

1 **Time-resolved comparative molecular evolution of oxygenic photosynthesis**

2

3 Thomas Oliver<sup>1</sup>, Patricia Sánchez-Baracaldo<sup>2</sup>, Anthony W. Larkum<sup>3</sup>, A. William  
4 Rutherford<sup>1</sup>, Tanai Cardona<sup>1\*</sup>

5

6 <sup>1</sup>Department of Life Sciences, Imperial College London, London, UK

7 <sup>2</sup>School of Geographical Sciences, University of Bristol, Bristol, UK

8 <sup>3</sup>University of Technology Sydney, Ultimo NSW, Australia

9

10 \*Correspondence to: [t.cardona@imperial.ac.uk](mailto:t.cardona@imperial.ac.uk)

11

12

13 **Abstract**

14 Oxygenic photosynthesis starts with the oxidation of water to O<sub>2</sub>, a light-driven reaction  
15 catalysed by photosystem II. Cyanobacteria are the only prokaryotes capable of water  
16 oxidation and therefore, it is assumed that relative to the origin of life and bioenergetics, the  
17 origin of oxygenic photosynthesis is a late innovation. However, when exactly water  
18 oxidation originated remains an unanswered question. Here we use relaxed molecular clocks  
19 to compare one of the two ancestral core duplications that are unique to water-oxidizing  
20 photosystem II, that leading to CP43 and CP47, with some of the oldest well-described events  
21 in the history of life. Namely, the duplication leading to the Alpha and Beta subunits of the  
22 catalytic head of ATP synthase, and the divergence of archaeal and bacterial RNA  
23 polymerases and ribosomes. We also compare it with more recent events such as the  
24 duplication of cyanobacteria-specific FtsH metalloprotease subunits, of CP43 variants used in  
25 a variety of photoacclimation responses, and the speciation events leading to  
26 Margulisbacteria, Sericytochromatia, Vampirovibrionia, and other clades containing  
27 anoxygenic phototrophs. We demonstrate that the ancestral core duplication of photosystem  
28 II exhibits patterns in the rates of protein evolution through geological time that are nearly  
29 identical to those of the ATP synthase, RNA polymerase, or the ribosome. Furthermore, we  
30 use ancestral sequence reconstruction in combination with comparative structural biology of  
31 photosystem subunits, to provide additional evidence supporting the premise that water  
32 oxidation had originated before the ancestral core duplications. Our work suggests that

33 photosynthetic water oxidation originated closer to the origin of life and bioenergetics than  
34 can be documented based on species trees alone.

35

## 36 **1. Introduction**

### 37 *1.1. Evolution of Cyanobacteria*

38 The origin of oxygenic photosynthesis is considered a turning point in the history of life,  
39 marking the transition from the ancient world of anaerobes into a productive aerobic world  
40 that permitted the emergence of complex life [1]. Oxygenic photosynthesis starts with  
41 photosystem II (PSII), the water-oxidising and O<sub>2</sub>-evolving enzyme of Cyanobacteria and  
42 photosynthetic eukaryotes. Today PSII is a highly conserved, multicomponent, membrane  
43 protein complex, which was inherited by the *most recent common ancestor* (MRCA) of  
44 Cyanobacteria in a form that is structurally and functionally quite similar to that found in all  
45 extant species [2]. Thus, the origin of oxygenic photosynthesis antedates the MRCA of  
46 Cyanobacteria by an undetermined amount of time.

47 Cyanobacteria's closest living relatives are the clades known as Vampirovibrionia  
48 (formerly Melainabacteria) [3, 4], followed by Sericytochromatia [5] and Margulisbacteria  
49 [6]. Currently, no photosynthetic strains have been described in these groups of uncultured  
50 bacteria and this has led to the hypothesis that oxygenic photosynthesis arose during the time  
51 spanning the divergence of Vampirovibrionia and the MRCA of Cyanobacteria, starting from  
52 an entirely non-photosynthetic ancestral state [5, 7]. Molecular clock studies have suggested  
53 that the span of time between the divergence of Cyanobacteria and Vampirovibrionia could  
54 be of several hundred million years [8, 9]. However, it is still unclear from molecular clock  
55 analyses and the microfossil record when exactly the MRCA of Cyanobacteria occurred [10].  
56 For example, two recent independent analysis placed the same cyanobacterial ancestor  
57 around a mean age younger than 1.5 Ga [11] and older than 3.5 Ga [12].

58

### 59 *1.2. Evolution of photosystem II*

60 The heart of PSII is made up of a heterodimeric reaction centre (RC) *core* coupled to a core  
61 *antenna*. The two subunits of the RC core of PSII are known as D1 and D2, and these are  
62 associated respectively with the antenna subunits known as CP43 and CP47. D1 and CP43  
63 make up one monomeric half of the RC, and D2 and CP47, the other half. Water oxidation is  
64 catalysed by a Mn<sub>4</sub>CaO<sub>5</sub> cluster coordinated by ligands from both D1 and CP43 [13, 14]. The  
65 cluster is functionally coupled to a redox active tyrosine-histidine pair (Yz-H190) also  
66 located in D1, which relays electrons from Mn to the oxidised chlorophyll pigments of the

67 RC during charge separation [15]. In a cycle of four consecutive light-driven charge  
68 separation events, O<sub>2</sub> is released in the decomposition of two water molecules.

69 Photosystems evolved first as homodimers [16, 17]: therefore, the core and the  
70 antenna of PSII originated from ancestral gene duplication events that antedated the MRCA  
71 of Cyanobacteria. In this way, CP43/D1 retain sequence and structural identity with  
72 CP47/D2. The conserved structural and functional traits between CP43/D1 and CP47/D2  
73 suggest that the ancestral PSII homodimer—prior to the duplication events—was not only  
74 structurally similar to heterodimeric PSII, but also that it split water and had evolved  
75 protective mechanisms against the formation of reactive oxygen species [17-19].

76 In a previous study, we attempted to measure the span of time between the duplication  
77 that led to D1 and D2 (*dD0*) and a point that approximated the MRCA of Cyanobacteria: a  
78 period of time that we called  $\Delta T$  [18]. We observed that the magnitude of  $\Delta T$  can be very  
79 large, well over a billion years. Such large  $\Delta T$  suggested that the origin of a water-oxidising  
80 photosystem antedated Cyanobacteria themselves and potentially other groups of Bacteria  
81 depending on how rapidly the domain diversified. This implies that an ancestral non-  
82 photosynthetic state at the divergence of Margulisbacteria, Sericytochromatia, and  
83 Vampirovibrionia (MSV) cannot be taken for granted despite their specialized heterotrophic  
84 lifestyles. However, our study neither provided an absolute age for the MRCA of  
85 Cyanobacteria nor the duplication event itself, as we simulated a comprehensive range of  
86 scenarios. Instead, we showed that even when  $\Delta T$  is over a billion years, the rate of protein  
87 evolution at the duplication point (*dD0*) needed to be over 40 times greater than any rate ever  
88 observed for D1 and D2, decreasing exponentially during the Archean and stabilising at  
89 current rates during the Proterozoic. Thus, the shorter the  $\Delta T$ , the faster the rate at *dD0*, with  
90 the rate increasing following a power law function and reaching unrealistic values even when  
91  $\Delta T$  is still in the order of several hundred million years [18]. It was still unclear if such  
92 patterns of evolution were unique to the core subunits of PSII or whether other systems have  
93 experienced similar evolutionary trajectories.

94 Here, to help in understanding the evolution of oxygenic photosynthesis,  
95 Cyanobacteria and MSV as a function of time, we compared the duplication leading to the  
96 RC antenna subunits, CP43 and CP47, to several well-defined ancient and more recent  
97 events: including, but not limited to, the duplication of the core catalytic subunits of ATP  
98 synthase, a very ancient event generally accepted to have occurred before the last universal  
99 common ancestor (LUCA) [20-24]; the evolution of RNA polymerase catalytic subunit  $\beta$   
100 (RpoB) and ribosomal proteins, which are universally conserved and widely accepted to have

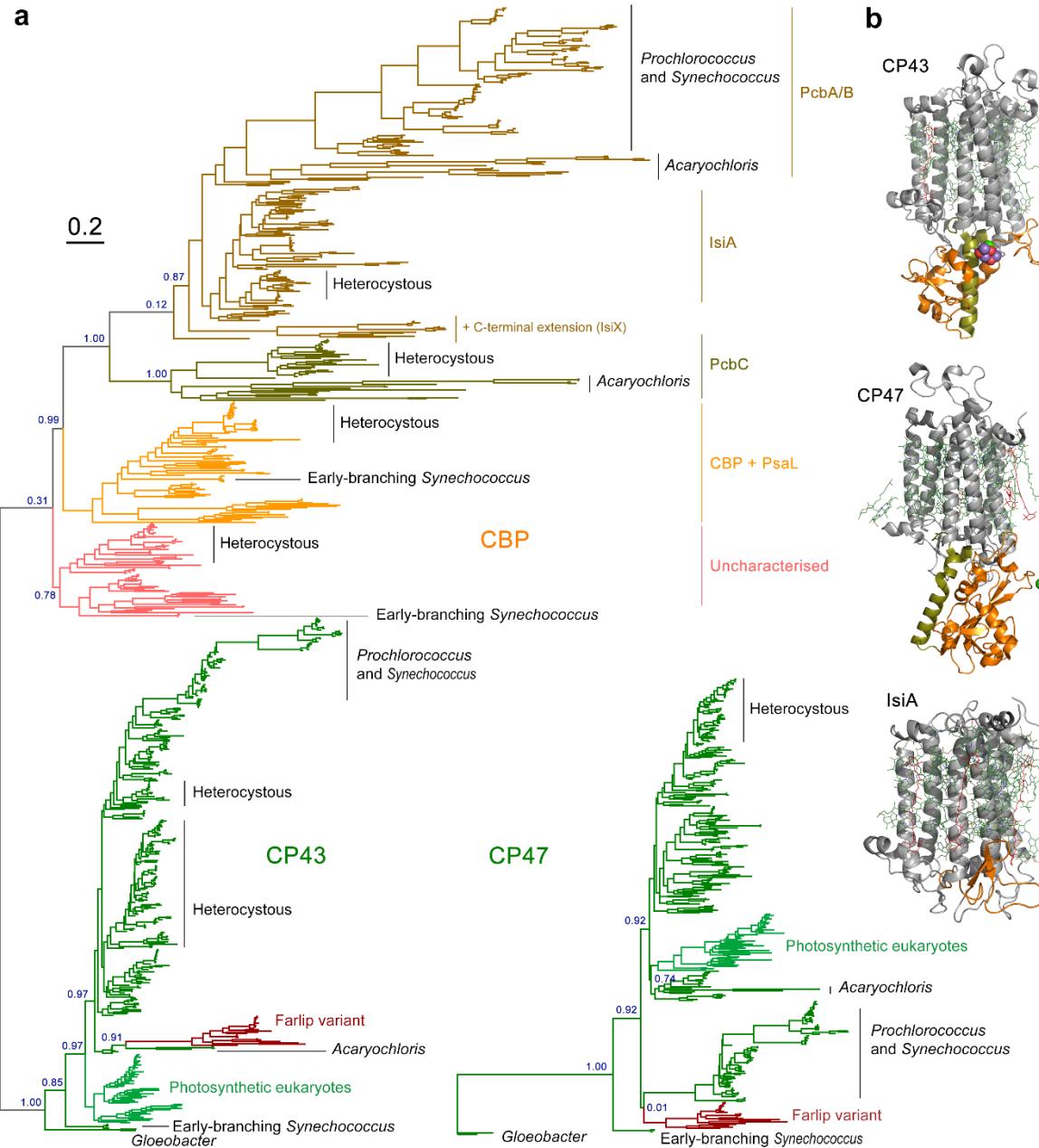
101 originated before the LUCA [25-28]. We further constrain our analysis using *in silico*  
102 ancestral sequence reconstruction of PSII and through strict structural and functional  
103 rationales. We show that the core subunits of PSII show molecular signatures that are usually  
104 associated with some of the oldest transitions in the molecular evolution of life. We also  
105 show that all events leading from the primordial homodimeric water-splitting photosystem to  
106 Cyanobacteria's heterodimeric PSII can be reconstructed from a comparison of available  
107 structural and genetic data.

108

## 109 **2. Results**

### 110 *2.1. Phylogenetic overview*

111 The phylogenies of CP43 and CP47 show that there is a much greater diversity of CP43 and  
112 CP43-derived subunits than CP47 (Figure 1). This difference is the result of a greater number  
113 of gene duplication in CP43 than CP47 and mirrors the evolution of D1 and D2 [2], in which  
114 D1 has undergone more duplication events than D2. CP43 can be divided into two major  
115 groups: those that are assembled into PSII and can bind the Mn<sub>4</sub>CaO<sub>5</sub> cluster, and those  
116 which have evolved to be used only as light harvesting complexes [29, 30], usually known as  
117 chlorophyll binding proteins (CBP). The CBP are characterized by the loss of the extrinsic  
118 loop between the 5<sup>th</sup> and 6<sup>th</sup> transmembrane helices, where the ligands to the cluster are  
119 located (Figure 1b). This large extrinsic loop is found in both CP43 and CP47 and interacts  
120 directly with the electron donor side of PSII, within D1 and D2 respectively. The unrooted  
121 tree of CP43 is consistent with CBP having a single origin likely occurring before the MRCA  
122 of Cyanobacteria (Supplementary Figure S1) but have undergone an extensive duplication-  
123 driven diversification process. It also mirrors the evolution of D1 in that duplications appear  
124 to have occurred before the MRCA of Cyanobacteria [2]. The earliest of these D1  
125 duplications also led to variants that have lost the capacity to bind the Mn<sub>4</sub>CaO<sub>5</sub> cluster [2,  
126 31], but are likely used in other supporting functions. Most notably, during chlorophyll *f*  
127 synthesis [32] in the far-red light acclimation response (farlip) [33].



128

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

138

139 CP43 and CP47 are also distantly related to the antenna domain of cyanobacterial PSI and  
140 that of the type I RCs of phototrophic Chlorobi, Acidobacteria and Heliobacteria  
141 (Supplementary Figure S2). We found that the level of sequence identity between any two  
142 type I RC proteins is always greater than between a type I RC protein and CP43/CP47  
143 (Supplementary Table S1). Therefore, the distance between CP43/CP47 and other type I  
144 antenna domains is the largest distance in the molecular evolution of RC proteins after that  
145 between type I and type II RC core domains. The phylogenetic relationships between type I  
146 and type II RCs have been reviewed in detail before [19, 34, 35]. These are briefly  
147 summarized and schematized in Supplementary Figure S3.

148 The phylogeny of Alpha and Beta subunits of the F-type ATP synthase showed that  
149 all Cyanobacteria have a F-type ATP synthase, and a fewer number of strains have an  
150 additional  $\text{Na}^+$ -translocating ATPase (N-ATPase) of the bacterial F-type, as had been  
151 reported before [36]. We found that MSV have a standard F-type ATP synthase  
152 (Supplementary Figure S4), but some N-ATPase Alpha and Beta sequences were also found  
153 in Vampirovibrionia and Sericytochromatia datasets, but not in Margulisbacteria. In this  
154 study we focused on the standard F-type ATP synthase of Cyanobacteria for further analysis.

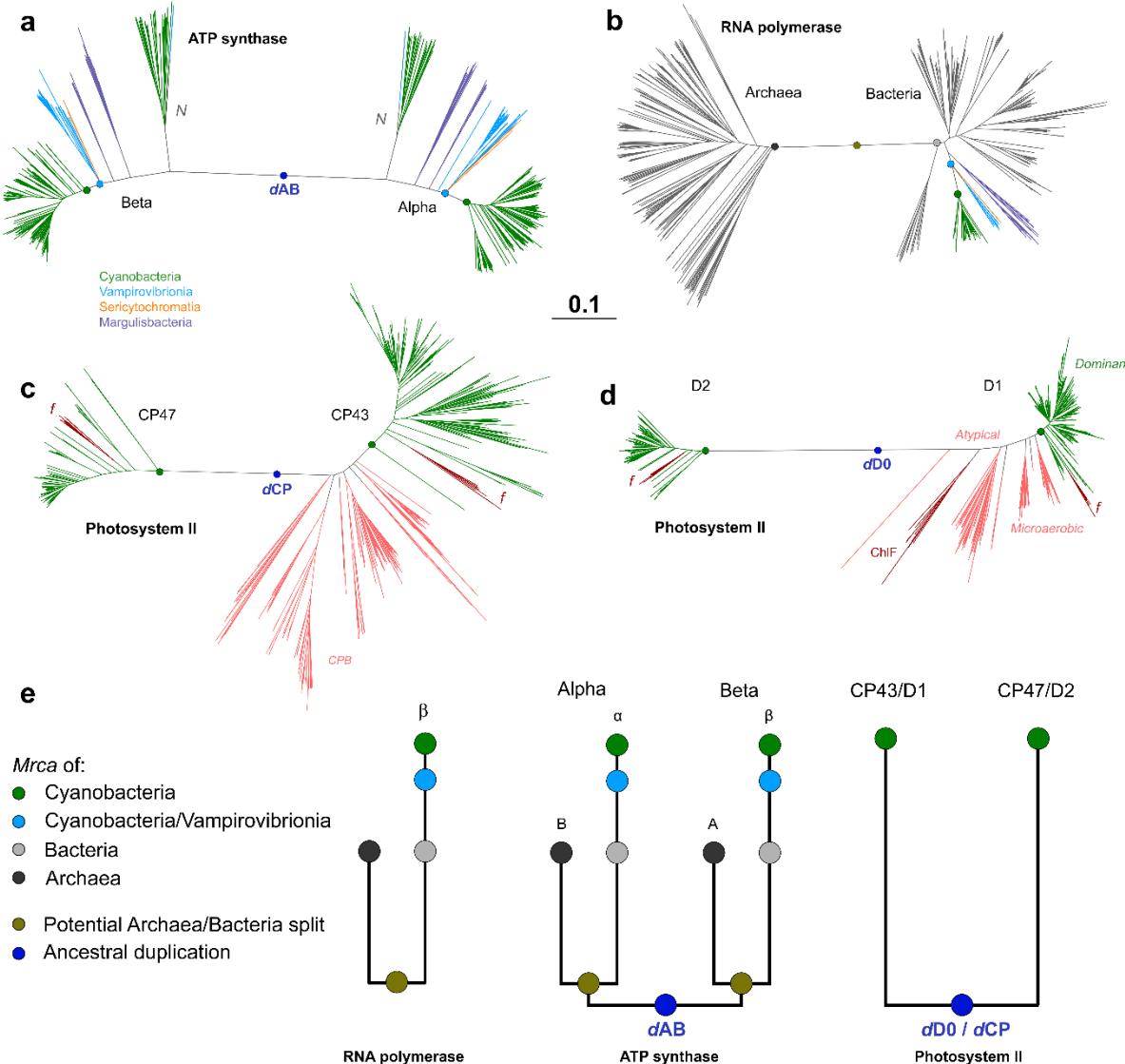
155 The phylogeny of bacterial RNA polymerase subunit  $\beta$  (RpoB), with the intention of  
156 molecular clock analysis, focused on Cyanobacteria and MSV, as well as phyla with known  
157 phototrophic representatives and included Thermotogae and Aquificae as potential roots  
158 (Supplementary Figure S5). The tree was largely consistent with previous observed  
159 relationships between the selected groups [37], within Cyanobacteria and MSV, and within  
160 other phototrophs and their non-phototrophic relatives [38, 39]. The only exception was  
161 Aquificae, which branched as a sister clade to Acidobacteria, a feature that had been reported  
162 before for RpoB [27], and likely represents an ancient horizontal gene transfer event.

163

## 164 2.2. Distances and rates

165 To gather temporal information, we compared the phylogenetic distances between CP43 and  
166 CP47, Alpha and Beta subunits of the F-type ATP synthase, and archaeal and bacterial RpoB  
167 (visualized in Figure 2, but see also Figures S1b, S2 and S3b). We found that the distances  
168 between Alpha and Beta, and the divergence of archaeal and bacterial RpoB, are very large  
169 relative to the distance between the divergence of Vampirovibrionia and Cyanobacteria. In  
170 the case of RpoB, the distance between Vampirovibrionia and Cyanobacteria is about a fifth  
171 of the distance between Archaea and Bacteria. However, the distance between CP43 and  
172 CP47 (but also between D1 and D2 [18]) is of similar magnitude to that between Alpha and

173 Beta, and to that between archaeal and bacterial RpoB, but substantially surpasses the  
 174 distance between MVS and Cyanobacteria (Figure 2). These observations suggest that  
 175 ancestral proteins to CP43/CP47 and D1/D2 existed before the divergences of MVS.  
 176



177  
 178  
 179 **Figure 2.** Distance comparison of the core subunits of ATP synthase, RNA polymerase, and PSII. **a** Alpha and  
 180 Beta subunits of the F-type ATP synthase from Cyanobacteria, Vampirovibrionia, Sericytochromatia and  
 181 Margulisbacteria. N denotes  $\text{Na}^+$  translocating N-type ATPase. dAB denotes the duplication event leading to  
 182 Alpha and Beta. The green dot marks the MRCA of Cyanobacteria, and the light-blue dot the MRCA of the  
 183 clade including Cyanobacteria and Vampirovibrionia. **b** Archaeal and Bacterial RpoB of RNA polymerase. **c**  
 184 CP43 and CP47 subunits of PSII. f denotes farlip variants. dCP marks the duplication event leading to CP43 and  
 185 CP47. Pink branches highlight the CPB subunits. **d** D1 and D2 of PSII showing a pattern that mimics that of  
 186 CP43 and CP47. Pink branches denote the atypical D1 forms and other variants that are thought to predate the  
 187 MRCA of Cyanobacteria. The sequences marked with *dominant* represent the standard D1 form of PSII,

188 inherited by the MRCA of Cyanobacteria and found in all oxygenic phototrophs [2, 18]. ChlF marks the atypical  
189 D1 form involved in the synthesis of chlorophyll *f* during farliph [32]. *dD0* marks the duplication leading to D1  
190 and D2. **e** Schematic representation of distance and distribution of these enzymes in Cyanobacteria and relatives  
191 in relation to the MRCA of Bacteria, of Archaea, and the LUCA.

192

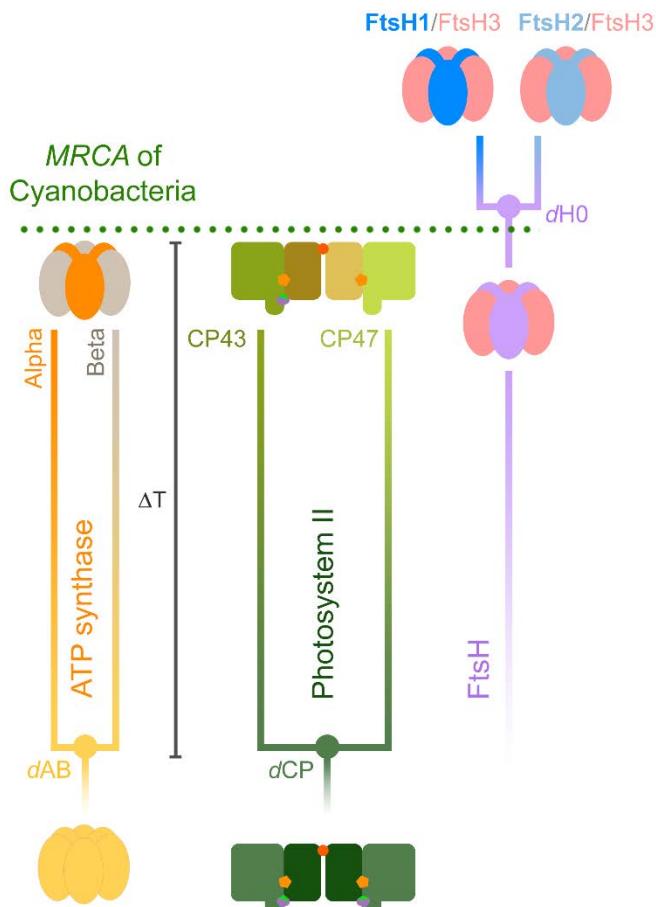
193 We compared the within-group mean distances for Alpha, Beta, RpoB, and a concatenated  
194 dataset of ribosomal proteins compiled in a previous independent study [38] (see  
195 Supplementary Table S2). We found consistently, that Vampirovibrionia and  
196 Margulisbacteria have larger within-group mean distances compared to Cyanobacteria, which  
197 suggests greater rates of evolution in the non-photosynthetic clades. These were consistently  
198 larger for Margulisbacteria relative to the other two groups. For example, RpoB in  
199 Vampirovibrionia and Margulisbacteria showed 1.6x and 4.0x larger corrected mean  
200 distances than Cyanobacteria, respectively (Supplementary Table S2). At the level of the  
201 concatenated ribosomal proteins dataset, Margulisbacteria showed an almost 2-fold larger  
202 within-group mean distance than Cyanobacteria.

203 We then compared the rates of evolution of CP43 and CP47 with those of Alpha and  
204 Beta, using a Bayesian relaxed molecular clock approach with identical calibrations,  
205 molecular clock parameters, and a simplified, but highly constrained sequence dataset (see  
206 Materials and Methods for an expanded rationale). The goal of these experiment is not to use  
207 the clock to estimate divergence times, but to measure and compare the rates of protein  
208 evolution that are required to simulate any chosen span of time between the ancestral  
209 duplications and the MRCA of Cyanobacteria. We used an autocorrelated log normal model  
210 of rate variation with a non-parametric CAT+Γ model of amino acid substitutions to extract  
211 rates of evolution. We will refer to the span of time between the duplication points leading to  
212 Alpha and Beta (**dAB**), or to CP43 and CP47 (**dCP**), and the MRCA of Cyanobacteria as  $\Delta T$   
213 (schematized in Figure 3).

214 In Figure 4a to d we examine the changes in the rate of evolution under specific  
215 evolutionary scenarios. In the case of ATP synthase, we first assumed that the MRCA of  
216 Cyanobacteria occurred after the GOE to simulate scenarios similar to those presented in [8]  
217 or [11], at about 1.7 Ga; and that *dAB* occurred at 3.5 Ga ( **$\Delta T = 1.8 \text{ Ga}$** ). Under these  
218 conditions the average rate of evolution of Alpha and Beta is calculated to be  $0.28 \pm 0.06$   
219 substitutions per site per Ga ( $\delta \text{ Ga}^{-1}$ ). We will refer to the average rate through the  
220 Proterozoic as  $v_{\min}$ . In this scenario, the rate of evolution at the point of duplication, which we  
221 denote  $v_{\max}$ , is  $7.32 \pm 1.00 \delta \text{ Ga}^{-1}$ , making  $v_{\max}/v_{\min} 26$ . In other words, when the span of time

222 between the ancestral pre-LUCA duplication and the MRCA of Cyanobacteria is 1.8 Ga, the  
223 rate of evolution at the point of duplication is about 26 times greater than any rate observed  
224 through the diversification of Cyanobacteria or photosynthetic eukaryotes. To place these  
225 rates in the larger context of protein evolution, we encourage the reader to refer to  
226 Supplementary Text S1.

227



228

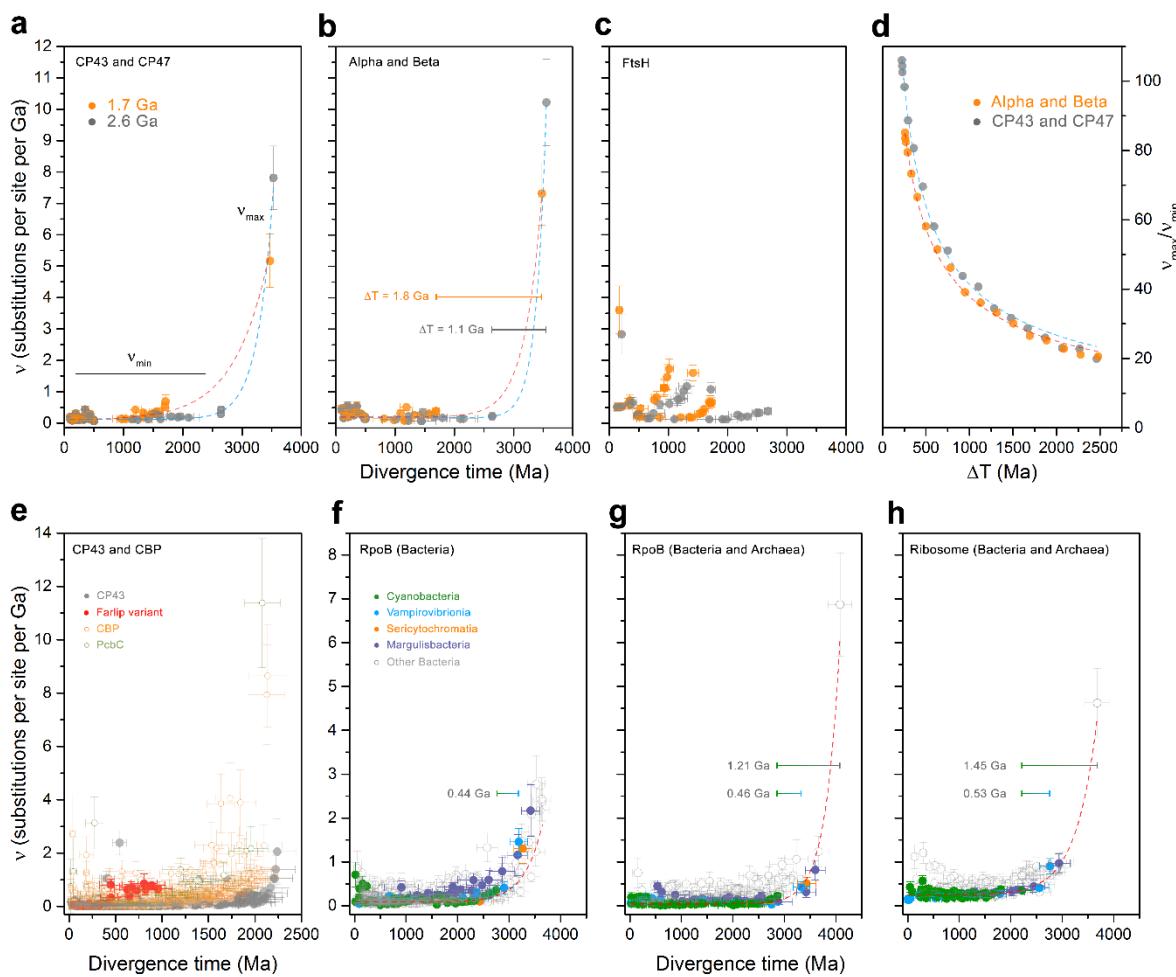
229 **Figure 3.** Schematic representation of ancestral duplication events. The MRCA of Cyanobacteria inherited a  
230 standard F-type ATP synthase, with a heterohexameric catalytic head (F<sub>1</sub>) made of alternating subunits Alpha  
231 and Beta; and a PSII with a heterodimeric core and antenna.  $\Delta T$  denotes the span of time between the ancestral  
232 duplication events and the MRCA of Cyanobacteria, which in the case of ATP synthase, is suspected to antedate  
233 the divergence of Bacteria and Archaea and their further diversification. FtsH contains an N-terminal  
234 membrane-spanning domain attached to a soluble domain consisting of an AAA (ATPase associated with  
235 diverse activities) module attached to a C-terminal protease domain. FtsH is universally conserved in Bacteria,  
236 has a hexameric structure like that of ATP synthase's catalytic head, and can be found usually as  
237 homohexamers, but also as heterohexamers. The MRCA of Cyanobacteria likely inherited three variant FtsH  
238 subunit forms, one of which appears to have duplicated after the divergence of the genus *Gloeobacter*, and  
239 possibly other early-branching Cyanobacteria [40]. This late duplication led to FtsH1 and FtsH2, which form  
240 heterohexamers with FtsH3, following the nomenclature of Shao, et al. [40] FtsH1/FtsH3 is found in the

241 cytoplasmic membrane of Cyanobacteria, while FtsH2/FtsH3 is involved in the degradation of PSII and other  
 242 thylakoid membrane proteins.

243

244 Now, if we consider a scenario in which  $dAB$  is 4.0 Ga and leaving all other constraints  
 245 unchanged,  $v_{\max}$  is  $6.02 \pm 0.9 \delta \text{ Ga}^{-1}$  resulting in a  $v_{\max}/v_{\min}$  of 21. If instead we keep the  
 246 duplication at 3.5 Ga but assume that the MRCA of Cyanobacteria occurred before the GOE  
 247 to simulate a more conservative scenario, at 2.6 Ga ( **$\Delta T = 1.1 \text{ Ga}$** ), we obtain that  $v_{\min}$  is  
 248 consequently slower,  $0.25 \pm 0.06 \delta \text{ Ga}^{-1}$ , when compared to a post-GOE ancestor. This older  
 249 MRCA (smaller  $\Delta T$ ) thus leads to a rise in  $v_{\max}$ , calculated to be  $10.22 \pm 1.37 \delta \text{ Ga}^{-1}$  and  
 250 leading to a  $v_{\max}/v_{\min}$  of 40. Given that the phylogenetic distance is a constant, the rate of  
 251 evolution increases with a decrease in  $\Delta T$  following a power law function. We had observed  
 252 nearly identical evolutionary patterns for the core RC proteins D1 and D2 of PSII [18]. The  
 253 change in  $v_{\max}/v_{\min}$  as a function of  $\Delta T$  is shown in Figure 4d.

254



255

256 **Figure 4.** Comparison of the changes in the rates of evolution as a function of time. **a** Rates of CP43 and CP47  
 257 modelled using two specific evolutionary scenarios. Orange trace was calculated under the assumption that the

258 MRCA of Cyanobacteria occurred after the GOE, at ~1.7 Ga, while the duplication of CP43 and CP47 occurred  
259 at ~3.5 Ga. The fastest rate seen at the point of duplication is denoted as  $v_{max}$ , and stabilizes during the  
260 Proterozoic,  $v_{min}$ . Grey dots represent a scenario in which the MRCA of Cyanobacteria is thought to have  
261 occurred before the GOE, at ~2.6 Ga instead. **b** Same calculations as **a** but performed on Alpha and Beta  
262 sequences of cyanobacterial and plastid ATP synthase. **c** Rates of evolution of cyanobacterial and plastid FtsH  
263 subunits assuming a MRCA of Cyanobacteria at 1.7 and 2.6 Ga. **d** Changes in the rates of evolution with  
264 varying  $\Delta T$  for ATP synthase (orange) and PSII subunits (grey). **e** Changes in the rates of evolution as a  
265 function of time for a relaxed molecular clock computed with the full dataset of CP43 (full circles) and CBP  
266 sequences (open circles). **f** Changes in the rate of evolution of bacterial RpoB showing an exponential decrease  
267 in the rate of evolution. The bar denotes the span of time between the MRCA of the clade containing  
268 Vampirovibrionia and Cyanobacteria and the MRCA of Cyanobacteria (0.44 Ga). **g** Changes in the rate of  
269 evolution of bacterial and archaeal RpoB. The longer bar represents the span of time between the divergence of  
270 Archaea and Bacteria and the MRCA of Cyanobacteria (1.21 Ga). **h** Changes in the rate of evolution of a dataset  
271 of concatenated ribosomal proteins showing similar patterns of evolution as RpoB. Error bars on each data point  
272 are standard errors on the rate and mean divergence time.

273

274 The core antenna of PSII, CP43 and CP47, showed patterns of divergence very similar to  
275 those of Alpha and Beta, both in terms of phylogenetic distances between paralogues and  
276 rates of evolution between orthologues (Figure 4a and b). The average rate of evolution of  
277 CP43 and CP47, assuming that the MRCA of Cyanobacteria occurred at 1.7 Ga, and the  
278 duplication ( $dCP$ ) at 3.5 Ga (**ΔT = 1.8 Ga**), is  $0.19 \pm 0.05 \delta \text{ Ga}^{-1}$ . Slower than for Alpha and  
279 Beta under the same condition. This slower rate is consistent with the fact that CP43 and  
280 CP47 show less sequence divergence between orthologues at all taxonomic ranks of oxygenic  
281 phototrophs when compared to Alpha and Beta (see Supplementary Table S3). Furthermore,  
282 the rate at  $dCP$ ,  $v_{max}$ , was  $5.17 \pm 0.84 \delta \text{ Ga}^{-1}$ , generating a  $v_{max}/v_{min}$  of 27, similar to Alpha  
283 and Beta (Figure 4). Thus, even when  $\Delta T$  is 1.8 Ga, the rate at duplication point needs to be  
284 27 times greater than the average rates observed during the Proterozoic. If we consider  
285 instead that the MRCA of Cyanobacteria occurred at 2.6 Ga and  $dCP$  at 3.5 Ga (**ΔT = 1.1**  
286 **Ga**), this would slowdown  $v_{min}$  to  $0.16 \pm 0.04 \delta \text{ Ga}^{-1}$ , while  $v_{max}$  would increase to  $7.81 \pm 1.01$   
287  $\delta \text{ Ga}^{-1}$  resulting in a  $v_{max}/v_{min}$  of 49. Therefore, the molecular evolution of the core subunits of  
288 PSII parallels that of ATP synthase both in terms of rates and distances through geological  
289 time.

290 We then studied a relatively recent gene duplication event (Figure 4c), which  
291 occurred long after the LUCA, but also after the MRCA of Cyanobacteria: that leading to  
292 Cyanobacteria-specific FtsH1 and FtsH2 ( $dH0$ ) [40]. This more recent duplication served as a  
293 point of comparison and control (see Figure 3 for a scheme). In marked contrast to  $dAB$ , the

294 rate at the point of duplication was  $0.66 \pm 0.21 \delta \text{ Ga}^{-1}$ . We found that FtsH1 is evolving at an  
295 average rate of  $1.42 \pm 0.29$ , while FtsH2 at a rate of  $0.24 \pm 0.06 \delta \text{ Ga}^{-1}$  under the assumption  
296 that MRCA of Cyanobacteria occurred at 1.7 Ga. Thus, under the assessed conditions, FtsH1  
297 is evolving about 5.3 times faster than FtsH2, while the latter is evolving at a rate similar to  
298 that of Alpha and Beta. If the MRCA of Cyanobacteria is assumed to have occurred at 2.6  
299 Ga, all rates slowdown respectively, but the rate of FtsH1 remains over five times faster than  
300 FtsH2. Unlike *dAB*, *dH0* is consistent with classical neofunctionalization, in which the copy  
301 that gains new function experiences an acceleration of the rate of evolution [41, 42]. Like  
302 PSII and ATP synthase, the calculated rates of evolution match observed distances as  
303 estimated by the change in the level of sequence identity as a function of time, in which the  
304 fastest evolving FtsH1 accumulated greater sequence change than FtsH2 in the same period  
305 (Supplementary Table S3).

306 Given that the complex evolution of CP43 and CBP involved several major  
307 duplication events and potentially large variations in the rate of evolution (Figure 1 and  
308 Supplementary Figure S1), we carried out a molecular clock analysis of a large dataset of 392  
309 CP43 and 465 CBP proteins using cross-calibrations across paralogues. Clocks were executed  
310 with no constraint on the MRCA of Cyanobacteria. The estimated mean divergence time for  
311 the oldest node, the duplication at the origin of CBP, is 2.23 Ga (95% confidence interval, CI:  
312 1.90 - 2.69 Ga) using an autocorrelated log normal model (see Figure 4e, Supplementary  
313 Figure S6 for a chronogram, and Supplementary Table S4 for a comparison of estimated ages  
314 under different models). The mean divergence time for the node representing the CP43  
315 inherited by the MRCA of Cyanobacteria was calculated to be 2.22 Ga (95% CI: 1.88 - 2.68  
316 Ga). Thus, a span of time of only 15 Ma is measured between these two mean ages. The  
317 average rate of evolution of CP43, not including CBP sequences, was found to be  $0.14 \pm 0.05$   
318  $\delta \text{ Ga}^{-1}$ , which is in the same range as determined in the simplified, but highly constrained  
319 experiment above. We noted a 6-fold increase in the rate of evolution associated with the  
320 duplication leading to the farlip-CP43 variant (Figure 4e). This duplication led to an  
321 acceleration of the rate similar in magnitude to that of FtsH1/FtsH2 and is consistent with a  
322 neofunctionalization as the photosystems evolved to use far-red light and bind chlorophyll *f*.

323 CBP sequences, on average, display rates of evolution about three times faster than  
324 CP43 (Figure 4e). However, the serial duplications that led to the evolution of CP43-derived  
325 light harvesting complexes resulted in accelerations in the rate of evolution of a similar  
326 magnitude as observed for *dAB* and *dCP*. The largest of these is associated with the origin of  
327 PcbC [30], a variant commonly found in heterocystous Cyanobacteria and Cyanobacteria that

328 use alternative pigments, such as chlorophyll *b*, *d* and *f*. The ancestral node of PcbC was  
329 timed at 2.07 Ga (95% CI: 1.76 - 2.50 Ga) with a rate of  $11.7 \pm 2.42 \text{ } \delta \text{ Ga}^{-1}$ , decelerating  
330 quickly, but stabilizing at about four times faster rates than the average rate of CP43. We find  
331 it noteworthy that the fast rates of evolution associated with the origin of CBP are not  
332 associated with very large spans of time between these and CP43, nor did it result in very old  
333 root node ages despite the use of very broad constraints.

334

### 335 2.3. *Species divergence*

336 To understand the evolution of MSV relative to Cyanobacteria we wished to apply a  
337 molecular clock to a system where the calculated rates could be compared to observed rates  
338 as determined by distances between species of known divergence times or at similar  
339 taxonomic ranks. We found RpoB to be suitable for this because it has been inherited  
340 vertically with few instances of horizontal gene transfer and had enough signal to resolve  
341 known phylogenetic relationships between and within clades. In collecting the RpoB  
342 sequences, we noted for the first time that Margulisbacteria and Vampirovibrionia share a  
343 comparatively greater level of divergence at similar taxonomic ranks than Cyanobacteria. For  
344 example, the level of sequence divergence of RpoB from two species of *Termititenax*  
345 (Margulisbacteria) [43] is about 40% greater than the distance between *Gloeobacter* spp. and  
346 any other cyanobacterium (not including gaps and insertions), the latter being the largest  
347 distance between oxygenic phototrophs. In the case of Gastranaerophilales  
348 (Vampirovibrionia) [5], which are specialised gut bacteria and should therefore not be much  
349 older than animals, the level of sequence identity of RpoB was found to be 70% for the two  
350 most distant strains in this group, contrasted to 84% for *Gloeobacter* spp. when compared to  
351 any other cyanobacterium. As listed in Supplementary Table S2, within-group mean  
352 distances suggest that faster rates are widespread and not just unique to RpoB.

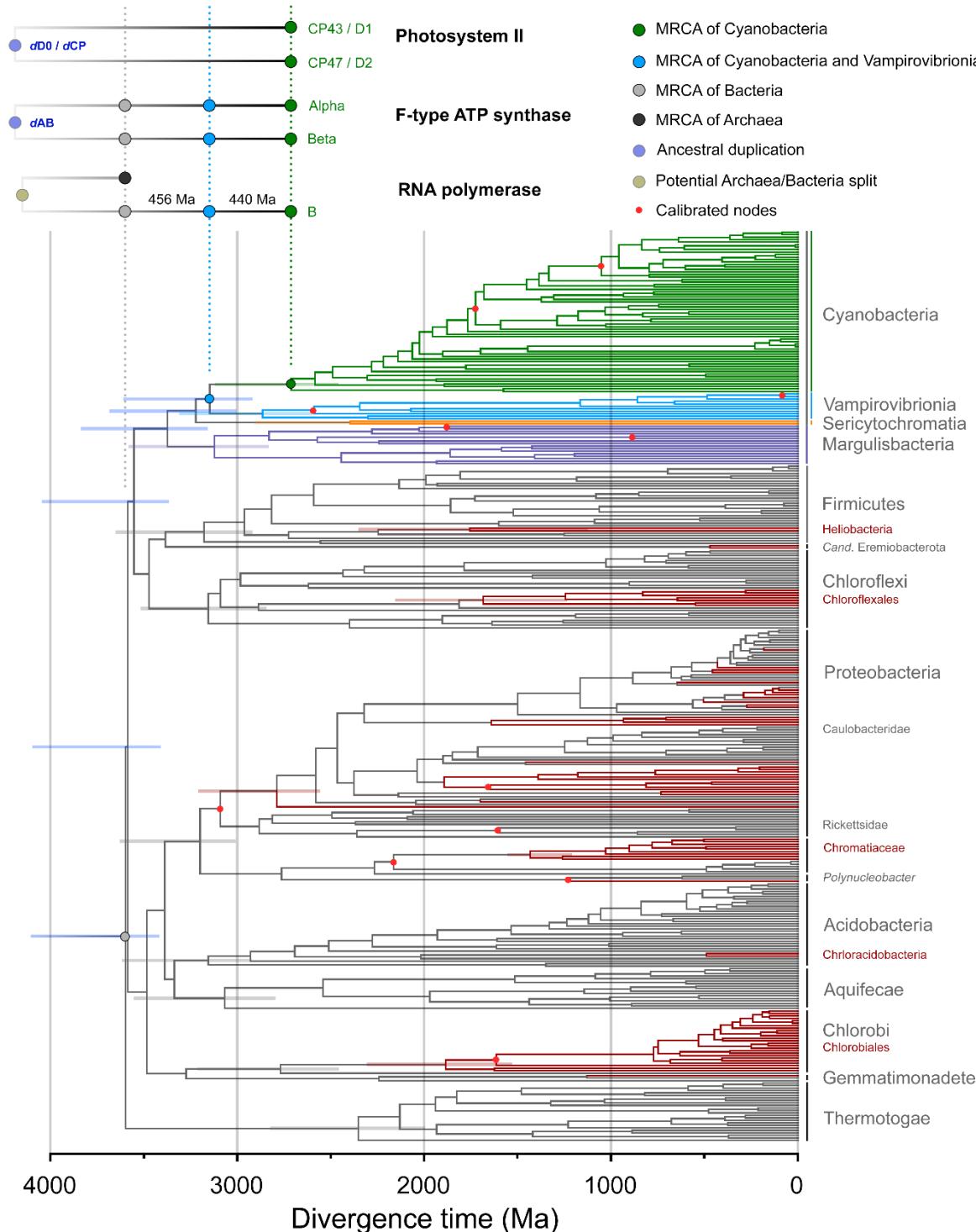
353 We implemented a set of 12 calibrations across bacteria, including two calibrations on  
354 Margulisbacteria and two in Vampirovibrionia with the aim of covering both slower and  
355 faster evolving lineages. The following results are based on an autocorrelated log normal  
356 molecular clock using CAT+Γ, a root constrained with a broad interval ranging from between  
357 4.52 and 3.41 Ga, and as described in Materials and Methods (Figure 5). We found this to  
358 perform well and provided results comparable to other independent studies that did not  
359 combine a full set of MVS sequences and other clades with phototrophs in a single tree  
360 (Table 1). Nonetheless, a pipeline of sensitivity experiments tested the dependency of these

361 results on models and prior assumptions: these are shown and described in Supplementary  
362 Figure S7 and S8.

363 The root of the tree (divergence of Thermotogae) was timed at 3.64 Ga (95% CI: 3.42  
364 - 4.11 Ga) and the divergence of Cyanobacteria at 2.74 Ga (95% CI: 2.46 - 3.12 Ga). Thus,  
365 the span of time between the mean age of the root and the mean age of the MRCA of  
366 Cyanobacteria was calculated to be 0.89 Ga. The span of time between Margulibacteria and  
367 Cyanobacteria was found to be 0.67 Ga; and between Vampirovibrionia and Cyanobacteria  
368 0.44 Ga (Figure 4f and 5). The latter is a value that is consistent with previous studies using  
369 entirely different rationales, datasets and calibrations [8, 9].

370 We also noted an exponential decrease in the rates of evolution of RpoB through the  
371 Archean, which stabilised at current levels in the Proterozoic (Figure 4f). The rate at the root  
372 node was calculated to be  $2.37 \pm 0.45 \delta \text{ Ga}^{-1}$  and the average rate of evolution of RpoB  
373 during the Proterozoic was found to be  $0.19 \pm 0.06 \delta \text{ Ga}^{-1}$ . The average rate of cyanobacterial  
374 RpoB was  $0.14 \pm 0.04 \delta \text{ Ga}^{-1}$ ; for Margulibacteria was  $0.44 \pm 0.17 \delta \text{ Ga}^{-1}$ , and for  
375 Vampirovibrionia  $0.19 \pm 0.05 \delta \text{ Ga}^{-1}$ : about 3.1x and 1.3x the mean cyanobacterial rate  
376 respectively. These rates agree reasonably well with the observed distances (Supplementary  
377 Table S2), further indicating that the calibrations used in these clades performed well and  
378 recapitulated patterns of evolution consistent with lifestyle and trophic modes. Nevertheless,  
379 we suspect that the values for MSV represent underestimations of the true rates of evolution  
380 (slower than they should be), as some of the clades that include symbionts still appear much  
381 older than anticipated from their hosts (Supplementary Text S2).

382 Furthermore, a more complex, but commonly used model like CAT+GTR+ $\Gamma$   
383 implementing a birth-death prior with ‘soft bounds’ on the calibrations, resulted in rates that  
384 were smoothed out, which translated into spread-out divergence times with Margulibacteria  
385 and Vampirovibrionia evolving at 1.9x and 0.7x times the cyanobacterial rates, respectively  
386 (Supplementary Figure S8, model **n** to **p**). These weird rate effects are thus translated into a  
387 Mesoproterozoic, very late, age for Cyanobacteria, and a relatively older divergence time for  
388 Vampirovibrionia: results that replicate those presented recently in ref. [11].



389

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

397

398 **Table 1.** Divergence time estimates

Event	RpoB (Ga) (95% CI)	Concatenated ribosomal proteins	TimeTree compilation <sup>a</sup>	Shih et al. <sup>b</sup>	Magnabosco et al. <sup>c</sup>
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 Sericytchromatia	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 <i>Chloracidobacterium</i>	2.04 (1.53-2.51)				

399 <sup>a</sup>Data from TimeTree compiling estimated divergence times from independent studies as described in ref. [44]  
400 For the divergence of Thermotogae the estimated age was taken from eight different studies (n=8). For the  
401 divergence of Melaibacteria (or divergence from Cyanobacteria's closest relatives), n=10. For stem  
402 Heliobacteria, n=2. For stem phototrophic Chloroflexi and Chlorobi, the values were reported from a single  
403 study, Marin, et al. [45] For earliest potential phototrophic Proteobacteria, n=3. This latter was taken as the node  
404 made by the clades including *Niveispirillum lacus* and *Rhodobacter sphaeroides* to match our RpoB tree. The  
405 independent studies used by TimeTree to generate the estimated ages are listed in Supplementary Table S8.

406 <sup>b</sup>Data taken from ref. [8] Model T68 for the cyanobacterial dates. That is, no GOE calibration with a calibration  
407 on *Bangiomorpha*. The dates on Chloroflexi were taken from ref. [46]

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

409  
410 To investigate the effect of the age of the root on the divergence time of MVS and  
411 Cyanobacteria we also varied the root prior from 3.2 to 4.4 Ga (Supplementary Figure S7).  
412 We noted that regardless of the time of origin of Bacteria (approximated by the divergence of  
413 Thermotogae in our analysis), a substantially faster rate is required during the earliest  
414 diversification events, decreasing through the Archaean and stabilizing in the Proterozoic.

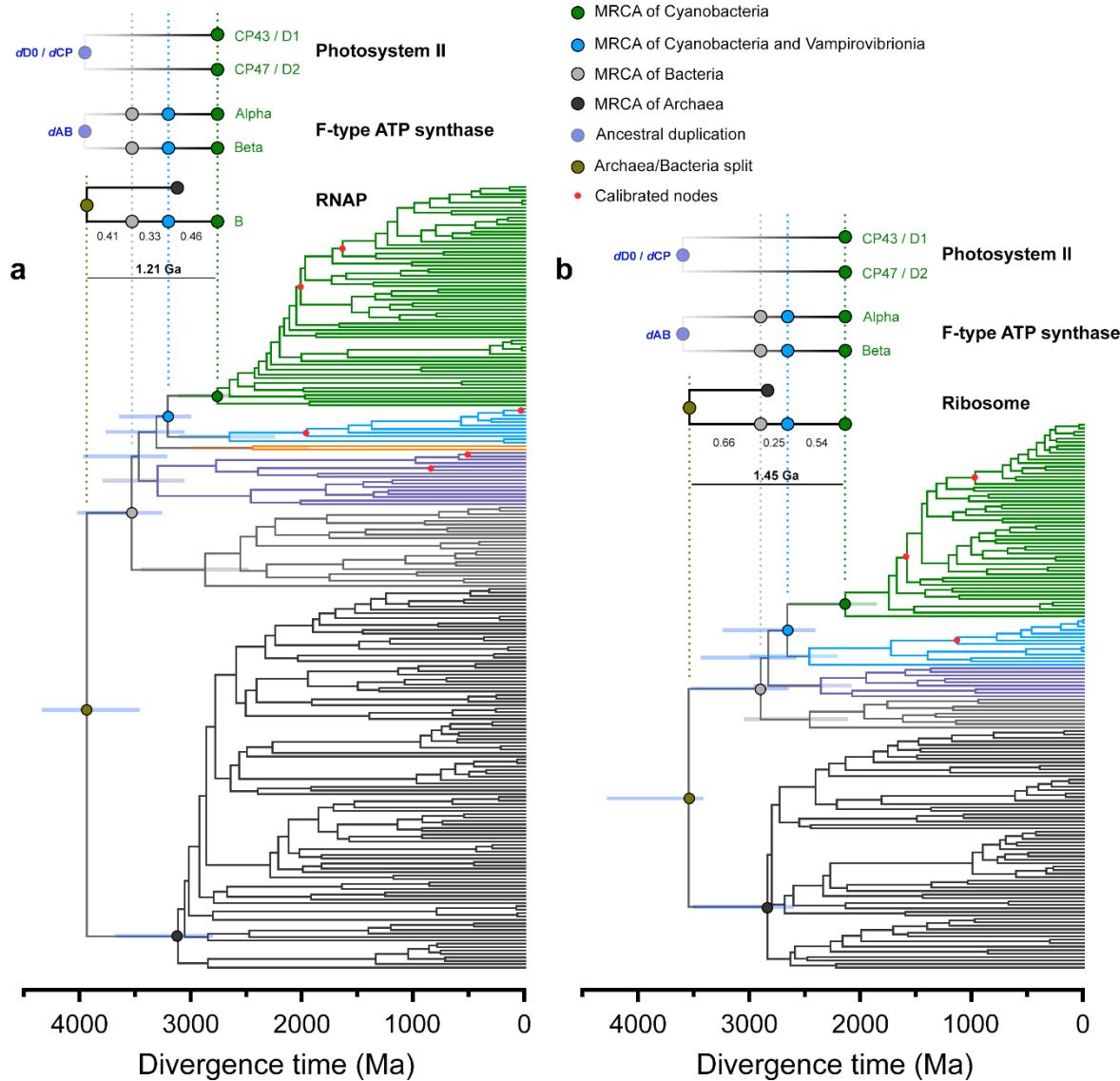
415 This matches well the patterns of evolution of ATP synthase and PSII core subunits as shown  
416 in the previous section.

417 We then compared the above RpoB molecular clock with a different clock that  
418 included a set of 112 diverse sequences from Archaea, in addition to the sequences from  
419 Thermotogae, MSV, and Cyanobacteria, but removing all other bacterial phyla (Figure 6a).  
420 We found that the calculated average rate of evolution of bacterial RpoB during the  
421 Proterozoic was slower ( $0.09 \pm 0.03 \delta \text{ Ga}^{-1}$ ) than in the absence of archaeal sequences ( $0.19 \pm$   
422  $0.06 \delta \text{ Ga}^{-1}$ ), resulting in overall older mean ages (see Figure 4g). However, the rate at the  
423 Bacteria/Archaea divergence point was  $6.87 \pm 1.17 \delta \text{ Ga}^{-1}$ , similar to the rate for *dAB* and  
424 *dCP*, requiring therefore an exponential decrease in the rates similar to that observed for ATP  
425 synthase and the PSII core subunits. Notably, this rate and exponential decrease is associated  
426 with a span of time between the mean estimated ages of the LUCA and the MRCA of  
427 Cyanobacteria of 1.21 Ga. The span between Vampirovibrionia and Cyanobacteria was found  
428 to be 0.46 Ga (Figure 6a).

429 Figure 5 also highlights that the MRCA of none of the groups containing anoxygenic  
430 phototrophs nor their divergence from their closest non-phototrophic relatives appears to be  
431 older than one of the oldest and best accepted geochemical evidence for photosynthesis at  
432 3.41 Ga [47] (Table 1). For example, the MRCA of Heliobacteria or of phototrophic  
433 Chlorobi, containing homodimeric type I RCs, which have traditionally been described as to  
434 harbour primitive forms of photosynthesis, are likely to have existed after the GOE, even  
435 allowing for large uncertainties in the calculations.

436 Finally, we performed a similar molecular clock analysis on a subset of concatenated  
437 ribosomal proteins published by Hug et al. [38] that included Archaea, Thermotogae,  
438 Margulisbacteria, Vampirovibrionia and Cyanobacteria. Even though this dataset generated  
439 younger ages compared to RpoB (see Table 1 and Figure 6), a similar exponential decrease in  
440 the rates was observed with a rate at the Bacteria/Archaea divergence point measured at  $4.62 \pm 0.71 \delta \text{ Ga}^{-1}$  and an average rate of bacterial ribosomal protein evolution of  $0.30 \pm 0.07 \delta \text{ Ga}^{-1}$  (Figure 4h). The span of time between the mean estimated ages of the LUCA and the  
442 MRCA of Cyanobacteria was 1.45 Ga; and between Vampirovibrionia and Cyanobacteria  
443 0.54 Ga (Figure 6b). If one assumes therefore that the span of time between the LUCA and  
444 the diversification of Bacteria is narrower, that would imply a faster rate at the root node and  
445 a steeper deceleration of the rates.

447



448  
449 **Figure 6.** Comparison of relaxed molecular clocks of RpoB (a) and a concatenated dataset of ribosomal proteins  
450 (b) that include sequences from Archaea and as described in the main text. RNAP stands for RNA polymerase.  
451 Clocks were calculated using a calibration on the root with a maximum of 4.52 Ga and a minimum of 3.41 Ga  
452 and applying a log normal autocorrelated clock with CAT+Γ model. Bars on selected nodes denote 95%  
453 confidence intervals.

454

#### 455 2.4. Structural analysis

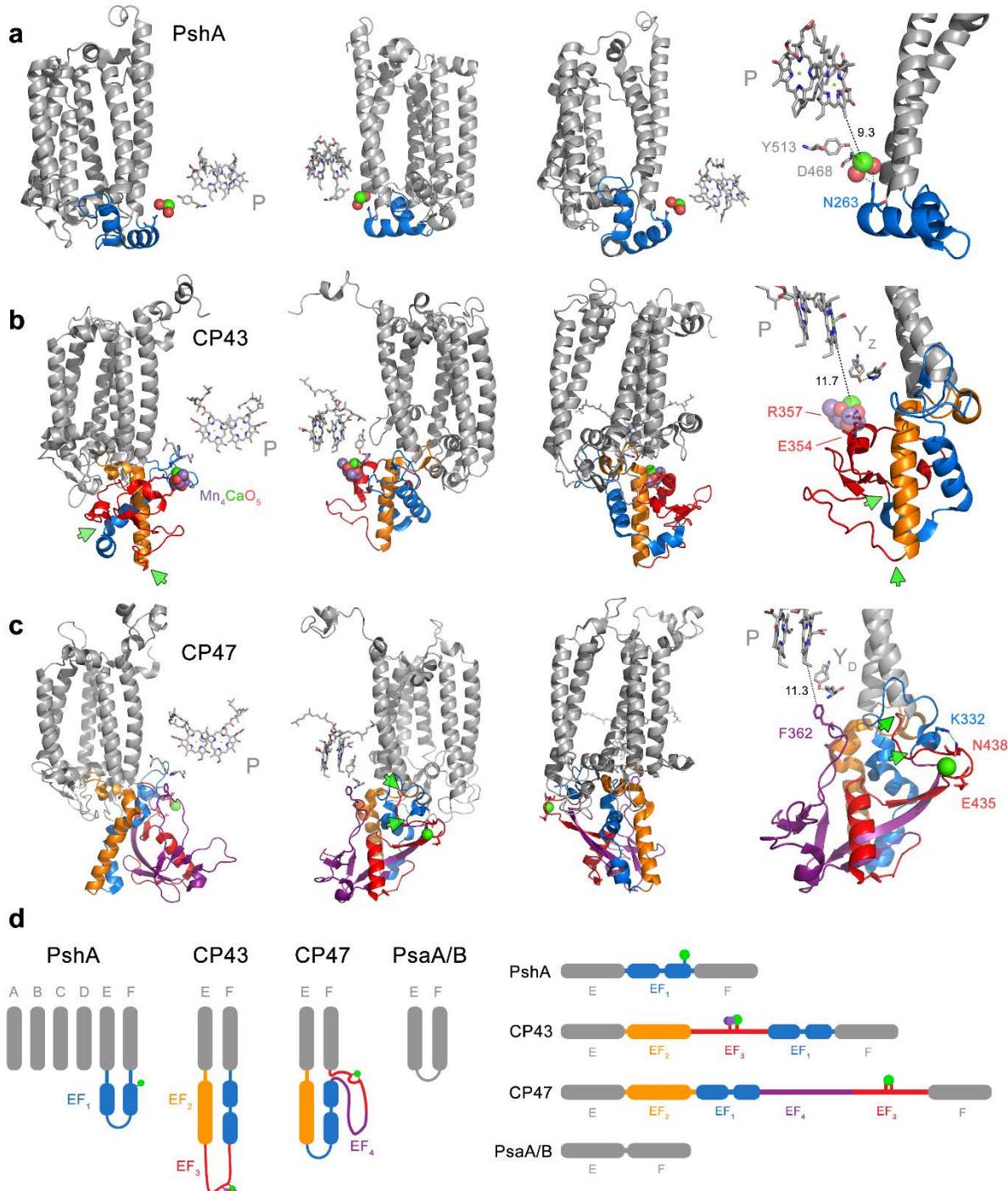
456 A fundamental premise of our investigation is that water oxidation started before the  
457 duplication of D1 and D2, and CP43 and CP47. The rationale behind this premise has been  
458 laid out before [17, 48], and more extensively recently [18]. This rationale raises the question  
459 of how the D2/CP47 side of the RC lost its capacity to carry out water-splitting catalysis. To  
460 gain further insight on the nature of the structural site around the water-oxidising complex in  
461 the ancestral photosystem, we used ancestral sequence reconstruction to predict the most

462 probable ancestral states. We will refer to the ancestral protein to D1 and D2 as D0  
463 (Supplementary Figure S9). We generated 14 predicted D0 sequences using a combination of  
464 three ASR methods and amino acid substitution models. On average the 14 D0 sequences had  
465  $87.12 \pm 0.55\%$  sequence identity indicating that the different algorithms provided largely  
466 consistent results. While the regions that include all transmembrane helices are aligned  
467 unambiguously, the N-terminal and C-terminal ends were aligned less confidently due to  
468 greater sequence variability at both ends. Nonetheless, we found that the predicted D0  
469 sequences retain more identity with D1 than D2 along the entire sequence. The level of  
470 sequence identity of D0 compared to the D1 (PsbA1) of *Thermosynechococcus vulcanus* was  
471 found to be  $69.58 \pm 0.55\%$ , and  $36.32 \pm 0.15\%$  compared to D2.

472 The ligands to the Mn<sub>4</sub>CaO<sub>5</sub> in PSII are provided from three different structural  
473 domains (Supplementary Figure S10): 1) the D1 ligands D170 and E189, located near the  
474 redox Y<sub>z</sub>. These are in the luminal loop between the 3<sup>rd</sup> and 4<sup>th</sup> transmembrane helices. 2)  
475 The D1 ligands H332, E333, D342, and A344 located at the C-terminus. 3) The CP43  
476 ligands, E354 and R357, located in the extrinsic loop between the 5<sup>th</sup> and 6<sup>th</sup> helices, with the  
477 latter residue less than 4 Å from the Ca. Remarkably, there is structural and sequence  
478 evidence supporting the loss of ligands in these three different regions of CP47/D2.

479 In all the D0 sequences, at position D1-170 and 189, located in the unambiguously  
480 aligned region, the calculated most likely ancestral states were E170 and E189, respectively.  
481 The mutation D170E results in a PSII phenotype with activity similar to that of the wild-type  
482 [49]. At position D1-170, a glutamate was predicted with average posterior probabilities  
483 (PPs) of 44.2% (FastML), 67.2% (MEGA) and 77.0% (PAML). At position D1-E189, the  
484 average PPs for glutamate were 31.1% (FastML), 35.2% (MEGA) and 40.2% (PAML). The  
485 distribution of PPs across a selection of D0 sequences and key sites is presented in  
486 Supplementary Figure S11 and Table S5. In contrast, D2 has strictly conserved phenylalanine  
487 residues at these positions, but the PP of phenylalanine being found at either of these  
488 positions was less than about 5% for all predicted D0 sequences. As a comparison, the redox  
489 active tyrosine residues Y<sub>z</sub> (D1-Y161) and Y<sub>D</sub> (D2-Y160), which are strictly conserved  
490 between D1 and D2, have a predicted average PPs of 68.8% (FastML), 98.8% (MEGA) and  
491 98.6% (PAML). Therefore, the ligands to the catalytic site in the ancestral protein leading to  
492 D2 were likely lost by direct substitutions to phenylalanine residues, while retaining the  
493 redox active D2-Y160 (Y<sub>D</sub>) and H189 pair (Supplementary Figure S10).

494



495

496 **Figure 7.** Structural rearrangements of the large extrinsic loop of CP43 and CP47. **a** The antenna domain of  
497 heliobacterial PshA is shown in three different rotated perspectives. Only the first six transmembrane helices of  
498 the antenna are shown for clarity. A Ca at the electron donor side is bound from an extrinsic loop between the  
499 5<sup>th</sup> (E) and 6<sup>th</sup> (F) helices. This extrinsic loop, EF<sub>1</sub> (blue), is made of two small alpha helices. The fourth  
500 molecular view furthest to the right shows the link between the electron donor site and EF<sub>1</sub> in closer detail. **b**  
501 The CP43 subunit of PSII with the extrinsic loop shown in colours. **c** The CP47 subunit of PSII. Immediately  
502 after the 5<sup>th</sup> helix (E), a long alpha helix protrudes outside the membrane in both CP43 and CP47 and showing  
503 structural and sequence identity (orange). We denote this helix EF<sub>2</sub>. After EF<sub>2</sub> structural differences are noticed  
504 between CP43 and CP47 as schematised in panel **d**. In CP43, after helix EF<sub>2</sub> a loop is found (shown in red

505 ribbons), which we denote EF<sub>3</sub>. This contains the residues that bind the Mn<sub>4</sub>CaO<sub>5</sub> cluster and it is followed by a  
506 domain that resembles EF<sub>1</sub> in the HbRC at a structural level. In CP47, EF<sub>3</sub> and EF<sub>1</sub> retain sequence identity with  
507 the respective regions in CP43. CP47 has additional sequence that is not found in CP43 (EF<sub>4</sub>, purple). The green  
508 arrows mark the position at which the domain swap occurred in CP43 relative to CP47. We found that the  
509 CP43-E354 and R357 are found in the equivalent domain in CP47 as E436 and N438 coordinating a Ca atom.  
510 N438 (EF<sub>3</sub>) links to EF<sub>1</sub> via K332. It is unclear if the EF<sub>1</sub> region in the HbRC is strictly homologous to that in  
511 CP43 and CP47 as very little sequence identity is found between the two: however, a couple of conserved  
512 residues between all EF<sub>1</sub> may suggest it emerged from structural domains present in the ancestral RC protein  
513 (see Supplementary Figure S13).

514

515 Prompted by the finding of a Ca-binding site at the electron donor site of the homodimeric  
516 type I RC of Helio**bacter**ia (Firmicutes) with several similarities to the Mn<sub>4</sub>CaO<sub>5</sub> cluster of  
517 PSII, including a link to the antenna domain and the C-terminus [19], we revisited the  
518 sequences and structural overlaps of CP43 and CP47. We found that a previously unnoticed  
519 structural rearrangement within the extrinsic loop occurred in one subunit relative to the other  
520 (marked EF<sub>3</sub> and EF<sub>4</sub> in Figure 7, Supplementary Figure S12 and S13). CP43 retains the  
521 simplest domain, being about 60 residues shorter than CP47. If CP43 retains the ancestral  
522 fold, the additional sequence in CP47's swapped domain (EF<sub>4</sub> in Figure 7d) would have  
523 contributed to the loss a catalytic cluster as it inserted one phenylalanine residue (CP47-  
524 F360) into the electron donor site, less than 4 Å from Y<sub>D</sub>. An equivalent residue does not exist  
525 in CP43.

526 We then noted that in the swapped region (EF<sub>3</sub> in Figure 7d), sequence identity is  
527 retained between CP43 and CP47 (Supplementary Figure S12). We found that CP43-E354  
528 and R357 are equivalent to CP47-E435 and N438. An inspection of the crystal structure of  
529 cyanobacterial PSII showed that these two residues specifically bind a Ca of unknown  
530 function in (Figure 7c and d). The presence of an equivalent glutamate to CP43-E354 in  
531 CP47 is consistent with this being already present before duplication.

532 Finally, a peculiar but well-known trait conserved across Cyanobacteria and  
533 photosynthetic eukaryotes is that the 5' end of the *psbC* gene (CP43) overlaps with the 3' end  
534 of the *psbD* gene (D2) usually over 16 bp (Supplementary Table S6). The *psbD* gene contains  
535 a well-defined Shine-Dalgarno ribosomal binding site downstream of the *psbC* gene and over  
536 the coded D2 C-terminal sequence [50-52]. The evolution of this unique gene overlap has no  
537 current explanation in the literature, but its origin could have disrupted the C-terminal ligands  
538 of the ancestral protein leading to the modern D2, an event that would have contributed to  
539 and favoured heterodimerization.

540

### 541 **3. Discussion**

#### 542 *3.1. Origin of oxygenic photosynthesis and rates of evolution*

543 The duplication of ATP synthase's ancestral catalytic subunit, and the archaeal/bacterial  
544 divergence of RNA polymerase and the ribosome, are some of the oldest evolutionary events  
545 known in biology [20-28]. When taken on their own, their phylogenetic features are often  
546 interpreted as evidence of the earliest origin. We have shown in here and in our previous  
547 work [18] that PSII show patterns of molecular evolution that closely parallel those of these  
548 very ancient systems and independently of the exact time of origin of Cyanobacteria or their  
549 closest relatives. Therefore, it cannot be taken for granted that there was ever a long period of  
550 time between the origin of life and the origin of anoxygenic photosynthesis, followed by  
551 another long period of time between the origin of anoxygenic photosynthesis and the origin  
552 of oxygenic photosynthesis.

553 The exponential decrease in the rates of evolution observed in the studied systems,  
554 even when the span of time between the ancestral duplications (or the LUCA) and the MRCA  
555 of Cyanobacteria was well over a billion years, are consistent with our assessment that a large  
556  $\Delta T$  for the evolution of the core subunits of ATP synthase and PSII cannot be explained  
557 entirely by duplication-driven effects. Instead, we speculate that these effects are the result of  
558 the faster rates of evolution that are expected to have occurred during the earliest history of  
559 life due to unrestricted positive selection at the origin of bioenergetics, accelerated rates of  
560 mutations caused by environmental factors [53-55], lack of sophisticated DNA repair  
561 mechanisms [56], or other genetic properties attributed to early life [57, 58], yet in  
562 combination with very large spans of time.

563 It is worth highlighting that ATP synthase, RNA polymerase, the ribosome, and the  
564 photosystems, are all complex molecular machines, with crucial functions and under strict  
565 regulation. These features largely explain their slow rates of evolution and the high degree of  
566 sequence (structural and functional) conservation through geological time. A similar level of  
567 complexity, however, can be traced back to the last common ancestor of each one of these  
568 molecular systems regardless of how ancient they truly are. Now, given that the rates of  
569 evolution of the core of PSII have remained slower than those of ATP synthase for billions of  
570 years, even through evolutionary processes that could be associated with acceleration in the  
571 rates of evolution such as: 1) primary endosymbiosis at the origin of photosynthetic  
572 eukaryotes [59], 2) radiations within clades of flowering plants [60], or 3) the radiation of  
573 marine *Prochlorococcus* and *Synechococcus* [61, 62], just to name a few; if those rates have

574 remained proportional since the origin of the enzymes, it could imply that the duplications of  
575 the core of PSII occurred before the duplication leading to Alpha and Beta.

576 In consequence, to argue that oxygenic photosynthesis is a late evolutionary  
577 innovation relative to the origin of life, one must first demonstrate that the rates of protein  
578 evolution of PSII and ATP synthase catalytic core subunits, RNA polymerase, and the  
579 ribosome, have not remained proportional to each other through geological time. In such a  
580 way that PSII—exclusively—experienced unprecedently faster rates of evolution. One  
581 could potentially argue that the origin of water oxidation itself could account for this  
582 hypothetical unprecedently faster rates. However, because of the relationships between  
583 *speed, distance, and time*, a decrease in  $\Delta T$  of about 60% (from 1.1 to 0.46 Ga for example),  
584 would result in a 30-fold additional increase in the rates at the earliest stages of  
585 diversification, resulting in photosystems that would be evolving at rates that surpass those of  
586 the fastest evolving peptide toxins (Figure 4d and Supplementary Text S1). It should be  
587 noted, that a  $\Delta T$  of over a billion years already accounts for a period of fast evolution at the  
588 origin of all these ancient systems.

589 The reader should also be reminded that these patterns of evolution, though they  
590 might appear somewhat unusual, do not emerge from the application of any particular  
591 molecular clock approach or computational analysis, but from the inherently long  
592 phylogenetic distance that separates Alpha and Beta, Archaea and Bacteria, or the core  
593 subunits of PSII, together with relatively slow rates of evolution throughout their entire  
594 multibillion-year diversification process.

595

### 596 *3.2. Diversification of Bacteria*

597 It has been postulated before that the diversification of the major phyla of Bacteria occurred  
598 very rapidly, starting at about 3.4 Ga and peaking at about 3.2 Ga, in what was dubbed as the  
599 Archean Genetic Expansion [63]. Another recent independent molecular clock analysis put  
600 the MRCA of Bacteria at about 3.4 Ga [11]; and Marin et al. [45] placed the major bacterial  
601 radiation, except for the divergence of Thermotogae and Aquifecae, starting at a mean  
602 divergence time of about 3.2 Ga too. Our clocks of RpoB or ribosomal proteins are also in  
603 agreement with these patterns. A recent study suggested that the major groups of Bacteria and  
604 Archaea diversified rapidly after the LUCA and hypothesized that the long distance between  
605 archaeal and bacterial ribosomal proteins could be attributed to fast rates of ribosome  
606 evolution during their divergence [64], but see also [65]. Our analyses suggest that the more  
607 explosive the diversification of prokaryotes, the greater the chance that photosynthetic water-

608 splitting is an ancestral trait of all life. And yet, if phototrophic communities—whether  
609 oxygenic or not—already existed 3.2 to 3.4 Ga ago, as it is supported by the geochemical  
610 record [47, 66], then the earliest bacteria were likely photosynthetic. That the earliest bacteria  
611 were photosynthetic is entirely consistent with the evolution of RC proteins, which indicates  
612 that the structural and functional specialization that led to the two photosystem types,  
613 antedated the diversification processes leading to the known phyla containing photosynthetic  
614 bacteria [34]. This is also consistent with the fact that none of the groups of extant  
615 photosynthetic bacteria appear to be older than the earliest geochemical evidence for  
616 photosynthesis.

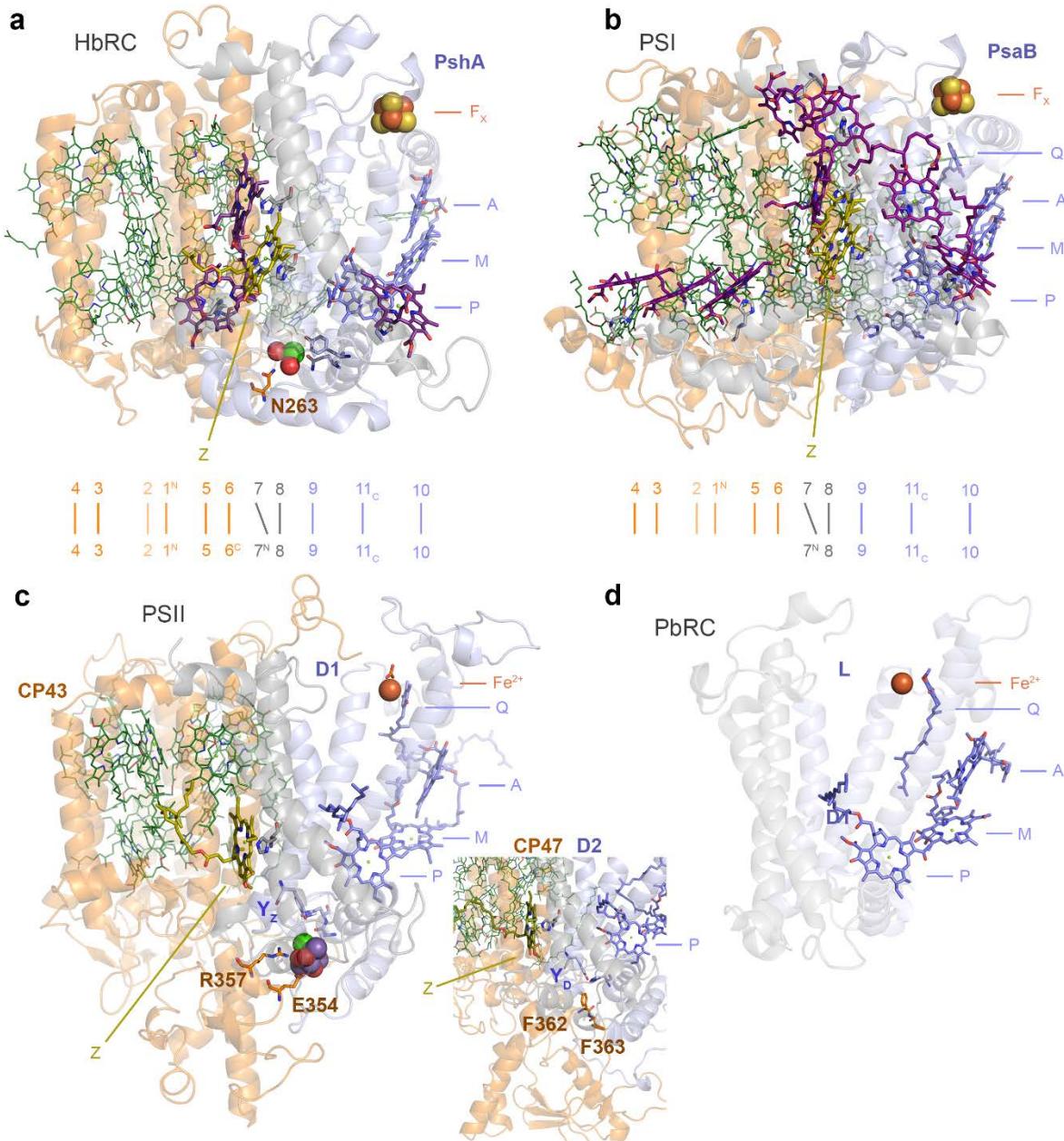
617 The presented data is also consistent with recent studies of the oxygenation of the  
618 planet, which suggests that even if photosynthetic O<sub>2</sub> evolution started as early as the oldest  
619 rocks, the properties of early Earth biogeochemistry would have maintained very low  
620 concentrations of O<sub>2</sub> over the Archean without the need to invoke any particular biological  
621 innovation or trigger to coincide with the GOE [67-69].

622

### 623 *3.2. Structural constraints*

624 We showed that the phylogenetic distance between CP43/CP47 and other type I RC proteins  
625 is the second largest distance after that between type I and type II (Supplementary Figure S2).  
626 It is conventionally considered that the first six transmembrane helices of the photosystems  
627 make up the antenna domain, while the photochemical core encompasses the last five helices.  
628 Structurally and functionally the antenna domain extends to the 8<sup>th</sup> helix, both in type I RCs  
629 and in PSII; with the latter retaining one antenna chlorophyll in the equivalent 8<sup>th</sup> helix  
630 (marked Z in Figure 8), as well as substantial sequence identity around this chlorophyll's  
631 binding site as shown before [34, 70]. These conserved features indicate that there is not a  
632 moment in time during the evolution of PSII in which it was devoid of core antenna proteins.  
633 In other words, CP43/CP47 are a descendant of the ancestral core antenna of type II RCs,  
634 now lost and replaced in anoxygenic type II RCs of phototrophic Proteobacteria, Chloroflexi  
635 and others. Both structural and phylogenetic evidence are in agreement and indicate that the  
636 ancestral type II RC, at the dawn of photosynthesis, was architecturally like water-splitting  
637 PSII. The above disproves and supersedes a previous hypothesis by Cardona [35] where the  
638 antenna of PSII was claimed to have originated from a refolding of an entire type I RC  
639 protein.

640



641  
642 **Figure 8.** Structural comparisons of the antenna and core domains of photosynthetic RCs. **a** A monomeric unit  
643 (PshA) of the homodimeric RC of Heliobacteria (HbRC, Firmicutes). The RC protein is made up of 11  
644 transmembrane helices. The first six N-terminal helices are traditionally considered as the antenna domain  
645 (orange ribbons), while the last five helices are the RC core domain (grey and light-blue ribbons). Below the  
646 structure, the organization of the 11 helices is laid down linearly for guidance: 1<sup>N</sup> denotes the first N-terminal  
647 helix and 11<sub>c</sub> the last C-terminal helix. P denotes the “special pair” pigment; M, the “monomeric”  
648 bacteriochlorophyll electron donor; and A, the primary electron acceptor.  $F_x$  is the  $Fe_4S_4$  cluster characteristic of  
649 type I RCs. Antenna pigments bound by the antenna domain are shown in green lines. Antenna pigments bound  
650 by the 7<sup>th</sup> and 8<sup>th</sup> helices are shown as purple sticks, except for the bacteriochlorophyll molecule denoted as Z, in  
651 olive green and bound by the 8<sup>th</sup> helix. The antenna domain connects the electron donor side of the RC via N263  
652 and a Ca-binding site. **b** A monomeric unit of PSI (PsaB, Cyanobacteria) following a similar structural  
653 organization and nomenclature as in panel **a**. Unlike the HbRC, PSI binds quinones (Q) as intermediary electron

654 acceptors between A and F<sub>X</sub>. **c** A monomeric unit of PSII (CP43/D1 and CP47/D2 in the inset). It has a  
655 structural organization similar to that of type I RCs. However, the monomeric unit is split into two proteins after  
656 the 6<sup>th</sup> transmembrane helix. The Mn<sub>4</sub>CaO<sub>5</sub> cluster is coordinated by D1, but is directly connected to the antenna  
657 domain via E354 and R357 in manner that resembles the HbRC. **d** A monomeric unit of an anoxygenic type II  
658 reaction (PbRC, Proteobacteria). Unlike type I RCs, type II RCs lack F<sub>X</sub>. Instead, a non-heme Fe<sup>2+</sup> is found  
659 linking the RC proteins. The PbRC lacks antenna domain and the antenna pigment Z bound at the equivalent 8<sup>th</sup>  
660 helix.

661  
662 Sequence reconstruction of the ancestral subunit to D1 and D2 is consistent with the  
663 existence of a highly oxidizing homodimeric photosystem, though transient and short-lived  
664 [18], that could split water in either side of the RC [48]. The structural comparisons between  
665 CP43/D1 and CP47/D2 suggest a mechanism for heterodimerization and loss of catalysis on  
666 one side that accounts for all ligands. These include direct amino acid substitutions, the  
667 domain swap within the extrinsic region of the ancestral protein to CP47, and the gene  
668 overlap between *psbC* (CP43) and *psbD* (D2). It would be difficult to reconcile these unique  
669 structural and genetic features with a scenario in which PSII evolved water oxidation at a late  
670 stage, or starting as a purple anoxygenic type II RC, once the heterodimerization process was  
671 well underway or completed, or in the absence of core antenna domains, given that these  
672 interact directly with the electron donor side of PSII, in a manner strikingly similar to that of  
673 the homodimeric type I RC of the Firmicutes [19]. What is more, these structural and  
674 functional features also suggest that the atypical forms of D1, lacking ligands to the Mn<sub>4</sub>CaO<sub>5</sub>  
675 cluster, such as the so-called chlorophyll *f* synthase [32, 71], or the so-called “rogue” or  
676 “sentinel” D1 [31, 72], diverged from forms of D1 that were able to support water oxidation.  
677 Even if they originated from early duplications that could have antedated the MRCA of  
678 Cyanobacteria [2] and as additionally supported by ASR analysis (Supplementary Figure S9  
679 and S10).

680 We have calculated that (oxygenic) PSII has experienced the slowest rates of  
681 evolution between type II RCs, with the core of the anoxygenic type II RC of Proteobacteria  
682 and Chloroflexi evolving approximately five times faster than the core of PSII [18]. That  
683 considerably faster rate has led to conspicuous structural changes of the anoxygenic RC  
684 relative to PSII and type I photosystems. The consequence of these differences in the rates are  
685 visually apparent (Figure 8 and Supplementary Figure S3) as the anoxygenic RC has  
686 experienced greater sequence and structural change than PSII. That is, PSII retains a greater  
687 number of ancestral traits that can be traced to before the ancestral core duplications. For  
688 example, like type I, PSII retains: 1) substantially greater structural symmetry at the core, 2)

689 the antenna functionally linked to the core and the electron donor side, 3) lack of histidine  
690 ligands to the monomeric photochemical chlorophylls (marked M in Figure 8), or 4) the use  
691 of a chlorophyll *a* derived pigment as the primary electron acceptor. This list is not meant to  
692 be exhaustive, but see refs. [19, 73] for additional detail. It follows then, that because these  
693 conserved traits can be traced to the common ancestor of type I and type II, then the rates of  
694 evolution of PSII should have remained slow and constrained relative to that of other  
695 photosystems, since before the core duplications. These structural constraints demonstrate  
696 that the core subunits of PSII did not experience unprecedently faster rates of evolution. In  
697 fact, and not surprisingly, the anoxygenic type II RC which experienced the highest rates of  
698 evolution, has accumulated greater change and greater loss of ancestral traits.

699

#### 700 **4. Conclusion**

701 Both phylogenetic and structural evidence converge towards a scenario in which  
702 photosynthetic water-splitting started at an early stage during the evolution of life and long  
703 before the rise of what we understand today as Cyanobacteria. Nonetheless, greater resolution  
704 of divergence time estimation in deep-time evolutionary studies could be achieved if  
705 molecular clocks are complemented with experimentally validated rates of protein and  
706 genome evolution across taxa of interest.

707 We think it is plausible that there was never a discrete origin of photosynthesis, but  
708 that the process may trace back to abiogenic photochemical reactions, some of which may  
709 have resulted in the oxidation of water, and at the interface of nascent membranes, membrane  
710 proteins, photoactive tetrapyrroles and other inorganic cofactors: much in the same way that  
711 ribosomes may have originated at the interface of nascent genetics and protein synthesis [74].  
712 A photosynthetic origin of life is not a new idea [75-77] and abiotic photosynthesis-like  
713 chemistry has been recently proposed to have occurred at Gale Crater on Mars [78], and to  
714 occur even on Earth [79].

715

#### 716 **5. Materials and methods**

##### 717 *5.1. Sequence alignments and phylogenetic analysis*

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

723           A second dataset was retrieved on the 10<sup>th</sup> of October 2018 from the same database.  
724    This dataset focused on CP43 and CP47 subunits and consisted of a total of 1232 sequences.  
725    It included 392 CP43, 465 CBP proteins, and 375 CP47 subunits. 40 CP43 and 40 CP47  
726    sequences from a diverse set of photosynthetic eukaryotes were selected manually and  
727    included. In addition, a selection of cyanobacterial and plastid CP43, CP47, AtpA (Alpha),  
728    AtpB (Beta), and FtsH were manually selected for an analysis of the rates of evolution as  
729    illustrated in Supplementary Figure S14 and described below.

730           A dataset of Alpha and Beta subunits belonging to cyanobacterial F-type ATP  
731    synthase were retrieved from the NCBI refseq on the 31<sup>st</sup> of August 2019. 507 and 529  
732    cyanobacterial Alpha and Beta sequences were obtained respectively. We retrieved Alpha  
733    and Beta homologous for Margulisbacteria (19 and 18 sequences respectively),  
734    Sericytochromatia (4 and 2), and Vampirovibronia (66 and 49) using AnnoTree [80] and  
735    searching with the KEGG codes K02111 (F-type Alpha) and K02112 (F-type Beta). A total  
736    of 111 AtpA (subunit A) and 176 AtpB (subunit B) from archaeal V-type ATP synthase were  
737    retrieved using the sequences from *Methanocaldococcus jannaschii* as BLAST queries.

738           A dataset of 366 bacterial RNA polymerase subunit β (RpoB) were collected from the  
739    NCBI refseq database on the same date as above. We focused on Cyanobacteria (65  
740    sequences) and other phyla known to contain phototrophic representatives, as well as for their  
741    potential for allocating calibration points as described below. These included sequences from  
742    Firmicutes (32), Chloroflexi (32), Proteobacteria (102), Acidobacteria (34), Chlorobi (26),  
743    Gemmatimonadetes (3), Aquifecae (12) and Thermotogae (24). Sequences for *Candidatus*  
744    Eremiobacterota (2) recently reported to include phototrophic representatives [81],  
745    Margulisbacteria (13), Sericytochromatia (2), and Vampirovibronia (11) were obtained from  
746    AnnoTree using KEGG code K03043 as query. In addition, three extra margulisbacterial  
747    RpoB sequences were collected: from *Termititenax persephonae* and *Termititenax aidoneus*  
748    retrieved from the refseq database, and from margulisbacterium *Candidatus Ruthmannia*  
749    eludens obtained from <https://www.ebi.ac.uk/ena/data/view/PRJEB30343> [82]. A second  
750    dataset containing only sequences from Cyanobacteria and MSV, in addition to sequences  
751    from Archaea was also used. To retrieve the archaeal sequences the subunit B' (RpoB1) of  
752    *Methanocaldococcus jannaschii* was used as a BLAST query. A total 213 sequences across  
753    the entire diversity of Archaea were obtained. Two version of these were observed, those  
754    with split B (RpoB1 and RpoB2) and those with full-length B subunits: from the latter, 112  
755    full-length B subunits were selected for molecular clock analysis.

756 All sequences were aligned with Clustal Omega using 5 combined guided trees and  
757 Hidden Markov Model Iterations [83]. Given that RpoB sequences are known to contain  
758 many clade specific indels [27], we further processed this particular dataset using Gblocks  
759 [84] to remove these indels and poorly aligned positions allowing smaller final blocks, gap  
760 positions within the final blocks, and less strict flanking positions. This procedure left a total  
761 of 903 well-aligned sites for the dataset with only bacterial sequences and 788 for the dataset  
762 with archaeal sequences.

763 Maximum Likelihood phylogenetic analysis was performed with the PhyML online  
764 service using the Smart Model Selection mode implementing the Bayesian Information  
765 Criterion for parameter selection [85, 86]. Tree searching operations were calculated with the  
766 Nearest Neighbour Interchange model. Support was computed using the average likelihood  
767 ratio test [87]. Trees were visualised with Dendroscope V3.5.9 [88].

768

### 769 *5.2. Distance estimation*

770 Distance estimation was performed as a straightforward and intuitive approach to detect  
771 variations in the rate of evolution across taxa with known divergence time of at similar  
772 taxonomic ranks. This was done in several ways. Distance trees were plotted using BioNJ  
773 [89] as implemented in Seaview V4.7 [90] using observed distances and 100 bootstrap  
774 replicates. Within-group mean distances were calculated using three different substitution  
775 models as implemented in the package MEGA-X [91]: no. of differences, a Poisson model,  
776 and a JTT model. A gamma distribution to model rates among sites was used with a  
777 parameter of 1.00 and 500 bootstrap replicates.

778 To compare changes in the rates of evolution as a function of time of key proteins  
779 used in this study, we compared the percentage of sequence identity between orthologues in a  
780 total of 20 pairs of sequences from representative photosynthetic eukaryotes and  
781 Cyanobacteria with known or approximate divergence times. These are listed in  
782 Supplementary Table S3. We compared CP43, CP47, Alpha, Beta and FtsH. Of these 17  
783 pairs, the first 15 were comparisons between land plants. The approximated divergence times  
784 were based on those recommended by Clarke, et al. [92] after their extensive review of the  
785 plant fossil record. These were taken as the average of the hard minimum and soft maximum  
786 fossil ages suggested by the authors. The soft maximum age for the earliest land plants was  
787 taken as 515 Ma as discussed extensively in Morris, et al. [93]

788 Another sequence comparison was made between the unicellular red alga  
789 *Cyanidioschyzon merolae* and *Arabidopsis thaliana*. The earliest well-accepted fossil

790 evidence for red algae is that of the multicellular *Bangiomorpha* [94, 95] dated to about 1.0  
791 Ga [96]. Recently fossils of multicellular eukaryotic algae have been reported at 1.6 Ga [97-  
792 99]. The distance between heterocystous Cyanobacteria and their closest non-heterocystous  
793 relatives is represented by a pairwise comparison between *Chroococcidiopsis thermalis* and  
794 *Nostoc* sp. PCC 7120. Excellently preserved heterocystous Cyanobacteria have been reported  
795 from Tonian deposits older than 0.72 Ga [100]. The largest distance is that between  
796 *Gloeobacter* spp. and any other cyanobacterium or photosynthetic eukaryote, representing the  
797 distance since the MRCA of Cyanobacteria of uncertain divergence time.

798 To compare the level of sequence identity between RC proteins, two datasets of 10  
799 random amino acid sequences were generated using the Sequence Manipulation Suit [101].  
800 The datasets contained sequences of 350 and 750 residues. These were independently aligned  
801 as described above, resulting in 45 pairwise sequence identity comparisons for each dataset.  
802 These random sequence datasets were used as a rough minimum threshold of identity.  
803 Alignments of RC proteins were generated using three representative sequences spanning  
804 known diversity. Cyanobacterial CP43, CP47, standard D1 and D2 sequences were from  
805 *Gloeobacter violaceus*, *Stanieria cyanosphaera*, and *Nostoc* sp. PCC 7120; Heliobacterial  
806 PshA from *Helio bacterium modesticaldum*, *Helio bacillus mobilis*, and *Helio restis convoluta*;  
807 PscA from *Chlorobium tepidum*, *Prosthecochloris aestuarii*, *Chlorobium* sp. GBchlB,  
808 *Chloracidobacterium thermophilum* and *Chloracidobacterium* sp. CP2-5A; Protoebacterial L  
809 and M from *Methylobacterium* sp. 88A, *Roseivivax halotolerans*, and *Blastochloris*  
810 *sulfoviridis*; and L and M from *Chloroflexus* sp. Y-400-fl, *Roseiflexus castenholzii*, and  
811 *Oscillochloris trichoides*.

812  
813 *5.3. Quantification of rates of evolution (rationale)*  
814 Molecular clocks are conventionally used to estimate divergence times. In general terms,  
815 given: 1) a tree topology, which sets the relationship between taxa; 2) a sequence alignment,  
816 which sets the phylogenetic distance between taxa; and 3) some known events (calibrations),  
817 which set the rates of evolution, the molecular clock can then estimate divergence times. This  
818 means that if the tree topology and divergence times for two sets of protein sequences are the  
819 same, any differences in phylogenetic distances between these two should only reflect  
820 differences in the rate of evolution. Thus, assuming that CP43/CP47 and Alpha/Beta have  
821 mainly been inherited vertically in Cyanobacteria and photosynthetic eukaryotes, any  
822 difference in phylogenetic distance between the two is the result of differences in the rates of  
823 evolution. For example, the level of sequence identity between CP43 in *Cyanidioschizon* and

824 *Arabidopsis* is 78%, and the level of sequence identity between Alpha in the same species is  
825 69%. Given that these plastid-encoded subunits have mostly been inherited vertically since  
826 the MRCA of Archaeplastida, then one can argue that Alpha is evolving somewhat faster  
827 than CP43. This is because faster rates of protein evolution should lead to faster rates of  
828 change resulting in a faster decrease in the level of sequence identity (increase in  
829 phylogenetic distance). Now, because CP43 and CP47 are paralogues, we can then use this  
830 approach to estimate the rates of evolution at the moment of duplication under a given  
831 number of specific scenarios. For example, assuming that the MRCA of Cyanobacteria  
832 occurred at 2.0 Ga, we can then model how the rates of evolution would change at the  
833 moment of duplication if this occurred at 2.5, 3.0, 3.5 Ga, or at any other time point of  
834 interest. Because these set of proteins have likely achieved mutational saturation at the largest  
835 distance, the rate of evolution at the point of duplication will be an underestimation (slower  
836 than it should be), given that saturation would hide substitution events. Similarly, keeping the  
837 time of duplication constant, we can then model how the rates of evolution would change  
838 across the tree if the MRCA of Cyanobacteria is assumed to have occurred at 2.0, 2.5, 3.0 Ga,  
839 or at any other time of interest. Therefore, even though it is difficult to determine the absolute  
840 time of origin of Cyanobacteria, or of the early duplications of the core subunits of ATP  
841 synthase, it is possible to determine what rates of evolution are required to fulfil any  
842 particular scenario.

843

#### 844 *5.4. Quantification of rates of evolution*

845 To measure rates of evolution of CP43/CP47 and Alpha/Beta a total of 19 sequences from  
846 photosynthetic eukaryotes and 4 cyanobacterial sequences per subunit were selected. A  
847 standardized tree topology was constructed from consensus evolutionary relationships as  
848 illustrated in Supplementary Figure S14a: 1) Relationships between land plants was taken  
849 from ref. [92] 2) It is well established that the divergence of red algae predates the MRCA of  
850 land plants. 3) Ponce-Toledo, Deschamps [102] recently suggested that *Gloeomargarita* is  
851 the closest living cyanobacterial relative to the plastid ancestor and predicated the emergence  
852 of heterocystous cyanobacteria, but see also [103]. 4) The clade containing  
853 *Chroococcidiopsis thermalis* PCC 7203 is one of heterocystous Cyanobacteria's closest non-  
854 heterocystous relatives [104, 105]. 5) *Gloeobacter* spp. is the earliest branching and well-  
855 described genus of Cyanobacteria capable of oxygenic photosynthesis [105-109].

856 Calibration points were allocated as shown in Supplementary Figure S14a and listed  
857 in Table 2. Nodes 1 to 15 were applied following the justifications and best practice listed in

858 ref. [92] and using the 515 Ma older calibration as suggested in ref. [93] Node 16 represents  
859 here the MRCA of red and green lineages of photosynthetic eukaryotes. The minimum age  
860 was set to 1.0 Ga based on the *Bangiomorpha* fossil as described above [94, 95]. The  
861 maximum age was set to 1.8 Ga, which is similar the oldest ages reported in recent molecular  
862 clock analyses for the MRCA of photosynthetic eukaryotes and it is also similar to the age of  
863 the earliest plausible fossilised unicellular eukaryotes [110, 111]. That said, compelling fossil  
864 evidence for eukaryotes could go as far back as 2.1 Ga [112]. Node 17 marks the divergence  
865 of heterocystous Cyanobacteria and it was given a minimum age of 0.72 Ga based on the  
866 recently described fossils of filaments bearing clear heterocysts from the Tonian period [100].  
867 A maximum age of 1.65 Ga was given to this node, based on the recent report of  
868 heterocystous cyanobacteria from the Gaoyuzhuang Formation [113]. Although, it is unclear  
869 if reported akinetes that are older than 1.65 Ga are truly that [10], we considered that the  
870 0.72-1.65 Ma range for the origin of heterocystous Cyanobacteria was a reasonably broad  
871 constraint for the purpose of this experiment, but see below.

872 Node 18 denotes the MRCA of Cyanobacteria. The age for this node is highly debated  
873 ranging from before to after the GOE. Node 19 represents the duplication events leading to  
874 CP43 and CP47, and to Alpha and Beta. To calculate the rates of evolution under different  
875 scenarios, node 18 and node 19 were varied. Firstly, a molecular clock was run using a  
876 scenario that assumed that the MRCA of Cyanobacteria postdated the GOE. To do this, node  
877 18 was set to be between 1.6 and 1.8 Ga, which emulates results reported in recent studies [8,  
878 11]. This was compared to a scenario that assumed that the MRCA of Cyanobacteria  
879 antedated the GOE, and thus node 18 was set to be between 2.6 and 2.8 Ga, which simulates  
880 other evolutionary scenarios [103, 114]. In both cases, the duplication event (node 19) was  
881 set to be 3.5 Ga old, or changed as stated in the main text, by assigning a gamma prior at the  
882 desired time fixed with a narrow standard deviation of 0.05 Ga. In a separate experiment, the  
883 age of the duplication was varied while maintaining node 18 restricted to between 1.6 and 1.8  
884 Ga, while node 19, the root, was set with a gamma prior with an average varied from 0.8 to  
885 4.2 and with a narrow standard deviation of 0.05 Ga.

886 The period of time between the duplication event (node 19), which led to the  
887 divergence of CP43 and CP47, and the MRCA of Cyanobacteria (node 18), we define as  $\Delta T$ .  
888  $\Delta T$  is calculated as the subtraction of the mean age of node 19 and node 18. For PSII, we  
889 used node 18 from the CP43 subunit and for ATP synthase we used node 18 from the Alpha  
890 subunit. In consequence, varying the age of the duplication from 0.8 to 4.2 Ga allows changes  
891 in the rate of evolution to be simulated with varying  $\Delta T$ , ranging from 0.2 to 2.5 Ga. That is

892 to say, it allows the rate of evolution at the point of duplication to be calculated if it occurred  
893 at any point in time before the MRCA of Cyanobacteria.

894

895 **Table 2.** Calibration points used in this study

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

896 Molecular clocks were calculated with Phylobayes 3.3f using a log normal autocorrelated  
897 molecular clock model under the CAT+Γ non-parametric model of amino acid substitution

898 and a uniform distribution of equilibrium frequencies. We preferred a CAT model instead of  
899 a CAT+GTR as the latter only outperforms the former on sequence alignments of over 1000  
900 sites and it is also much less computationally expensive. Four discrete categories for the  
901 gamma distribution were used and four chains were executed in parallel through all  
902 experiments carried out in this study. The instant rates of evolution, which are the rates at  
903 each internal node in the tree, were retrieved from the output files of Phylobayes. These rates  
904 are calculated by the software as described by the developers elsewhere [115, 116] and are  
905 expressed as amino acid changes per site per unit of time.

906 As an additional comparison, we calculated rates of evolution for cyanobacterial FtsH  
907 subunits, which AAA ATPase domain is structurally similar to the catalytic head of ATP  
908 synthase [117]. Cyanobacterial FtsH are involved in membrane protein quality control, with  
909 some isoforms targeting specifically PSII subunits. It features a late duplication event leading  
910 to cyanobacterial FtsH1 and FtsH2 subunits (using the nomenclature of Shao, et al. [40]). It  
911 was shown previously that the duplication leading to FtsH1 and FtsH2 occurred after the  
912 divergence of *Gloeobacter* spp. Genes encoding plastid FtsH subunits are encoded in the  
913 nuclear genome of photosynthetic eukaryotes. Because not all the selected species had fully  
914 sequenced nuclear genomes, only those with available FtsH sequences at the time of this  
915 study were used. The tree topology shown in Supplementary Figure S14b was used as  
916 template for the calculation of the rate of evolution and was based on the topology presented  
917 by Shao, et al. [40] It was concluded that the Cyanobacteria-inherited closest paralog to  
918 FtsH1 and FtsH2 in photosynthetic eukaryotes was also acquired before their initial  
919 duplication. Therefore, from all FtsH paralogs in photosynthetic eukaryote genomes, those  
920 with greater sequence identity to cyanobacterial FtsH1/2 were used. Because this duplication  
921 is specific to Cyanobacteria, a few additional strains were included in this tree following  
922 well-established topologies [103, 105]. Calibrations were placed as indicated in  
923 Supplementary Figure S14b. To test the change in the rate of evolution at the time of  
924 duplication in comparison with CP43/CP47, node 19 was set to 1.6-1.8 Ga or 2.6-2.8 Ga and  
925 molecular clocks were run as described in the preceding paragraph.

926 Finally, we conducted a large molecular clock using the combined 897 CP43 and CBP  
927 sequences, including 40 eukaryotic CP43 sequences, to test whether using a more complex  
928 phylogeny would result in rates of evolution substantially different to those calculated with  
929 the method described above. Calibrations were assigned as illustrated in Supplementary  
930 Figure S15. Cross-calibrations were used across paralogs constraining the origin of  
931 heterocystous Cyanobacteria. In this case, only the minimum constraint of 0.72 Ga was used

932 with no maximum constraint to allow greater flexibility. Additional calibrations were  
933 assigned also across paralogs (point 20 in Supplementary Figure S15), this was considered as  
934 the node made by *Richelia intracellularis* and its closest sister sequence, as implemented in  
935 ref. [114] This strain is a specific endosymbiont of a diatom and its divergence was set to be  
936 no older than the earliest discussed age for diatoms [118]. The root equivalent to the MRCA  
937 of Cyanobacteria (divergence of *Gloeobacter* in CP43) was not calibrated. The root of the  
938 tree was varied: first it was given a maximum age of 4.52 Ga as recently implemented and  
939 justified by Betts, et al. [11] as the earliest plausible time in which the planet was inhabitable  
940 after the moon forming impact [119], and no minimum age was used. A second tree was  
941 executed with no constraint on the root and no root prior. A third root was implemented  
942 constrained to be between 2.3 Ga (the GOE) and 3.2 Ga. The latter date represents the age of  
943 the cyanobacteria-like well-preserved microbial mats of the Berberton Greenstone Belt in  
944 South Africa and neighbouring Eswatini [66]. Rates were obtained using the autocorrelated  
945 CAT+ $\Gamma$  model as described above. Because these root constraints did not have a strong effect  
946 in the overall estimated rates, we carried out an additional control applying an uncorrelated  
947 gamma clock model [120] with a root constrained at 4.52 Ga and no minimum age.

948

#### 949 5.5. Molecular clock of *RpoB* and concatenated ribosomal proteins

950 The primary objective of this experiment was not to determine the absolute time of origin of  
951 Cyanobacteria, but to understand the spans of time between Cyanobacteria and their relatives.  
952 We also wanted to understand what rates of evolution are associated with those spans of time  
953 and how these change under different evolutionary scenarios. To do this, we applied a  
954 molecular clock to the phylogeny of *RpoB* sequences described above. We implemented 12  
955 calibrations. The calibrations were assigned on the phylogeny as shown in Supplementary  
956 Figure S16 and listed in Table 2. A set of calibrations consisted of the earliest unambiguous  
957 evidence for Chroococcales Cyanobacteria of the Belcher group (point 21), the age of which  
958 has been recently revisited to 2.01 Ga [121]. This was assigned to the younger node from  
959 where Chroococcales strains branch out in the tree, with no maximum restrictions. The  
960 appearance of heterocystous Cyanobacteria were restricted from 0.72 Ga and 1.56 Ga as  
961 described above. No constraints on the node representing the MRCA of Cyanobacteria were  
962 used. However, for rigor, we also tested an alternative single calibration on the node  
963 representing the MRCA of Cyanobacteria with a maximum of 2.01 Ga and no minimum, and  
964 with no other calibrations in the clade. This considered a scenario in which crown group  
965 Cyanobacteria are younger than the Belcher fossils.

966 In addition to cyanobacterial calibrations, we also applied the often-used biomarker  
967 evidence for phototrophic Chlorobi and Chromatiaceae at 1.64 Ga [122], see for example  
968 refs. [9, 63] These were used as a minimum with no maximum constraints (node 30 and 31  
969 respectively in Supplementary Figure S16).

970 A set of calibrations were chosen from well-described symbiotic relationships. Two of  
971 these are known in the phylum Margulisbacteria. The first one is *Termititenax*, these are  
972 specific ectosymbionts of spirochetes that live within oxymonad protists in the gut of diverse  
973 termites and cockroaches [43]. The two *Termititenax* sequences used in this analysis  
974 clustered together and therefore we calibrated this node to be between 396 Ma, some of the  
975 oldest fossil evidence of insects [123] and 140 Ma for *Mastotermes nepropadyom*, a Jurassic  
976 fossil termite [124] (node 22). This was done under the assumption that this symbiotic  
977 relationship may have started before the divergence of termites and cockroaches, as  
978 suggested in ref. [43] The second one is *Candidatus Ruthmannia eludens*, a cell-type specific  
979 endosymbiont of placozoans, early-branching metazoans. This symbiont has been detected in  
980 all haplotypes examined, regardless of geographical location or sampling time [82, 125],  
981 which suggest that this association may be as old as placozoans. Therefore, we used a  
982 minimum calibration of 542 Ma as the Cambrian explosion of animal diversity and 660 Ma,  
983 the earliest biomarker evidence for desmosponges [126] (node 23), which should antedate the  
984 divergence of placozoans. This calibration was assigned to the node separating *Candidatus*  
985 *Ruthmannia eludens* from its closest sister sequence.

986 Similar to Margulisbacteria, members of the clade Vampirovibrionia have been  
987 reported to form close associations with eukaryotes. The clade Gastranaerophilales is thought  
988 to be composed mostly of strains that inhabit the animal gut. Thus, we calibrated the node  
989 separating two strains isolated from human and koala faeces that clustered together with a  
990 maximum age of 201 Ma, representing the Jurassic split of marsupials and placental  
991 mammals [127] (node 24a). However, the koala and human sequences were embedded in the  
992 Gastranaerophilales clade within other sequences from the human gut. Because of this, we  
993 trialled changing this calibration to 55 Ma instead, the oldest primate fossil [128] and  
994 assuming that the retrieved sequences from the human gut had a common ancestor younger  
995 than the MRCA of primates. Alternatively, we tested moving this calibration to the ancestral  
996 nodes of the clade that included all the human gut sequences (node 24b). Gastranaerophilales  
997 is closely related to the order Vampirovibrionales, which include *Vampirovibrio*  
998 *chlorellavorus*. This strain is a predator of the eukaryotic green algae *Chlorella* [129], and  
999 therefore we trialled a calibration assuming that Gastranaerophilales and Vampirovibrionales

1000 radiated after the MRCA of eukaryotes (node 25). We thus assigned a maximum calibration  
1001 to this node of 1.8 Ga representing the earliest described plausible eukaryote fossils [110] and  
1002 no minimum age.

1003 Another highly specific obligate symbiosis is that of the betaproteobacterium  
1004 *Polynucleobacter necessarius* and ciliates of the genus *Euplotes* (Spirotrichea) [130].  
1005 *Polynucleobacter* has close free-living phototrophic relatives within the same genus [130].  
1006 We set the node separating the phototrophic and non-phototrophic *Polynucleobacter* (node  
1007 26) a maximum age of 444 Ma for the oldest fossil evidence of spirotrichs, as implemented in  
1008 Parfrey et al. [131], and which predates the radiation of the genus *Euplotes* [132].

1009 Another well-known association is that of the soil bacteria *Bradyrhizobium* and  
1010 legumes. Thus we gave the node separating *Bradyrhizobium* spp. from its closest relative in  
1011 the RpoB tree, *Xanthobacter autotrophicus*, a maximum age of 86 Ma for Rosids, which  
1012 contain legumes [93] (node 27).

1013 The Rickettsiales are Alphaproteobacteria that exists in very close association with  
1014 eukaryotes [133]. An association that may reach to the lineage leading to the origin of  
1015 mitochondria [134]. Therefore, we assumed that the divergence of Rickettsiales occurred  
1016 before the MRCA of eukaryotes and gave this node a minimum age of 1.8 Ga [110] (node  
1017 28). Finally, the family Anaplasmataceae contains bacteria that exists in close association  
1018 with insects as endosymbionts (e.g. *Wolbachia*) or as parasite vectors (e.g. *Anaplasma*).  
1019 Therefore, we set a maximum constraint for the MRCA of *Wolbachia* and *Anaplasma* (node  
1020 29), excluding *Neorickettsia*, to be as old as the earliest evidence for insects about 395 Ma  
1021 ago [123].

1022 To constrain the age of the root, we first set a broad gamma prior with an average of  
1023 3.8 Ga and a standard deviation of 0.5 Ga. We found this to perform well and used it as  
1024 benchmark to compare with a range of evolutionary models and the effects of key  
1025 calibrations (Supplementary Figure S8). Alternatively, we applied a broad calibration on the  
1026 root with a maximum of 4.52 Ga as described above and a minimum of 3.41 Ga, which is the  
1027 earliest well-accepted evidence for photosynthesis [47]. This evidence was hypothesized to  
1028 be anoxygenic in ref. [47] Therefore, Bacteria should be at least as old as the earliest  
1029 evidence for photosynthesis under conventional evolutionary scenarios.

1030 To further understand the effect that a distant outgroup would have on the estimated  
1031 divergence time, and to measure the rate of evolution of RpoB at the divergence of Bacteria  
1032 and Archaea, we repeated the clock including 112 diverse archaeal sequences, in addition to  
1033 Thermotogae, MSV, and Cyanobacteria, but removing all other clades as a compromise

1034 between robustness and computing time. Calibrations were assigned on MSV and  
1035 Cyanobacteria as described above. Relaxed molecular clocks were computed using an  
1036 autocorrelated log normal clock and applying a CAT+ $\Gamma$  model. This was compared against a  
1037 CAT+ $\Gamma$  and a CAT+GTR+ $\Gamma$  using birth-death priors and soft bounds on the calibrations  
1038 allowing for 2.5% tail probability falling outside the minimum and maximum boundary, or  
1039 5% in the case of a single boundary. In addition, we also compared to an uncorrelated gamma  
1040 model (Supplementary Figure S8).

1041 We compared the RpoB molecular clock (CAT+ $\Gamma$ ) with that of a clock executed using  
1042 a dataset of concatenated ribosomal proteins of 2596 aligned sites, obtained from an  
1043 independent study [38]. In total, 157 sequences were used including 56 cyanobacterial  
1044 sequences, 14 sequences from Vampirovibrionia, 9 from Margulisbacteria, 9 from  
1045 Thermotogae and 78 from Archaea. A set of calibrations were used including those two in  
1046 Cyanobacteria (point 17 and 21), a calibration on the MRCA of Gastranaerophilales with a  
1047 maximum 201 Ma (point 24b), and a root calibration between 4.52 and 3.41 Ga as described  
1048 above.

1049

### 1050 5.6 Ancestral sequence reconstruction

1051 Ancestral sequence reconstruction (ASR) of D1 and D2 amino acid sequences was carried  
1052 out with a dataset collected on the 17<sup>th</sup> of November 2017. Duplicates and partial sequences  
1053 were removed, leaving 755 D1 and 248 D2 sequences. CD-HIT [135] was used to remove  
1054 sequences with greater than 92% sequence identity to create a representative sample. The L  
1055 and M RC subunits from 5 strains of Proteobacteria were used as outgroup. The final  
1056 alignment did not include the atypical variant D1 sequence from *Gloeobacter kilaueensis*  
1057 (NCBI accession AGY58976.1) as this showed an unstable phylogenetic position in this  
1058 dataset. Maximum likelihood trees used as input for ASR were computed with PhyML using  
1059 Smart Model Selection [86]. The LG substitution model with observed amino acid  
1060 frequencies and four gamma rate categories exhibited the best log likelihood (LG+ $\Gamma$ +F)  
1061 (Supplementary Table S7). We used the top four models for tree reconstruction (LG+ $\Gamma$ +F,  
1062 LG+ $\Gamma$ +I+F, LG+ $\Gamma$  and LG+ $\Gamma$ +I) in addition to another tree computed using the WAG  
1063 substitution model with observed amino acid frequencies and four gamma rate categories  
1064 (WAG+ $\Gamma$ +F). These trees were used as input trees to calculate maximum likelihood ancestral  
1065 states at each site for the node corresponding to the homodimeric reaction protein, D0. Three  
1066 ASR programs were used for the reconstructions: FastML [136], Lazarus [137] (a set of  
1067 Python scripts which wraps PAML [138]) and MEGA-7 [139]. The substitution model used

1068 by all three programs corresponded to the substitution model used for the specific input tree  
1069 in PhyML and the branch lengths of the trees were fixed. In FastML, a maximum likelihood  
1070 method of indel reconstruction using a probability cut-off of 0.7 was used. In MEGA-7, all  
1071 sites were used for analysis with no branch swap filter. In Lazarus, the ‘—gapcorrect’ option  
1072 was used to parsimoniously place indels.

1073

#### 1074 *5.7. Structural analysis*

1075 The following crystal structures were used in this work: the crystal structure of PSII from  
1076 *Thermosynechococcus vulcanus*, PDB ID: 3wu2 [13]; the anoxygenic type II RC of  
1077 *Thermochromatium tepidum*, PDB ID: 5y5s [140]; the homodimeric type I RC from  
1078 *Helio bacterium modesticaldum*, 8v5k [141]; PSI from *Thermosynechococcus elongatus*, PDB  
1079 ID: 1jb0 [142]; and the cryo-EM IsiA structure from *Synechocystis* sp. PCC 6803 [143].  
1080 Structures were visualised using Pymol™ V. 1.8.2.2 (Schrodinger, LLC) and structural  
1081 overlaps were carried out with the CEAlign plugin [144].

1082

#### 1083 **6. Acknowledgements**

1084 This work was done with the support of a Leverhulme Trust grant (RPG-2017-223 to T.C.  
1085 and A.W.R.), a UKRI Future Leaders Fellowship (MR/T017546/1 to T.C.) a Royal Society  
1086 University Research Fellowship (P.S.-B.), and the Biotechnology and Biological Sciences  
1087 Research Council (BB/K002627/1 and BB/L011206/1 to A.W.R.). The authors are grateful to  
1088 Dr. Travis J. Lawrence and Prof. Christoph Heubeck for feedback on a version of this  
1089 manuscript. Computing resources were provided by the HPC facility at Imperial College  
1090 London. The authors declare no conflict of interests.

1091

#### 1092 **7. References**

1. Sánchez-Baracaldo, P. and T. Cardona, *On the origin of oxygenic photosynthesis and Cyanobacteria*. New Phytol, 2020. **225**: 1440-1446. DOI: 10.1111/nph.16249.
2. Cardona, T., J.W. Murray, and A.W. Rutherford, *Origin and evolution of water oxidation before the last common ancestor of the Cyanobacteria*. Mol. Biol. Evol., 2015. **32**: 1310-1328. DOI: 10.1093/molbev/msv024.
3. Di Rienzi, S.C., I. Sharon, K.C. Wrighton, O. Koren, L.A. Hug, B.C. Thomas, J.K. Goodrich, J.T. Bell, T.D. Spector, J.F. Banfield, and R.E. Ley, *The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria*. Elife, 2013. **2**: e01102. DOI: 10.7554/elife.01102.
4. Soo, R.M., J. Hemp, and P. Hugenholtz, *Evolution of photosynthesis and aerobic respiration in the cyanobacteria*. Free Radical Biol. Med., 2019. **140**: 200-205. DOI: 10.1016/j.freeradbiomed.2019.03.029.

1105 5. Soo, R.M., J. Hemp, D.H. Parks, W.W. Fischer, and P. Hugenholtz, *On the origins of*  
1106 *oxygenic photosynthesis and aerobic respiration in Cyanobacteria*. *Science*, 2017.  
1107 **355**: 1436-1440. DOI: 10.1126/science.aal3794.

1108 6. Anantharaman, K., C.T. Brown, L.A. Hug, I. Sharon, C.J. Castelle, A.J. Probst, B.C.  
1109 Thomas, A. Singh, M.J. Wilkins, U. Karaoz, E.L. Brodie, K.H. Williams, S.S.  
1110 Hubbard, and J.F. Banfield, *Thousands of microbial genomes shed light on*  
1111 *interconnected biogeochemical processes in an aquifer system*. *Nat. Commun.*, 2016.  
1112 **7**: 13219. DOI: 10.1038/ncomms13219.

1113 7. Carnevali, P.B.M., F. Schulz, C.J. Castelle, R.S. Kantor, P.M. Shih, I. Sharon, J.M.  
1114 Santini, M.R. Olm, Y. Amano, B.C. Thomas, K. Anantharaman, D. Burstein, E.D.  
1115 Becroft, R. Stepanauskas, T. Woyke, and J.F. Banfield, *Hydrogen-based metabolism*  
1116 *as an ancestral trait in lineages sibling to the Cyanobacteria*. *Nat. Commun.*, 2019.  
1117 **10**: 463. DOI: 10.1038/s41467-018-08246-y.

1118 8. Shih, P.M., J. Hemp, L.M. Ward, N.J. Matzke, and W.W. Fischer, *Crown group*  
1119 *Oxyphotobacteria postdate the rise of oxygen*. *Geobiology*, 2017. **15**: 19-29. DOI:  
1120 10.1111/gbi.12200.

1121 9. Magnabosco, C., K.R. Moore, J.M. Wolfe, and G.P. Fournier, *Dating phototrophic*  
1122 *microbial lineages with reticulate gene histories*. *Geobiology*, 2018. **16**: 179-189.  
1123 DOI: 10.1111/gbi.12273.

1124 10. Schirrmeister, B.E., P. Sánchez-Baracaldo, and D. Wacey, *Cyanobacterial evolution*  
1125 *during the Precambrian*. *Int. J. Astrobiol.*, 2016. **15**: 187-204. DOI:  
1126 10.1017/S1473550415000579.

1127 11. Betts, H.C., M.N. Puttik, J.W. Clark, T.A. Williams, P.C.J. Donoghue, and D. Pisani,  
1128 *Integrated genomic and fossil evidence illuminates life's early evolution and*  
1129 *eukaryote origin*. *Nat. Ecol. Evol.*, 2018. **2**: 1556-1562. DOI: 10.1038/s41559-018-  
1130 0644-x.

1131 12. Garcia-Pichel, F., J. Lombard, T. Soule, S. Dunaj, S.H. Wu, and M.F. Wojciechowski,  
1132 *Timing the Evolutionary Advent of Cyanobacteria and the Later Great Oxidation*  
1133 *Event Using Gene Phylogenies of a Sunscreen*. *Mbio*, 2019. **10**. DOI:  
1134 10.1128/mBio.00561-19.

1135 13. Umena, Y., K. Kawakami, J.R. Shen, and N. Kamiya, *Crystal structure of oxygen-*  
1136 *evolving Photosystem II at a resolution of 1.9 Å*. *Nature*, 2011. **473**: 55-60. DOI:  
1137 10.1038/nature09913.

1138 14. Ferreira, K.N., T.M. Iverson, K. Maghlaoui, J. Barber, and S. Iwata, *Architecture of*  
1139 *the photosynthetic oxygen-evolving center*. *Science*, 2004. **303**: 1831-1838. DOI:  
1140 10.1126/science.1093087.

1141 15. Cardona, T., A. Sedoud, N. Cox, and A.W. Rutherford, *Charge separation in*  
1142 *Photosystem II: A comparative and evolutionary overview*. *Biochim. Biophys. Acta*,  
1143 2012. **1817**: 26-43. DOI: 10.1016/j.bbabi.2011.07.012.

1144 16. Rutherford, A.W., T. Mattioli, and W. Nitschke, *The FeS-type photosystems and the*  
1145 *evolution of photosynthetic reaction centers*, in *Origin and evolution of biological*  
1146 *energy conversion*, H. Baltscheffsky, Editor. 1996, VCH: New York, N. Y. 177-203.

1147 17. Rutherford, A.W. and W. Nitschke, *Photosystem II and the quinone-iron-containing*  
1148 *reaction centers*, in *Origin and evolution of biological energy conversion*, H.  
1149 Baltscheffsky, Editor. 1996, VCH: New York, N. Y. 143-175.

1150 18. Cardona, T., P. Sánchez-Baracaldo, A.W. Rutherford, and A.W.D. Larkum, *Early*  
1151 *Archean origin of Photosystem II*. *Geobiology*, 2019. **17**: 127-150. DOI:  
1152 10.1111/gbi.12322.

1153 19. Cardona, T. and A.W. Rutherford, *Evolution of photochemical reaction centres: more*  
1154 *twists?* *Trends Plant Sci.*, 2019. **24**: 1008-1021. DOI: 10.1016/j.tplants.2019.06.016.

1155 20. Mulkidjanian, A.Y., K.S. Makarova, M.Y. Galperin, and E.V. Koonin, *Inventing the*  
1156 *dynamo machine: the evolution of the F-type and V-type ATPases*. *Nat. Rev.*  
1157 *Microbiol.*, 2007. **5**: 892-9. DOI: 10.1038/nrmicro1767.

1158 21. Gogarten, J.P., H. Kibak, P. Dittrich, L. Taiz, E.J. Bowman, B.J. Bowman, M.F.  
1159 Manolson, R.J. Poole, T. Date, T. Oshima, J. Konishi, K. Denda, and M. Yoshida,  
1160 *Evolution of the vacuolar H<sup>+</sup>-ATPase: Implications for the origin of eukaryotes*. *Proc.*  
1161 *Natl. Acad. Sci. U.S.A.*, 1989. **86**: 6661-6665. DOI: 10.1073/pnas.86.17.6661.

1162 22. Lane, N., J.F. Allen, and W. Martin, *How did LUCA make a living? Chemiosmosis in*  
1163 *the origin of life*. *Bioessays*, 2010. **32**: 271-280. DOI: 10.1002/bies.200900131.

1164 23. Ouzounis, C.A., V. Kunin, N. Darzentas, and L. Goldovsky, *A minimal estimate for*  
1165 *the gene content of the last universal common ancestor: Exobiology from a terrestrial*  
1166 *perspective*. *Res. Microbiol.*, 2006. **157**: 57-68. DOI: 10.1016/j.resmic.2005.06.015.

1167 24. Muller, V. and G. Gruber, *ATP synthases: structure, function and evolution of unique*  
1168 *energy converters*. *Cell. Mol. Life Sci.*, 2003. **60**: 474-494. DOI:  
1169 10.1007/s000180300040.

1170 25. Werner, F. and D. Grohmann, *Evolution of multisubunit RNA polymerases in the three*  
1171 *domains of life*. *Nat. Rev. Microbiol.*, 2011. **9**: 85-98. DOI: 10.1038/nrmicro2507.

1172 26. Kyrpides, N., R. Overbeek, and C. Ouzounis, *Universal protein families and the*  
1173 *functional content of the Last Universal Common Ancestor*. *J. Mol. Evol.*, 1999. **49**:  
1174 413-423. DOI: 10.1007/Pl00006564.

1175 27. Lane, W.J. and S.A. Darst, *Molecular evolution of multisubunit RNA polymerases:*  
1176 *sequence analysis*. *J. Mol. Biol.*, 2010. **395**: 671-85. DOI: 10.1016/j.jmb.2009.10.062.

1177 28. Harris, J.K., S.T. Kelley, G.B. Spiegelman, and N.R. Pace, *The genetic core of the*  
1178 *universal ancestor*. *Genome Res.*, 2003. **13**: 407-412. DOI: 10.1101/gr.652803.

1179 29. Chen, M., Y. Zhang, and R.E. Blankenship, *Nomenclature for membrane-bound light-*  
1180 *harvesting complexes of cyanobacteria*. *Photosynth. Res.*, 2008. **95**: 147-154. DOI:  
1181 10.1007/s11120-007-9255-0.

1182 30. Chen, M., R.G. Hiller, C.J. Howe, and A.W.D. Larkum, *Unique origin and lateral*  
1183 *transfer of prokaryotic chlorophyll-b and chlorophyll-d light-harvesting systems*. *Mol.*  
1184 *Biol. Evol.*, 2005. **22**: 21-28. DOI: 10.1093/molbev/msh250.

1185 31. Murray, J.W., *Sequence variation at the oxygen-evolving centre of Photosystem II: a*  
1186 *new class of 'rogue' cyanobacterial D1 proteins*. *Photosynth. Res.*, 2012. **110**: 177-84.  
1187 DOI: 10.1007/s11120-011-9714-5.

1188 32. Ho, M.Y., G. Shen, D.P. Canniffe, C. Zhao, and D.A. Bryant, *Light-dependent*  
1189 *chlorophyll f synthase is a highly divergent paralog of PsbA of Photosystem II*.  
1190 *Science*, 2016. **353**: aaf9178. DOI: 10.1126/science.aaf9178.

1191 33. Gan, F., G. Shen, and D.A. Bryant, *Occurrence of far-red light photoacclimation*  
1192 *(FarLiP) in diverse cyanobacteria*. *Life (Basel)*, 2014. **5**: 4-24. DOI:  
1193 10.3390/life5010004.

1194 34. Cardona, T., *A fresh look at the evolution and diversification of photochemical*  
1195 *reaction centers*. *Photosynth. Res.*, 2015. **126**: 111-134. DOI: 10.1007/s11120-014-  
1196 0065-x.

1197 35. Cardona, T., *Reconstructing the origin of oxygenic photosynthesis: Do assembly and*  
1198 *photoactivation recapitulate evolution?* *Front. Plant Sci.*, 2016. **7**: 257. DOI:  
1199 10.3389/fpls.2016.00257.

1200 36. Dibrova, D.V., M.Y. Galperin, and A.Y. Mulkidjanian, *Characterization of the N-*  
1201 *ATPase, a distinct, laterally transferred Na<sup>+</sup>-translocating form of the bacterial F-*  
1202 *type membrane ATPase*. *Bioinformatics*, 2010. **26**: 1473-1476. DOI:  
1203 10.1093/bioinformatics/btq234.

1204 37. Battistuzzi, F.U., A. Feijao, and S.B. Hedges, *A genomic timescale of prokaryote*  
1205 *evolution: insights into the origin of methanogenesis, phototrophy, and the*  
1206 *colonization of land*. *BMC Evol. Biol.*, 2004. **4**: 44. DOI: 10.1186/1471-2148-4-44.

1207 38. Hug, L.A., B.J. Baker, K. Anantharaman, C.T. Brown, A.J. Probst, C.J. Castelle, C.N.  
1208 Butterfield, A.W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D.A. Relman,  
1209 K.M. Finstad, R. Amundson, B.C. Thomas, and J.F. Banfield, *A new view of the tree*  
1210 *of life*. *Nat. Microbiol.*, 2016. **1**: 16048. DOI: 10.1038/nmicrobiol.2016.48.

1211 39. Parks, D.H., C. Rinke, M. Chuvochina, P.A. Chaumeil, B.J. Woodcroft, P.N. Evans,  
1212 P. Hugenholtz, and G.W. Tyson, *Recovery of nearly 8,000 metagenome-assembled*  
1213 *genomes substantially expands the tree of life*. *Nat. Microbiol.*, 2017. **2**: 1533-1542.  
1214 DOI: 10.1038/s41564-017-0012-7.

1215 40. Shao, S., T. Cardona, and P.J. Nixon, *Early emergence of the FtsH proteases involved*  
1216 *in photosystem II repair*. *Photosynthetica*, 2018. **56**: 163-177. DOI: 10.1007/s11099-  
1217 018-0769-9.

1218 41. Innan, H. and F. Kondrashov, *The evolution of gene duplications: classifying and*  
1219 *distinguishing between models*. *Nature Reviews Genetics*, 2010. **11**: 97-108. DOI:  
1220 10.1038/nrg2689.

1221 42. Lynch, M. and J.S. Conery, *The evolutionary fate and consequences of duplicate*  
1222 *genes*. *Science*, 2000. **290**: 1151-1155. DOI: 10.1126/science.290.5494.1151.

1223 43. Utami, Y.D., H. Kuwahara, K. Igai, T. Murakami, K. Sugaya, T. Morikawa, Y.  
1224 Nagura, M. Yuki, P. Deevong, T. Inoue, K. Kihara, N. Lo, A. Yamada, M. Ohkuma,  
1225 and Y. Hongoh, *Genome analyses of uncultured TG2/ZB3 bacteria in*  
1226 *'Margulisbacteria' specifically attached to ectosymbiotic spirochetes of protists in the*  
1227 *termite gut*. *ISME J.*, 2019. **13**: 455-467. DOI: 10.1038/s41396-018-0297-4.

1228 44. Kumar, S., G. Stecher, M. Suleski, and S.B. Hedges, *TimeTree: A resource for*  
1229 *timelines, timetrees, and divergence times*. *Mol. Biol. Evol.*, 2017. **34**: 1812-1819.  
1230 DOI: 10.1093/molbev/msx116.

1231 45. Marin, J., F.U. Battistuzzi, A.C. Brown, and S.B. Hedges, *The timetree of*  
1232 *prokaryotes: New insights into their evolution and speciation*. *Mol. Biol. Evol.*, 2017.  
1233 **34**: 437-446. DOI: 10.1093/molbev/msw245.

1234 46. Shih, P.M., L.M. Ward, and W.W. Fischer, *Evolution of the 3-hydroxypropionate*  
1235 *bicycle and recent transfer of anoxygenic photosynthesis into the Chloroflexi*. *Proc.*  
1236 *Natl. Acad. Sci. U.S.A.*, 2017. **114**: 10749-10754. DOI: 10.1073/pnas.1710798114.

1237 47. Tice, M.M. and D.R. Lowe, *Photosynthetic microbial mats in the 3,416-Myr-old*  
1238 *ocean*. *Nature*, 2004. **431**: 549-52. DOI: 10.1038/nature02888.

1239 48. Rutherford, A.W. and P. Faller, *Photosystem II: evolutionary perspectives*. *Philos.*  
1240 *Trans. Royal Soc. B*, 2003. **358**: 245-253. DOI: 10.1098/rstb.2002.1186.

1241 49. Nixon, P.J. and B.A. Diner, *Aspartate 170 of the Photosystem II reaction center*  
1242 *polypeptide D1 is involved in the assembly of the oxygen-evolving manganese cluster*.  
1243 *Biochemistry*, 1992. **31**: 942-948. DOI: 10.1021/Bi00118a041.

1244 50. Adachi, Y., H. Kuroda, Y. Yukawa, and M. Sugiura, *Translation of partially*  
1245 *overlapping psbD-psbC mRNAs in chloroplasts: the role of 5'-processing and*  
1246 *translational coupling*. *Nucleic Acids Res.*, 2012. **40**: 3152-8. DOI:  
1247 10.1093/nar/gkr1185.

1248 51. Carpenter, S.D., J. Charite, B. Eggers, and W.F. Vermaas, *The psbC start codon in*  
1249 *Synechocystis sp. PCC 6803*. *FEBS Lett.*, 1990. **260**: 135-7. DOI: 10.1016/0014-  
1250 5793(90)80085-w.

1251 52. Chisholm, D. and J.G. Williams, *Nucleotide sequence of psbC, the gene encoding the*  
1252 *CP-43 chlorophyll a-binding protein of Photosystem II, in the cyanobacterium*  
1253 *Synechocystis 6803*. *Plant Mol. Biol.*, 1988. **10**: 293-301. DOI: 10.1007/BF00029879.

1254 53. Cockell, C.S., *Ultraviolet radiation and the photobiology of earth's early oceans*.  
1255 Orig. Life Evol. Biospheres, 2000. **30**: 467-99. DOI: 10.1023/a:1006765405786.  
1256 54. Lewis, C.A., J. Crayle, S.T. Zhou, R. Swanstrom, and R. Wolfenden, *Cytosine*  
1257 *deamination and the precipitous decline of spontaneous mutation during Earth's*  
1258 *history*. Proc. Natl. Acad. Sci. U.S.A., 2016. **113**: 8194-8199. DOI:  
1259 10.1073/pnas.1607580113.  
1260 55. Levy, M. and S.L. Miller, *The stability of the RNA bases: implications for the origin*  
1261 *of life*. Proc. Natl. Acad. Sci. U.S.A., 1998. **95**: 7933-8. DOI:  
1262 10.1073/pnas.95.14.7933.  
1263 56. Eisen, J.A. and P.C. Hanawalt, *A phylogenomic study of DNA repair genes, proteins,*  
1264 *and processes*. Mutat. Res./DNA Repair, 1999. **435**: 171-213. DOI: 10.1016/s0921-  
1265 8777(99)00050-6.  
1266 57. Koonin, E.V. and M.Y. Galperin, *The major transitions in evolution: A comparative*  
1267 *genomic perspective in Sequence — Evolution — Function: Computational*  
1268 *Approaches in Comparative Genomics*. 2003, Springer-Science+Business Media,  
1269 B.V>. 252-292. DOI: 10.1007/978-1-4757-3783-7.  
1270 58. Woese, C., *The universal ancestor*. Proc. Natl. Acad. Sci. U.S.A., 1998. **95**: 6854-9.  
1271 DOI: 10.1073/pnas.95.12.6854.  
1272 59. Moore, K.R., C. Magnabosco, L. Momper, D.A. Gold, T. Bosak, and G.P. Fournier,  
1273 *An Expanded Ribosomal Phylogeny of Cyanobacteria Supports a Deep Placement of*  
1274 *Plastids*. Front. Microbiol., 2019. **10**. DOI: 10.3389/fmicb.2019.01612.  
1275 60. Soltis, P.S., R.A. Folk, and D.E. Soltis, *Darwin review: angiosperm phylogeny and*  
1276 *evolutionary radiations*. P Roy Soc B-Biol Sci, 2019. **286**. DOI:  
1277 10.1098/rspb.2019.0099.  
1278 61. McCutcheon, J.P. and N.A. Moran, *Extreme genome reduction in symbiotic bacteria*.  
1279 Nat. Rev. Microbiol., 2011. **10**: 13-26. DOI: 10.1038/nrmicro2670.  
1280 62. van der Staay, G.W., S.Y. Moon-van der Staay, L. Garczarek, and F. Partensky, *Rapid*  
1281 *evolutionary divergence of Photosystem I core subunits PsaA and PsaB in the marine*  
1282 *prokaryote Prochlorococcus*. Photosynth Res, 2000. **65**: 131-9. DOI:  
1283 10.1023/A:1006445810996.  
1284 63. David, L.A. and E.J. Alm, *Rapid evolutionary innovation during an Archaean genetic*  
1285 *expansion*. Nature, 2011. **469**: 93-96. DOI: 10.1038/Nature09649.  
1286 64. Zhu, Q., U. Mai, W. Pfeiffer, S. Janssen, F. Asnicar, J.G. Sanders, P. Belda-Ferre,  
1287 G.A. Al-Ghalith, E. Kopylova, D. McDonald, T. Kosciolek, J.B. Yin, S. Huang, N.  
1288 Salam, J.-Y. Jiao, Z. Wu, Z.Z. Xu, K. Cantrell, Y. Yang, E. Sayyari, M. Rabiee, J.T.  
1289 Morton, S. Podell, D. Knights, W.-J. Li, C. Huttenhower, N. Segata, L. Smarr, S.  
1290 Mirarab, and R. Knight, *Phylogenomics of 10,575 genomes reveals evolutionary*  
1291 *proximity between domains Bacteria and Archaea*. Nat. Commun., 2019. **10**: 5477.  
1292 DOI: 10.1038/s41467-019-13443-4.  
1293 65. Berkemer, S.J. and S.E. McGlynn, *A New Analysis of Archaea–Bacteria Domain*  
1294 *Separation: Variable Phylogenetic Distance and the Tempo of Early Evolution*. Mol.  
1295 *Biol. Evol.*, 2020. DOI: 10.1093/molbev/msaa089.  
1296 66. Homann, M., C. Heubeck, A. Airo, and M.M. Tice, *Morphological adaptations of*  
1297 *3.22 Ga-old tufted microbial mats to Archean coastal habitats (Moodies Group,*  
1298 *Barberton Greenstone Belt, South Africa)*. Precambrian Res., 2015. **266**: 47-64. DOI:  
1299 10.1016/j.precamres.2015.04.018.  
1300 67. Alcott, L.J., B.J.W. Mills, and S.W. Poulton, *Stepwise Earth oxygenation is an*  
1301 *inherent property of global biogeochemical cycling*. Science, 2019. **366**: 1333-1337.  
1302 DOI: 10.1126/science.aax6459.

1303 68. Kadoya, S., D.C. Catling, R.W. Nicklas, I.S. Puchtel, and A.D. Anbar, *Mantle data*  
1304 *imply a decline of oxidizable volcanic gases could have triggered the Great*  
1305 *Oxidation*. Nat. Commun., 2020. **11**: 2774. DOI: 10.1038/s41467-020-16493-1.

1306 69. Reinhard, C.T. and N.J. Planavsky, *Biogeochemical Controls on the Redox Evolution*  
1307 *of Earth's Oceans and Atmosphere*. Elements, 2020. **16**: 191-196. DOI:  
1308 10.2138/gselements.16.3.191.

1309 70. Cardona, T., *Photosystem II is a chimera of reaction centers*. J. Mol. Evol., 2017. **84**:  
1310 149-151. DOI: 10.1007/s00239-017-9784-x.

1311 71. Trinugroho, J.P., M. Beckova, S. Shao, J. Yu, Z. Zhao, J.W. Murray, R. Sobotka, J.  
1312 Komenda, and P.J. Nixon, *Chlorophyll f synthesis by a super-rogue photosystem II*  
1313 *complex*. Nat Plants, 2020. **6**: 238-244. DOI: 10.1038/s41477-020-0616-4.

1314 72. Wegener, K.M., A. Nagarajan, and H.B. Pakrasi, *An atypical psbA gene encodes a*  
1315 *sentinel D1 protein to form a physiologically relevant inactive Photosystem II*  
1316 *complex in Cyanobacteria*. J. Biol. Chem., 2015. **290**: 3764-74. DOI:  
1317 10.1074/jbc.M114.604124.

1318 73. Cardona, T., *Thinking twice about the evolution of photosynthesis*. Open Biology,  
1319 2019. **9**: 180246. DOI: 10.1098/rsob.180246.

1320 74. Petrov, A.S., C.R. Bernier, C.L. Hsiao, A.M. Norris, N.A. Kovacs, C.C. Waterbury,  
1321 V.G. Stepanov, S.C. Harvey, G.E. Fox, R.M. Wartell, N.V. Hud, and L.D. Williams,  
1322 *Evolution of the ribosome at atomic resolution*. Proc. Natl. Acad. Sci. U.S.A., 2014.  
1323 **111**: 10251-10256. DOI: 10.1073/pnas.1407205111.

1324 75. Granick, S., *Speculations on the origins and evolution of photosynthesis*. Ann. N. Y.  
1325 Acad. Sci., 1957. **69**: 292-308. DOI: 10.1111/j.1749-6632.1957.tb49665.x.

1326 76. Mauzerall, D., *Light, iron, Sam Granick and the origin of life*. Photosynth. Res., 1992.  
1327 **33**: 163-170. DOI: 10.1007/BF00039178.

1328 77. Mulkidjanian, A.Y., A.Y. Bychkov, D.V. Dibrova, M.Y. Galperin, and E.V. Koonin,  
1329 *Origin of first cells at terrestrial, anoxic geothermal fields*. Proc. Natl. Acad. Sci.  
1330 U.S.A., 2012. **109**: E821-E830. DOI: 10.1073/pnas.1117774109.

1331 78. Franz, H.B., P.R. Mahaffy, C.R. Webster, G.J. Flesch, E. Raaen, C. Freissinet, S.K.  
1332 Atreya, C.H. House, A.C. McAdam, C.A. Knudson, P.D. Archer, J.C. Stern, A.  
1333 Steele, B. Sutter, J.L. Eigenbrode, D.P. Glavin, J.M.T. Lewis, C.A. Malespin, M.  
1334 Millan, D.W. Ming, R. Navarro-Gonzalez, and R.E. Summons, *Indigenous and*  
1335 *exogenous organics and surface-atmosphere cycling inferred from carbon and oxygen*  
1336 *isotopes at Gale crater*. Nat Astron, 2020. DOI: 10.1038/s41550-019-0990-x.

1337 79. Lu, A.H., Y. Li, H.R. Ding, X.M. Xu, Y.Z. Li, G.P. Ren, J. Liang, Y.W. Liu, H.  
1338 Hong, N. Chen, S.Q. Chu, F.F. Liu, H.R. Wang, C. Ding, C.Q. Wang, Y. Lai, J. Liu,  
1339 J. Dick, K.H. Liu, and M.F. Hochella, *Photoelectric conversion on Earth's surface via*  
1340 *widespread Fe- and Mn-mineral coatings*. Proc. Natl. Acad. Sci. U.S.A., 2019. **116**:  
1341 9741-9746. DOI: 10.1073/pnas.1902473116.

1342 80. Mendler, K., H. Chen, D.H. Parks, B. Lobb, L.A. Hug, and A.C. Doxey, *AnnoTree:*  
1343 *visualization and exploration of a functionally annotated microbial tree of life*.  
1344 Nucleic Acids Res., 2019. **47**: 4442-4448. DOI: 10.1093/nar/gkz246.

1345 81. Ward, L.M., T. Cardona, and H. Holland-Moritz, *Evolutionary implications of*  
1346 *an oxygenic phototrophy in the bacterial phylum Candidatus Eremiobacterota (WPS-*  
1347 *2)*. Front. Microbiol., 2019. **10**: 1658. DOI: 10.3389/fmicb.2019.01658.

1348 82. Gruber-Vodicka, H.R., N. Leisch, M. Kleiner, T. Hinzke, M. Liebeke, M. McFall-  
1349 Ngai, M.G. Hadfield, and N. Dubilier, *Two intracellular and cell type-specific*  
1350 *bacterial symbionts in the placozoan Trichoplax H2*. Nat. Microbiol., 2019. **4**: 1465-  
1351 1474. DOI: 10.1038/s41564-019-0475-9.

1352 83. Sievers, F., A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W.Z. Li, R. Lopez, H.  
1353 McWilliam, M. Remmert, J. Soding, J.D. Thompson, and D.G. Higgins, *Fast,*  
1354 *scalable generation of high-quality protein multiple sequence alignments using*  
1355 *Clustal Omega.* Mol. Syst. Biol., 2011. **7**: 539. DOI: 10.1038/msb.2011.75.

1356 84. Castresana, J., *Selection of conserved blocks from multiple alignments for their use in*  
1357 *phylogenetic analysis.* Mol. Biol. Evol., 2000. **17**: 540-552. DOI:  
1358 10.1093/oxfordjournals.molbev.a026334.

1359 85. Guindon, S., J.F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel,  
1360 *New algorithms and methods to estimate maximum-likelihood phylogenies: assessing*  
1361 *the performance of PhyML 3.0.* Syst. Biol., 2010. **59**: 307-21. DOI:  
1362 10.1093/sysbio/syq010.

1363 86. Lefort, V., J.E. Longueville, and O. Gascuel, *SMS: Smart model selection in PhyML.*  
1364 Mol. Biol. Evol., 2017. **34**: 2422-2424. DOI: 10.1093/molbev/msx149.

1365 87. Anisimova, M. and O. Gascuel, *Approximate likelihood-ratio test for branches: A*  
1366 *fast, accurate, and powerful alternative.* Syst. Biol., 2006. **55**: 539-52. DOI:  
1367 10.1080/10635150600755453.

1368 88. Huson, D.H. and C. Scornavacca, *Dendroscope 3: an interactive tool for rooted*  
1369 *phylogenetic trees and networks.* Syst. Biol., 2012. **61**: 1061-7. DOI:  
1370 10.1093/sysbio/sys062.

1371 89. Gascuel, O., *BIONJ: An improved version of the NJ algorithm based on a simple*  
1372 *model of sequence data.* Mol. Biol. Evol., 1997. **14**: 685-695. DOI:  
1373 10.1093/oxfordjournals.molbev.a025808.

1374 90. Gouy, M., S. Guindon, and O. Gascuel, *SeaView version 4: A multiplatform graphical*  
1375 *user interface for sequence alignment and phylogenetic tree building.* Mol. Biol.

1376 Evol., 2010. **27**: 221-224. DOI: 10.1093/molbev/msp259.

1377 91. Kumar, S., G. Stecher, M. Li, C. Knyaz, and K. Tamura, *MEGA X: Molecular*  
1378 *evolutionary genetics analysis across computing platforms.* Mol. Biol. Evol., 2018.  
1379 **35**: 1547-1549. DOI: 10.1093/molbev/msy096.

1380 92. Clarke, J.T., R.C.M. Warnock, and P.C.J. Donoghue, *Establishing a time-scale for*  
1381 *plant evolution.* New Phytol, 2011. **192**: 266-301. DOI: 10.1111/j.1469-  
1382 8137.2011.03794.x.

1383 93. Morris, J.L., M.N. Puttik, J.W. Clark, D. Edwards, P. Kenrick, S. Pressel, C.H.  
1384 Wellman, Z.H. Yang, H. Schneider, and P.C.J. Donoghue, *The timescale of early land*  
1385 *plant evolution.* Proc. Natl. Acad. Sci. U.S.A., 2018. **115**: E2274-E2283. DOI:  
1386 10.1073/pnas.1719588115.

1387 94. Butterfield, N.J., *Bangiomorpha pubescens n. gen., n. sp.: implications for the*  
1388 *evolution of sex, multicellularity, and the Mesoproterozoic/Neoproterozoic radiation*  
1389 *of eukaryotes.* Paleobiology, 2000. **26**: 386-404. DOI: 10.1666/0094-  
1390 8373(2000)026<0386:Bpngns>2.0.Co;2.

1391 95. Knoll, A.H., S. Worndle, and L.C. Kah, *Covariance of microfossil assemblages and*  
1392 *microbialite textures across an upper mesoproterozoic carbonate platform.* Palaios,  
1393 2013. **28**: 453-470. DOI: 10.2110/palo.2013.p13-005r.

1394 96. Gibson, T.M., P.M. Shih, V.M. Cumming, W.W. Fischer, P.W. Crockford, M.S.W.  
1395 Hodgsiss, S. Worndle, R.A. Creaser, R.H. Rainbird, T.M. Skulski, and G.P.  
1396 Halverson, *Precise age of Bangiomorpha pubescens dates the origin of eukaryotic*  
1397 *photosynthesis.* Geology, 2017. **46**: 135-138. DOI: 10.1130/G39829.1.

1398 97. Bengtson, S., T. Sallstedt, V. Belivanova, and M. Whitehouse, *Three-dimensional*  
1399 *preservation of cellular and subcellular structures suggests 1.6 billion-year-old*  
1400 *crown-group red algae.* Plos Biol, 2017. **15**: e2000735. DOI:  
1401 10.1371/journal.pbio.2000735.

1402 98. Sallstedt, T., S. Bengtson, C. Broman, P.M. Crill, and D.E. Canfield, *Evidence of*  
1403 *oxygenic phototrophy in ancient phosphatic stromatolites from the Paleoproterozoic*  
1404 *Vindhyan and Aravalli Supergroups, India*. *Geobiology*, 2018. **16**: 139-159. DOI:  
1405 10.1111/gbi.12274.

1406 99. Qu, Y.G., S.X. Zhu, M. Whitehouse, A. Engdahl, and N. McLoughlin, *Carbonaceous*  
1407 *biosignatures of the earliest putative macroscopic multicellular eukaryotes from 1630*  
1408 *Ma Tuanshanzi Formation, north China*. *Precambrian Res.*, 2018. **304**: 99-109. DOI:  
1409 10.1016/j.precamres.2017.11.004.

1410 100. Pang, K., Q. Tang, L. Chen, B. Wan, C. Niu, X. Yuan, and S. Xiao, *Nitrogen-fixing*  
1411 *heterocystous cyanobacteria in the tonian period*. *Curr Biol*, 2018. **28**: 616-622 e1.  
1412 DOI: 10.1016/j.cub.2018.01.008.

1413 101. Stothard, P., *The sequence manipulation suite: JavaScript programs for analyzing*  
1414 *and formatting protein and DNA sequences*. *BioTechniques*, 2000. **28**: 1102-1104.  
1415 DOI: 10.2144/00286ir01.

1416 102. Ponce-Toledo, R.I., P. Deschamps, P. Lopez-Garcia, Y. Zivanovic, K. Benzerara, and  
1417 D. Moreira, *An early-branching freshwater cyanobacterium at the origin of plastids*.  
1418 *Curr Biol*, 2017. **27**: 386-391. DOI: 10.1016/j.cub.2016.11.056.

1419 103. Sánchez-Baracaldo, P., J.A. Raven, D. Pisani, and A.H. Knoll, *Early photosynthetic*  
1420 *eukaryotes inhabited low-salinity habitats*. *Proc. Natl. Acad. Sci. U.S.A.*, 2017. **114**:  
1421 E7737-E7745. DOI: 10.1073/pnas.1620089114.

1422 104. Fewer, D., T. Friedl, and B. Budel, *Chroococcidiopsis and heterocyst-differentiating*  
1423 *cyanobacteria are each other's closest living relatives*. *Mol. Phylogen. Evol.*, 2002.  
1424 **23**: 82-90. DOI: 10.1006/mpev.2001.1075.

1425 105. Shih, P.M., D. Wu, A. Latifi, S.D. Axen, D.P. Fewer, E. Talla, A. Calteau, F. Cai, N.  
1426 Tandeau de Marsac, R. Rippka, M. Herdman, K. Sivonen, T. Coursin, T. Laurent, L.  
1427 Goodwin, M. Nolan, K.W. Davenport, C.S. Han, E.M. Rubin, J.A. Eisen, T. Woyke,  
1428 M. Gugger, and C.A. Kerfeld, *Improving the coverage of the cyanobacterial phylum*  
1429 *using diversity-driven genome sequencing*. *Proc. Natl. Acad. Sci. U.S.A.*, 2013. **110**:  
1430 1053-8. DOI: 10.1073/pnas.1217107110.

1431 106. Schirrmeister, B.E., A. Antonelli, and H.C. Bagheri, *The origin of multicellularity in*  
1432 *cyanobacteria*. *BMC Evol. Biol.*, 2011. **11**: 45. DOI: 10.1186/1471-2148-11-45.

1433 107. Ciccarelli, F.D., T. Doerks, C. von Mering, C.J. Creevey, B. Snel, and P. Bork,  
1434 *Toward automatic reconstruction of a highly resolved tree of life*. *Science*, 2006. **311**:  
1435 1283-7. DOI: 10.1126/science.1123061.

1436 108. Nakamura, Y., T. Kaneko, S. Sato, M. Mimuro, H. Miyashita, T. Tsuchiya, S.  
1437 Sasamoto, A. Watanabe, K. Kawashima, Y. Kishida, C. Kiyokawa, M. Kohara, M.  
1438 Matsumoto, A. Matsuno, N. Nakazaki, S. Shimpo, C. Takeuchi, M. Yamada, and S.  
1439 Tabata, *Complete genome structure of Gloeobacter violaceus PCC 7421, a*  
1440 *cyanobacterium that lacks thylakoids*. *DNA Res*, 2003. **10**: 137-145. DOI:  
1441 10.1093/dnares/10.4.137.

1442 109. Honda, D., A. Yokota, and J. Sugiyama, *Detection of seven major evolutionary*  
1443 *lineages in cyanobacteria based on the 16S rRNA gene sequence analysis with new*  
1444 *sequences of five marine Synechococcus strains*. *J. Mol. Biol.*, 1999. **48**: 723-739.  
1445 DOI: 10.1007/Pl00006517.

1446 110. Knoll, A.H., E.J. Javaux, D. Hewitt, and P. Cohen, *Eukaryotic organisms in*  
1447 *Proterozoic oceans*. *Philos. Trans. Royal Soc. B*, 2006. **361**: 1023-1038. DOI:  
1448 10.1098/rstb.2006.1843.

1449 111. Han, T.M. and B. Runnegar, *Megascopic eukaryotic algae from the 2.1-billion-year-*  
1450 *old Negaunee iron-formation, Michigan*. *Science*, 1992. **257**: 232-235. DOI:  
1451 10.1126/science.1631544.

1452 112. El Albani, A., M.G. Mangano, L.A. Buatois, S. Bengtson, A. Riboulleau, A. Bekker,  
1453 K. Konhauser, T. Lyons, C. Rollion-Bard, O. Bankole, S.G.L. Baghekema, A.  
1454 Meunier, A. Trentesaux, A. Mazurier, J. Aubineau, C. Laforest, C. Fontaine, P.  
1455 Recourt, E.C. Fru, R. Macchiarelli, J.Y. Reynaud, F. Gauthier-Lafaye, and D.E.  
1456 Canfield, *Organism motility in an oxygenated shallow-marine environment 2.1 billion*  
1457 *years ago*. Proc. Natl. Acad. Sci. U.S.A., 2019. **116**: 3431-3436. DOI:  
1458 10.1073/pnas.1815721116.

1459 113. Shi, M., Q.L. Feng, M.Z. Khan, and S.X. Zhu, *An eukaryote-bearing microbiota from*  
1460 *the early mesoproterozoic Gaoyuzhuang Formation, Tianjin, China and its*  
1461 *significance*. Precambrian Res., 2017. **303**: 709-726. DOI:  
1462 10.1016/j.precamres.2017.09.013.

1463 114. Sánchez-Baracaldo, P., *Origin of marine planktonic cyanobacteria*. Sci. Rep., 2015.  
1464 **5**: 17418. DOI: 10.1038/srep17418.

1465 115. Lepage, T., D. Bryant, H. Philippe, and N. Lartillot, *A general comparison of relaxed*  
1466 *molecular clock models*. Mol. Biol. Evol., 2007. **24**: 2669-80. DOI:  
1467 10.1093/molbev/msm193.

1468 116. Kishino, H., J.L. Thorne, and W.J. Bruno, *Performance of a divergence time*  
1469 *estimation method under a probabilistic model of rate evolution*. Mol. Biol. Evol.,  
1470 2001. **18**: 352-361. DOI: 10.1093/oxfordjournals.molbev.a003811.

1471 117. Langklotz, S., U. Baumann, and F. Narberhaus, *Structure and function of the*  
1472 *bacterial AAA protease FtsH*. Biochim. Biophys. Acta, 2012. **1823**: 40-8. DOI:  
1473 10.1016/j.bbamcr.2011.08.015.

1474 118. Medlin, L.K., W.H.C.F. Kooistra, R. Gersonde, P.A. Sims, and U. Wellbrock, *Is the*  
1475 *origin of the diatoms related to the end-Permian mass extinction?* Nova Hedwigia,  
1476 1997. **65**: 1-11.

1477 119. Barboni, M., P. Boehnke, B. Keller, I.E. Kohl, B. Schoene, E.D. Young, and K.D.  
1478 McKeegan, *Early formation of the Moon 4.51 billion years ago*. Science Advances,  
1479 2017. **3**: e1602365. DOI: 10.1126/sciadv.1602365.

1480 120. Drummond, A.J., S.Y.W. Ho, M.J. Phillips, and A. Rambaut, *Relaxed phylogenetics*  
1481 *and dating with confidence*. Plos Biol, 2006. **4**: 699-710. DOI:  
1482 10.1371/journal.pbio.0040088.

1483 121. Hodgskiss, M.S.W., O.M.J. Dagnaud, J.L. Frost, G.P. Halverson, M.D. Schmitz, N.L.  
1484 Swanson-Hysell, and E.A. Sperling, *New insights on the Orosirian carbon cycle,*  
1485 *early Cyanobacteria, and the assembly of Laurentia from the Paleoproterozoic*  
1486 *Belcher Group*. Earth Planet. Sci. Lett., 2019. **520**: 141-152. DOI:  
1487 10.1016/j.epsl.2019.05.023.

1488 122. Brocks, J.J., G.D. Love, R.E. Summons, A.H. Knoll, G.A. Logan, and S.A. Bowden,  
1489 *Biomarker evidence for green and purple sulphur bacteria in a stratified*  
1490 *Palaeoproterozoic sea*. Nature, 2005. **437**: 866-870. DOI: 10.1038/Nature04068.

1491 123. Engel, M.S. and D.A. Grimaldi, *New light shed on the oldest insect*. Nature, 2004.  
1492 **427**: 627-630. DOI: 10.1038/nature02291.

1493 124. Legendre, F., A. Nel, G.J. Svenson, T. Robillard, R. Pellens, and P. Grandcolas,  
1494 *Phylogeny of Dictyoptera: Dating the origin of cockroaches, praying mantises and*  
1495 *termites with molecular data and controlled fossil evidence*. PloS one, 2015. **10**:  
1496 e0130127. DOI: 10.1371/journal.pone.0130127.

1497 125. Eitel, M., H.J. Osigus, R. DeSalle, and B. Schierwater, *Global diversity of the*  
1498 *Placozoa*. PloS one, 2013. **8**: e57131. DOI: 10.1371/journal.pone.0057131.

1499 126. Zumberge, J.A., G.D. Love, P. Cardenas, E.A. Sperling, S. Gunasekera, M. Rohrssen,  
1500 E. Grosjean, J.P. Grotzinger, and R.E. Summons, *Demosponge steroid biomarker 26-*

1501      *methylstigmastane provides evidence for Neoproterozoic animals*. *Nat. Ecol. Evol.*,  
 1502      2018. **2**: 1709-1714. DOI: 10.1038/s41559-018-0676-2.

1503      127. Grossnickle, D.M., S.M. Smith, and G.P. Wilson, *Untangling the multiple ecological*  
 1504      *radiations of early mammals*. *Trends Ecol Evol*, 2019. **34**: 936-949. DOI:  
 1505      10.1016/j.tree.2019.05.008.

1506      128. Williams, B.A., R.F. Kay, and E.C. Kirk, *New perspectives on anthropoid origins*.  
 1507      *Proc. Natl. Acad. Sci. U.S.A.*, 2010. **107**: 4797-804. DOI: 10.1073/pnas.0908320107.

1508      129. Soo, R.M., B.J. Woodcroft, D.H. Parks, G.W. Tyson, and P. Hugenholtz, *Back from*  
 1509      *the dead; the curious tale of the predatory cyanobacterium Vampirovibrio*  
 1510      *chlorellavorus*. *PeerJ*, 2015. **3**: e968. DOI: 10.7717/peerj.968.

1511      130. Hahn, M.W., J. Schmidt, A. Pitt, S.J. Taipale, and E. Lang, *Reclassification of four*  
 1512      *Polynucleobacter necessarius* strains as representatives of *Polynucleobacter*  
 1513      *asymbioticus* comb. nov., *Polynucleobacter duraquae* sp. nov., *Polynucleobacter*  
 1514      *yangtzensis* sp. nov and *Polynucleobacter sinensis* sp.s nov., and emended description  
 1515      of *Polynucleobacter necessarius*. *Int. J. Syst. Evol. Microbiol.*, 2016. **66**: 2883-2892.  
 1516      DOI: 10.1099/ijsem.0.001073.

1517      131. Parfrey, L.W., D.J. Lahr, A.H. Knoll, and L.A. Katz, *Estimating the timing of early*  
 1518      *eukaryotic diversification with multigene molecular clocks*. *Proc. Natl. Acad. Sci.*  
 1519      *U.S.A.*, 2011. **108**: 13624-9. DOI: 10.1073/pnas.1110633108.

1520      132. Fernandes, N.M. and C.G. Schrago, *A multigene timescale and diversification*  
 1521      *dynamics of Ciliophora evolution*. *Mol. Phylogen. Evol.*, 2019. **139**: 106521. DOI:  
 1522      10.1016/j.ympev.2019.106521.

1523      133. Yu, X.-J. and D.H. Walker, *The Order Rickettsiales*, in *The prokaryotes: A handbook*  
 1524      *on the biology of bacteria*, M. Dworkin, Editor. 2006, Springer: Singapore. 493-528.  
 1525      DOI: 10.1007/0-387-30745-1\_20.

1526      134. Roger, A.J., S.A. Munoz-Gomez, and R. Kamikawa, *The Origin and Diversification*  
 1527      *of Mitochondria*. *Curr Biol*, 2017. **27**: R1177-R1192. DOI:  
 1528      10.1016/j.cub.2017.09.015.

1529      135. Li, W. and A. Godzik, *Cd-hit: a fast program for clustering and comparing large sets*  
 1530      *of protein or nucleotide sequences*. *Bioinformatics*, 2006. **22**: 1658-9. DOI:  
 1531      10.1093/bioinformatics/btl158.

1532      136. Ashkenazy, H., O. Penn, A. Doron-Faigenboim, O. Cohen, G. Cannarozzi, O. Zomer,  
 1533      and T. Pupko, *FastML: a web server for probabilistic reconstruction of ancestral*  
 1534      *sequences*. *Nucleic Acids Res.*, 2012. **40**: W580-W584. DOI: 10.1093/nar/gks498.

1535      137. Hanson-Smith, V., B. Kolaczkowski, and J.W. Thornton, *Robustness of ancestral*  
 1536      *sequence reconstruction to phylogenetic uncertainty*. *Mol. Biol. Evol.*, 2010. **27**:  
 1537      1988-1999. DOI: 10.1093/molbev/msq081.

1538      138. Yang, Z.H., *PAML 4: Phylogenetic analysis by maximum likelihood*. *Mol. Biol. Evol.*,  
 1539      2007. **24**: 1586-1591. DOI: 10.1093/molbev/msm088.

1540      139. Kumar, S., G. Stecher, and K. Tamura, *MEGA7: Molecular evolutionary genetics*  
 1541      *analysis version 7.0 for bigger datasets*. *Mol. Biol. Evol.*, 2016. **33**: 1870-1874. DOI:  
 1542      10.1093/molbev/msw054.

1543      140. Yu, L.J., M. Suga, Z.Y. Wang-Otomo, and J.R. Shen, *Structure of photosynthetic*  
 1544      *LH1-RC supercomplex at 1.9 Å resolution*. *Nature*, 2018. **556**: 209-213. DOI:  
 1545      10.1038/s41586-018-0002-9.

1546      141. Gisriel, C., I. Sarrou, B. Ferlez, J.H. Golbeck, K.E. Redding, and R. Fromme,  
 1547      *Structure of a symmetric photosynthetic reaction center-photosystem*. *Science*, 2017.  
 1548      **357**: 1021-1025. DOI: 10.1126/science.aan5611.

1549 142. Jordan, P., P. Fromme, H.T. Witt, O. Klukas, W. Saenger, and N. Krauss, *Three-*  
1550 *dimensional structure of cyanobacterial Photosystem I at 2.5 Å resolution*. *Nature*,  
1551 2001. **411**: 909-17. DOI: 10.1038/35082000.

1552 143. Toporik, H., J. Li, D. Williams, P.L. Chiu, and Y. Mazor, *The structure of the stress-*  
1553 *induced photosystem I-IsiA antenna supercomplex*. *Nat Struct Mol Biol*, 2019. **26**:  
1554 443-449. DOI: 10.1038/s41594-019-0228-8.

1555 144. Jia, Y.T., G. Dewey, I.N. Shindyalov, and P.E. Bourne, *A new scoring function and*  
1556 *associated statistical significance for structure alignment by CE*. *J Comput Biol*,  
1557 2004. **11**: 787-799. DOI: Doi 10.1089/1066527042432260.

1558