

To determine the extent of molecular evolution between orthologous Calvin-Benson-Bassham cycle gene isoforms, the same method as described above was followed. For this purpose, the phylogenies inferred previously which describe the evolutionary history of Calvin-Benson-Bassham cycle gene families were interrogated. In contrast to the above however, the extent of molecular evolution measured across Calvin-Benson-Bassham cycle genes for each pairwise combination of species was expressed as a percentage ratio of that which has been measured in the *rbcL* to enable direct comparisons in molecular evolution rates to be made. The full set of data generated from all analyses described here to quantify the relative extent of rubisco molecular evolution, as broken down by species pairwise comparison, is available in Supplemental File 3.

Analysis of RbcL molecular evolution and rubisco kinetics

To interrogate the relationship between the molecular and kinetic evolution of extant Form I rubisco, a dataset of rubisco kinetic traits was downloaded from (Bouvier *et al.*, 2021), as modified from that originally compiled by (Flamholz *et al.*, 2019). 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 (Bouvier *et al.*, 2021). 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 inferred. Measurements of the Michaelis constant for RuBP (K_{RuBP}) in the dataset were not considered owing to a limited sample size (n = 19). Further, all *Limonium* species in the dataset were ignored on the basis that trait values obtained across different studies are not consistent (Galmés *et al.*, 2005; Galmés, Andralojc, *et al.*, 2014). In total, this left a dataset of rubisco kinetic trait measurements for 123 angiosperms (including 93 C_3 species) from which to base the analysis herein.

Coding sequences of the *rbcL* gene were obtained from (Bouvier *et al.*, 2021) for each of the species in the kinetic dataset used for analysis in this study. In order to facilitate more accurate downstream phylogenetic tree inference across these gene sequences and to minimize the impact of long-branch effects (Su and Townsend, 2015), the complete set of publicly available *rbcL* coding sequences in land plants were also acquired from NCBI (https://www.ncbi.nlm.nih.gov/) using the query term “rbcL[Gene Name] AND “plants”[porgn:_txid3193]” and were included in this analysis. These sequences thus obtained were subject to the exact same data processing steps to remove ambiguous, partial or chimeric sequences, as performed previously in (Bouvier *et al.*, 2021). In total, this resulted in an additional set of 29218 full-length *rbcL* coding sequences. Protein

sequences were inferred from each *rbcL* coding sequences via *in silico* translation. RbcL protein sequences were respectively aligned using MAFFT L-INS-I (Katoh and Standley, 2013), and the resulting sequence alignment was trimmed to remove non-aligned residue positions. A bootstrapped phylogenetic tree was inferred using IQ-TREE (Nguyen *et al.*, 2015) following the method described above, and was rooted manually using Dendroscope (Huson and Scornavacca, 2012). The extent of protein evolution which has occurred in each angiosperm in the kinetic dataset was measured 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. The kinetic trait values and extent of molecular evolution for each rubisco in this study is found in Supplemental File 4.

References

Aigner, H. *et al.* (2017) 'Plant RuBisCo assembly in E. coli with five chloroplast chaperones including BSD2', *Science*, 358(6368), pp. 1272–1278. doi: 10.1126/science.aap9221.

Andersson, I. and Backlund, A. (2008) 'Structure and function of Rubisco', *Plant Physiology and Biochemistry*, 46(3), pp. 275–291. doi: 10.1016/j.plaphy.2008.01.001.

Andrews, T. J. (1988) 'Catalysis by cyanobacterial ribulose-bisphosphate carboxylase large subunits in the complete absence of small subunits.', *The Journal of biological chemistry*, 263(25), pp. 12213–12219. doi: 10.1016/s0021-9258(18)37741-x.

Andrews, T. J. and Lorimer, G. H. (1985) 'Catalytic properties of a hybrid between cyanobacterial large subunits and higher plant small subunits of ribulosebisphosphate carboxylase-oxygenase', *Journal of Biological Chemistry*. Elsevier, 260(8), pp. 4632–4636. doi: 10.1016/s0021-9258(18)89117-7.

Andrews, T. J. and Whitney, S. M. (2003) 'Manipulating ribulose bisphosphate carboxylase/oxygenase in the chloroplasts of higher plants', *Archives of Biochemistry and Biophysics*, 414(2), pp. 159–169. doi: 10.1016/S0003-9861(03)00100-0.

APG (1998) *An Ordinal Classification for the Families of Flowering Plants*, *Annals of the Missouri Botanical Garden*. doi: 10.2307/2992015.

APG (2016) 'An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV', *Botanical Journal of the Linnean Society*. Oxford University Press, 181(1), pp. 1–20. doi: 10.1111/boj.12385.

Armenteros, J. J. A. *et al.* (2019) 'Detecting sequence signals in targeting peptides using deep learning', *Life Science Alliance*. Life Science Alliance LLC, 2(5). doi: 10.26508/lsa.201900429.

Ashida, H. *et al.* (2008) 'RuBisCO-like proteins as the enolase enzyme in the methionine salvage pathway: Functional and evolutionary relationships between RuBisCO-like proteins and photosynthetic RuBisCO', *Journal of Experimental Botany*. J Exp Bot, 59(7), pp. 1543–1554. doi: 10.1093/jxb/ern104.

Ashida, H., Danchin, A. and Yokota, A. (2005) 'Was photosynthetic RuBisCO recruited by acquisitive evolution from RuBisCO-like proteins involved in sulfur metabolism?', *Research in Microbiology*. Res Microbiol, 156(5–6), pp. 611–618. doi: 10.1016/j.resmic.2005.01.014.

Atkinson, N. *et al.* (2019) 'The pyrenoidal linker protein EPYC1 phase separates with hybrid Arabidopsis-Chlamydomonas Rubisco through interactions with the algal Rubisco small subunit', *Journal of Experimental Botany*, 70(19), pp. 5271–5285. doi: 10.1093/jxb/erz275.

Badger, M. R. *et al.* (1998) 'The diversity and coevolution of Rubisco, plastids, pyrenoids, and chloroplast-based CO2-concentrating mechanisms in algae', *Canadian Journal of Botany*. National Research Council of Canada, 76(6), pp. 1052–1071. doi: 10.1139/b98-074.

Banda, D. M. *et al.* (2020) 'Novel bacterial clade reveals origin of form I Rubisco', *Nature Plants*, 6(9), pp. 1158–1166. doi: 10.1038/s41477-020-00762-4.

Bar-Even, A. *et al.* (2011) 'The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters', *Biochemistry*, 50(21), pp. 4402–4410. doi: 10.1021/bi2002289.

Bar-On, Y. M. and Milo, R. (2019) 'The global mass and average rate of rubisco', *Proceedings of the National Academy of Sciences of the United States of America*. PNAS, 116(10), pp. 4738–4743. doi: 10.1073/pnas.1816654116.

Bloom, J. D. *et al.* (2006) 'Structural determinants of the rate of protein evolution in yeast', *Molecular Biology and Evolution*. Oxford Academic, 23(9), pp. 1751–1761. doi: 10.1093/molbev/msl040.

Bousquet, J. *et al.* (1992) 'Extensive variation in evolutionary rate of rbcL gene sequences among seed plants', *Proceedings of the National Academy of Sciences of the United States of America*, 89(16), pp. 7844–7848. doi: 10.1073/pnas.89.16.7844.

Bouvier, J. W. *et al.* (2021) 'Rubisco Adaptation Is More Limited by Phylogenetic Constraint Than by Catalytic Trade-off', *Molecular Biology and Evolution*, 38(7), pp. 2880–2896. doi: 10.1093/molbev/msab079.

Bowes, G., Ogren, W. L. and Hageman, R. H. (1971) *Phosphoglycolate production catalyzed by ribulose diphosphate carboxylase*, *Biochemical and Biophysical Research Communications*. doi: 10.1016/0006-291X(71)90475-X.

Busch, F. A. (2020) 'Photorespiration in the context of Rubisco biochemistry, CO2 diffusion and metabolism', *Plant Journal*, 101(4), pp. 919–939. doi: 10.1111/tpj.14674.

Cagan, A. *et al.* (2021) 'Somatic mutation rates scale with lifespan across mammals', *bioRxiv*, p. 2021.08.19.456982. Available at: https://www.biorxiv.org/content/10.1101/2021.08.19.456982v1%0Ahttps://www.biorxiv.org/content/10.1101/2021.08.19.456982v1.abstract.

Canfield, D. E. (2005) 'The early history of atmospheric oxygen: Homage to Robert M. Garrels', *Annual Review of Earth and Planetary Sciences*, 33, pp. 1–36. doi: 10.1146/annurev.earth.33.092203.122711.

Carmo-Silva, E. *et al.* (2015) 'Optimizing Rubisco and its regulation for greater resource use efficiency', *Plant, Cell and Environment*, pp. 1817–1832. doi: 10.1111/pce.12425.

Chatterjee, A. and Anindya, B. (2011) 'Rubisco: Limitations and Re-engineering for a Better Enzyme', *International Research Journal of Plant Science*, 2(2), pp. 022–024. Available at: http://interesjournals.org/IRJPS/Pdf/2011/Feb/Chatterjee and Basu.pdf (Accessed: 1 February 2022).

Chollet, R. (1977) *The biochemistry of photorespiration*, *Trends in Biochemical Sciences*. doi: 10.1016/0968-0004(77)90364-4.

Christin, P. A. *et al.* (2008) 'Evolutionary switch and genetic convergence on rbcL following the evolution of C4 photosynthesis', *Molecular Biology and Evolution*. Oxford Academic, 25(11), pp. 2361–2368. doi: 10.1093/molbev/msn178.

Cummins, P. L. (2021) 'The Coevolution of RuBisCO, Photorespiration, and Carbon Concentrating Mechanisms in Higher Plants', *Frontiers in Plant Science*, 12. doi: 10.3389/fpls.2021.662425.

Cummins, P. L., Kannappan, B. and Gready, J. E. (2018) 'Directions for optimization of photosynthetic carbon fixation: Rubisco's efficiency may not be so constrained after all', *Frontiers in Plant Science*, 9. doi: 10.3389/fpls.2018.00183.

Cummins, P. L., Kannappan, B. and Gready, J. E. (2019) 'Response: Commentary: Directions for Optimization of Photosynthetic Carbon Fixation: RuBisCO's Efficiency May Not Be So Constrained After All', *Frontiers in Plant Science*. Frontiers Media S.A., 10, p. 1426. doi: 10.3389/fpls.2019.01426.

Davidi, D. *et al.* (2020) 'Highly active rubiscos discovered by systematic interrogation of natural

sequence diversity', *The EMBO Journal*. John Wiley & Sons, Ltd, 39(18). doi: 10.15252/embj.2019104081.

Drummond, D. A. *et al.* (2005) 'Why highly expressed proteins evolve slowly', *Proceedings of the National Academy of Sciences of the United States of America*, 102(40), pp. 14338–14343. doi: 10.1073/pnas.0504070102.

Duraõ, P. *et al.* (2015) 'Opposing effects of folding and assembly chaperones on evolvability of Rubisco', *Nature Chemical Biology*, 11(2), pp. 148–155. doi: 10.1038/nchembio.1715.

Echave, J. (2019) 'Beyond Stability Constraints: A Biophysical Model of Enzyme Evolution with Selection on Stability and Activity', *Molecular Biology and Evolution*. Oxford Academic, 36(3), pp. 613–620. doi: 10.1093/molbev/msy244.

Eckardt, N. A. (2005) 'Photorespiration revisited', *Plant Cell*. American Society of Plant Biologists, pp. 2139–2141. doi: 10.1105/tpc.105.035873.

Ellis, R. J. (1979) *The most abundant protein in the world*, *Trends in Biochemical Sciences*. doi: 10.1016/0968-0004(79)90212-3.

Emanuelsson, O. *et al.* (2000) 'Predicting subcellular localization of proteins based on their N-terminal amino acid sequence', *Journal of Molecular Biology*. Academic Press, 300(4), pp. 1005–1016. doi: 10.1006/jmbi.2000.3903.

Emms, D. M. and Kelly, S. (2015) 'OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy', *Genome Biology*. BioMed Central Ltd., 16(1), pp. 1–14. doi: 10.1186/s13059-015-0721-2.

Emms, D. M. and Kelly, S. (2019) 'OrthoFinder: Phylogenetic orthology inference for comparative genomics', *Genome Biology*. BioMed Central Ltd., 20(1), pp. 1–14. doi: 10.1186/s13059-019-1832-y.

Erb, T. J. and Zarzycki, J. (2018) *A short history of RubisCO: the rise and fall (?) of Nature's predominant CO2 fixing enzyme*, *Current Opinion in Biotechnology*. doi: 10.1016/j.copbio.2017.07.017.

Felsenstein, J. (1973) 'Maximum likelihood estimation of evolutionary trees from continuous characters', *American Journal of Human Genetics*, 25(5), pp. 471–492.

Felsenstein, J. (1985) *Phylogenies and the comparative method.*, *American Naturalist*. doi: 10.1086/284325.

Field, C. B. *et al.* (1998) 'Primary production of the biosphere: Integrating terrestrial and oceanic components', *Science*. American Association for the Advancement of Science, 281(5374), pp. 237–240. doi: 10.1126/science.281.5374.237.

Fitchen, J. H. *et al.* (1990) 'Residues in three conserved regions of the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase are required for quaternary structure', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 87(15), pp. 5768–5772. doi: 10.1073/pnas.87.15.5768.

Flamholz, A. I. *et al.* (2019) 'Revisiting Trade-offs between Rubisco Kinetic Parameters', *Biochemistry*, pp. 3365–3376. doi: 10.1021/acs.biochem.9b00237.

Flamholz, A. and Shih, P. M. (2020) 'Cell biology of photosynthesis over geologic time', *Current Biology*, 30(10), pp. R490–R494. doi: 10.1016/j.cub.2020.01.076.

Fraser, H. B. and Hirsh, A. E. (2004) 'Evolutionary rate depends on number of protein-protein interactions independently of gene expression level', *BMC Evolutionary Biology*. BioMed Central, 4(1), pp. 1–5. doi: 10.1186/1471-2148-4-13.

Fraser, H. B., Wall, D. P. and Hirsh, A. E. (2003) 'A simple dependence between protein evolution rate and the number of protein-protein interactions', *BMC Evolutionary Biology*. BioMed Central, 3(1), pp. 1–6. doi: 10.1186/1471-2148-3-11.

Galmés, J. *et al.* (2005) 'Rubisco specificity factor tends to be larger in plant species from drier habitats and in species with persistent leaves', *Plant, Cell and Environment*, 28(5), pp. 571–579. doi: 10.1111/j.1365-3040.2005.01300.x.

Galmés, J., Andralojc, P. J., *et al.* (2014) 'Environmentally driven evolution of Rubisco and improved photosynthesis and growth within the C3 genus Limonium (Plumbaginaceae)', *New*

Phytologist. New Phytol, 203(3), pp. 989–999. doi: 10.1111/nph.12858.

Galmés, J., Kapralov, M. V., *et al.* (2014) 'Expanding knowledge of the Rubisco kinetics variability in plant species: Environmental and evolutionary trends', *Plant, Cell and Environment*. Plant Cell Environ, 37(9), pp. 1989–2001. doi: 10.1111/pce.12335.

Genkov, T. and Spreitzer, R. J. (2009) 'Highly conserved small subunit residues influence Rubisco large subunit catalysis', *Journal of Biological Chemistry*. Elsevier, 284(44), pp. 30105–30112. doi: 10.1074/jbc.M109.044081.

Getzoff, T. P. *et al.* (1998) 'Chimeric Arabidopsis thaliana Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase Containing a Pea Small Subunit Protein Is Compromised in Carbamylation', *Plant Physiology*. Plant Physiol, 116(2), pp. 695–702. doi: 10.1104/pp.116.2.695.

Gielly, L. and Taberlet, P. (1994) *The use of chloroplast DNA to resolve plant phylogenies: Noncoding versus rbcL sequences*, *Molecular Biology and Evolution*. doi: 10.1093/oxfordjournals.molbev.a040157.

Goodstein, D. M. *et al.* (2012) 'Phytozome: A comparative platform for green plant genomics', *Nucleic Acids Research*, 40(D1). doi: 10.1093/nar/gkr944.

Gout, J. F., Kahn, D. and Duret, L. (2010) 'The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution', *PLoS Genetics*. Public Library of Science, 6(5), p. 20. doi: 10.1371/journal.pgen.1000944.

Grabsztunowicz, M. *et al.* (2015) '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', *Physiologia Plantarum*. Physiol Plant, 154(4), pp. 591–608. doi: 10.1111/ppl.12322.

Gunn, L. H. *et al.* (2020) 'The dependency of red Rubisco on its cognate activase for enhancing plant photosynthesis and growth', *Proceedings of the National Academy of Sciences of the United States of America*, 117(41), pp. 25890–25896. doi: 10.1073/pnas.2011641117.

He, S. *et al.* (2020) 'The structural basis of Rubisco phase separation in the pyrenoid', *Nature Plants*. Springer US, 6(12), pp. 1480–1490. doi: 10.1038/s41477-020-00811-y.

Heureux, A. M. C. *et al.* (2017) 'The role of Rubisco kinetics and pyrenoid morphology in shaping the CCM of haptophyte microalgae', *Journal of Experimental Botany*, 68(14), pp. 3959–3969. doi: 10.1093/jxb/erx179.

Holland, H. D. (2002) 'Volcanic gases, black smokers, and the great oxidation event', *Geochimica et Cosmochimica Acta*, 66(21), pp. 3811–3826. doi: 10.1016/S0016-7037(02)00950-X.

Horken, K. M. and Tabita, F. R. (1999) 'The "green" form I ribulose 1,5-bisphosphate carboxylase/oxygenase from the nonsulfur purple bacterium Rhodobacter capsulatus', *Journal of Bacteriology*. J Bacteriol, 181(13), pp. 3935–3941. doi: 10.1128/jb.181.13.3935-3941.1999.

Huson, D. H. and Scornavacca, C. (2012) 'Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks', *Systematic Biology*. Oxford University Press, pp. 1061–1067. doi: 10.1093/sysbio/sys062.

Iñiguez, C. *et al.* (2020) 'Evolutionary trends in RuBisCO kinetics and their co-evolution with CO2 concentrating mechanisms', *Plant Journal*, 101(4), pp. 897–918. doi: 10.1111/tpj.14643.

Iñiguez, C. *et al.* (2021) 'Analyzing the causes of method-to-method variability among Rubisco kinetic traits: From the first to the current measurements', *Journal of Experimental Botany*, 72(22), pp. 7846–7862. doi: 10.1093/jxb/erab356.

Iqbal, W. A. *et al.* (2021) 'Rubisco substitutions predicted to enhance crop performance through carbon uptake modelling', *Journal of Experimental Botany*. Oxford Academic, 72(17), pp. 6066–6075. doi: 10.1093/jxb/erab278.

Ishikawa, C. *et al.* (2011) 'Functional incorporation of sorghum small subunit increases the catalytic turnover rate of rubisco in transgenic rice', *Plant Physiology*. Plant Physiol, 156(3), pp. 1603–1611. doi: 10.1104/pp.111.177030.

Joshi, J. *et al.* (2015) 'Role of small subunit in mediating assembly of red-type Form I Rubisco', *Journal of Biological Chemistry*, 290(2), pp. 1066–1074. doi: 10.1074/jbc.M114.613091.

Kanevski, I. *et al.* (1999) 'Plastome engineering of ribulose-1,5-bisphosphate

carboxylase/oxygenase in tobacco to form a sunflower large subunit and tobacco small subunit hybrid', *Plant Physiology*. American Society of Plant Biologists, 119(1), pp. 133–141. doi: 10.1104/pp.119.1.133.

Kapralov, M. V. *et al.* (2011) 'Changes in Rubisco kinetics during the evolution of C4 Photosynthesis in Flaveria (Asteraceae) are associated with positive selection on genes encoding the enzyme', *Molecular Biology and Evolution*, 28(4), pp. 1491–1503. doi: 10.1093/molbev/msq335.

Kapralov, M. V. and Filatov, D. A. (2007) 'Widespread positive selection in the photosynthetic Rubisco enzyme', *BMC Evolutionary Biology*, 7. doi: 10.1186/1471-2148-7-73.

Kasting, J. F. (1993) 'Earth's early atmosphere', *Science*. American Association for the Advancement of Science, 259(5097), pp. 920–926. doi: 10.1126/science.11536547.

Katoh, K. and Standley, D. M. (2013) 'MAFFT multiple sequence alignment software version 7: Improvements in performance and usability', *Molecular Biology and Evolution*. Oxford University Press, pp. 772–780. doi: 10.1093/molbev/mst010.

Kaufman, A. J. *et al.* (2007) 'Late archean biospheric oxygenation and atmospheric evolution', *Science*, 317(5846), pp. 1900–1903. doi: 10.1126/science.1138700.

Kostov, R. V., Small, C. L. and McFadden, B. A. (1997) 'Mutations in a sequence near the N-terminus of the small subunit alter the CO2/O2 specificity factor for ribulose bisphosphate carboxylase/oxygenase', *Photosynthesis Research*. Springer, 54(2), pp. 127–134. doi: 10.1023/A:1005967106993.

Krylov, D. M. *et al.* (2003) 'Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution', *Genome Research*. Cold Spring Harbor Laboratory Press, 13(10), pp. 2229–2235. doi: 10.1101/gr.1589103.

Kumar, S. *et al.* (2017) 'TimeTree: A Resource for Timelines, Timetrees, and Divergence Times', *Molecular biology and evolution*, 34(7), pp. 1812–1819. doi: 10.1093/molbev/msx116.

Lamesch, P. *et al.* (2012) 'The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools', *Nucleic Acids Research*. Oxford Academic, 40(D1), pp. D1202–D1210. doi: 10.1093/nar/gkr1090.

Lee, B., Berka, R. M. and Tabita, F. R. (1991) 'Mutations in the small subunit of cyanobacterial ribulose-bisphosphate carboxylase/oxygenase that modulate interactions with large subunits', *Journal of Biological Chemistry*, 266(12), pp. 7417–7422. doi: 10.1016/s0021-9258(20)89463-0.

Lee, B., Read, B. A. and Tabita, F. R. (1991) 'Catalytic properties of recombinant octameric, hexadecameric, and heterologous cyanobacterial/ bacterial ribulose-1,5-bisphosphate carboxylase/oxygenase', *Archives of Biochemistry and Biophysics*. Arch Biochem Biophys, 291(2), pp. 263–269. doi: 10.1016/0003-9861(91)90133-4.

Lee, B. and Tabita, F. R. (1990) 'Purification of Recombinant Ribulose-l,5-bisphosphate Carboxylase/Oxygenase Large Subunits Suitable for Reconstitution and Assembly of Active L8S8 Enzyme', *Biochemistry*, 29(40), pp. 9352–9357. doi: 10.1021/bi00492a007.

Lin, M. T. *et al.* (2021) 'A procedure to introduce point mutations into the Rubisco large subunit gene in wild-type plants', *Plant Journal*. John Wiley & Sons, Ltd, 106(3), pp. 876–887. doi: 10.1111/tpj.15196.

Lobkovsky, A. E., Wolf, Y. I. and Koonin, E. V. (2010) 'Universal distribution of protein evolution rates as a consequence of protein folding physics', *Proceedings of the National Academy of Sciences of the United States of America*, 107(7), pp. 2983–2988. doi: 10.1073/pnas.0910445107.

Van Lun, M. *et al.* (2014) 'CO2 and O2 distribution in rubisco suggests the small subunit functions as a CO2 reservoir', *Journal of the American Chemical Society*. American Chemical Society, 136(8), pp. 3165–3171. doi: 10.1021/ja411579b.

Matsumura, H. *et al.* (2020) 'Hybrid Rubisco with Complete Replacement of Rice Rubisco Small Subunits by Sorghum Counterparts Confers C4 Plant-like High Catalytic Activity', *Molecular Plant*. Cell Press, 13(11), pp. 1570–1581. doi: 10.1016/j.molp.2020.08.012.

Matsuzaki, R. *et al.* (2021) 'The Rubisco small subunits in the green algal genus Chloromonas provide insights into evolutionary loss of the eukaryotic carbon-concentrating organelle, the pyrenoid', *BMC ecology and evolution*. BioMed Central, 21(1), p. 11. doi: 10.1186/s12862-020-

01733-1.

McFadden, B. A. and Small, C. L. (1988) ‘Cloning, expression and directed mutagenesis of the genes for ribulose bisphosphate carboxylase/oxygenase’, *Photosynthesis Research*. Springer, 18(1–2), pp. 245–260. doi: 10.1007/BF00042987.

Meyer, M. and Griffiths, H. (2013) ‘Origins and diversity of eukaryotic CO2-concentrating mechanisms: Lessons for the future’, *Journal of Experimental Botany*, pp. 769–786. doi: 10.1093/jxb/ers390.

Meyer, M. T. *et al.* (2012) ‘Rubisco small-subunit α-helices control pyrenoid formation in Chlamydomonas’, *Proceedings of the National Academy of Sciences of the United States of America*, 109(47), pp. 19474–19479. doi: 10.1073/pnas.1210993109.

Mitchell, M. C. *et al.* (2017) ‘Pyrenoid loss impairs carbon-concentrating mechanism induction and alters primary metabolism in chlamydomonas reinhardtii’, *Journal of Experimental Botany*. Oxford Academic, 68(14), pp. 3891–3902. doi: 10.1093/jxb/erx121.

Nguyen, L. T. *et al.* (2015) ‘IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies’, *Molecular Biology and Evolution*, pp. 268–274. doi: 10.1093/molbev/msu300.

Nisbet, E. G. *et al.* (2007) ‘The age of Rubisco: The evolution of oxygenic photosynthesis’, *Geobiology*, 5(4), pp. 311–335. doi: 10.1111/j.1472-4669.2007.00127.x.

Nisbet, E. G. and Nisbet, R. E. R. (2008) ‘Methane, oxygen, photosynthesis, rubisco and the regulation of the air through time’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1504), pp. 2745–2754. doi: 10.1098/rstb.2008.0057.

Ogren, W. L. (1984) ‘Photorespiration: Pathways, Regulation, and Modification’, *Annual Review of Plant Physiology*. Annual Reviews, 35(1), pp. 415–442. doi: 10.1146/annurev.pp.35.060184.002215.

Ogren, W. L. and Bowes, G. (1971) ‘Ribulose diphosphate carboxylase regulates soybean photorespiration’, *Nature New Biology*. Nat New Biol, 230(13), pp. 159–160. doi: 10.1038/newbio230159a0.

Orr, D. J. *et al.* (2017) ‘Engineering photosynthesis: Progress and perspectives’, *F1000Research*. doi: 10.12688/f1000research.12181.1.

Pál, C., Papp, B. and Lercher, M. J. (2006) ‘An integrated view of protein evolution’, *Nature Reviews Genetics*, 7(5), pp. 337–348. doi: 10.1038/nrg1838.

Paradis, E., Claude, J. and Strimmer, K. (2004) ‘APE: Analyses of phylogenetics and evolution in R language’, *Bioinformatics*, 20(2), pp. 289–290. doi: 10.1093/bioinformatics/btg412.

Paradis, E. and Schliep, K. (2019) ‘Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R’, *Bioinformatics*, 35(3), pp. 526–528. doi: 10.1093/bioinformatics/bty633.

Parry, M. A. J. *et al.* (2003) ‘Manipulation of Rubisco: The amount, activity, function and regulation’, *Journal of Experimental Botany*. Oxford Academic, 54(386), pp. 1321–1333. doi: 10.1093/jxb/erg141.

Parry, M. A. J. *et al.* (2007) ‘Prospects for increasing photosynthesis by overcoming the limitations of Rubisco’, in *Journal of Agricultural Science*, pp. 31–43. doi: 10.1017/S0021859606006666.

Parry, M. A. J. *et al.* (2013) ‘Rubisco activity and regulation as targets for crop improvement’, *Journal of Experimental Botany*, 64(3), pp. 717–730. doi: 10.1093/jxb/ers336.

Read, B. A. and Tabita, F. R. (1992a) ‘A Hybrid Ribulosebisphosphate Carboxylase/Oxygenase Enzyme Exhibiting a Substantial Increase in Substrate Specificity Factor’, *Biochemistry*, 31(24), pp. 5553–5560. doi: 10.1021/bi00139a018.

Read, B. A. and Tabita, F. R. (1992b) ‘Amino Acid Substitutions in the Small Subunit of Ribulose-1, 5-bisphosphate Carboxylase/Oxygenase That Influence Catalytic Activity of the Holoenzyme’, *Biochemistry*, 31(2), pp. 519–525. doi: 10.1021/bi00117a031.

Rickaby, R. E. M. and Eason Hubbard, M. R. (2019) ‘Upper ocean oxygenation, evolution of RuBisCO and the Phanerozoic succession of phytoplankton’, *Free Radical Biology and Medicine*. Elsevier Inc., 140, pp. 295–304. doi: 10.1016/j.freeradbiomed.2019.05.006.

Ryu, J. Y., Kim, H. U. and Lee, S. Y. (2019) ‘Deep learning enables high-quality and high-

throughput prediction of enzyme commission numbers', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 116(28), pp. 13996–14001. doi: 10.1073/pnas.1821905116.

Saeed, R. and Deane, C. M. (2006) 'Protein protein interactions, evolutionary rate, abundance and age', *BMC Bioinformatics*, 7. doi: 10.1186/1471-2105-7-128.

Sakoda, K. *et al.* (2021) 'Effects of introduction of sorghum RbcS with rice RbcS knockdown by RNAi on photosynthetic activity and dry weight in rice', *Plant Production Science*. Taylor and Francis Ltd., 24(3), pp. 346–353. doi: 10.1080/1343943X.2020.1847669.

Savir, Y. *et al.* (2010) 'Cross-species analysis traces adaptation of Rubisco toward optimality in a low-dimensional landscape', *Proceedings of the National Academy of Sciences of the United States of America*, 107(8), pp. 3475–3480. doi: 10.1073/pnas.0911663107.

Schluter, D. *et al.* (1997) 'Likelihood of ancestor states in adaptive radiation', *Evolution*, 51(6), pp. 1699–1711. doi: 10.1111/j.1558-5646.1997.tb05095.x.

Schneider, G. *et al.* (1990) 'Comparison of the crystal structures of L2 and L8S8 Rubisco suggests a functional role for the small subunit', *EMBO Journal*. John Wiley & Sons, Ltd, 9(7), pp. 2045–2050. doi: 10.1002/j.1460-2075.1990.tb07371.x.

Schneider, G., Lindqvist, Y. and Brändén, C. I. (1992) 'RUBISCO: Structure and mechanism', *Annual Review of Biophysics and Biomolecular Structure*, pp. 119–143. doi: 10.1146/annurev.bb.21.060192.001003.

Seward, E. A. and Kelly, S. (2018) 'Selection-driven cost-efficiency optimization of transcripts modulates gene evolutionary rate in bacteria', *Genome Biology*, 19(1). doi: 10.1186/s13059-018-1480-7.

Sharkey, T. D. (2020) 'Emerging research in plant photosynthesis', *Emerging Topics in Life Sciences*. Portland Press Ltd., 4(2), pp. 137–150. doi: 10.1042/ETLS20200035.

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

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

Shih, P. M. *et al.* (2016) 'Biochemical characterization of predicted Precambrian RuBisCO', *Nature Communications*. Nature Publishing Group, 7. doi: 10.1038/ncomms10382.

Ślesak, I., Ślesak, H. and Kruk, J. (2017) 'RubisCO Early Oxygenase Activity: A Kinetic and Evolutionary Perspective', *BioEssays*. John Wiley and Sons Inc., 39(11). doi: 10.1002/bies.201700071.

Small, I. *et al.* (2004) 'Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences', *Proteomics*, 4(6), pp. 1581–1590. doi: 10.1002/pmic.200300776.

Smith, D. R. (2015) 'Mutation rates in plastid genomes: They are lower than you might think', *Genome Biology and Evolution*, 7(5), pp. 1227–1234. doi: 10.1093/gbe/evv069.

Soria-Hernanz, D. F. *et al.* (2008) 'Reconsidering the generation time hypothesis based on nuclear ribosomal ITS sequence comparisons in annual and perennial angiosperms', *BMC Evolutionary Biology*, 8(1). doi: 10.1186/1471-2148-8-344.

Soskine, M. and Tawfik, D. S. (2010) 'Mutational effects and the evolution of new protein functions', *Nature Reviews Genetics*, 11(8), pp. 572–582. doi: 10.1038/nrg2808.

Spreitzer, R. J. (2003) 'Role of the small subunit in ribulose-1,5-bisphosphate carboxylase/oxygenase', *Archives of Biochemistry and Biophysics*. Academic Press, 414(2), pp. 141–149. doi: 10.1016/S0003-9861(03)00171-1.

Studer, R. A. *et al.* (2014) 'Stability-activity tradeoffs constrain the adaptive evolution of RubisCO', *Proceedings of the National Academy of Sciences of the United States of America*, 111(6), pp. 2223–2228. doi: 10.1073/pnas.1310811111.

Su, Z. and Townsend, J. P. (2015) '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, p. 86. doi: 10.1186/s12862-015-0364-7.

Suyama, M., Torrents, D. and Bork, P. (2006) 'PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments', *Nucleic Acids Research*. Oxford University Press, 34(WEB. SERV. ISS.), p. W609. doi: 10.1093/nar/gkl315.

Swarbreck, D. *et al.* (2008) 'The Arabidopsis Information Resource (TAIR): Gene structure and function annotation', *Nucleic Acids Research*, 36(SUPPL. 1), pp. 1009–1014. doi: 10.1093/nar/gkm965.

Tabita, F. R. (1999) 'Microbial ribulose 1,5-bisphosphate carboxylase/oxygenase: A different perspective', *Photosynthesis Research*. Springer, 60(1), pp. 1–28. doi: 10.1023/A:1006211417981.

Tabita, F. R., Satagopan, S., *et al.* (2008) 'Distinct form I, II, III, and IV Rubisco proteins from the three kingdoms of life provide clues about Rubisco evolution and structure/function relationships', *Journal of Experimental Botany*, 59(7), pp. 1515–1524. doi: 10.1093/jxb/erm361.

Tabita, F. R., Hanson, T. E., *et al.* (2008) 'Phylogenetic and evolutionary relationships of RubisCO and the RubisCO-like proteins and the functional lessons provided by diverse molecular forms', *Philosophical Transactions of the Royal Society B: Biological Sciences*. Philos Trans R Soc Lond B Biol Sci, 363(1504), pp. 2629–2640. doi: 10.1098/rstb.2008.0023.

Tcherkez, G. G. B., Farquhar, G. D. and Andrews, T. J. (2006) 'Despite slow catalysis and confused substrate specificity, all ribulose bisphosphate carboxylases may be nearly perfectly optimized', *Proceedings of the National Academy of Sciences of the United States of America*, 103(19), pp. 7246–7251. doi: 10.1073/pnas.0600605103.

Tóth-Petróczy, Á. and Tawfik, D. S. (2011) 'Slow protein evolutionary rates are dictated by surface - core association', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 108(27), pp. 11151–11156. doi: 10.1073/pnas.1015994108.

Tourasse, N. J. and Li, W. H. (2000) 'Selective constraints, amino acid composition, and the rate of protein evolution', *Molecular Biology and Evolution*. Oxford Academic, 17(4), pp. 656–664. doi: 10.1093/oxfordjournals.molbev.a026344.

Uemura, K. *et al.* (1997) 'Ribulose-1,5-bisphosphate carboxylase/oxygenase from thermophilic red algae with a strong specificity for CO2 fixation', *Biochemical and Biophysical Research Communications*, 233(2), pp. 568–571. doi: 10.1006/bbrc.1997.6497.

Ungerer, J. *et al.* (2018) 'Adjustments to photosystem stoichiometry and electron transfer proteins are key to the remarkably fast growth of the Cyanobacterium synechococcus elongatus UTEX 2973', *mBio*, 9(1). doi: 10.1128/mBio.02327-17.

van der Vies, S. M., Bradley, D. and Gatenby, A. A. (1986) ' Assembly of cyanobacterial and higher plant ribulose bisphosphate carboxylase subunits into functional homologous and heterologous enzyme molecules in Escherichia coli ', *The EMBO Journal*, 5(10), pp. 2439–2444. doi: 10.1002/j.1460-2075.1986.tb04519.x.

VOORDOUW, G. *et al.* (1987) 'Site-directed mutagenesis of the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase from Anacystis nidulans', *European Journal of Biochemistry*. Eur J Biochem, 163(3), pp. 591–598. doi: 10.1111/j.1432-1033.1987.tb10908.x.

Wall, D. P. *et al.* (2005) 'Functional genomic analysis of the rates of protein evolution', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 102(15), pp. 5483–5488. doi: 10.1073/pnas.0501761102.

Wang, Y. L. *et al.* (2001) 'Properties of hybrid enzymes between Synechococcus large subunits and higher plant small subunits of ribulose-1,5-bisphosphate carboxylase/oxygenase in Escherichia coli', *Archives of Biochemistry and Biophysics*. Arch Biochem Biophys, 396(1), pp. 35–42. doi: 10.1006/abbi.2001.2555.

Weller, C. and Wu, M. (2015) 'A generation-time effect on the rate of molecular evolution in bacteria', *Evolution*, 69(3), pp. 643–652. doi: 10.1111/evo.12597.

Whitney, S. M. *et al.* (2001) 'Form I Rubiscos from non-green algae are expressed abundantly but not assembled in tobacco chloroplasts', *Plant Journal*, 26(5), pp. 535–547. doi: 10.1046/j.1365-313X.2001.01056.x.

Whitney, S. M. *et al.* (2011) 'Isoleucine 309 acts as a C 4 catalytic switch that increases ribulose-

1,5-bisphosphate carboxylase/oxygenase (rubisco) carboxylation rate in flaveria', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 108(35), pp. 14688–14693. doi: 10.1073/PNAS.1109503108/-/DCSUPPLEMENTAL.

Whitney, S. M., Houtz, R. L. and Alonso, H. (2011) 'Advancing our understanding and capacity to engineer nature's CO2-sequestering enzyme, Rubisco', *Plant Physiology*, 155(1), pp. 27–35. doi: 10.1104/pp.110.164814.

Wilson, R. H. *et al.* (2018) 'An improved Escherichia coli screen for Rubisco identifies a protein-protein interface that can enhance CO2-fixation kinetics', *Journal of Biological Chemistry*, 293(1), pp. 18–27. doi: 10.1074/jbc.M117.810861.

Wilson, R. H., Alonso, H. and Whitney, S. M. (2016) 'Evolving Methanococcoides burtonii archaeal Rubisco for improved photosynthesis and plant growth', *Scientific Reports*, 6. doi: 10.1038/srep22284.

Wolfe, K. H., Li, W. H. and Sharp, P. M. (1987) 'Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs.', *Proceedings of the National Academy of Sciences of the United States of America*, 84(24), pp. 9054–9058. doi: 10.1073/pnas.84.24.9054.

Wunder, T. *et al.* (2018) 'The phase separation underlying the pyrenoid-based microalgal Rubisco supercharger', *Nature Communications*. Springer US, 9(1), pp. 1–10. doi: 10.1038/s41467-018-07624-w.

Young, J. N. *et al.* (2016) 'Large variation in the Rubisco kinetics of diatoms reveals diversity among their carbon-concentrating mechanisms', *Journal of Experimental Botany*, 67(11), pp. 3445–3456. doi: 10.1093/jxb/erw163.

Yu, J. *et al.* (2015) 'Synechococcus elongatus UTEX 2973, a fast growing cyanobacterial chassis for biosynthesis using light and CO_2', *Scientific reports*, 5, p. 8132. doi: 10.1038/srep08132.

Zhou, Y. and Whitney, S. (2019) 'Directed evolution of an improved Rubisco; in vitro analyses to decipher fact from fiction', *International Journal of Molecular Sciences*, 20(20). doi: 10.3390/ijms20205019.

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

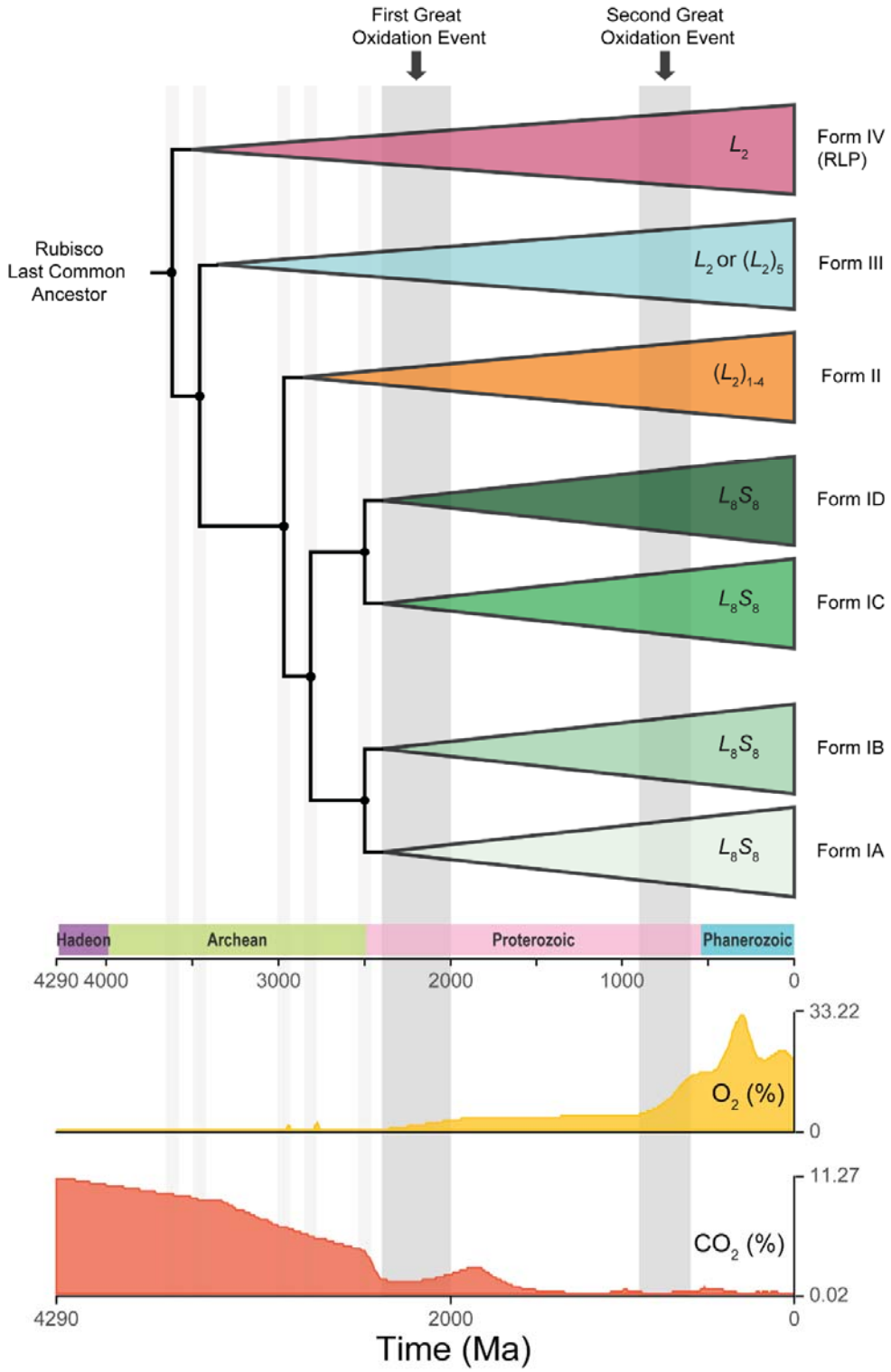

Figure 1. The evolutionary history of rubisco in the context of atmospheric CO_2 (%) and O_2 (%) following divergence from the ancestral rubisco-like protein (RLP). For ease of visualisation, branch points in the phylogeny are indicated by grey vertical bars. The First and Second Great Oxidation events 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; (Kumar *et al.*, 2017)).

Figure 2

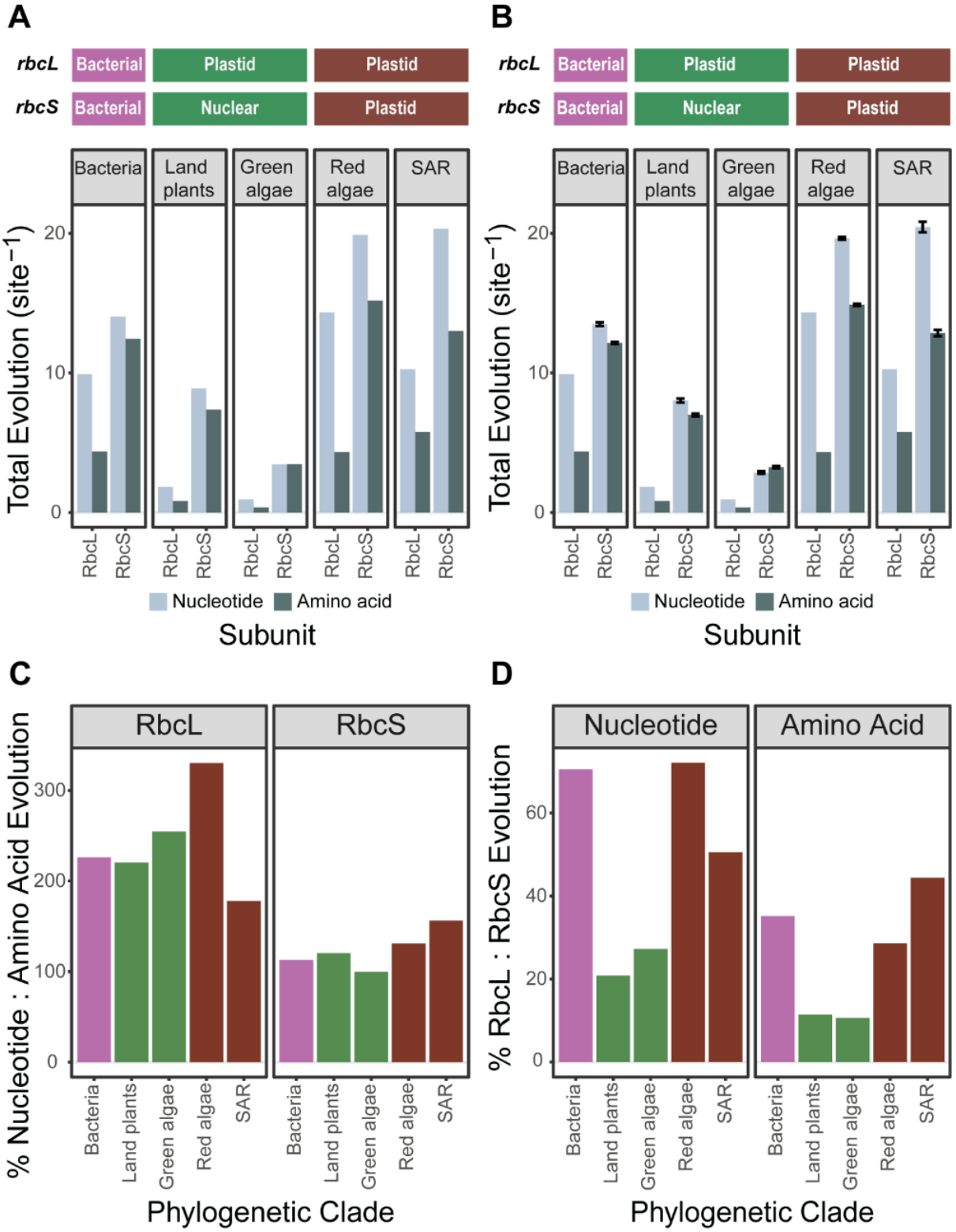
A
rbcL
Bacterial
Plastid
Plastid
rbcS
Bacterial
Nuclear
Plastid
Bacteria
Land plants
Green algae
Red algae
SAR
Total Evolution (site^{-1})
20
10
0
RbcL
RbcS
Nucleotide
Amino acid
Subunit
B
C
RbcL
RbcS
% Nucleotide : Amino Acid Evolution
300
200
100
Phylogenetic Clade
D
Nucleotide
Amino Acid
% RbcL : RbcS Evolution
60
40

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 1000 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 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 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 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 the same as described in (C).

Figure 3

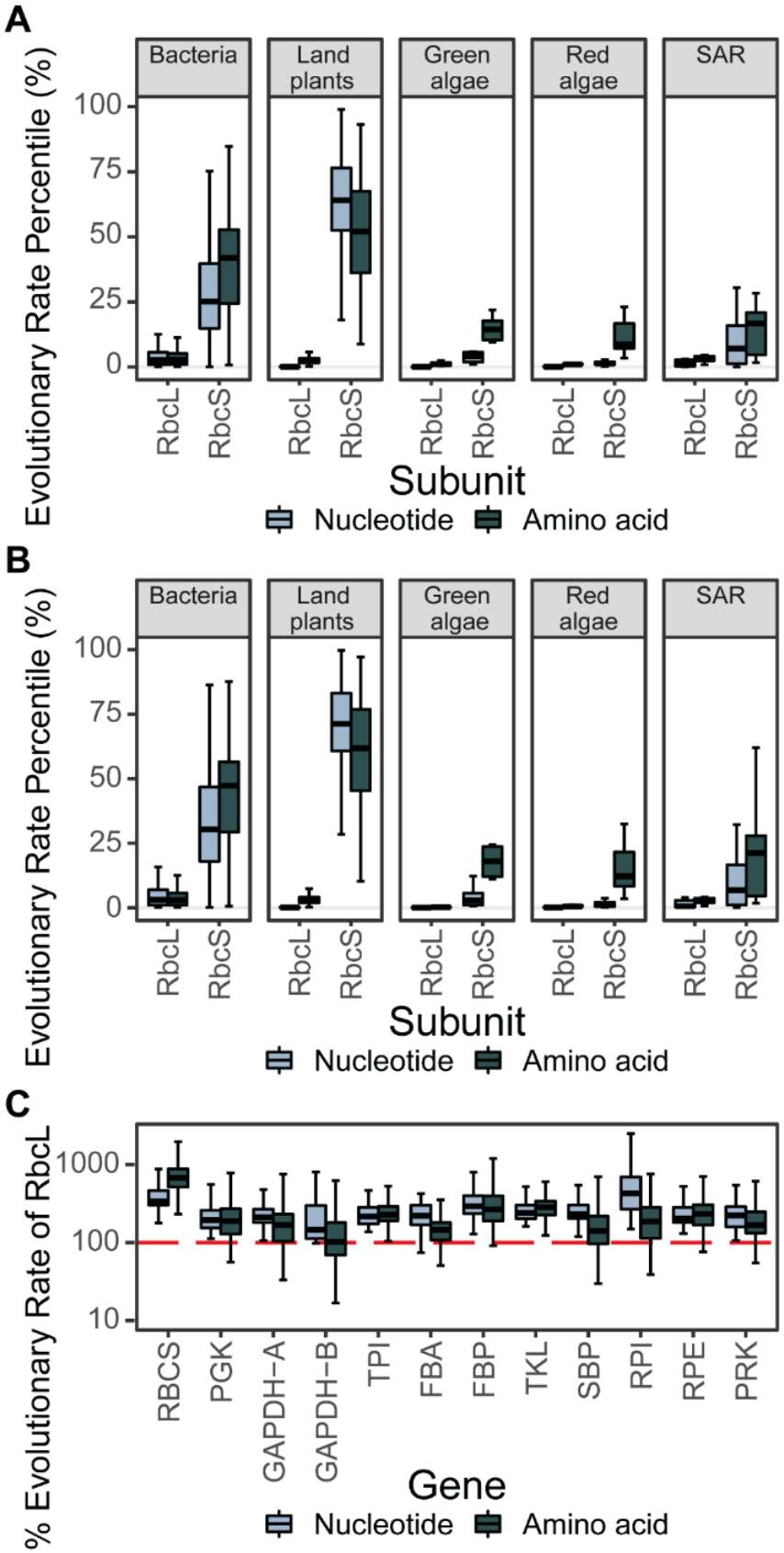
A
Bacteria
Land plants
Green algae
Red algae
SAR
Evolutionary Rate Percentile (%)
100
75
50
25
0
RbcL
RbcS
Subunit
Nucleotide
Amino acid
B
Bacteria
Land plants
Green algae
Red algae
SAR
Evolutionary Rate Percentile (%)
100
75
50
25
0
RbcL
RbcS
Subunit
Nucleotide
Amino acid
C
% Evolutionary Rate of RbcL
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 of all other genes in each taxonomic group (A and B), or in context of the Calvin-Benson-Bassham cycle enzymes in land plants (C). **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. Within each taxonomic group, percentile extents of rubisco evolution were calculated for each pairwise comparison of species for which whole genome data could be acquired. Summary data for this figure can be found in Supplemental File 1, table S1. **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. Summary data for this figure can be found in Supplemental File 1, table S2. **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). Percentage extents of evolution were calculated between each possible pairwise comparison of land plant species for which a nuclear genome could be acquired (see Methods). Calvin-Bensen-Bassham cycle enzymes are listed in the order in which their reactions occur. 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. The y axis is on a log scale. Summary data for this figure can be found in Supplemental File 1, table S3.

Figure 4

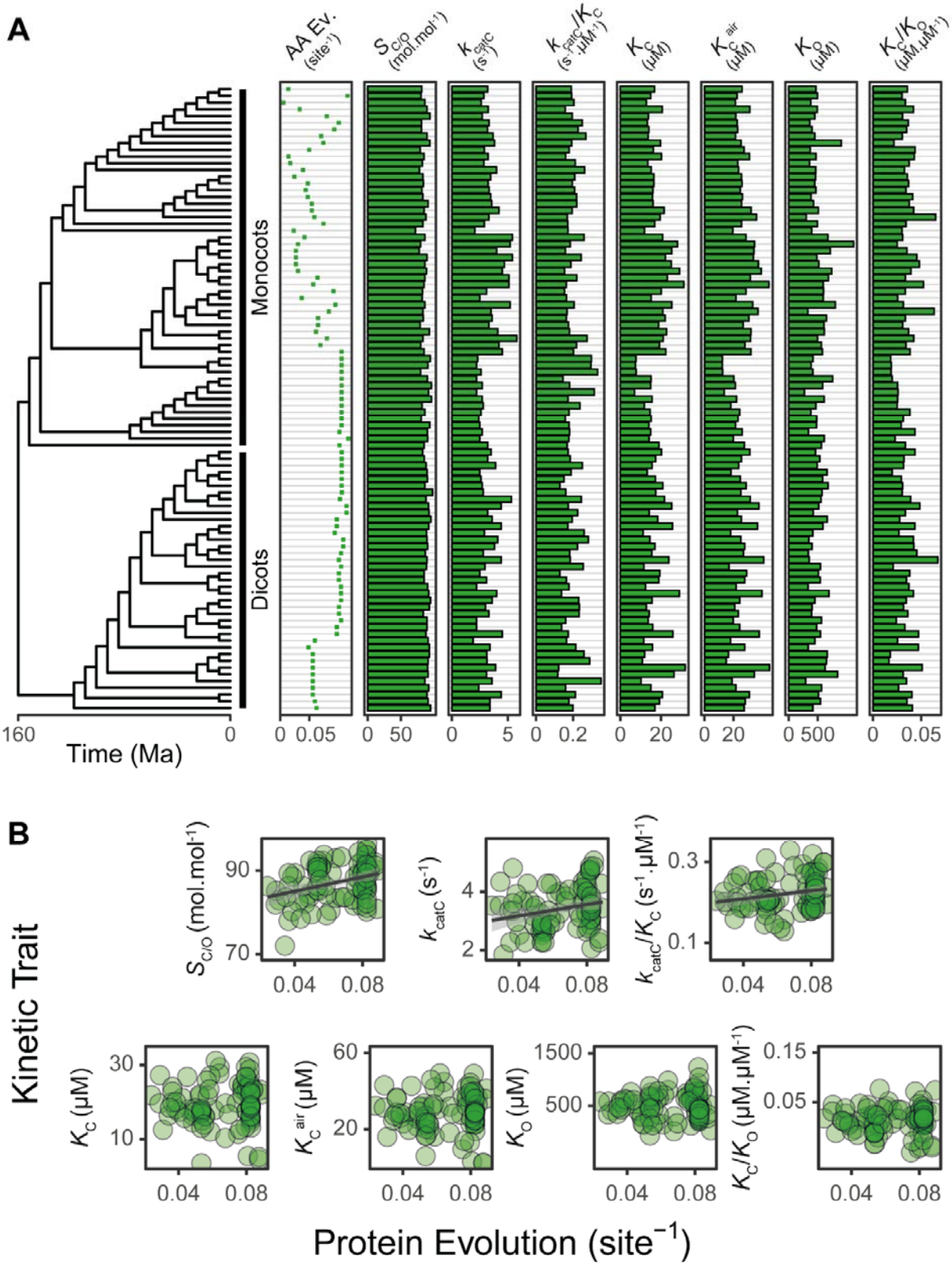
A
AA Ev. (site⁻¹)
$S_{C/O}$ (mol.mol⁻¹)
k_{catC} (s⁻¹)
k_{catC}/K_C (s⁻¹.µM⁻¹)
K_C (µM)
K_C^{air} (µM)
K_O (µM)
K_C/K_O (µM.µM⁻¹)
Monocots
Dicots
160
0
Time (Ma)
0 0.05
0 50
0 5
0 0.2
0 20
0 20
0 500
0 0.05
B
Kinetic Trait
$S_{C/O}$ (mol.mol⁻¹)
k_{catC} (s⁻¹)
k_{catC}/K_C (s⁻¹.µM⁻¹)
K_C (µM)
K_C^{air} (µM)
K_O (µM)
K_C/K_O (µM.µM⁻¹)
Protein Evolution (site⁻¹)

Figure 4. The relationship between rubisco molecular and kinetic evolution in C_3 angiosperms. **A)** The extent of protein evolution that has occurred in the rubisco large subunit (RbcL) and the corresponding kinetic trait values of the enzyme across angiosperm species. 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. Respective units for each trait are shown in column labels. The phylogeny has been inferred based on similarity of RbcL protein sequence and is displayed as a cladogram. The time point of angiosperm divergence (160 million years) was estimated using the *TimeTree* resource (http://www.timetree.org; (Kumar *et al.*, 2017)). **B)** The correlations 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 pairwise correlation coefficients and associated P-values for each regression model are summarised in table 2. The y axis on each plot is displayed on a log scale.

Tables

Table 1

Table 1. Summary of the sequence dataset used for analysis of rubisco molecular evolution in this study.

Taxonomic Group	*n*			
	Species	**Species Representative Gene Models**	***rbcL*/RbcL Sequences**	***rbcS*/RbcS Sequences**
Bacteria	78	33	78	86
Land plants	68	27	68	172
Green algae	12	4	12	24
Red algae	201	7	201	595
SAR	129	9	129	283
Total	**488**	**80**	**488**	**1140**

Table 2

Table 2. The pairwise correlation coefficient (% variation explained) and associated significance from the results of the linear regression analysis between the extent of RbcL protein evolution and each rubisco kinetic trait. Significance values are represented as α levels, where α = 0.001 if P < 0.001, α = 0.01 if 0.001 < P < 0.01, α = 0.05 if 0.01 < P < 0.05, and α = ns if P > 0.05.

Kinetic Trait	% Var Explained	*α*
$S_{C/O}$	10.1	0.01
k_{catC}	4.6	0.05
K_{catC}/k_C	3.8	0.05
K_C	ns	ns
K_C^{air}	ns	ns
K_O	ns	ns
K_C/K_O	ns	ns

Table 3

Table 3. 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. Mean values of rubisco kinetic traits and associated variation (± 1 S.E.) in extant C_3 species are shown for comparison.

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	81.0 ± 1.0	2.6 ± 1.1	0.16 ± 1.1	17.6 ± 0.5	26.4 ± 0.7	517.2 ± 14.7	0.035 ± 0.001
Extant Angiosperms	87.1 ± 0.5	3.4 ± 0.1	0.20 ± 0.005	17.6 ± 0.5	26.4 ± 0.7	517.2 ± 14.7	0.035 ± 0.001