Origin of photosynthetic water oxidation at the dawn of life

Thomas Oliver[1], Patricia Sánchez-Baracaldo[2], Anthony W. Larkum[3], A. William Rutherford[1], Tanai Cardona[1]*

[1]Department of Life Sciences, Imperial College London, London, UK
[2]School of Geographical Sciences, University of Bristol, Bristol, UK
[3]University of Technology Sydney, Ultimo NSW, Australia

*Correspondence to: t.cardona@imperial.ac.uk

Abstract

Oxygenic photosynthesis starts with the oxidation of water to O_2. Cyanobacteria are the only known prokaryotes capable of oxygenic photosynthesis and therefore, it is assumed that water oxidation is a late innovation relative to the origin of life. However, when exactly oxygenic photosynthesis originated remains lively debated. Here we show that the origin of photosystem II, the water-splitting enzyme, occurred at an early stage during the evolution of life and long before the origin of Cyanobacteria. We use relaxed molecular clocks, ancestral sequence reconstruction, and comparative structural biology to demonstrate that photosystem II exhibits patterns of evolution through geological time that are indistinguishable from those of ATP synthase, RNA polymerase, or the ribosome—some of the oldest known enzymes. Our work suggests that water oxidation originated during the establishment of bioenergetics reaching farther into the past than can be documented based on species trees alone.

Introduction

The origin of oxygenic photosynthesis is considered a turning point in the history of life marking the transition from the ancient world of anaerobes into a productive aerobic world that permitted the emergence of complex life[1]. Oxygenic photosynthesis starts with photosystem II (PSII), the water-oxidising and O_2-evolving enzyme of Cyanobacteria and photosynthetic eukaryotes. Today PSII is a highly conserved, multicomponent, membrane protein complex, which was inherited by the *most recent common ancestor* (MRCA) of Cyanobacteria in a form that is structurally and functionally quite similar to that found in all

extant species[2]. Thus, the origin of oxygenic photosynthesis antedates the MRCA of Cyanobacteria by an undetermined amount of time. Cyanobacteria's closest living relatives are the clades known as Vampirovibrionia (formerly Melainabacteria)[3,4], followed by Sericytochromatia[5] and Margulisbacteria[6]. Currently, no photosynthetic strains have been described in these groups of uncultured bacteria and this has led to the hypothesis that oxygenic photosynthesis arose during the time spanning the divergence of Vampirovibrionia and the MRCA of Cyanobacteria starting from an entirely non-photosynthetic ancestral state[5,7]. Molecular clock studies have suggested that the span of time between Cyanobacteria and Vampirovibrionia could be of several hundred million years[8,9]. However, it is still unclear from molecular clock analyses and the microfossil record when exactly the MRCA of Cyanobacteria started to diversify[10]. Alternatively, it is possible that these novel clades emerged from photosynthetic ancestors. In agreement with this, we recently showed that the span of time between the origin of a water-splitting photosystem and the MRCA of Cyanobacteria, a period of time that we refer to as ΔT, could be well over a billion years[11]. This would imply that Margulisbacteria, Sericytochromatia, and Vampirovibrionia (MSV) could have been ancestrally more like Cyanobacteria than is apparent from their specialized heterotrophic lifestyles. Hence, to time the origin of oxygenic photosynthesis one should first resolve when PSII evolved water-oxidation photochemistry relative to the MRCA of Cyanobacteria.

The heart of PSII is made up of a heterodimeric reaction centre (RC) *core* coupled to a core *antenna*. The two subunits of the RC core of PSII are known as D1 and D2, and these are associated respectively with the antenna subunits known as CP43 and CP47. D1 and CP43 make up one monomeric half of the RC and D2 and CP47 the other half. Water oxidation is catalysed by a Mn_4CaO_5 cluster coordinated by ligands from both D1 and CP43[12,13]. The cluster is functionally coupled to a redox active tyrosine-histidine pair (Y_Z-H190) also located in D1, which relays electrons from Mn to the oxidised chlorophyll pigments of the RC during charge separation[14]. In a cycle of four consecutive light-driven charge separation events, O_2 is released in the decomposition of two water molecules. Photosystems evolved first as homodimers[15,16]: therefore, the core and the antenna of PSII originated from ancestral gene duplication events that antedated the MRCA of Cyanobacteria. In this way CP43/D1 retain sequence and structural identity with CP47/D2. The conserved structural and functional traits between CP43/D1 and CP47/D2 suggest that the ancestral PSII homodimer—prior to the duplication events—was not only structurally similar to

heterodimeric PSII, but also that it split water and had evolved protective mechanisms against the formation of reactive oxygen species[11,16,17].

Here, to help in timing the evolution of oxygenic photosynthesis, Cyanobacteria and MSV, we compared the duplication leading to the RC antenna subunits, CP43 and CP47, to several well-defined ancient and more recent events: including, but not limited to, the duplication of the core catalytic subunits of ATP synthase, a very ancient event generally accepted to have occurred before the last universal common ancestor (LUCA)[18-22]; the evolution of RNA polymerase catalytic subunit β (RpoB) and ribosomal proteins, which are universally conserved and widely accepted to have originated before the LUCA[23-26]. We further constrain our analysis using *in silico* ancestral sequence reconstruction of PSII and through strict structural and functional rationales. We show that the duplications of the core subunits of PSII, which were driven by the selective pressures imposed by water oxidation, have patterns of molecular evolution that are nearly indistinguishable to those of ATP synthase or the early divergence of bacterial and archaeal RNA polymerase. It is suggested that the duplications of the core subunits of PSII, antedate the rise of Cyanobacteria and could have occurred during the establishment of the earliest bioenergetic processes.

Results

Phylogenetic overview

The phylogenies of CP43 and CP47 show that there is a much greater diversity of CP43 and CP43-derived subunits than CP47 (Figure 1). This difference is the result of a greater number of gene duplication in CP43 than CP47 and mirrors the evolution of D1 and D2[2], in which D1 has undergone more duplication events than D2. CP43 can be divided into two major groups: those that are assembled into PSII and can bind the Mn_4CaO_5 cluster, and those which have evolved to be used only as light harvesting complexes[27,28], usually known as chlorophyll binding proteins (CBP). The CBP are characterized by the loss of the extrinsic loop between the 5th and 6th transmembrane helix, where the ligands to the cluster are located (Figure 1b). This large extrinsic loop is found in both CP43 and CP47 and interacts directly with the electron donor side of PSII. The unrooted tree of CP43 is consistent with CBP having a single origin likely occurring before the MRCA of Cyanobacteria (Supplementary Figure S1) but have undergone an extensive duplication-driven diversification process. It also mirrors the evolution of D1 in that duplications appear to have occurred before the MRCA of Cyanobacteria[2]. The earliest of these D1 duplications also led to variants that have lost the capacity to bind the Mn_4CaO_5 cluster[2,29], but are likely used in other supporting functions.

Most notably, during chlorophyll *f* synthesis[30] in the far-red light acclimation response (farlip)[31].

CP43 and CP47 are also distantly related to the antenna domain of cyanobacterial PSI and that of the type I RCs of phototrophic Chlorobi, Acidobacteria and Heliobacteria (Supplementary Figure S2). We found that the level of sequence identity between any two type I RC proteins is always greater than between a type I RC protein and CP43/CP47 (Supplementary Table S1). Therefore, the distance between CP43/CP47 and other type I antenna domains is the largest distance in the molecular evolution of RC proteins after that between type I and type II RC core domains. The phylogenetic relationships between type I and type II RCs have been reviewed in detail before[17,32,33]. These are briefly summarized and schematized in Supplementary Figure S3.

The phylogeny of Alpha and Beta subunits of the F-type ATP synthase showed that all Cyanobacteria have a F-type ATP synthase, and a fewer number of strains have an additional Na^+-translocating ATPase (N-ATPase) of the bacterial F-type, as had been reported before[34]. We found that MSV have a standard F-type ATP synthase (Supplementary Figure S4), but some N-ATPase Alpha and Beta sequences were also found in Vampirovibrionia and Sericytochromatia datasets, but not in Margulisbacteria. In this study we focused on the standard F-type ATP synthase of Cyanobacteria for further analysis.

The phylogeny of bacterial RNA polymerase subunit β (RpoB), with the intention of molecular clock analysis, focused on Cyanobacteria and MSV, as well as phyla with known phototrophic representatives and included Thermotogae and Aquificae as potential roots (Supplementary Figure S5). The tree was largely consistent with previous observed relationships between the selected groups[35], within Cyanobacteria and MSV, and within other phototrophs and their non-phototrophic relatives[36,37]. The only exception was Aquificae, which branched as a sister clade to Acidobacteria, a feature that had been reported before[25], and likely represents an ancient horizontal gene transfer event.

Distances and rates

To gather temporal information, we compared the phylogenetic distances between CP43 and CP47, Alpha and Beta subunits of the F-type ATP synthase, and archaeal and bacterial RpoB (visualized in Figure 2, but see also Figures S1b, S2 and S3b). We found that the distances between Alpha and Beta, and the divergence of archaeal and bacterial RpoB, are very large relative to the distance between the divergence of Vampirovibrionia and Cyanobacteria. In the case of RpoB, the distance between Vampirovibrionia and Cyanobacteria is about a fifth

of the distance between Archaea and Bacteria. However, the distance between CP43 and CP47 (but also between D1 and D2[11]) is of similar magnitude to that between Alpha and Beta, and to that between archaeal and bacterial RpoB, but substantially surpasses the distance between MVS and Cyanobacteria (Figure 2). These observations suggest that ancestral proteins to CP43/CP47 and D1/D2 existed before the divergences of MVS.

We compared the within-group mean distances for Alpha, Beta, RpoB, and a concatenated dataset of ribosomal proteins compiled in a previous independent study[36] (see Supplementary Table S2). We found consistently, that Vampirovibrionia and Margulisbacteria have larger within-group mean distances compared to Cyanobacteria, which suggests greater rates of evolution in the non-photosynthetic clades. These were consistently larger for Margulisbacteria relative to the other two groups. For example, RpoB in Vampirovibrionia and Margulisbacteria showed 1.6 and 4.0 times larger corrected mean distances than Cyanobacteria, respectively (Supplementary Table S2). At the level of the concatenated ribosomal proteins dataset, Margulisbacteria showed an almost 2-fold larger within-group mean distance than Cyanobacteria.

We then compared the rates of evolution of CP43 and CP47 with those of Alpha and Beta using a Bayesian relaxed molecular clock approach with identical calibrations, molecular clock parameters, and a simplified, but highly constrained sequence dataset (see Materials and Methods for an expanded rationale). We used an autocorrelated log normal model of rate variation with a non-parametric CAT+Γ model of amino acid substitutions to extract rates of evolution. We will refer to the span of time between the duplication points leading to Alpha and Beta (*d*AB), or to CP43 and CP47 (*d*CP), and the MRCA of Cyanobacteria as ΔT (schematized in Figure 3).

In Figure 4a to d we examine the changes in the rate of evolution under specific evolutionary scenarios. In the case of ATP synthase, we first assumed that the MRCA of Cyanobacteria occurred after the GOE, at about 1.7 Ga, and that *d*AB occurred at 3.5 Ga (**ΔT = 1.8 Ga**). Under these conditions the average rate of evolution of Alpha and Beta is calculated to be 0.28 ± 0.06 substitutions per site per Ga (δ Ga^{-1}). We will refer to the average rate through the Proterozoic as ν_{min}. In this scenario, the rate of evolution at the point of duplication, which we denote ν_{max}, is 7.32 ± 1.00 δ Ga^{-1} making ν_{max}/ν_{min} 26. In other words, when the span of time between the ancestral pre-LUCA duplication and the MRCA of Cyanobacteria is 1.8 Ga, the rate of evolution at the point of duplication is about 26 times greater than any rate observed through the diversification of Cyanobacteria or photosynthetic eukaryotes.

Now, if we consider a scenario in which *d*AB is 4.0 Ga and leaving all other constraints unchanged, ν_{max} is 6.02 ± 0.9 δ Ga^{-1} resulting in a ν_{max}/ν_{min} of 21. If instead we keep the duplication at 3.5 Ga but assume that the MRCA of Cyanobacteria occurred before the GOE at 2.6 Ga (**ΔT = 1.1 Ga**), we obtain that ν_{min} is consequently slower, 0.25 ± 0.06 δ Ga^{-1}, when compared to a post-GOE ancestor. This older MRCA (smaller ΔT) thus leads to a rise in ν_{max}, calculated to be 10.22 ± 1.37 δ Ga^{-1} and leading to a ν_{max}/ν_{min} of 40. Given that the phylogenetic distance is a constant, the rate of evolution increases with a decrease in ΔT following a power law function. The change in ν_{max}/ν_{min} as a function of ΔT is shown in Figure 4d. We had observed nearly identical evolutionary patterns for the core RC proteins D1 and D2 of PSII[11].

The core antenna of PSII, CP43 and CP47, showed patterns of divergence very similar to those of Alpha and Beta, both in terms of phylogenetic distances between paralogues and rates of evolution between orthologues (Figure 4a and b). The average rate of evolution of CP43 and CP47, assuming that the MRCA of Cyanobacteria occurred at 1.7 Ga, and the duplication (*d*CP) at 3.5 Ga (**ΔT = 1.8 Ga**), is 0.19 ± 0.05 δ Ga^{-1}. Slightly slower than for Alpha and Beta under the same condition. This slower rate is consistent with the fact that CP43 and CP47 show less sequence divergence between orthologues at all taxonomic ranks of oxygenic phototrophs when compared to Alpha and Beta (see Supplementary Table S3). Furthermore, the rate at *d*CP, ν_{max}, was 5.17 ± 0.84 δ Ga^{-1}, generating a ν_{max}/ν_{min} of 27, similar to Alpha and Beta (Figure 4). Thus, even when ΔT is 1.8 Ga, the rate at duplication point needs to be 27 times greater than the average rates observed during the Proterozoic. If we consider instead that the MRCA of Cyanobacteria occurred at 2.6 Ga and *d*CP at 3.5 Ga (**ΔT = 1.1 Ga**), this would slowdown ν_{min} to 0.16 ± 0.04 δ Ga^{-1}, while ν_{max} would increase to 7.81 ± 1.01 δ Ga^{-1} resulting in a ν_{max}/ν_{min} of 49. Therefore, the molecular evolution of the core subunits of PSII parallels that of ATP synthase both in terms of rates and distances through geological time.

We then studied a relatively recent gene duplication event (Figure 4c), which occurred long after the LUCA, but also after the MRCA of Cyanobacteria: that leading to Cyanobacteria-specific FtsH1 and FtsH2 (*d*H0)[38]. This more recent duplication served as a point of comparison and control (see Figure 3 for a scheme). In marked contrast to *d*AB, the rate at the point of duplication was 0.66 ± 0.21 δ Ga^{-1}. We found that FtsH1 is evolving at an average rate of 1.42 ± 0.29, while FtsH2 at a rate of 0.24 ± 0.06 δ Ga^{-1} under the assumption that MRCA of Cyanobacteria occurred at 1.7 Ga. Thus, under the assessed conditions, FtsH1 is evolving about 5.3 times faster than FtsH2, while the latter is evolving at a rate similar to

that of Alpha and Beta. If the MRCA of Cyanobacteria is assumed to have occurred at 2.6 Ga, all rates slowdown respectively, but the rate of FtsH1 remains over five times faster than FtsH2. Unlike *d*AB, *d*H0 is consistent with classical neofunctionalization, in which the copy that gains new function experiences an acceleration of the rate of evolution[39,40]. Like PSII and ATP synthase, the calculated rates of evolution match observed distances as estimated by the change in the level of sequence identity as a function of time, in which the fastest evolving FtsH1 accumulated greater sequence change than FtsH2 in the same period (Supplementary Table S3).

Given that the complex evolution of CP43 and CBP involved several major duplication events and potentially large variations in the rate of evolution (Figure 1 and Supplementary Figure S1), we carried out a molecular clock analysis of a large dataset of 392 CP43 and 465 CBP proteins using cross-calibrations across paralogues. Clocks were executed with no constraint on the MRCA of Cyanobacteria. The estimated mean divergence time for the oldest node, the duplication at the origin of CBP, is 2.23 Ga (95% confidence interval, CI: 1.90 - 2.69 Ga) using an autocorrelated log normal model (see Figure 4e, Supplementary Figure S6 for a chronogram, and Supplementary Table S4 for a comparison of estimated ages under different models). The mean divergence time for the node representing the CP43 inherited by the MRCA of Cyanobacteria was calculated to be 2.22 Ga (95% CI: 1.88 - 2.68 Ga). Thus, a span of time of only 15 Ma is seen between these two mean ages. The average rate of evolution of CP43, not including CBP sequences, was found to be 0.14 ± 0.05 δ Ga^{-1}, which is in the same range as determined in the simplified, but highly constrained experiment above. We noted a 6-fold increase in the rate of evolution associated with the duplication leading to the farlip-CP43 variant (Figure 4e). This duplication led to an acceleration of the rate similar in magnitude to that of FtsH1/FtsH2 and is consistent with a neofunctionalization process as the photosystems evolved to use far-red light to drive water-splitting.

CBP sequences, on average, display rates of evolution about three times faster than CP43 (Figure 4e). However, the serial duplications that led to the evolution of CP43-derived light harvesting complexes resulted in accelerations in the rate of evolution of a similar magnitude as observed for *d*AB and *d*CP. The largest of these is associated with the origin of PcbC[28], a variant commonly found in heterocystous Cyanobacteria and Cyanobacteria that use alternative pigments, such as chlorophyll *b*, *d* and *f*. The ancestral node of PcbC was timed at 2.07 Ga (95% CI: 1.76 - 2.50 Ga) with a rate of 11.7 ± 2.42 δ Ga^{-1}, decelerating quickly, but stabilizing at about four times faster rates than the average rate of CP43. We find it noteworthy that the fast rates of evolution associated with the origin of CBP are not

associated with very large spans of time between these and CP43, nor did it result in very old root node ages despite the use of very broad constraints.

Species divergence

To understand the evolution of MSV relative to Cyanobacteria we wished to apply a molecular clock to a system where the calculated rates could be compared to observed rates as determined by distances between species of known divergence times or at similar taxonomic ranks. We found RpoB to be suitable for this because it has been inherited vertically with few instances of horizontal gene transfer and had enough signal to resolve known phylogenetic relationships between and within clades. In collecting the RpoB sequences, we noted for the first time that Margulisbacteria and Vampirovibrionia share a comparatively greater level of divergence at similar taxonomic ranks than Cyanobacteria. For example, the level of sequence divergence of RpoB from two species of *Termititenax* (Margulisbacteria)[41] is about 40% greater than the distance between *Gloeobacter* spp. and any other cyanobacterium, the latter being the largest distance between oxygenic phototrophs. In the case of Gastranaerophilales (Vampirovibrionia)[5], which are specialised gut bacteria and should therefore not be much older than animals, the level of sequence identity of RpoB was found to be 70% for the two most distant strains in this group, contrasted to 84% for *Gloeobacter* spp. when compared to any other cyanobacterium. As listed in Supplementary Table S2, within-group mean distances suggest that faster rates are widespread and not just unique to RpoB.

We implemented a set of 12 calibrations across bacteria, including two calibrations on Margulisbacteria and two in Vampirovibrionia with the aim of covering both slower and faster evolving lineages. The following results are based on an autocorrelated log normal molecular clock using CAT+Γ, a root with a broad interval ranging from between 4.52 and 3.41 Ga, and as described in Materials and Methods (Figure 5). We found this to perform well and better than other more complex models (e.g. CAT+GTR+Γ with birth-death priors and soft bounds on calibrations). In addition, it provided results comparable to other independent studies that did not combine a full set of MVS sequences and other clades with phototrophs in a single tree (Table 1). Nonetheless, a pipeline of sensitivity experiments tested the dependency of these results on models and prior assumptions: these are shown and described in Supplementary Figure S7 and S8.

The root of the tree (divergence of Thermotogae) was timed at 3.64 Ga (95% CI: 3.42 - 4.11 Ga) and the divergence of Cyanobacteria at 2.74 Ga (95% CI: 2.46 - 3.12 Ga). Thus,

the span of time between the mean age of the root and the MRCA of Cyanobacteria was calculated to be 0.89 Ga. The span of time between Margulisbacteria and Cyanobacteria was found to be 0.67 Ga; and between Vampirovibrionia and Cyanobacteria 0.44 Ga (Figure 4f and 5). The latter is a value that is consistent with previous studies using entirely different rationales, datasets and calibrations[8,9]. However, we note that the lower bound of the confidence interval between Sericytochromatia and Vampirovibrionia overlaps with the upper bound of Cyanobacteria, by over 200 Ma in the latter case.

We also noted an exponential decrease in the rates of evolution of RpoB through the Archean, which stabilised at current levels in the Proterozoic (Figure 4f). The rate at the root node was calculated to be 2.37 ± 0.45 δ Ga^{-1} and the average rate of evolution of RpoB during the Proterozoic was found to be 0.19 ± 0.06 δ Ga^{-1}. The average rate of cyanobacterial RpoB was 0.14 ± 0.04 δ Ga^{-1}; for Margulisbacteria was 0.44 ± 0.17 δ Ga^{-1}, and for Vampirovibrionia 0.19 ± 0.05 δ Ga^{-1}: about 3.1 and 1.3 times the mean cyanobacterial rate respectively. These rates agree reasonably well with the observed distances, further indicating that the calibrations used in these clades performed well (Supplementary Table S2). Nevertheless, we suspect that the values for MSV represent underestimations of the true rates of evolution (slower than they should be), as some of the clades that include symbionts still appear much older than anticipated from their hosts (Supplementary Text S1). In contrast, the more complex CAT+GTR+Γ model implementing a birth-death prior with soft bounds on the calibrations resulted in smoothed out rates, which translated into unrealistically spread out divergence times (Supplementary Figure S8, models **n** to **p**) with Margulisbacteria and Vampirovibrionia evolving at 1.9 and 0.7 times the cyanobacterial rates, respectively.

To investigate the effect of the estimated divergence time of the root on the divergence time of MVS and Cyanobacteria we also varied the root prior from 3.2 to 4.4 Ga (Supplementary Figure S7). We noted that regardless of the time of origin of Bacteria (approximated by the divergence of Thermotogae in our analysis), a substantially faster rate is required during the earliest diversification events, decreasing through the Archaean and stabilizing in the Proterozoic. This matches well the patterns of evolution of ATP synthase and PSII core subunits as shown in the previous section.

We then compared the above RpoB molecular clock with a different clock that included a set of 112 diverse sequences from Archaea, in addition to the sequences from Thermotogae, MSV, and Cyanobacteria, but removing all other bacterial phyla (Supplementary Figure S9). We found that the calculated average rate of evolution of bacterial RpoB during the Proterozoic was slower (0.09 ± 0.03 δ Ga^{-1}) than in the absence of

archaeal sequences (0.19 ± 0.06 δ Ga^{-1}), resulting in overall older mean ages (see Figure 4g and Supplementary Figure S8a and b). However, the rate at the Bacteria/Archaea divergence point was 6.87 ± 1.17 δ Ga^{-1}, similar to the rate for *d*AB and *d*CP, requiring therefore an exponential decrease in the rates similar to that observed for ATP synthase and the PSII core subunits. Remarkably, this rate and exponential decrease is associated with a span of time between the mean estimated ages of the LUCA and the MRCA of Cyanobacteria of 1.21 Ga. The span between Vampirovibrionia and Cyanobacteria was found to be 0.46 Ga with an overlap in the confidence intervals of about 120 Ma (Supplementary Figure S9a).

Figure 5 also highlights that the MRCA of none of the groups containing phototrophs nor their divergence from their closest non-phototrophic relatives appears to be older than one of the oldest and best accepted geochemical evidence for photosynthesis at 3.41 Ga[42] (Table 1). This is consistent with the molecular evolution of photosynthetic RCs, which indicates that the early evolution of RCs proteins, including the divergence of type I and type II RCs, antedated the diversification of the known groups of phototrophs[32] (see also Supplementary Figure S3). It suggests that losses of photosynthesis and/or extinctions of phototrophs have been a common and continuous process through time.

Finally, we performed a similar molecular clock analysis on a subset of concatenated ribosomal proteins published by Hug et al.,[36] that included Archaea, Thermotagae, MSV and Cyanobacteria. Despite the fact that this dataset generated younger ages compared to RpoB (see Table 1 and Supplementary Figure S7), a similar exponential decrease in the rates was observed with a rate at the Bacteria/Archaea divergence point measured at 4.62 ± 0.71 δ Ga^{-1} and an average rate of bacterial ribosomal protein evolution of 0.30 ± 0.07 δ Ga^{-1} (Figure 4h). The estimated mean age for the MRCA of Cyanobacteria was found to be 2.22 Ga (95% CI: 1.86 - 2.64 Ga), which is identical to the estimated dates determined with the dataset of CP43 and CBP. The span of time between the mean estimated ages of the LUCA and the MRCA of Cyanobacteria was 1.45 Ga; and between Vampirovibrionia and Cyanobacteria 0.54 Ga, with an overlap of the confidence interval of over 200 Ma (Supplementary Figure S9b). It is unclear why the RpoB clock generated an older MRCA of Cyanobacteria than the CP43 and ribosomal clocks. We attribute this difference to the sparser set of calibrations and broader constraints used on the latter two relative to the former. It is clear however that a narrower span of time between the LUCA and the diversification of bacteria would imply a faster rate at the root node and a steeper deceleration of the rates.

Structural analysis

A fundamental premise of our investigation is that water oxidation started before the duplication of D1 and D2, and CP43 and CP47. The rationale behind this premise has been laid out before[16,43], and more extensively recently[11]. This rationale raises the question of how the D2/CP47 side of the RC lost its capacity to carry out water-splitting catalysis. To gain further insight on the nature of the structural site around the water-oxidising complex in the ancestral photosystem, we used ancestral sequence reconstruction to predict the most probable ancestral states. We will refer to the ancestral protein to D1 and D2 as D0 (Supplementary Figure S10). We generated 14 predicted D0 sequences using a combination of three ASR methods and amino acid substitution models. On average the 14 D0 sequences had 87.12 ± 0.55% sequence identity indicating that the different algorithms provided largely consistent results. While the regions that include all transmembrane helices are aligned unambiguously, the N-terminal and C-terminal ends were aligned less confidently due to greater sequence variability at both ends. Nonetheless, we found that the predicted D0 sequences retain more identity with D1 than D2 along the entire sequence. The level of sequence identity of D0 compared to the D1 (PsbA1) of *Thermosynechococcus vulcanus* was found to be 69.58 ± 0.55% and 36.32 ± 0.15% compared to D2.

The ligands to the Mn_4CaO_5 in PSII are provided from three different structural domains (Supplementary Figure S11): 1) the D1 ligands D170 and E189, located near the redox Y_Z. These are in the lumenal loop between the 3rd and 4th transmembrane helices. 2) The D1 ligands H332, E333, D342, and A344 located at the C-terminus. 3) The CP43 ligands, E354 and R357, located in the extrinsic loop between the 5th and 6th helices, with the latter residue less than 4 Å from the Ca. Remarkably, there is structural and sequence evidence supporting the loss of ligands in these three different regions of CP47/D2.

In all the D0 sequences, at position D1-170 and 189, located in the unambiguously aligned region, the calculated most likely ancestral states were E170 and E189, respectively. The mutation D170E results in a PSII phenotype with activity similar to that of the wild-type[44]. At position D1-170, a glutamate was predicted with average posterior probabilities (PPs) of 44.2% (FastML), 67.2% (MEGA) and 77.0% (PAML). At position D1-E189, the average PPs for glutamate were 31.1% (FastML), 35.2% (MEGA) and 40.2% (PAML). The distribution of PPs across a selection of D0 sequences and key sites is presented in Supplementary Figure S12 and Table S5. In contrast, D2 has strictly conserved phenylalanine residues at these positions, but the PP of phenylalanine being found at either of these positions was less than about 5% for all predicted D0 sequences. As a comparison, the redox active tyrosine residues Y_Z (D1-Y161) and Y_D (D2-Y160), which are strictly conserved

between D1 and D2 have a predicted average PPs of 68.8% (FastML), 98.8% (MEGA) and 98.6% (PAML). Therefore, the ligands to the catalytic site in the ancestral protein leading to D2 were likely lost by direct substitutions to phenylalanine residues, while retaining the redox active D2-Y160 (Y_D) and H189 pair (Supplementary Figure S11).

Prompted by the finding of a Ca-binding site at the electron donor site of the homodimeric type I RC of Heliobacteria (Firmicutes) with several similarities to the Mn_4CaO_5 cluster of PSII, including a link to the antenna domain and the C-terminus[17], we revisited the sequences and structural overlaps of CP43 and CP47. We found that a previously unnoticed structural rearrangement within the extrinsic loop occurred in one subunit relative to the other (marked EF_3 and EF_4 in Figure 6, Supplementary Figure S13 and S14). CP43 retains the simplest domain, being about 60 residues shorter than CP47. If CP43 retains the ancestral fold, the additional sequence in CP47's swapped domain (EF_4 in Figure 6d) would have contributed to the loss a catalytic cluster as it inserted one phenylalanine residue (CP47-F360) into the electron donor site, less than 4Å from Y_D. An equivalent residue does not exist in CP43.

We then noted that in the swapped region (EF_3 in Figure 6d), sequence identity is retained between CP43 and CP47 (Supplementary Figure S13). We found that CP43-E354 and R357 are equivalent to CP47-E435 and N438. An inspection of the crystal structure of cyanobacterial PSII showed that these two residues specifically bind a Ca of unknown function in (Figure 6c and d). The presence of an equivalent glutamate to CP43-E354 in CP47 is consistent with this being already present before duplication.

Finally, a peculiar but well-known trait conserved across Cyanobacteria and photosynthetic eukaryotes is that the 5' end of the *psbC* gene (CP43) overlaps with the 3' end of the *psbD* gene (D2) usually over 16 bp (Supplementary Table S6). The *psbD* gene containing a well-defined Shine-Dalgarno ribosomal binding site downstream of the *psbC* gene and over the coded D2 C-terminal sequence[45-47]. The evolution of this unique gene overlap has no current explanation in the literature, but its origin could have disrupted the C-terminal ligands of the ancestral protein leading to the modern D2, an event that would have contributed to and favoured heterodimerization.

Discussion

Rates of evolution

The duplication of the subunits of the catalytic head of ATP synthase, and the archaeal/bacterial divergence of RNA polymerase and the ribosome are some of the oldest

evolutionary events known in biology[18-26]. We have shown in this study that the antenna subunits of PSII, as well as D1 and D2[11], show patterns of molecular evolution that parallel those of these very ancient systems. This is true regardless of the exact time of origin of Cyanobacteria or their closest relatives. The exponential decreases in the rates of evolution observed in all of them, even when the span of time between the ancestral duplications, or the LUCA, and the MRCA of Cyanobacteria was well over a billion years, are consistent with our assessment that a large ΔT for the evolution of the core subunits of ATP synthase and PSII cannot be explained entirely by duplication-driven neofunctionalization. Instead, they may be the result of the faster rates of evolution that are expected to have occurred during early evolution due to positive selection at the origin of bioenergetics, accelerated rates of mutations caused by environmental factors[48-50], lack of sophisticated DNA repair mechanisms[51], or other genetic properties attributed to early life[52,53], yet in combination with very large spans of time.

It has been postulated before that the diversification of the known phyla of bacteria occurred very rapidly peaking at about 3.2 Ga[54]. A recent study suggested that the major groups of Bacteria and Archaea diversified rapidly after the LUCA and attributed the long distance between archaeal and bacterial ribosomal proteins to fast rates of ribosome evolution during their divergence[55]. Our analyses suggest that the more explosive the diversification of Bacteria, the greater the chance that photosynthetic water-splitting is an ancestral trait of all life, and the greater the chance that Margulisbacteria, Sericytochromatia, Vampirovibrionia, and perhaps other phyla of Bacteria, originated from losses of early versions of oxygenic photosynthesis. Nevertheless, the present work cannot—and it is not designed to—resolve when exactly the MRCA of Cyanobacteria existed. Instead, we wish to reiterate that our results place emphasis on the predicted large span of time between the emergence of a photosystem that was capable of water oxidation and the MRCA of Cyanobacteria from which all known extant clades emerged. This large span of time necessarily implies losses, of oxygenic photosynthesis before, and leading to, the MRCA of Cyanobacteria.

There is an emerging consensus that oxygenic photosynthesis was occurring already at 3.0 Ga[56-58]. Well-preserved Cyanobacteria-like microbial mats are suggestive, although not entirely conclusive, of oxygenic photosynthesis at 3.2 Ga[59,60]. The possibility of oxygenic photosynthesis existing as early as the Eoarchean has been discussed before, but the evidence is contentious. If the MRCA of Cyanobacteria occurred 3.0 to 3.2 Ga ago, or even before that, it would imply slower rates of evolution within Cyanobacteria than currently anticipated from most molecular clock analyses (Table 1). It would also imply slower rates of evolution

of PSII than reported in this work pushing the ancestral duplications towards even earlier times. If that is case, it becomes even more likely that water oxidation is a trait primordial to life given the large ΔT and the constraints imposed by the rates of evolution[11], but would still necessitate losses of oxygenic photosynthesis across Bacteria or extinction events. Alternatively, it may be that ancestral forms of Bacteria had more in common with Cyanobacteria that it is apparent based on today's biodiversity of prokaryotes[36,37].

Our data is consistent with recent models of the stepwise oxygenation of the planet, which suggests that even if relatively high fluxes of O_2 (4.5×10^{13} mol O_2 equivalent per year) started about 4.0 Ga ago, the inherent properties of global biogeochemical cycling would result in a "great" oxygenation at around 2.5 to 2 Ga, while maintaining low concentrations of O_2 over the Archean, and without involving any particular triggers[61]. In this study, we found no evidence to justify an origin of oxygenic photosynthesis that coincided with the GOE. For this to be the case, exceedingly high rates of amino acid substitutions of the PSII core subunits would be required relative to other RC proteins[11], occurring at a late stage in their evolution, and followed by a precipitous decline in the rate. While we cannot completely reject the possibility that this sudden spike occurred, neither phylogenetic data nor the comparative structural biology of the photosystems supports a scenario in which PSII experienced rates of evolution greatly superior to those of any other RC. In fact, structural constraints indicate that the opposite is likely to be the case.

Structural constraints

We have calculated that (oxygenic) PSII has experienced the slowest rates of evolution between type II RCs, with the core of the anoxygenic type II RC of Proteobacteria and Chloroflexi evolving approximately five times faster than the core of PSII[11]. That considerably faster rate has led to conspicuous structural changes of the anoxygenic RC relative to PSII and type I RCs. This is not only visually apparent (Figure 7), but also PSII retains greater structural symmetry at its core and a greater number of conserved structural traits with type I RCs not found in its anoxygenic type II RC cousin[17,62]. It follows then that because these conserved traits can be traced to the homodimeric stage of the earliest RC, the rates of evolution of PSII should have remained slow relative to that of other RCs since before the core duplications and as it has been the case for well over the past two billion years. However, the large distance between the core subunits of PSII not only already accounts for a period of fast evolution at its origin, but it also requires a large span of time between the core duplications and the MRCA of Cyanobacteria.

One of the earliest events in the evolution of photosynthesis is the structural and functional specialisation that led to type I and type II RCs. It is conventionally considered that the first six transmembrane helices of RC proteins make up the antenna domain, while the photochemical core encompasses the last five helices. In actuality, the antenna domain extends to the 8th helix, both in type I RCs and in PSII; with the latter retaining one antenna chlorophyll in the equivalent 8th helix (marked Z in Figure 7), as well as substantial sequence identity around this chlorophyll's binding site[32,63]. These conserved features indicate that there is not a moment in time during the evolution of PSII in which it was devoid of core antenna proteins. In other words, CP43 and CP47 are likely the ancestral core antenna of type II RCs, now lost in anoxygenic type II RCs. These structural considerations are consistent with the large phylogenetic distance observed between CP43/CP47 and other type I RC proteins (Supplementary Figure S2) and disproves a previous hypothesis by Cardona[33] where the antenna of PSII was claimed to have originated from a refolding of an entire type I RC protein. Nevertheless, it is consistent with the premise that the ancestral type II RC, at the dawn of photosynthesis, was architecturally, and perhaps also functionally[17], rather similar to cyanobacterial PSII.

Sequence reconstruction of the ancestral subunit to D1 and D2 is also consistent with a highly oxidizing homodimeric photosystem, though transient and short-lived[11], that could generate a tyrosyl radical and split water in either side of the RC[43]: not necessarily at the same time or over as many catalytic turnovers. The structural comparisons between CP43/D1 and CP47/D2 suggest a mechanism for heterodimerization and loss of catalysis on one side that accounts for all ligands. These include direct substitution of ligands for phenylalanine residues in the ancestral protein to D2, the domain swap within the extrinsic region of the ancestral protein to CP47, and the gene overlap between CP43 and D2. It would be difficult to reconcile these unique structural and genetic features with a scenario in which PSII evolved water oxidation at a late stage, once the heterodimerization process was well underway or completed: as would be expected if oxygenic photosynthesis originated not long before the MRCA of Cyanobacteria.

Conclusion

Both phylogenetic and structural evidence converge towards a scenario in which photosynthetic water-splitting started at an early stage during the evolution of life and long before the rise of Cyanobacteria as we understand them today. Nonetheless, greater resolution of divergence time estimation in deep-time evolutionary studies could be achieved if

molecular clocks are complemented with experimentally validated rates of protein and genome evolution across taxa of interest. A photosynthetic origin of life is not a new idea[64-66] and abiotic photosynthetic chemistry has been recently proposed to have occurred at Gale Crater on Mars[67], and to occur even on Earth[68]. We think it is plausible that there was never a discrete origin of photosynthesis, but that the process may trace back to abiogenic photochemical reactions, some of which may have resulted in the oxidation of water, and at the interface of nascent membranes, membrane proteins, photoactive tetrapyrroles and other inorganic cofactors: much in the same way that ribosomes originated at the interface of nascent RNA-dominated genetics and protein synthesis[69].

Materials and methods

Sequence alignments and phylogenetic analysis

The first dataset was retrieved on the 31st of October 2017 to initiate this project. It included a total of 1389 type I RC, CP43 and CP47 protein sequences from the NCBI refseq database using BLAST. From Cyanobacteria, 675 PsaA and PsaB and 685 CP43 and CP47 subunits were retrieved. From anoxygenic phototrophs, 24 PscA and 4 PshA were obtained. This dataset did not contain CBP proteins.

A second dataset was retrieved on the 10th of October 2018 from the same database. This dataset focused on CP43 and CP47 subunits and consisted of a total of 1232 sequences. It included 392 CP43, 465 CBP proteins, and 375 CP47 subunits. 40 CP43 and 40 CP47 sequences from a diverse set of photosynthetic eukaryotes were selected manually and included. In addition, a selection of cyanobacterial and plastid CP43, CP47, AtpA (Alpha), AtpB (Beta), and FtsH were manually selected for an analysis of the rates of evolution as illustrated in Supplementary Figure S15 and described below.

A dataset of Alpha and Beta subunits belonging to cyanobacterial F-type ATP synthase were retrieved from the NCBI refseq on the 31st of August 2019. 507 and 529 cyanobacterial Alpha and Beta sequences were obtained respectively. We retrieved Alpha and Beta homologous for Margulisbacteria (19 and 18 sequences respectively), Sericytochromatia (4 and 2), and Vampirovibrionia (66 and 49) using AnnoTree[70] and searching with the KEGG codes K02111 (F-type Alpha) and K02112 (F-type Beta). A total of 111 AtpA (subunit A) and 176 AtpB (subunit B) from archaeal V-type ATP synthase were retrieved using the sequences from *Methanocaldococcus jannaschii* as BLAST queries.

A dataset of 366 bacterial RNA polymerase subunit β (RpoB) were collected from the NCBI refseq database on the same date as above. We focused on Cyanobacteria (65

sequences) and other phyla known to contain phototrophic representatives, as well as for their potential for allocating calibration points as described below. These included sequences from Firmicutes (32), Chloroflexi (32), Proteobacteria (102), Acidobacteria (34), Chlorobi (26), Gemmatimonadetes (3), Aquifecae (12) and Thermotogae (24). Sequences for *Candidatus* Eremiobacterota (2) recently reported to include phototrophic representatives[71], Margulisbacteria (13), Sericytochromatia (2), and Vampirovibrionia (11) were obtained from AnnoTree using KEGG code K03043 as query. In addition, three extra margulisbacterial RpoB sequences were collected: from *Termititenax persephonae* and *Termititenax aidoneus* retrieved from the refseq database, and from margulisbacterium *Candidatus* Ruthmannia eludens obtained from https://www.ebi.ac.uk/ena/data/view/PRJEB30343[72]. A second dataset containing only sequences from Cyanobacteria and MSV, in addition to sequences from Archaea was also used. To retrieve the archaeal sequences the subunit B' (RpoB1) of *Methanocaldococcus jannaschii* was used as a BLAST query. A total 213 sequences across the entire diversity of Archaea were obtained. Two version of these were observed, those with split B (RpoB1 and RpoB2) and those with full-length B subunits: from the latter, 112 full-length B subunits were selected for molecular clock analysis.

All sequences were aligned with Clustal Omega using 5 combined guided trees and Hidden Markov Model Iterations[73]. Given that RpoB sequences are known to contain many clade specific indels[25], we further processed this particular dataset using Gblocks[74] to remove these indels and poorly aligned positions allowing smaller final blocks, gap positions within the final blocks, and less strict flanking positions. This procedure left a total of 903 well-aligned sites for the dataset with only bacterial sequences and 788 for the dataset with archaeal sequences.

Maximum Likelihood phylogenetic analysis was performed with the PhyML online service using the Smart Model Selection mode implementing the Bayesian Information Criterion for parameter selection[75,76]. Tree searching operations were calculated with the Nearest Neighbour Interchange model. Support was computed using the average likelihood ratio test[77]. Trees were visualised with Dendroscope V3.5.9[78].

Distance estimation

Distance estimation was performed as a straightforward and intuitive approach to detect variations in the rate of evolution across taxa with known divergence time of at similar taxonomic ranks. This was done in several ways. Distance trees were plotted using BioNJ[79] as implemented in Seaview V4.7[80] using observed distances and 100 bootstrap replicates.

Within-group mean distances were calculated using three different substitution models as implemented in the package MEGA-X[81]: no. of differences, a Poisson model, and a JTT model. A gamma distribution to model rates among sites was used with a parameter of 1.00 and 500 bootstrap replicates.

To compare changes in the rates of evolution as a function of time of key proteins used in this study, we compared the percentage of sequence identity between orthologues in a total of 20 pairs of sequences from representative photosynthetic eukaryotes and Cyanobacteria with known or approximate divergence times. These are listed in Supplementary Table S3. We compared CP43, CP47, Alpha, Beta and FtsH. Of these 17 pairs, the first 15 were comparisons between land plants. The approximated divergence times were based on those recommended by Clarke, et al.[82] after their extensive review of the plant fossil record. These were taken as the average of the hard minimum and soft maximum fossil ages suggested by the authors. The soft maximum age for the earliest land plants was taken as 515 Ma as discussed extensively in Morris, et al.[83]

Another sequence comparison was made between the unicellular red alga *Cyanidioschyzon merolae* and *Arabidopsis thaliana*. The earliest well-accepted fossil evidence for red algae is that of the multicellular *Bangiomorpha*[84,85] dated to about 1.0 Ga[86]. Recently fossils of multicellular eukaryotic algae have been reported at 1.6 Ga[87-89]. The distance between heterocystous Cyanobacteria and their closest non-heterocystous relatives is represented by a pairwise comparison between *Chroococcidiopsis thermalis* and *Nostoc* sp. PCC 7120. Excellently preserved heterocystous Cyanobacteria have been reported from Tonian deposits older than 0.72 Ga[90]. The largest distance is that between *Gloeobacter* spp. and any other cyanobacterium or photosynthetic eukaryote, representing the distance since the MRCA of Cyanobacteria of uncertain divergence time.

To compare the level of sequence identity between RC proteins, two datasets of 10 random amino acid sequences were generated using the Sequence Manipulation Suit[91]. The datasets contained sequences of 350 and 750 residues. These were independently aligned as described above, resulting in 45 pairwise sequence identity comparisons for each dataset. These random sequence datasets were used as a rough minimum threshold of identity. Alignments of RC proteins were generated using three representative sequences spanning known diversity. Cyanobacterial CP43, CP47, standard D1 and D2 sequences were from *Gloeobacter violaceus*, *Stanieria cyanosphaera*, and *Nostoc* sp. PCC 7120; Heliobacterial PshA from *Heliobacterium modesticaldum*, *Heliobacillus mobilis*, and *Heliorestis convoluta*; PscA from *Chlorobium tepidum*, *Prosthecochloris aestuarii*, *Chlorobium* sp. GBchlB,

Chloracidobacterium thermophilum and *Chloracidobacterium* sp. CP2-5A; Protoebacterial L and M from *Methylobacterium* sp. 88A, *Roseivivax halotolerans*, and *Blastochloris sulfoviridis*; and L and M from *Chloroflexus* sp. Y-400-fl, *Roseiflexus castenholzii*, and *Oscillochloris trichoides*.

Quantification of rates of evolution (rationale)

Molecular clocks are conventionally used to estimate divergence times. In general terms, given: 1) a tree topology, which sets the relationship between taxa; 2) a sequence alignment, which sets the phylogenetic distance between taxa; and 3) some known events (calibrations), which set the rates of evolution, the molecular clock can then estimate divergence times. This means that if the tree topology and divergence times for two sets of protein sequences are the same, any differences in phylogenetic distances between these two should only reflect differences in the rate of evolution. Thus, assuming that CP43/CP47 and Alpha/Beta have mainly been inherited vertically in Cyanobacteria and photosynthetic eukaryotes, any difference in phylogenetic distance between the two is the result of differences in the rates of evolution. For example, the level of sequence identity between CP43 in *Cyanidioschizon* and *Arabidopsis* is 78%, and the level of sequence identity between Alpha in the same species is 69%. Given that these plastid-encoded subunits have mostly been inherited vertically since the MRCA of Archaeplastida, then one can argue that Alpha is evolving somewhat faster than CP43. This is because faster rates of protein evolution should lead to faster rates of change resulting in a faster decrease in the level of sequence identity (increase in phylogenetic distance). Now, because CP43 and CP47 are paralogues, we can then use this approach to estimate the rates of evolution at the moment of duplication under a given number of specific scenarios. For example, assuming that the MRCA of Cyanobacteria occurred at 2.0 Ga, we can then model how the rates of evolution would change at the moment of duplication if this occurred at 2.5, 3.0, 3.5 Ga, or at any other time point of interest. Because these set of proteins have likely achieved mutational saturation at the largest distance, the rate of evolution at the point of duplication will be an underestimation (slower than it should be), given that saturation would hide substitution events. Similarly, keeping the time of duplication constant, we can then model how the rates of evolution would change across the tree if the MRCA of Cyanobacteria is assumed to have occurred at 2.0, 2.5, 3.0 Ga, or at any other time of interest. Therefore, even though it is difficult to determine the absolute time of origin of Cyanobacteria, or of the early duplications of the core subunits of ATP

synthase, it is possible to determine what rates of evolution are required to fulfil any particular scenario.

Quantification of rates of evolution

To measure rates of evolution of CP43/CP47 and Alpha/Beta a total of 19 sequences from photosynthetic eukaryotes and 4 cyanobacterial sequences per subunit were selected. A standardized tree topology was constructed from consensus evolutionary relationships as illustrated in Supplementary Figure S15a: 1) Relationships between land plants was taken from ref.[82] 2) It is well established that the divergence of red algae predates the MRCA of land plants. 3) Ponce-Toledo, et al. [92] recently suggested that *Gloeomargarita* is the closest living cyanobacterial relative to the plastid ancestor and predated the emergence of heterocystous cyanobacteria, but see also[93]. 4) The clade containing *Chroococcidiopsis thermalis* PCC 7203 is one of heterocystous Cyanobacteria's closest non-heterocystous relatives[94,95]. 5) *Gloeobacter* spp. is the earliest branching and well-described genus of Cyanobacteria capable of oxygenic photosynthesis[95-99].

Calibration points were allocated as shown in Supplementary Figure S15a and listed in Table 2. Nodes 1 to 15 were applied following the justifications and best practice listed in ref.[82] and using the 515 Ma older calibration as suggested in ref.[83] Node 16 represents here the MRCA of red and green lineages of photosynthetic eukaryotes. The minimum age was set to 1.0 Ga based on the *Bangiomorpha* fossil as described above[84,85]. The maximum age was set to 1.8 Ga, which is similar the oldest ages reported in recent molecular clock analyses for the MRCA of photosynthetic eukaryotes and it is also similar to the age of the earliest plausible fossilised unicellular eukaryotes[100,101]. That said, compelling fossil evidence for eukaryotes could go as far back as 2.1 Ga[102]. Node 17 marks the divergence of heterocystous Cyanobacteria and it was given a minimum age of 0.72 Ga based on the recently described fossils of filaments bearing clear heterocysts from the Tonian period[90]. A maximum age of 1.65 Ga was given to this node, based on the recent report of heterocystous cyanobacteria from the Gaoyuzhuang Formation[103]. Although, it is unclear if reported akinetes that are older than 1.65 Ga are truly that[10], we considered that the 0.72-1.65 Ma range for the origin of heterocystous Cyanobacteria was a reasonably broad constraint for the purpose of this experiment, but see below.

Node 18 denotes the MRCA of Cyanobacteria. The age for this node is highly debated ranging from before to after the GOE. Node 19 represents the duplication events leading to CP43 and CP47, and to Alpha and Beta. To calculate the rates of evolution under different

scenarios, node 18 and node 19 were varied. Firstly, a molecular clock was run using a scenario that assumed that the MRCA of Cyanobacteria postdated the GOE. To do this, node 18 was set to be between 1.6 and 1.8 Ga, which emulates results reported in recent studies[8,104]. This was compared to a scenario that assumed that the MRCA of Cyanobacteria antedated the GOE, and thus node 18 was set to be between 2.6 and 2.8 Ga, which simulates other evolutionary scenarios[93,105]. In both cases, the duplication event (node 19) was set to be 3.5 Ga old, or changed as stated in the main text, by assigning a gamma prior at the desired time fixed with a narrow standard deviation of 0.05 Ga. In a separate experiment, the age of the duplication was varied while maintaining node 18 restricted to between 1.6 and 1.8 Ga, while node 19, the root, was set with a gamma prior with an average varied from 0.8 to 4.2 and with a narrow standard deviation of 0.05 Ga.

The period of time between the duplication event (node 19), which led to the divergence of CP43 and CP47, and the MRCA of Cyanobacteria (node 18), we define as ΔT. ΔT is calculated as the subtraction of the mean age of node 19 and node 18. For PSII, we used node 18 from the CP43 subunit and for ATP synthase we used node 18 from the Alpha subunit. In consequence, varying the age of the duplication from 0.8 to 4.2 Ga allows changes in the rate of evolution to be simulated with varying ΔT, ranging from 0.2 to 2.5 Ga. That is to say, it allows the rate of evolution at the point of duplication to be calculated if it occurred at any point in time before the MRCA of Cyanobacteria.

Molecular clocks were calculated with Phylobayes 3.3f using a log normal autocorrelated molecular clock model under the CAT+Γ non-parametric model of amino acid substitution and a uniform distribution of equilibrium frequencies. We preferred a CAT model instead of a CAT+GTR as the latter only outperforms the former on sequence alignments of over 1000 sites and it is also much less computationally expensive. Four discrete categories for the gamma distribution were used and four chains were executed in parallel through all experiments carried out in this study. The instant rates of evolution, which are the rates at each internal node in the tree, were retrieved from the output files of Phylobayes. These rates are calculated by the software as described by the developers elsewhere[106,107] and are expressed as amino acid changes per site per unit of time.

As an additional comparison, we calculated rates of evolution for cyanobacterial FtsH subunits, which AAA ATPase domain is structurally similar to the catalytic head of ATP synthase[108]. Cyanobacterial FtsH are involved in membrane protein quality control, with some isoforms targeting specifically PSII subunits. It features a late duplication event leading to cyanobacterial FtsH1 and FtsH2 subunits (using the nomenclature of Shao, et al.[38]). It was

shown previously that the duplication leading to FtsH1 and FtsH2 occurred after the divergence of *Gloeobacter* spp. Genes encoding plastid FtsH subunits are encoded in the nuclear genome of photosynthetic eukaryotes. Because not all the selected species had fully sequenced nuclear genomes, only those with available FtsH sequences at the time of this study were used. The tree topology shown in Supplementary Figure S15b was used as template for the calculation of the rate of evolution and was based on the topology presented by Shao, et al.[38] It was concluded that the Cyanobacteria-inherited closest paralog to FtsH1 and FtsH2 in photosynthetic eukaryotes was also acquired before their initial duplication. Therefore, from all FtsH paralogs in photosynthetic eukaryote genomes, those with greater sequence identity to cyanobacterial FtsH1/2 were used. Because this duplication is specific to Cyanobacteria, a few additional strains were included in this tree following well-established topologies[93,95]. Calibrations were placed as indicated in Supplementary Figure S15b. To test the change in the rate of evolution at the time of duplication in comparison with CP43/CP47, node 19 was set to 1.6-1.8 Ga or 2.6-2.8 Ga and molecular clocks were run as described in the preceding paragraph.

Finally, we conducted a large molecular clock using the combined 897 CP43 and CBP sequences, including 40 eukaryotic CP43 sequences, to test whether using a more complex phylogeny would result in rates of evolution substantially different to those calculated with the method described above. Calibrations were assigned as illustrated in Supplementary Figure S16. Cross-calibrations were used across paralogs constraining the origin of heterocystous Cyanobacteria. In this case, only the minimum constraint of 0.72 Ga was used with no maximum constraint to allow greater flexibility. Additional calibrations were assigned also across paralogs (point 20 in Supplementary Figure S16), this was considered as the node made by *Richelia intracellularis* and its closest sister sequence, as implemented in ref.[105] This strain is a specific endosymbiont of a diatom and its divergence was set to be no older than the earliest discussed age for diatoms[109]. The root equivalent to the MRCA of Cyanobacteria (divergence of *Gloeobacter* in CP43) was not calibrated. The root of the tree was varied: first it was given a maximum age of 4.52 Ga as recently implemented and justified by Betts, et al.[104] as the earliest plausible time in which the planet was inhabitable after the moon forming impact[110], and no minimum age was used. A second tree was executed with no constraint on the root and no root prior. A third root was implemented constrained to be between 2.3 Ga (the GOE) and 3.2 Ga. The latter date represents the age of the cyanobacteria-like well-preserved microbial mats of the Berbeton Greenston Belt in South Africa[111]. Rates were obtained using the autocorrelated CAT+Γ model as described

above. Because these root constraints did not have a strong effect in the overall estimated rates, we carried out an additional control applying an uncorrelated gamma clock model[112] with a root constrained at 4.52 Ga and no minimum age.

Molecular clock of RpoB and concatenated ribosomal proteins

The primary objective of this experiment was not to determine the absolute time of origin of Cyanobacteria, but to understand the spans of time between Cyanobacteria and their relatives. We also wanted to understand what rates of evolution are associated with those spans of time and how these change under different evolutionary scenarios. To do this, we applied a molecular clock to the phylogeny of RpoB sequences described above. We implemented 12 calibrations. The calibrations were assigned on the phylogeny as shown in Supplementary Figure S17 and listed in Table 2. A set of calibrations consisted of the earliest unambiguous evidence for Chroococcales Cyanobacteria of the Belcher group (point 21), the age of which has been recently revisited to 2.01 Ga[113]. This was assigned to the younger node from where Chroococcales strains branch out in the tree, with no maximum restrictions. The appearance of heterocystous Cyanobacteria were restricted from 0.72 Ga and 1.56 Ga as described above. No constraints on the node representing the MRCA of Cyanobacteria were used. However, for rigor, we also tested an alternative single calibration on the node representing the MRCA of Cyanobacteria with a maximum of 2.01 Ga and no minimum, and with no other calibrations in the clade. This considered a scenario in which crown group Cyanobacteria are younger than the Belcher fossils.

In addition to cyanobacterial calibrations, we also applied the often-used biomarker evidence for phototrophic Chlorobi and Chromatiaceae at 1.64 Ga[114], see for example refs.[9,54] These were used as a minimum with no maximum constraints (node 30 and 31 respectively in Supplementary Figure S17).

A set of calibrations were chosen from well-described symbiotic relationships. Two of these are known in the phylum Margulisbacteria. The first one is *Termititenax*, these are specific ectosymbionts of spirochetes that live within oxymonad protists in the gut of diverse termites and cockroaches[41]. The two *Termititenax* sequences used in this analysis clustered together and therefore we calibrated this node to be between 396 Ma, some of the oldest fossil evidence of insects[115] and 140 Ma for *Mastotermes nepropadyom*, a Jurassic fossil termite[116] (node 22). This was done under the assumption that this symbiotic relationship may have started before the divergence of termites and cockroaches, as suggested in ref.[41] The second one is *Candidatus* Ruthmannia eludens, a cell-type specific endosymbiont of

placozoans, early-branching metazoans. This symbiont has been detected in all haplotypes examined, regardless of geographical location or sampling time[72,117], which suggest that this association may be as old as placozoans. Therefore, we used a minimum calibration of 542 Ma as the Cambrian explosion of animal diversity and 660 Ma, the earliest biomarker evidence for desmosponges[118] (node 23), which should antedate the divergence of placozoans. This calibration was assigned to the node separating *Candidatus* Ruthmannia eludens from its closest sister sequence.

Similar to Margulisbacteria, members of the clade Vampirovibrionia have been reported to form close associations with eukaryotes. The clade Gastranaerophilales is thought to be composed mostly of strains that inhabit the animal gut. Thus, we calibrated the node separating two strains isolated from human and koala faeces that clustered together with a maximum age of 201 Ma, representing the Jurassic split of marsupials and placental mammals[119] (node 24a). However, the koala and human sequences were embedded in the Gastranaerophilales clade within other sequences from the human gut. Because of this, we trialled changing this calibration to 55 Ma instead, the oldest primate fossil[120] and assuming that the retrieved sequences from the human gut had a common ancestor younger than the MRCA of primates. Alternatively, we tested moving this calibration to the ancestral nodes of the clade that included all the human gut sequences (node 24b). Gastranaerophilales is closely related to the order Vampirovibrionales, which include *Vampirovibrio chlorellavorus*. This strain is a predator of the eukaryotic green algae *Chlorella*[121], and therefore we trialled a calibration assuming that Gastranaerophilales and Vampirovibrionales radiated after the MRCA of eukaryotes (node 25). We thus assigned a maximum calibration to this node of 1.8 Ga representing the earliest described plausible eukaryote fossils[100] and no minimum age.

Another highly specific obligate symbiosis is that of the betaproteobacterium *Polynucleobacter necessarius* and ciliates of the genus *Euplotes* (Spirotrichea)[122]. *Polynucleobacter* has close free-living phototrophic relatives within the same genus[122]. We set the node separating the phototrophic and non-phototrophic *Polynucleobacter* (node 26) a maximum age of 444 Ma for the oldest fossil evidence of spirotrichs, as implemented in Parfrey et al.[123], and which predates the radiation of the genus *Euplotes*[124].

Another well-known association is that of the soil bacteria *Bradyrhizobium* and legumes. Thus we gave the node separating *Bradyrhizobium* spp. from its closest relative in the RpoB tree, *Xanthobacter autotrophicus*, a maximum age of 86 Ma for Rosids, which contain legumes[83] (node 27).

The Rickettsiales are Alphaproteobacteria that exists in very close association with eukaryotes[125]. An association that may reach to the lineage leading to the origin of mitochondria[126]. Therefore, we assumed that the divergence of Rickettsiales occurred before the MRCA of eukaryotes and gave this node a minimum age of 1.8 Ga[100] (node 28). Finally, the family Anaplasmataceae contains bacteria that exists in close association with insects as endosymbionts (e.g. *Wolbachia*) or as parasite vectors (e.g. *Anaplasma*). Therefore, we set a maximum constraint for the MRCA of *Wolbachia* and *Anaplasma* (node 29), excluding *Neorickettsia*, to be as old as the earliest evidence for insects about 395 Ma ago[115].

To constrain the age of the root, we first set a broad gamma prior with an average of 3.8 Ga and a standard deviation of 0.5 Ga. We found this to perform well and used it as benchmark to compare with a range of evolutionary models and the effects of key calibrations (Supplementary Figure S8). Alternatively, we applied a broad calibration on the root with a maximum of 4.52 Ga as described above and a minimum of 3.41 Ga, which is the earliest well-accepted evidence for photosynthesis[42]. This evidence was hypothesized to be anoxygenic in ref.[42] Therefore, Bacteria should be at least as old as the earliest evidence for photosynthesis under conventional evolutionary scenarios.

To further understand the effect that a distant outgroup would have on the estimated divergence time, and to measure the rate of evolution of RpoB at the divergence of Bacteria and Archaea, we repeated the clock including 112 diverse archaeal sequences, in addition to Thermotogae, MSV, and Cyanobacteria, but removing all other clades as a compromise between robustness and computing time. Calibrations were assigned on MSV and Cyanobacteria as described above. Relaxed molecular clocks were computed using an autocorrelated log normal clock and applying a CAT+Γ model. This was compared against a CAT+Γ and a CAT+GTR+Γ using birth-death priors and soft bounds on the calibrations, in addition to an uncorrelated gamma model (Supplementary Figure S8).

We compared the RpoB molecular clock (CAT+Γ) with that of a clock executed using a dataset of concatenated ribosomal proteins of 2596 aligned sites, obtained from an independent study[36]. In total, 157 sequences were used including 56 cyanobacterial sequences, 14 sequences from Vampirovibrionia, 9 from Margulisbacteria, 9 from Thermotogae and 78 from Archaea. A set of calibrations were used including those two in Cyanobacteria (point 17 and 21), a calibration on the MRCA of Gastranaerophilales with a maximum 201 Ma (point 24b), and a root calibration between 4.52 and 3.41 Ga as described above.

Ancestral sequence reconstruction

Ancestral sequence reconstruction (ASR) of D1 and D2 amino acid sequences was carried out with a dataset collected on the 17th of November 2017. Duplicates and partial sequences were removed, leaving 755 D1 and 248 D2 sequences. CD-HIT[127] was used to remove sequences with greater than 92% sequence identity to create a representative sample. The L and M RC subunits from 5 strains of Proteobacteria were used as outgroup. The final alignment did not include the atypical variant D1 sequence from *Gloeobacter kilaueensis* (NCBI accession AGY58976.1) as this showed an unstable phylogenetic position in this dataset. Maximum likelihood trees used as input for ASR were computed with PhyML using Smart Model Selection[76]. The LG substitution model with observed amino acid frequencies and four gamma rate categories exhibited the best log likelihood (LG+Γ+F) (Supplementary Table S7). We used the top four models for tree reconstruction (LG+Γ+F, LG+Γ+I+F, LG+Γ and LG+ Γ+I) in addition to another tree computed using the WAG substitution model with observed amino acid frequencies and four gamma rate categories (WAG+Γ+F). These trees were used as input trees to calculate maximum likelihood ancestral states at each site for the node corresponding to the homodimeric reaction protein, D0. Three ASR programs were used for the reconstructions: FastML[128], Lazarus[129] (a set of Python scripts which wraps PAML[130]) and MEGA-7[131]. The substitution model used by all three programs corresponded to the substitution model used for the specific input tree in PhyML and the branch lengths of the trees were fixed. In FastML, a maximum likelihood method of indel reconstruction using a probability cut-off of 0.7 was used. In MEGA-7, all sites were used for analysis with no branch swap filter. In Lazarus, the '—gapcorrect' option was used to parsimoniously place indels.

Structural analysis

The following crystal structures were used in this work: the crystal structure of PSII from *Thermosynechococcus vulcanus*, PDB ID: 3wu2[12]; the anoxygenic type II RC of *Thermochromatium tepidum*, PDB ID: 5y5s[132]; the homodimeric type I RC from *Heliobacterium modesticaldum*, 8v5k[133]; PSI from *Thermosynechococcus elongatus*, PDB ID: 1jb0[134]; and the cryo-EM IsiA structure from *Synechocystis* sp. PCC 6803[135]. Structures were visualised using Pymol™ V. 1.8.2.2 (Schrodinger, LLC) and structural overlaps were carried out with the CEAlign plugin[136].

Acknowledgements

This work was done with the support of a Leverhulme Trust grant (RPG-2017-223 to T.C and A.W.R.), a Royal Society University Research Fellowship (P.S.-B.), and the Biotechnology and Biological Sciences Research Council (BB/K002627/1 and BB/L011206/1 to A.W.R.). The authors are grateful to Dr. Travis J. Lawrence for feedback on a version of this manuscript. Computing resources were provided by the HPC facility of Imperial College London. The authors declare no conflict of interests.

References

1 Sánchez-Baracaldo, P. & Cardona, T. On the origin of oxygenic photosynthesis and Cyanobacteria. *New Phytol* **225**, 1440-1446, doi:10.1111/nph.16249 (2020).

2 Cardona, T., Murray, J. W. & Rutherford, A. W. Origin and evolution of water oxidation before the last common ancestor of the Cyanobacteria. *Mol. Biol. Evol.* **32**, 1310-1328, doi:10.1093/molbev/msv024 (2015).

3 Di Rienzi, S. C. *et al.* The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *Elife* **2**, e01102, doi:10.7554/eLife.01102 (2013).

4 Soo, R. M., Hemp, J. & Hugenholtz, P. Evolution of photosynthesis and aerobic respiration in the cyanobacteria. *Free Radical Biol. Med.* **140**, 200-205, doi:10.1016/j.freeradbiomed.2019.03.029 (2019).

5 Soo, R. M., Hemp, J., Parks, D. H., Fischer, W. W. & Hugenholtz, P. On the origins of oxygenic photosynthesis and aerobic respiration in Cyanobacteria. *Science* **355**, 1436-1440, doi:10.1126/science.aal3794 (2017).

6 Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219, doi:10.1038/ncomms13219 (2016).

7 Carnevali, P. B. M. *et al.* Hydrogen-based metabolism as an ancestral trait in lineages sibling to the Cyanobacteria. *Nat. Commun.* **10**, 463, doi:10.1038/s41467-018-08246-y (2019).

8 Shih, P. M., Hemp, J., Ward, L. M., Matzke, N. J. & Fischer, W. W. Crown group Oxyphotobacteria postdate the rise of oxygen. *Geobiology* **15**, 19-29, doi:10.1111/gbi.12200 (2017).

9 Magnabosco, C., Moore, K. R., Wolfe, J. M. & Fournier, G. P. Dating phototrophic microbial lineages with reticulate gene histories. *Geobiology* **16**, 179-189, doi:10.1111/gbi.12273 (2018).

10 Schirrmeister, B. E., Sánchez-Baracaldo, P. & Wacey, D. Cyanobacterial evolution during the Precambrian. *Int. J. Astrobiol.* **15**, 187-204, doi:10.1017/S1473550415000579. (2016).

11 Cardona, T., Sánchez-Baracaldo, P., Rutherford, A. W. & Larkum, A. W. D. Early Archean origin of Photosystem II. *Geobiology* **17**, 127-150, doi:10.1111/gbi.12322 (2019).

12 Umena, Y., Kawakami, K., Shen, J. R. & Kamiya, N. Crystal structure of oxygen-evolving Photosystem II at a resolution of 1.9 Å. *Nature* **473**, 55-60, doi:10.1038/nature09913 (2011).

13 Ferreira, K. N., Iverson, T. M., Maghlaoui, K., Barber, J. & Iwata, S. Architecture of the photosynthetic oxygen-evolving center. *Science* **303**, 1831-1838, doi:10.1126/science.1093087 (2004).

14 Cardona, T., Sedoud, A., Cox, N. & Rutherford, A. W. Charge separation in Photosystem II: A comparative and evolutionary overview. *Biochim. Biophys. Acta* **1817**, 26-43, doi:10.1016/j.bbabio.2011.07.012 (2012).

15 Rutherford, A. W., Mattioli, T. & Nitschke, W. in *Origin and evolution of biological energy conversion* (ed H Baltscheffsky) 177-203 (VCH, 1996).

16 Rutherford, A. W. & Nitschke, W. in *Origin and evolution of biological energy conversion* (ed H Baltscheffsky) 143–175 (VCH, 1996).

17 Cardona, T. & Rutherford, A. W. Evolution of photochemical reaction centres: more twists? *Trends Plant Sci.* **24**, 1008-1021, doi:10.1016/j.tplants.2019.06.016 (2019).

18 Mulkidjanian, A. Y., Makarova, K. S., Galperin, M. Y. & Koonin, E. V. Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nat. Rev. Microbiol.* **5**, 892-899, doi:10.1038/nrmicro1767 (2007).

19 Gogarten, J. P. *et al.* Evolution of the vacuolar H^+-ATPase: Implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* **86**, 6661-6665, doi:10.1073/pnas.86.17.6661 (1989).

20 Lane, N., Allen, J. F. & Martin, W. How did LUCA make a living? Chemiosmosis in the origin of life. *Bioessays* **32**, 271-280, doi:10.1002/bies.200900131 (2010).

21 Ouzounis, C. A., Kunin, V., Darzentas, N. & Goldovsky, L. A minimal estimate for the gene content of the last universal common ancestor: Exobiology from a terrestrial perspective. *Res. Microbiol.* **157**, 57-68, doi:10.1016/j.resmic.2005.06.015 (2006).

22 Muller, V. & Gruber, G. ATP synthases: structure, function and evolution of unique energy converters. *Cell. Mol. Life Sci.* **60**, 474-494, doi:10.1007/s000180300040 (2003).

23 Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three domains of life. *Nat. Rev. Microbiol.* **9**, 85-98, doi:10.1038/nrmicro2507 (2011).

24 Kyrpides, N., Overbeek, R. & Ouzounis, C. Universal protein families and the functional content of the Last Universal Common Ancestor. *J. Mol. Evol.* **49**, 413-423, doi:10.1007/Pl00006564 (1999).

25 Lane, W. J. & Darst, S. A. Molecular evolution of multisubunit RNA polymerases: sequence analysis. *J. Mol. Biol.* **395**, 671-685, doi:10.1016/j.jmb.2009.10.062 (2010).

26 Harris, J. K., Kelley, S. T., Spiegelman, G. B. & Pace, N. R. The genetic core of the universal ancestor. *Genome Res.* **13**, 407-412, doi:10.1101/gr.652803 (2003).

27 Chen, M., Zhang, Y. & Blankenship, R. E. Nomenclature for membrane-bound light-harvesting complexes of cyanobacteria. *Photosynth. Res.* **95**, 147-154, doi:10.1007/s11120-007-9255-0 (2008).

28 Chen, M., Hiller, R. G., Howe, C. J. & Larkum, A. W. D. Unique origin and lateral transfer of prokaryotic chlorophyll-*b* and chlorophyll-*d* light-harvesting systems. *Mol. Biol. Evol.* **22**, 21-28, doi:10.1093/molbev/msh250 (2005).

29 Murray, J. W. Sequence variation at the oxygen-evolving centre of Photosystem II: a new class of 'rogue' cyanobacterial D1 proteins. *Photosynth. Res.* **110**, 177-184, doi:10.1007/s11120-011-9714-5 (2012).

30 Ho, M. Y., Shen, G., Canniffe, D. P., Zhao, C. & Bryant, D. A. Light-dependent chlorophyll *f* synthase is a highly divergent paralog of PsbA of Photosystem II. *Science* **353**, aaf9178, doi:10.1126/science.aaf9178 (2016).

31 Gan, F., Shen, G. & Bryant, D. A. Occurrence of far-red light photoacclimation (FaRLiP) in diverse cyanobacteria. *Life (Basel)* **5**, 4-24, doi:10.3390/life5010004 (2014).

32 Cardona, T. A fresh look at the evolution and diversification of photochemical reaction centers. *Photosynth. Res.* **126**, 111-134, doi:10.1007/s11120-014-0065-x (2015).

33 Cardona, T. Reconstructing the origin of oxygenic photosynthesis: Do assembly and photoactivation recapitulate evolution? *Front. Plant Sci.* **7**, 257, doi:10.3389/fpls.2016.00257 (2016).

34 Dibrova, D. V., Galperin, M. Y. & Mulkidjanian, A. Y. Characterization of the N-ATPase, a distinct, laterally transferred Na^+-translocating form of the bacterial F-type membrane ATPase. *Bioinformatics* **26**, 1473-1476, doi:10.1093/bioinformatics/btq234 (2010).

35 Battistuzzi, F. U., Feijao, A. & Hedges, S. B. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evol. Biol.* **4**, 44, doi:10.1186/1471-2148-4-44 (2004).

36 Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**, 16048, doi:10.1038/nmicrobiol.2016.48 (2016).

37 Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533-1542, doi:10.1038/s41564-017-0012-7 (2017).

38 Shao, S., Cardona, T. & Nixon, P. J. Early emergence of the FtsH proteases involved in photosystem II repair. *Photosynthetica* **56**, 163-177, doi:10.1007/s11099-018-0769-9 (2018).

39 Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics* **11**, 97-108, doi:10.1038/nrg2689 (2010).

40 Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151-1155, doi:10.1126/science.290.5494.1151 (2000).

41 Utami, Y. D. *et al.* Genome analyses of uncultured TG2/ZB3 bacteria in 'Margulisbacteria' specifically attached to ectosymbiotic spirochetes of protists in the termite gut. *ISME J.* **13**, 455-467, doi:10.1038/s41396-018-0297-4 (2019).

42 Tice, M. M. & Lowe, D. R. Photosynthetic microbial mats in the 3,416-Myr-old ocean. *Nature* **431**, 549-552, doi:10.1038/nature02888 (2004).

43 Rutherford, A. W. & Faller, P. Photosystem II: evolutionary perspectives. *Philos. Trans. Royal Soc. B* **358**, 245-253, doi:10.1098/rstb.2002.1186 (2003).

44 Nixon, P. J. & Diner, B. A. Aspartate 170 of the Photosystem II reaction center polypeptide D1 is involved in the assembly of the oxygen-evolving manganese cluster. *Biochemistry* **31**, 942-948, doi:10.1021/Bi00118a041 (1992).

45 Adachi, Y., Kuroda, H., Yukawa, Y. & Sugiura, M. Translation of partially overlapping *psbD-psbC* mRNAs in chloroplasts: the role of 5'-processing and translational coupling. *Nucleic Acids Res.* **40**, 3152-3158, doi:10.1093/nar/gkr1185 (2012).

46 Carpenter, S. D., Charite, J., Eggers, B. & Vermaas, W. F. The *psbC* start codon in *Synechocystis* sp. PCC 6803. *FEBS Lett.* **260**, 135-137, doi:10.1016/0014-5793(90)80085-w (1990).

47 Chisholm, D. & Williams, J. G. Nucleotide sequence of *psbC*, the gene encoding the CP-43 chlorophyll *a*-binding protein of Photosystem II, in the cyanobacterium *Synechocystis* 6803. *Plant Mol. Biol.* **10**, 293-301, doi:10.1007/BF00029879 (1988).

48 Cockell, C. S. Ultraviolet radiation and the photobiology of earth's early oceans. *Orig. Life Evol. Biospheres* **30**, 467-499, doi:10.1023/a:1006765405786 (2000).

49 Lewis, C. A., Crayle, J., Zhou, S. T., Swanstrom, R. & Wolfenden, R. Cytosine deamination and the precipitous decline of spontaneous mutation during Earth's history. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 8194-8199, doi:10.1073/pnas.1607580113 (2016).

50 Levy, M. & Miller, S. L. The stability of the RNA bases: implications for the origin of life. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 7933-7938, doi:10.1073/pnas.95.14.7933 (1998).
51 Eisen, J. A. & Hanawalt, P. C. A phylogenomic study of DNA repair genes, proteins, and processes. *Mutat. Res./DNA Repair* **435**, 171-213, doi:10.1016/s0921-8777(99)00050-6 (1999).
52 Koonin, E. V. & Galperin, M. Y. in *Sequence — Evolution — Function: Computational Approaches in Comparative Genomics* Ch. 6.4, 252-292 (Springer-Science+Business Media, B.V>, 2003).
53 Woese, C. The universal ancestor. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6854-6859, doi:10.1073/pnas.95.12.6854 (1998).
54 David, L. A. & Alm, E. J. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* **469**, 93-96, doi:10.1038/Nature09649 (2011).
55 Zhu, Q. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.* **10**, 5477, doi:10.1038/s41467-019-13443-4 (2019).
56 Catling, D. C. & Zahnle, K. J. The Archean atmosphere. *Science Advances* **6**, eaax1420, doi:10.1126/sciadv.aax1420 (2020).
57 Havig, J. R., Hamilton, T. L., Bachan, A. & Kump, L. R. Sulfur and carbon isotopic evidence for metabolic pathway evolution and a four-stepped Earth system progression across the Archean and Paleoproterozoic. *Earth-Sci. Rev.* **174**, 1-21, doi:10.1016/j.earscirev.2017.06.014 (2017).
58 Lyons, T. W., Reinhard, C. T. & Planavsky, N. J. The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**, 307-315, doi:10.1038/nature13068 (2014).
59 Homann, M. Earliest life on Earth: Evidence from the Barberton Greenstone Belt, South Africa. *Earth-Sci. Rev.* **196**, doi:10.1016/j.earscirev.2019.102888 (2019).
60 Homann, M. *et al.* Microbial life and biogeochemical cycling on land 3,220 million years ago. *Nat. Geosci.* **11**, 665-671, doi:10.1038/s41561-018-0190-9 (2018).
61 Alcott, L. J., Mills, B. J. W. & Poulton, S. W. Stepwise Earth oxygenation is an inherent property of global biogeochemical cycling. *Science* **366**, 1333-1337, doi:10.1126/science.aax6459 (2019).
62 Cardona, T. Thinking twice about the evolution of photosynthesis. *Open Biology* **9**, 180246, doi:10.1098/rsob.180246 (2019).
63 Cardona, T. Photosystem II is a chimera of reaction centers. *J. Mol. Evol.* **84**, 149-151, doi:10.1007/s00239-017-9784-x (2017).
64 Granick, S. Speculations on the origins and evolution of photosynthesis. *Ann. N. Y. Acad. Sci.* **69**, 292-308, doi:10.1111/j.1749-6632.1957.tb49665.x (1957).
65 Mauzerall, D. Light, iron, Sam Granick and the origin of life. *Photosynth. Res.* **33**, 163-170, doi:10.1007/BF00039178 (1992).
66 Mulkidjanian, A. Y., Bychkov, A. Y., Dibrova, D. V., Galperin, M. Y. & Koonin, E. V. Origin of first cells at terrestrial, anoxic geothermal fields. *Proc. Natl. Acad. Sci. U.S.A.* **109**, E821-E830, doi:10.1073/pnas.1117774109 (2012).
67 Franz, H. B. *et al.* Indigenous and exogenous organics and surface-atmosphere cycling inferred from carbon and oxygen isotopes at Gale crater. *Nat Astron*, doi:10.1038/s41550-019-0990-x (2020).
68 Lu, A. H. *et al.* Photoelectric conversion on Earth's surface via widespread Fe- and Mn-mineral coatings. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 9741-9746, doi:10.1073/pnas.1902473116 (2019).
69 Petrov, A. S. *et al.* Evolution of the ribosome at atomic resolution. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 10251-10256, doi:10.1073/pnas.1407205111 (2014).

70 Mendler, K. *et al.* AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Res.* **47**, 4442-4448, doi:10.1093/nar/gkz246 (2019).

71 Ward, L. M., Cardona, T. & Holland-Moritz, H. Evolutionary implications of anoxygenic phototrophy in the bacterial phylum *Candidatus* Eremiobacterota (WPS-2). *Front. Microbiol.* **10**, 1658, doi:10.3389/fmicb.2019.01658 (2019).

72 Gruber-Vodicka, H. R. *et al.* Two intracellular and cell type-specific bacterial symbionts in the placozoan Trichoplax H2. *Nat. Microbiol.* **4**, 1465-1474, doi:10.1038/s41564-019-0475-9 (2019).

73 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539, doi:10.1038/msb.2011.75 (2011).

74 Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540-552, doi:10.1093/oxfordjournals.molbev.a026334 (2000).

75 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307-321, doi:10.1093/sysbio/syq010 (2010).

76 Lefort, V., Longueville, J. E. & Gascuel, O. SMS: Smart model selection in PhyML. *Mol. Biol. Evol.* **34**, 2422-2424, doi:10.1093/molbev/msx149 (2017).

77 Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.* **55**, 539-552, doi:10.1080/10635150600755453 (2006).

78 Huson, D. H. & Scornavacca, C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* **61**, 1061-1067, doi:10.1093/sysbio/sys062 (2012).

79 Gascuel, O. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**, 685-695, doi:10.1093/oxfordjournals.molbev.a025808 (1997).

80 Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221-224, doi:10.1093/molbev/msp259 (2010).

81 Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547-1549, doi:10.1093/molbev/msy096 (2018).

82 Clarke, J. T., Warnock, R. C. M. & Donoghue, P. C. J. Establishing a time-scale for plant evolution. *New Phytol* **192**, 266-301, doi:10.1111/j.1469-8137.2011.03794.x (2011).

83 Morris, J. L. *et al.* The timescale of early land plant evolution. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E2274-E2283, doi:10.1073/pnas.1719588115 (2018).

84 Butterfield, N. J. *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation of eukaryotes. *Paleobiology* **26**, 386-404, doi:10.1666/0094-8373(2000)026<0386:Bpngns>2.0.Co;2 (2000).

85 Knoll, A. H., Worndle, S. & Kah, L. C. Covariance of microfossil assemblages and microbialite textures across an upper mesoproterozoic carbonate platform. *Palaios* **28**, 453-470, doi:10.2110/palo.2013.p13-005r (2013).

86 Gibson, T. M. *et al.* Precise age of *Bangiomorpha pubescens* dates the origin of eukaryotic photosynthesis. *Geology* **46**, 135-138, doi:10.1130/G39829.1 (2017).

87 Bengtson, S., Sallstedt, T., Belivanova, V. & Whitehouse, M. Three-dimensional preservation of cellular and subcellular structures suggests 1.6 billion-year-old crown-group red algae. *Plos Biol* **15**, e2000735, doi:10.1371/journal.pbio.2000735 (2017).

88 Sallstedt, T., Bengtson, S., Broman, C., Crill, P. M. & Canfield, D. E. Evidence of oxygenic phototrophy in ancient phosphatic stromatolites from the Paleoproterozoic Vindhyan and Aravalli Supergroups, India. *Geobiology* **16**, 139-159, doi:10.1111/gbi.12274 (2018).

89 Qu, Y. G., Zhu, S. X., Whitehouse, M., Engdahl, A. & McLoughlin, N. Carbonaceous biosignatures of the earliest putative macroscopic multicellular eukaryotes from 1630 Ma Tuanshanzi Formation, north China. *Precambrian Res.* **304**, 99-109, doi:10.1016/j.precamres.2017.11.004 (2018).

90 Pang, K. *et al.* Nitrogen-fixing heterocystous cyanobacteria in the tonian period. *Curr Biol* **28**, 616-622 e611, doi:10.1016/j.cub.2018.01.008 (2018).

91 Stothard, P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques* **28**, 1102-1104, doi:10.2144/00286ir01 (2000).

92 Ponce-Toledo, R. I. *et al.* An early-branching freshwater cyanobacterium at the origin of plastids. *Curr Biol* **27**, 386-391, doi:10.1016/j.cub.2016.11.056 (2017).

93 Sánchez-Baracaldo, P., Raven, J. A., Pisani, D. & Knoll, A. H. Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E7737-E7745, doi:10.1073/pnas.1620089114 (2017).

94 Fewer, D., Friedl, T. & Budel, B. Chroococcidiopsis and heterocyst-differentiating cyanobacteria are each other's closest living relatives. *Mol. Phylogen. Evol.* **23**, 82-90, doi:10.1006/mpev.2001.1075 (2002).

95 Shih, P. M. *et al.* Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 1053-1058, doi:10.1073/pnas.1217107110 (2013).

96 Schirrmeister, B. E., Antonelli, A. & Bagheri, H. C. The origin of multicellularity in cyanobacteria. *BMC Evol. Biol.* **11**, 45, doi:10.1186/1471-2148-11-45 (2011).

97 Ciccarelli, F. D. *et al.* Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283-1287, doi:10.1126/science.1123061 (2006).

98 Nakamura, Y. *et al.* Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. *DNA Res* **10**, 137-145, doi:10.1093/dnares/10.4.137 (2003).

99 Honda, D., Yokota, A. & Sugiyama, J. Detection of seven major evolutionary lineages in cyanobacteria based on the 16S rRNA gene sequence analysis with new sequences of five marine Synechococcus strains. *J. Mol. Biol.* **48**, 723-739, doi:10.1007/Pl00006517 (1999).

100 Knoll, A. H., Javaux, E. J., Hewitt, D. & Cohen, P. Eukaryotic organisms in Proterozoic oceans. *Philos. Trans. Royal Soc. B* **361**, 1023-1038, doi:10.1098/rstb.2006.1843 (2006).

101 Han, T. M. & Runnegar, B. Megascopic eukaryotic algae from the 2.1-billion-year-old Negaunee iron-formation, Michigan. *Science* **257**, 232-235, doi:10.1126/science.1631544 (1992).

102 El Albani, A. *et al.* Organism motility in an oxygenated shallow-marine environment 2.1 billion years ago. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 3431-3436, doi:10.1073/pnas.1815721116 (2019).

103 Shi, M., Feng, Q. L., Khan, M. Z. & Zhu, S. X. An eukaryote-bearing microbiota from the early mesoproterozoic Gaoyuzhuang Formation, Tianjin, China and its

significance. *Precambrian Res.* **303**, 709-726, doi:10.1016/j.precamres.2017.09.013 (2017).

104 Betts, H. C. *et al.* Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat. Ecol. Evol.* **2**, 1556-1562, doi:10.1038/s41559-018-0644-x (2018).

105 Sánchez-Baracaldo, P. Origin of marine planktonic cyanobacteria. *Sci. Rep.* **5**, 17418, doi:10.1038/srep17418 (2015).

106 Lepage, T., Bryant, D., Philippe, H. & Lartillot, N. A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.* **24**, 2669-2680, doi:10.1093/molbev/msm193 (2007).

107 Kishino, H., Thorne, J. L. & Bruno, W. J. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* **18**, 352-361, doi:10.1093/oxfordjournals.molbev.a003811 (2001).

108 Langklotz, S., Baumann, U. & Narberhaus, F. Structure and function of the bacterial AAA protease FtsH. *Biochim. Biophys. Acta* **1823**, 40-48, doi:10.1016/j.bbamcr.2011.08.015 (2012).

109 Medlin, L. K., Kooistra, W. H. C. F., Gersonde, R., Sims, P. A. & Wellbrock, U. Is the origin of the diatoms related to the end-Permian mass extinction? *Nova Hedwigia* **65**, 1-11 (1997).

110 Barboni, M. *et al.* Early formation of the Moon 4.51 billion years ago. *Science Advances* **3**, e1602365, doi:10.1126/sciadv.1602365 (2017).

111 Homann, M., Heubeck, C., Airo, A. & Tice, M. M. Morphological adaptations of 3.22 Ga-old tufted microbial mats to Archean coastal habitats (Moodies Group, Barberton Greenstone Belt, South Africa). *Precambrian Res.* **266**, 47-64, doi:10.1016/j.precamres.2015.04.018 (2015).

112 Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *Plos Biol* **4**, 699-710, doi:10.1371/journal.pbio.0040088 (2006).

113 Hodgskiss, M. S. W. *et al.* New insights on the Orosirian carbon cycle, early Cyanobacteria, and the assembly of Laurentia from the Paleoproterozoic Belcher Group. *Earth Planet. Sci. Lett.* **520**, 141-152, doi:10.1016/j.epsl.2019.05.023 (2019).

114 Brocks, J. J. *et al.* Biomarker evidence for green and purple sulphur bacteria in a stratified Palaeoproterozoic sea. *Nature* **437**, 866-870, doi:10.1038/Nature04068 (2005).

115 Engel, M. S. & Grimaldi, D. A. New light shed on the oldest insect. *Nature* **427**, 627-630, doi:10.1038/nature02291 (2004).

116 Legendre, F. *et al.* Phylogeny of Dictyoptera: Dating the origin of cockroaches, praying mantises and termites with molecular data and controlled fossil evidence. *PloS one* **10**, e0130127, doi:10.1371/journal.pone.0130127 (2015).

117 Eitel, M., Osigus, H. J., DeSalle, R. & Schierwater, B. Global diversity of the Placozoa. *PloS one* **8**, e57131, doi:10.1371/journal.pone.0057131 (2013).

118 Zumberge, J. A. *et al.* Demosponge steroid biomarker 26-methylstigmastane provides evidence for Neoproterozoic animals. *Nat. Ecol. Evol.* **2**, 1709-1714, doi:10.1038/s41559-018-0676-2 (2018).

119 Grossnickle, D. M., Smith, S. M. & Wilson, G. P. Untangling the multiple ecological radiations of early mammals. *Trends Ecol Evol* **34**, 936-949, doi:10.1016/j.tree.2019.05.008 (2019).

120 Williams, B. A., Kay, R. F. & Kirk, E. C. New perspectives on anthropoid origins. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 4797-4804, doi:10.1073/pnas.0908320107 (2010).

121 Soo, R. M., Woodcroft, B. J., Parks, D. H., Tyson, G. W. & Hugenholtz, P. Back from the dead; the curious tale of the predatory cyanobacterium Vampirovibrio chlorellavorus. *PeerJ* **3**, e968, doi:10.7717/peerj.968 (2015).

122 Hahn, M. W., Schmidt, J., Pitt, A., Taipale, S. J. & Lang, E. Reclassification of four *Polynucleobacter necessarius* strains as representatives of *Polynucleobacter asymbioticus* comb. nov., *Polynucleobacter duraquae* sp. nov., *Polynucleobacter yangtzensis* sp. nov and *Polynucleobacter sinensis* sp.s nov., and emended description of *Polynucleobacter necessarius*. *Int. J. Syst. Evol. Microbiol.* **66**, 2883-2892, doi:10.1099/ijsem.0.001073 (2016).

123 Parfrey, L. W., Lahr, D. J., Knoll, A. H. & Katz, L. A. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 13624-13629, doi:10.1073/pnas.1110633108 (2011).

124 Fernandes, N. M. & Schrago, C. G. A multigene timescale and diversification dynamics of Ciliophora evolution. *Mol. Phylogen. Evol.* **139**, 106521, doi:10.1016/j.ympev.2019.106521 (2019).

125 Yu, X.-J. & Walker, D. H. in *The prokaryotes: A handbook on the biology of bacteria* Vol. 5 (ed M. Dworkin) Ch. 3.1.19, 493-528 (Springer, 2006).

126 Roger, A. J., Munoz-Gomez, S. A. & Kamikawa, R. The Origin and Diversification of Mitochondria. *Curr Biol* **27**, R1177-R1192, doi:10.1016/j.cub.2017.09.015 (2017).

127 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659, doi:10.1093/bioinformatics/btl158 (2006).

128 Ashkenazy, H. *et al.* FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res.* **40**, W580-W584, doi:10.1093/nar/gks498 (2012).

129 Hanson-Smith, V., Kolaczkowski, B. & Thornton, J. W. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol. Biol. Evol.* **27**, 1988-1999, doi:10.1093/molbev/msq081 (2010).

130 Yang, Z. H. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586-1591, doi:10.1093/molbev/msm088 (2007).

131 Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33**, 1870-1874, doi:10.1093/molbev/msw054 (2016).

132 Yu, L. J., Suga, M., Wang-Otomo, Z. Y. & Shen, J. R. Structure of photosynthetic LH1-RC supercomplex at 1.9 Å resolution. *Nature* **556**, 209-213, doi:10.1038/s41586-018-0002-9 (2018).

133 Gisriel, C. *et al.* Structure of a symmetric photosynthetic reaction center-photosystem. *Science* **357**, 1021-1025, doi:10.1126/science.aan5611 (2017).

134 Jordan, P. *et al.* Three-dimensional structure of cyanobacterial Photosystem I at 2.5 Å resolution. *Nature* **411**, 909-917, doi:10.1038/35082000 (2001).

135 Toporik, H., Li, J., Williams, D., Chiu, P. L. & Mazor, Y. The structure of the stress-induced photosystem I-IsiA antenna supercomplex. *Nat Struct Mol Biol* **26**, 443-449, doi:10.1038/s41594-019-0228-8 (2019).

136 Jia, Y. T., Dewey, G., Shindyalov, I. N. & Bourne, P. E. A new scoring function and associated statistical significance for structure alignment by CE. *J Comput Biol* **11**, 787-799, doi:Doi 10.1089/1066527042432260 (2004).

137 Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812-1819, doi:10.1093/molbev/msx116 (2017).

138 Marin, J., Battistuzzi, F. U., Brown, A. C. & Hedges, S. B. The timetree of prokaryotes: New insights into their evolution and speciation. *Mol. Biol. Evol.* **34**, 437-446, doi:10.1093/molbev/msw245 (2017).

139 Shih, P. M., Ward, L. M. & Fischer, W. W. Evolution of the 3-hydroxypropionate bicycle and recent transfer of anoxygenic photosynthesis into the Chloroflexi. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 10749-10754, doi:10.1073/pnas.1710798114 (2017).

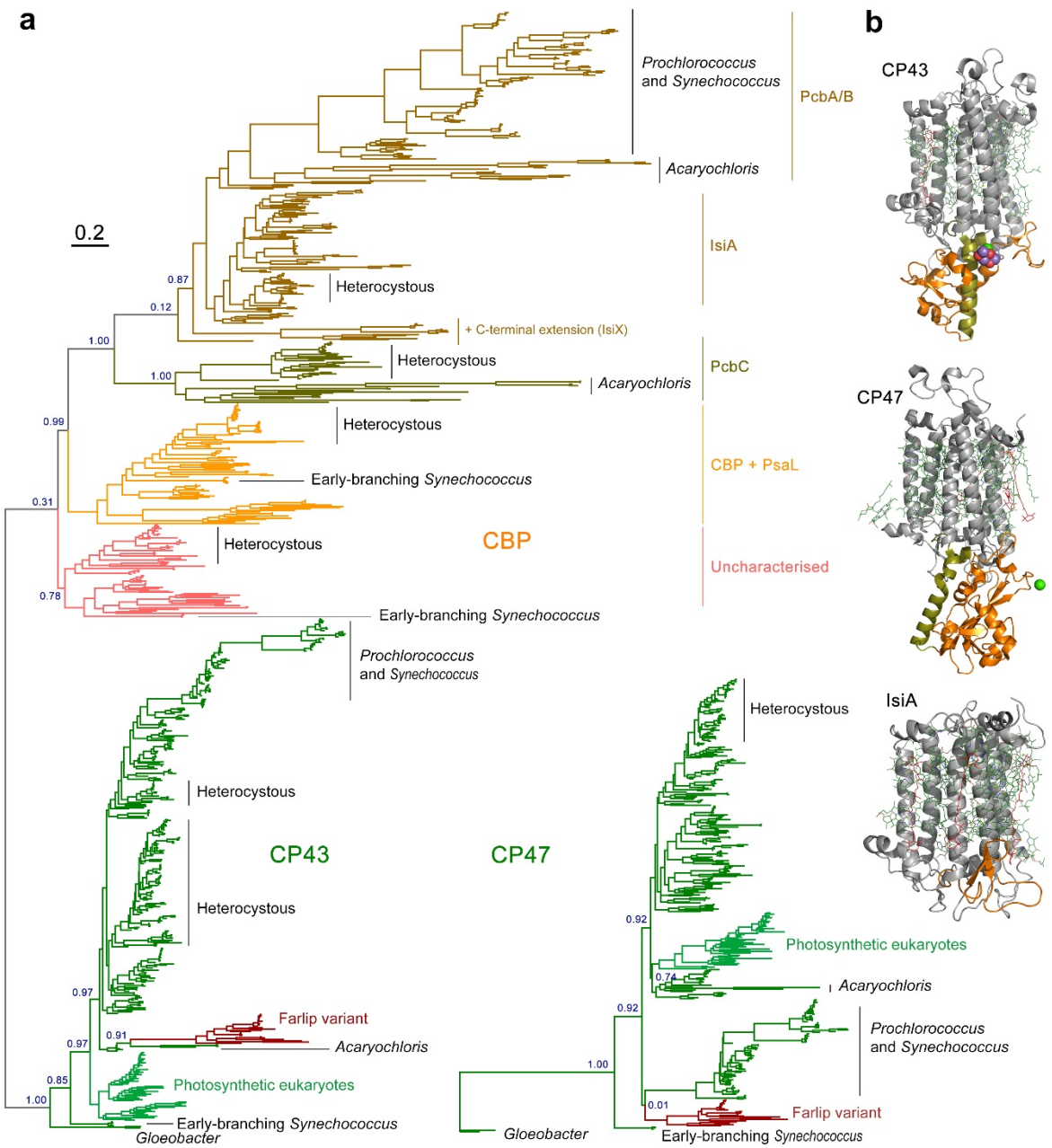

Figure 1. Maximum Likelihood trees of PSII core antenna subunits and derived light harvesting proteins. **a** A tree of CP43 and chlorophyll binding proteins (CBP). The unrooted tree splits into CP43 and CBP, with CP43 displaying a phylogeny roughly consistent with the evolution of Cyanobacteria, although several potential duplications of CP43 are noticeable within heterocystous Cyanobacteria and closer relatives. Farlip variant denotes the isoform used in the far-red light acclimation response. CBP shows a complex phylogeny strongly driven by gene duplication events and fast evolution. The tree of CP47 was rooted at the divergence of *Gloeobacter spp.* Scale bar represents number of substitutions per site. **b** Structural comparison of CP43, CP47, and IsiA. The latter lacks the characteristic extrinsic loop (orange) of CP43 and CP47 that links the antenna with the electron donor side of PSII.

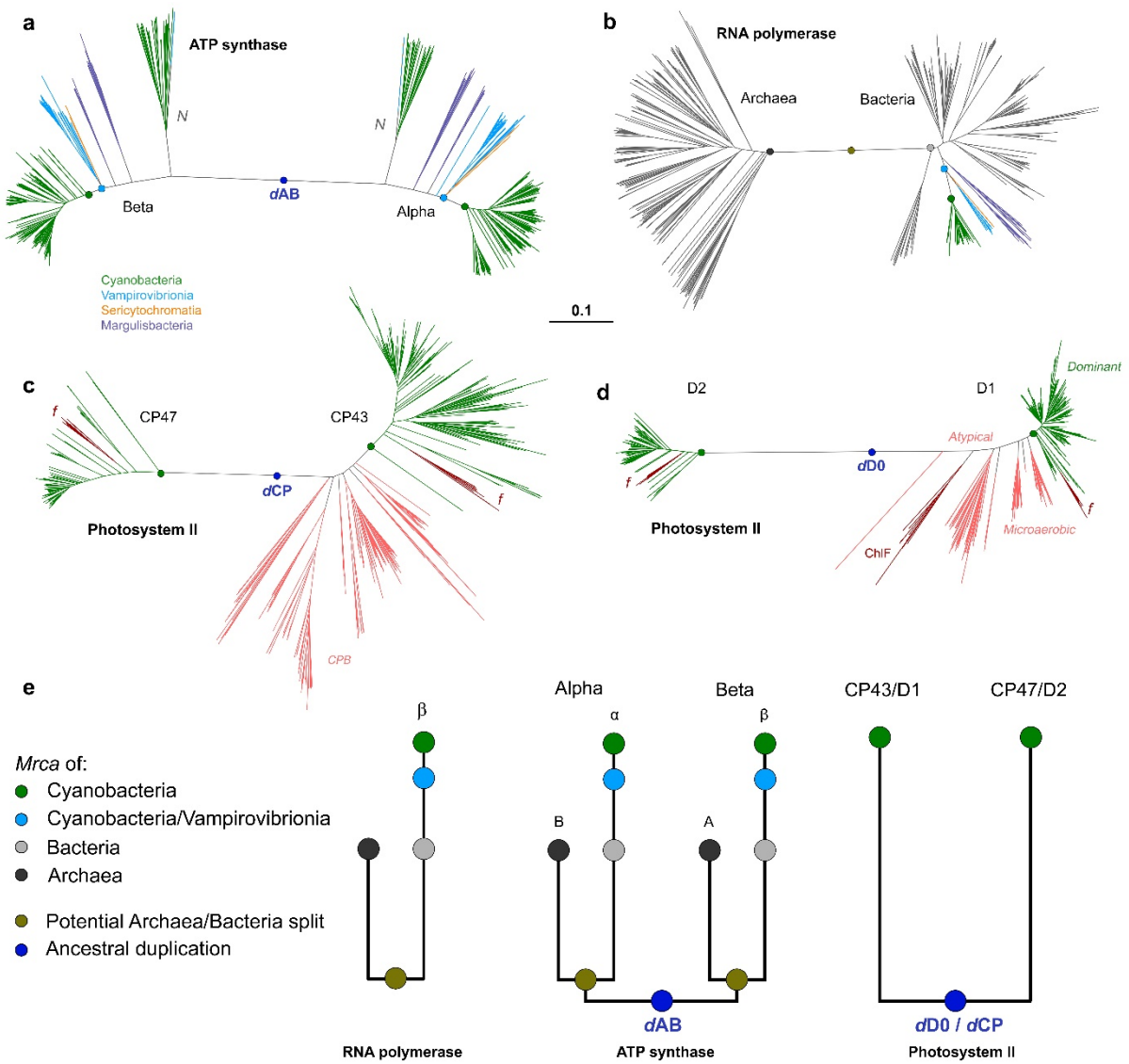

Figure 2. Distance comparison of the core subunits of ATP synthase, RNA polymerase, and PSII. **a** Alpha and Beta subunits of the F-type ATP synthase from Cyanobacteria, Vampirovibrionia, Sericytochromatia and Margulisbacteria. *N* denotes Na^+ translocating N-type ATPase. *d*AB denotes the duplication event leading to Alpha and Beta. The green dot marks the MRCA of Cyanobacteria, and the light-blue dot the MRCA of the clade including Cyanobacteria and Vampirovibrionia. **b** Archaeal and Bacterial RpoB of RNA polymerase. **c** CP43 and CP47 subunits of PSII. *f* denotes farlip variants. *d*CP marks the duplication event leading to CP43 and CP47. Pink branches highlight the CPB subunits. **d** D1 and D2 of PSII showing a pattern that mimics that of CP43 and CP47. Pink branches denote the atypical D1 forms and other variants that are thought to predate the MRCA of Cyanobacteria. The sequences marked with *dominant* represent the standard D1 form of PSII, inherited by the MRCA of Cyanobacteria and found in all oxygenic phototrophs[2,11]. ChlF marks the atypical D1 form involved in the synthesis of chlorophyll *f* during farlip[30]. *d*D0 marks the duplication leading to D1 and D2. **e** Schematic representation of distance and distribution of these enzymes in Cyanobacteria and relatives in relation to the MRCA of Bacteria, of Archaea, and the LUCA.

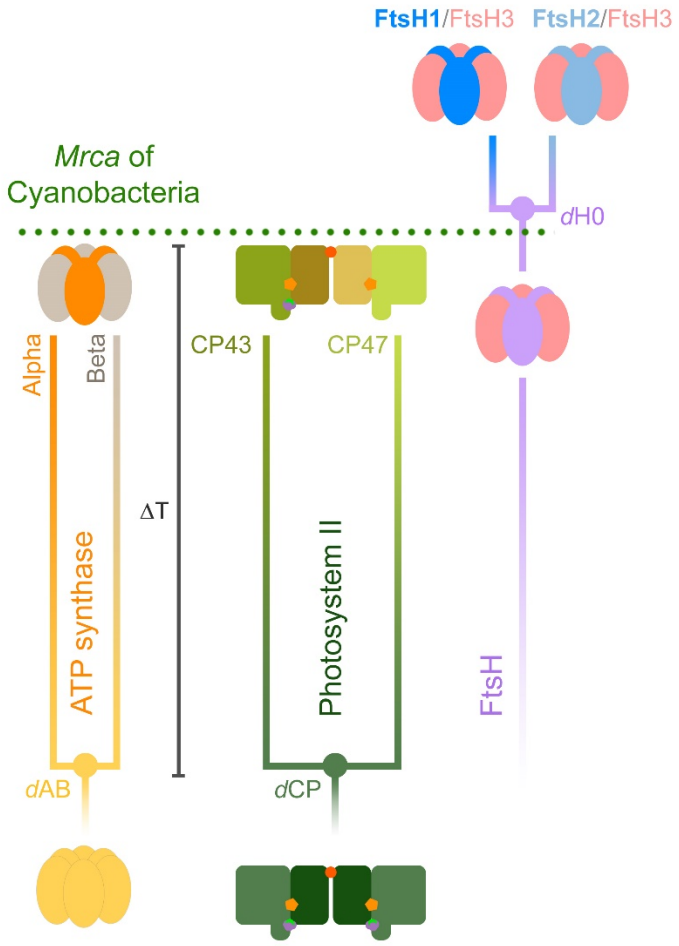

Figure 3. Schematic representation of ancestral duplication events. The MRCA of Cyanobacteria inherited a standard F-type ATP synthase, with a heterohexameric catalytic head (F_1) made of alternating subunits Alpha and Beta; and a PSII with a heterodimeric core and antenna. ΔT denotes the span of time between the ancestral duplication events and the MRCA of Cyanobacteria, which in the case of ATP synthase, is suspected to antedate the divergence of Bacteria and Archaea and their further diversification. FtsH contains an N-terminal membrane-spanning domain attached to a soluble domain consisting of an AAA (ATPase associated with diverse activities) module attached to a C-terminal protease domain. FtsH is universally conserved in Bacteria, has a hexameric structure like that of ATP synthase's catalytic head, and can be found usually as homohexamers, but also as heterohexamers. The MRCA of Cyanobacteria likely inherited three variant FtsH subunit forms, one of which appears to have duplicated after the divergence of the genus *Gloeobacter*, and possibly other early-branching Cyanobacteria[38]. This late duplication led to FtsH1 and FtsH2, which form heterohexamers with FtsH3, following the nomenclature of Shao, et al.[38] FtsH1/FtsH3 is found in the cytoplasmic membrane of Cyanobacteria, while FtsH2/FtsH3 is involved in the degradation of PSII and other thylakoid membrane proteins.

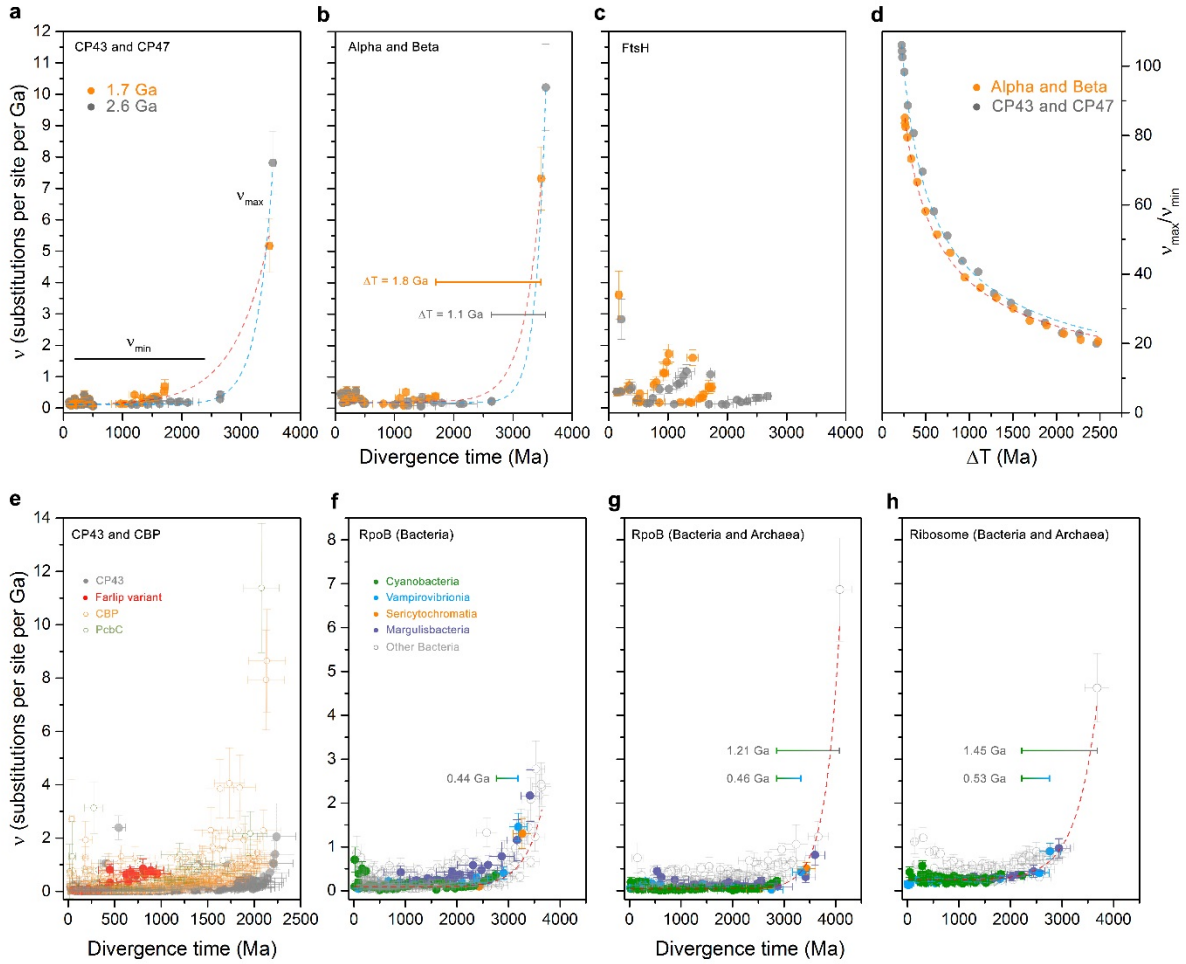

Figure 4. Comparison of the changes in the rates of evolution as a function of time. **a** Rates of CP43 and CP47 modelled using two specific evolutionary scenarios. Orange trace was calculated under the assumption that the MRCA of Cyanobacteria occurred after the GOE, at ~1.7 Ga, while the duplication of CP43 and CP47 occurred at ~3.5 Ga. The fastest rate seen at the point of duplication is denoted as ν_{max}, and stabilizes during the Proterozoic, ν_{min}. Grey dots represent a scenario in which the MRCA of Cyanobacteria is thought to have occurred before the GOE, at ~2.6 Ga instead. **b** Same calculations as **a** but performed on Alpha and Beta sequences of cyanobacterial and plastid ATP synthase. **c** Rates of evolution of cyanobacterial and plastid FtsH subunits assuming a MRCA of Cyanobacteria at 1.7 and 2.6 Ga. **d** Changes in the rates of evolution with varying ΔT for ATP synthase (orange) and PSII subunits (grey). **e** Changes in the rates of evolution as a function of time for a relaxed molecular clock computed with the full dataset of CP43 (full circles) and CBP sequences (open circles). **f** Changes in the rate of evolution of bacterial RpoB showing an exponential decrease in the rate of evolution. The bar denotes the span of time between the MRCA of the clade containing Vampirovibrionia and Cyanobacteria and the MRCA of Cyanobacteria (0.44 Ga). **g** Changes in the rate of evolution of bacterial and archaeal RpoB. The longer bar represents the span of time between the divergence of Archaea and Bacteria and the MRCA of Cyanobacteria (1.21 Ga). **h** Changes in the rate of evolution of a dataset of concatenated ribosomal proteins showing similar patterns of evolution as RpoB. Error bars on each data point are standard errors on the rate and mean divergence time.

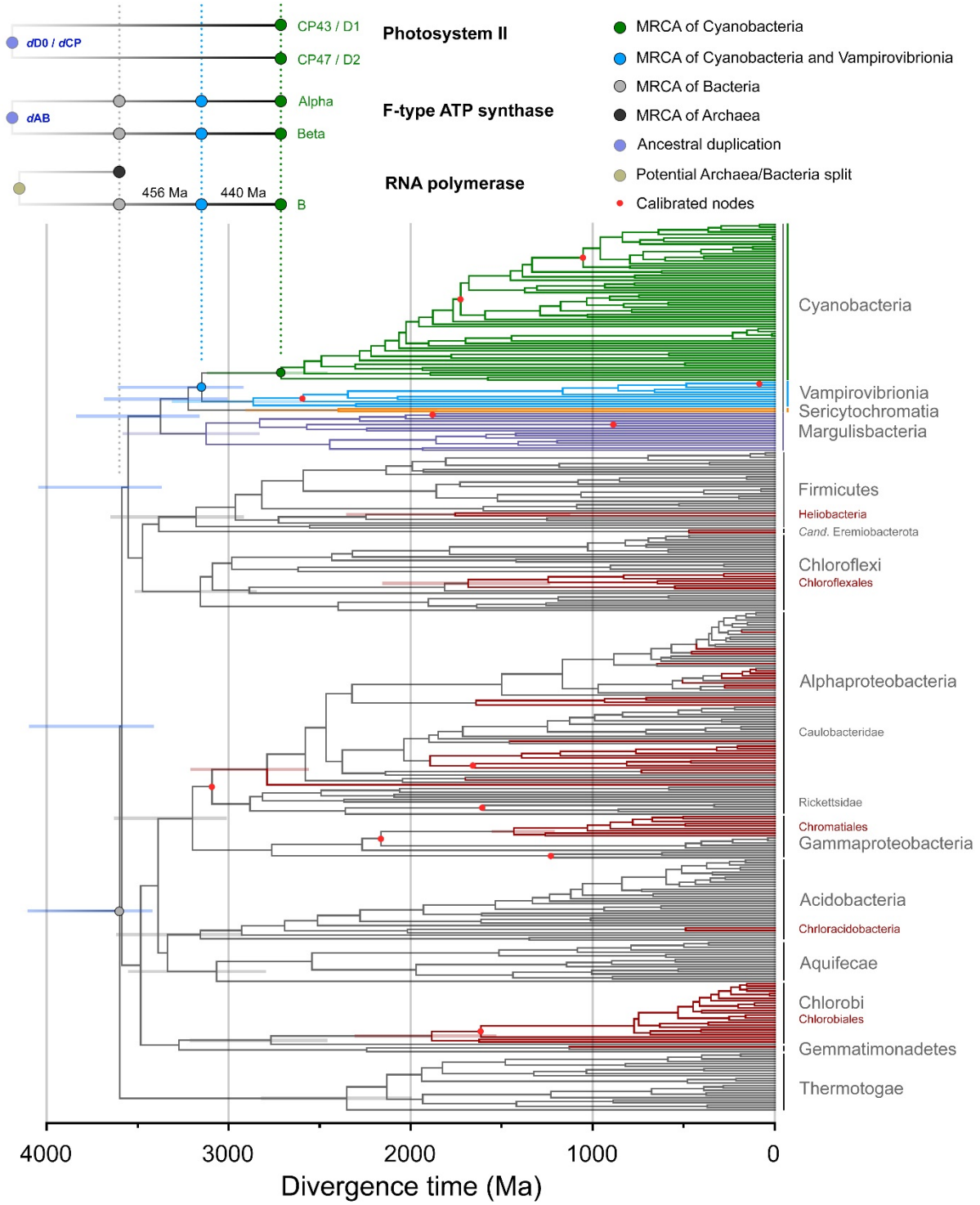

Figure 5. Bayesian relaxed molecular clock of bacterial RpoB. The tree highlights the spans of time between the MRCA of Cyanobacteria (green dot) and their closest relatives. Calibrated nodes are marked with red dots. Anoxygenic phototrophs are highlighted in red branches and non-phototrophic bacteria in grey, with the exception of Margulisbacteria, Sericytochromatia and Vampirovibronia, which are coloured as indicated in the figure. Superimposed at the top are the implied distribution and divergence time for ATP synthase and PSII. Horizontal bars within the tree mark 95% confidence intervals. These are shown in selected nodes of interest for clarity but see Table 1.

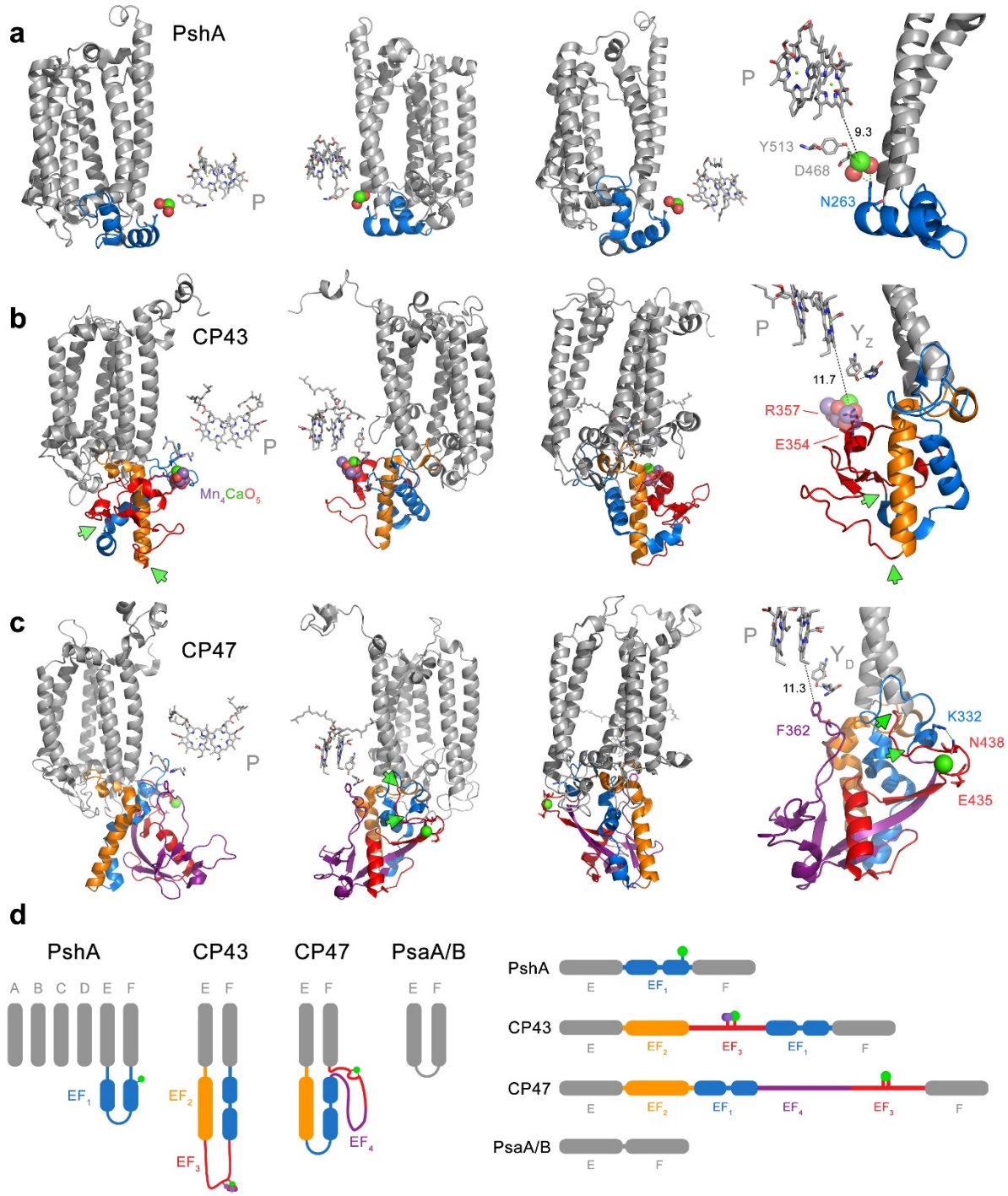

Figure 6. Structural rearrangements of the large extrinsic loop of CP43 and CP47. **a** The antenna domain of heliobacterial PshA is shown in three different rotated perspectives. Only the first six transmembrane helices of the antenna are shown for clarity. A Ca at the electron donor side is bound from an extrinsic loop between the 5th (E) and 6th (F) helices. This extrinsic loop, EF_1 (blue), is made of two small alpha helices. The fourth molecular view furthest to the right shows the link between the electron donor site and EF_1 in closer detail. **b** The CP43 subunit of PSII with the extrinsic loop shown in colours. **c** The CP47 subunit of PSII. Immediately after the 5th helix (E), a long alpha helix protrudes outside the membrane in both CP43 and CP47 and showing structural and sequence identity (orange). We denote this helix EF_2. After EF_2 structural differences are noticed between CP43 and CP47 as

schematised in panel **d**. In CP43, after helix EF_2 a loop is found (shown in red ribbons), which we denote EF_3. This contains the residues that bind the Mn_4CaO_5 cluster and it is followed by a domain that resembles EF_1 in the HbRC at a structural level. In CP47, EF_3 and EF_1 retain sequence identity with the respective regions in CP43. CP47 has additional sequence that is not found in CP43 (EF_4, purple). The green arrows mark the position at which the domain swap occurred in CP43 relative to CP47. We found that the CP43-E354 and R357 are found in the equivalent domain in CP47 as E436 and N438 coordinating a Ca atom. N438 (EF_3) links to EF_1 via K332. It is unclear if the EF_1 region in the HbRC is strictly homologous to that in CP43 and CP47 as very little sequence identity is found between the two: however, a couple of conserved residues between all EF_1 may suggest it emerged from structural domains present in the ancestral RC protein (see Supplementary Figure S14).

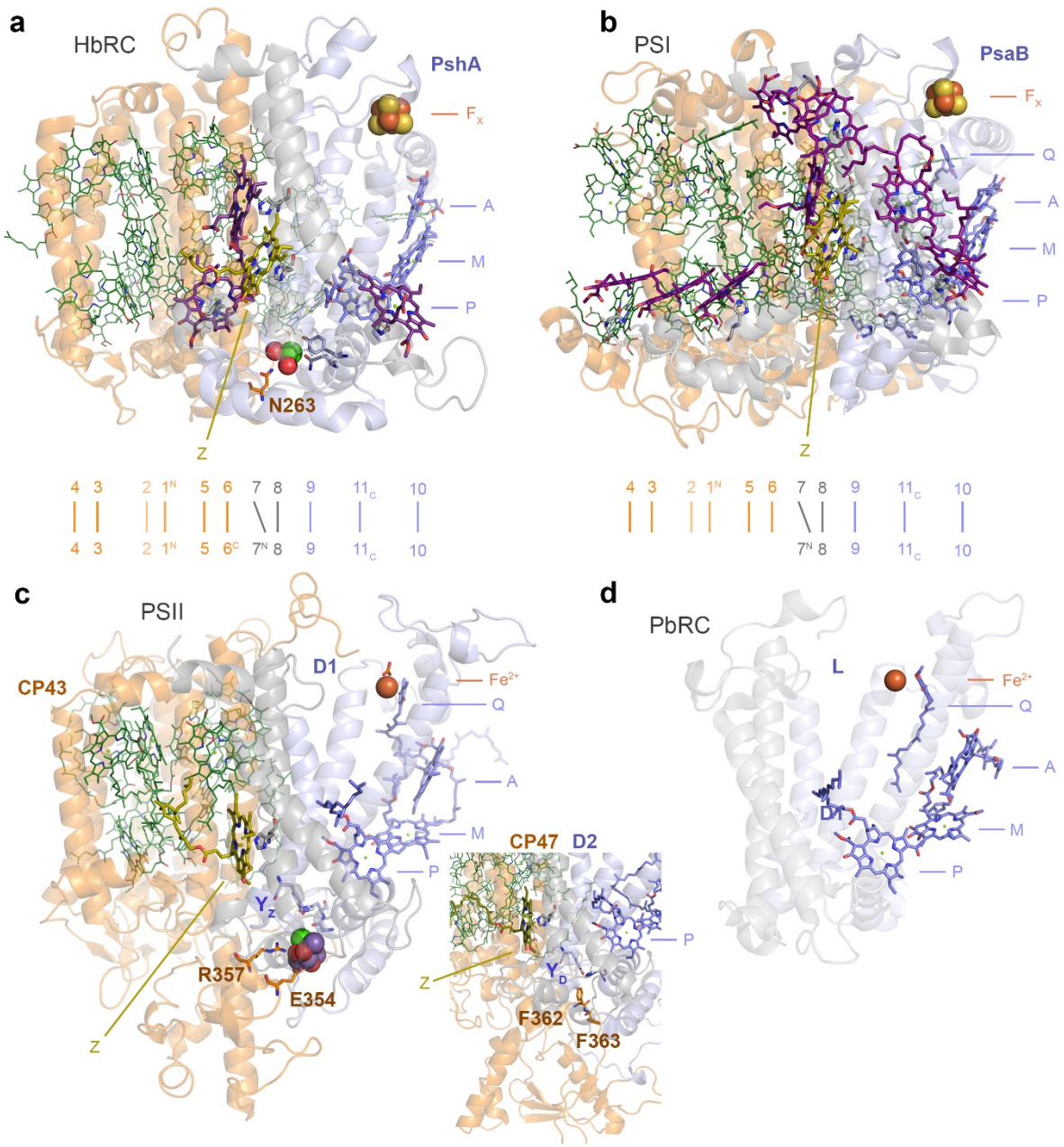

Figure 7. Structural comparisons of the antenna and core domains of photosynthetic RCs. **a** A monomeric unit (PshA) of the homodimeric RC of Heliobacteria (HbRC, Firmicutes). The RC protein is made up of 11 transmembrane helices. The first six N-terminal helices are traditionally considered as the antenna domain (orange ribbons), while the last five helices are the RC core domain (grey and light-blue ribbons). Below the structure, the organization of the 11 helices is laid down linearly for guidance: 1^N denotes the first N-terminal helix and 11_C the last C-terminal helix. P denotes the "special pair" pigment; M, the "monomeric" bacteriochlorophyll electron donor; and A, the primary electron acceptor. F_X is the Fe_4S_4 cluster characteristic of type I RCs. Antenna pigments bound by the antenna domain are shown in green lines. Antenna pigments bound by the 7th and 8th helices are shown as purple sticks, with the exception of the bacteriochlorophyll molecule denoted as Z, in olive green and bound by the 8th helix. The antenna domain connects the electron donor side of the RC via N263 and a Ca-binding site. **b** A monomeric unit of PSI (PsaB, Cyanobacteria) following a similar structural organization and nomenclature as in panel **a**. Unlike the HbRC, PSI binds

quinones (Q) as intermediary electron acceptors between A and F_X. **c** A monomeric unit of PSII (CP43/D1 and CP47/D2 in the inset). It has a structural organization similar to that of type I RCs. However, the monomeric unit is split into two proteins after the 6th transmembrane helix. The Mn_4CaO_5 cluster is coordinated by D1, but is directly connected to the antenna domain via E354 and R357 in manner that resembles the HbRC. **d** A monomeric unit of an anoxygenic type II reaction (PbRC, Proteobacteria). Unlike type I RCs, type II RCs lack F_X. Instead, a non-heme Fe^{2+} is found linking the RC proteins. The PbRC lacks antenna domain and the antenna pigment Z bound at the equivalent 8th helix.

Table 1. Divergence time estimates

Event	RpoB (Ga) (95% CI)	Concatenated ribosomal proteins	TimeTree compilation[a]	Shih et al.[b]	Magnabosco et al.[c]
Divergence of Thermotogae	3.64 (3.43 - 4.11)	3.01 (2.64 - 3. 51)	3.81 (3.51 - 4.16)		
Divergence of Margulisbacteria	3.42 (3.17 - 3.86)	2.93 (2.58 - 3.44)			
Divergence of Sericytochromatia	3.26 (3.01 - 3.68)				
Divergence of Vampirovibrionia/Stem Cyanobacteria	3.19 (2.92 - 3.61)	2.76 (2.40 - 3.23)	3.19 (2.26 - 3.55)	2.54 (2.09 - 3.06)	2.77 (2.39 -2.93)
MRCA of Cyanobacteria	2.75 (2.46 - 3.13)	2.22 (1.86 - 2.64)	2.24 (1.81 - 2.82)	2.02 (1.72 - 2.37)	2.24 (1.91 -2.41)
Divergence of Heliobacteria	2.28 (1.74 - 2.78)		2.23 (1.99 - 2.48)		
MRCA of Heliobacteria	1.78 (1.13 - 2.36)				
Divergence of phototrophic Chloroflexi	1.83 (1.41 - 2.25)		1076	1.098 (0.80 - 1.41)	2.86 (2.49 - 3.00)
MRCA of phototrophic Chloroflexi	1.71 (1.24 - 2.16)			0.86 (0.61 - 1.14)	1.98 (1.68 - 2.43)
Divergence of phototrophic Chlorobi	2.80 (2.45 - 3.20)				2.56 (2.26 - 2.85)
MRCA of phototrophic Chlorobi	1.91 (1.53 - 2.31)		525		1.71 (1.64 -1.95)
Earliest phototrophic Proteobacteria	2.82 (2.55 - 3.21)		2.48 (2.04 - 2.62)		
Divergence of *Chloracidobacterium*	2.04 (1.53-2.51)				

[a]Data from TimeTree compiling estimated divergence times from independent studies as described in ref.[137] For the divergence of Thermotogae the estimated age was taken from eight different studies (n=8). For the divergence of Melaibacteria (or divergence from Cyanobacteria's closest relatives), n=10. For stem Heliobacteria, n=2. For stem phototrophic Chloroflexi and Chlorobi, the values were reported from a single study, Marin, et al.[138] For earliest potential phototrophic Protepbacteria, n=3. This latter was taken as the node made by the clades including *Niveispirillum lacus* and *Rhodobacter sphaeroides* to match our RpoB tree. The independent studies used by TimeTree to generate the estimated ages are listed in Supplementary Table S8.

[b]Data taken from ref. [8] Model T68 for the cyanobacterial dates. That is, no GOE calibration with a calibration on *Bangiomorpha*. The dates on Chloroflexi were taken from ref. [139]

[c]Data taken from ref.[9] Model A. This used a 1.2 Ga calibration on heterocystous Cyanobacteria.

Table 2. Calibration points used in this study

Node label	Relationship	Max age (Ma)	Min age (Ma)
1	*Arabidopsis* v *Populus*	127	82
2	*Arabidopsis* v *Sorghum*	248	124
3	*Chloranthus* v *Liriodendron*	248	98
4	*Arabidopsis* v *Illicium*	248	124
5	*Arabidopsis* v *Nymphaea*	248	124
6	*Arabidopsis* v *Amborella*	248	124
7	*Welwitschia* v *Pinus*	309	121
8	*Welwitschia* v *Cryptomeria*	309	147
9	*Welwitschia* v *Ginkgo*	366	107
10	*Welwitschia* v *Cycas*	366	306
11	*Arabidopsis* v *Cycas*	366	306
12	*Arabidopsis* v *Psilotun*	454	388
13	*Arabidopsis* v *Anthoceros*	515	420
14	*Arabidopsis* v *Physcomitrella*	515	420
15	*Arabidopsis* v *Marchantia*	515	449
16	*Arabidopsis* v *Cyanidioschyzon*	—	1000
17	*Nostoc* v closest non-heterocystous strain	1560 (or none)	720
18	*Arabidopsis* v *Gloeobacter*	Variable (or none)	Variable (or none)
19	CP43 v CP47 or AtpA v AtpB duplications	Variable	Variable
20	*Richelia intracelullaris* v closest neighbouring sequence	250	100
21	Chroococcales cyanobacteria	—	2016
22	*Termititenax*	396	140
23	*Candidatus* Ruthmannia eludens v GCA 002716725.1	660	542
24a	GB GCA 001899365.1 Koala v GB GCA 001917115.1 Homo	201 (or 55)	—
24b	GB GCA 001917115.1 *Homo* v GB GCA 000980455.1 Homo	201	—
25	GB GCA 001917115.1 *Homo* v *Vampirovibrio chlorellavorus*	1800	—
26	*Polynucleobacter necessarius*	444	—
27	*Bradyrhizobium*	86	—
28	Rickettsias	—	1800
29	*Wolbachia* and *Anaplasma*	395	—
30	Phototrophic Chlorobi	—	1640
31	Chromatiaceae	—	1640