Article:
Sheridan, KJ, Duncan, EJ orcid.org/0000-0002-1841-504X, Eaton-Rye, JJ et al. (1 more author) (2020) The diversity and distribution of D1 proteins in cyanobacteria. Photosynthesis Research, 145 (2). pp. 111-128. ISSN 0166-8595

https://doi.org/10.1007/s11120-020-00762-7

The diversity and distribution of D1 proteins in cyanobacteria

Kevin J. Sheridan[1,2] • Elizabeth J. Duncan[3] • Julian J. Eaton-Rye[2] • Tina C. Summerfield[1]

[1]Department of Botany, University of Otago, Dunedin, New Zealand

[2]Department of Biochemistry, University of Otago, Dunedin, New Zealand

[3]Department of Biological Sciences, School of Biology, University of Leeds, Leeds, UK

Corresponding Author Tina C. Summerfield

tina.summerfield@otago.ac.nz

Abstract

The *psbA* gene family in cyanobacteria encodes different forms of the D1 protein that is part of the Photosystem II reaction centre. We have identified a phylogenetically distinct D1 group that is intermediate between previously identified G3 and G4 D1 proteins [Cardona T, Murray JW, Rutherford AW (2015) Mol Biol Evol 32: 1310–1328]. This new group contained two subgroups: $D1^{INT}$, which was frequently in the genomes of heterocystous cyanobacteria and $D1^{FR}$ that was part of the far-red light photoacclimation gene cluster of cyanobacteria. In addition, we have identified subgroups within G3, the micro-aerobically expressed D1 protein. There are amino acid changes associated with each of the subgroups that might affect function of Photosystem II. We show a phylogenetically broad range of cyanobacteria have these D1 types, as well as the genes encoding the G2 protein and chlorophyll *f* synthase. We suggest identification of additional D1 isoforms and the presence of multiple D1 isoforms in phylogenetically diverse cyanobacteria supports the role of these proteins in conferring a selective advantage under specific conditions.

Keywords Cyanobacteria • D1 • Evolution • Photosystem II • Phylogenetics • *psbA*

Introduction

Photosystem II (PS II) catalyses the light-driven splitting of water at the start of the photosynthetic electron transport chain in the thylakoid membrane of oxygenic phototrophs (Vinyard and Brudvig 2018). High resolution PS II structures (~1.9 – 2.1 Å) have been obtained from thermophilic cyanobacteria (Umena et al. 2011; Suga et al. 2015, 2017; Kern et al. 2018) and detailed structures confirming a high degree of conservation in eukaryotes have been obtained (Ago et al. 2016; Wei et al. 2016). The major polypeptides of the PS II reaction centre are referred to as D1 and D2 and these proteins provide the majority of the ligands to the redox active cofactors. In particular, the D1 protein provides the majority of the ligands to the Mn_4CaO_5 oxygen-evolving complex (OEC) with the remainder coming from the chlorophyll-binding CP43 protein of the core antenna (Ferreira et al. 2004; Shen 2015). Although D1 and D2 form a heterodimer, only the D1 branch is active in the reduction of the primary and secondary plastoquinone electron acceptors Q_A and Q_B (Cardona et al. 2012). In addition, the oxidative chemistry and photochemistry associated with water splitting results in light-induced photodamage that preferentially targets the D1 protein and subsequently D1 has a higher turnover rate than the other PS II proteins (Mulo et al. 2009).

Many cyanobacteria contain multiple copies of the *psbA* gene which encodes the D1 protein (Mulo et al. 2012), with some cyanobacteria containing as many as eight copies. A survey of 360 cyanobacterial D1 proteins supported the previous identification of several distinct types of the D1 protein (G0-G4), with the majority of cyanobacteria having between two and four isoforms encoded by three to six copies of *psbA* (Cardona et al. 2015). The G4 type (G4-D1) is the most prevalent form of D1 that supports oxygen evolution and this is the D1 type found in higher plants. It has been suggested that plastids evolved from an ancestor of extant cyanobacterium *Gloeomargarita lithophora* which has only *psbA* genes encoding

68 the G4-D1 unlike other deeply branching cyanobacteria (Ponce-Toledo et al. 2017). All
69 cyanobacteria investigated to date contain at least one gene encoding a G4-D1 and some
70 strains contain multiple copies of *psbA* encoding G4-D1. Two variants of G4-D1 have been
71 designated as D1:1 and D1:2 (Cardona et al. 2015). When environmental conditions result in
72 increased turnover of D1, multiple copies of *psbA* encoding G4-D1 can benefit cyanobacteria
73 in one of two ways. Firstly, the up-regulation of genes encoding identical copies of D1 (D1:1)
74 increases both the *psbA* transcript pool and the D1 protein level, facilitating rapid
75 replacement of photodamaged D1, thereby reducing photoinhibition (El Bissati and
76 Kirilovsky 2001). In the second mechanism, the *psbA* gene encoding D1:2 is up-regulated.
77 The alternative D1:2 copy is characterized by an amino acid substitution from glutamine to
78 glutamate at position 130. This amino acid change decreases photoinhibition under high light
79 by reducing the formation of triplet state chlorophyll species and singlet oxygen by favouring
80 direct recombination (Vinyard et al. 2014). However, further amino acid differences between
81 D1:1 and D1:2 appear to the impact PS II efficiency (Vinyard et al. 2014).
82 Evidence for additional roles of D1 proteins includes the up-regulation of a *psbA* gene
83 under low-oxygen conditions in several cyanobacteria: *Synechocystis* sp. PCC 6803,
84 *Thermosynechococcus elongatus* BP-1, *Cyanothece* sp. ATCC 51142 and *Anabaena* sp. PCC
85 7120 (Summerfield et al. 2008; Sicora et al. 2009). The D1′ proteins encoded by these low-
86 oxygen-induced *psbA* genes share three amino acid substitutions, Gly80Ala, Phe158Leu,
87 Thr286Ala (Sicora et al. 2009). Furthermore, PS II centres containing the D1′ in
88 *Synechocystis* sp. PCC 6803 produced higher rates of oxygen than centres containing D1:1
89 from *psbA2* when expressed under the low-oxygen promoter (Crawford et al. 2016). A
90 conserved role for these micro-aerobic D1′ proteins is supported by the finding that they
91 were part of a monophyletic group of sequences (G3) from 39 cyanobacterial strains
92 (Cardona et al. 2015).

Phylogenetic analysis of D1 proteins identified three groups lacking residues that provide ligands to the OEC (Cardona et al. 2015). One group (G2) contained 36 proteins (G2-D1), including the rogue D1 identified by Murray (2012), also named sentinel D1 by Wegener et al. (2015). The *psbA* gene encoding the G2-D1 from *Cyanothece* sp. ATCC 51142 was up-regulated in the subjective dark and it has been proposed that this copy of D1 is incorporated into inactive PS II centres to protect oxygen sensitive enzymes such as nitrogenase (Toepel et al. 2008). Wegener et al. (2015) demonstrated that expression of the *psbA* gene encoding G2-D1 from *Cyanothece* sp. ATCC 51142 in *Synechocystis* sp. PCC 6803 resulted in inactive PS II centres when G2-D1 was present. In the unicellular diazotroph *Crocosphaera watsonii* WH8501, during the dark period, G2-D1-containing PS II centres were detected in low numbers consistent with a regulatory role (Masuda et al. 2018). Signalling from the small numbers of G2-D1 PS II centres was part of a proposed two-step mechanism for the inactivation of PS II to protect nitrogenase activity in *Cyanothece* sp. ATCC 51142 (Sicora et al. 2019).

The second phylogenetic group of D1 proteins lacking ligands to the OEC (designated as G1 in Cardona et al. 2015) contained the super rogue class of D1 reported by Murray (2012), and this isoform has subsequently been identified as a chlorophyll *f* synthase that catalyses the production of a far-red / near-infrared absorbing chlorophyll *f* (Ho et al. 2016). The chlorophyll *f* synthase gene is in a far-red-inducible gene cluster (FaRLiP) that is up-regulated under prolonged exposure to far-red / near infrared wavelengths of light. Genes in this cluster encode alternative Photosystem I (PS I), PS II and phycobilisome proteins, along with regulatory proteins, that modify the photosynthetic electron transport chain as a part of a far-red photoacclimation process (Gan et al. 2014; Ho et al. 2016; Nürnberg et al. 2018; Shen et al. 2019). The final phylogenetic group of D1, G0, contained a single sequence from *Gloeobacter kilaueensis* JS-1 (Cardona *et al*. 2015). This sequence also lacks the ligands to

bind the OEC, having a C-terminus which is more similar to D2 than D1 and has an unknown function.

To further investigate the possible roles and extent of D1 diversity in cyanobacteria, we expanded the phylogenetic analyses of D1 proteins using 206 cyanobacterial genomes. We have identified two additional phylogenetically distinct groups of D1 proteins and identified distinct subgroups within the G3 D1 sequences. Our approach has shown the distribution of *psbA* genes is highly varied among the cyanobacteria, likely reflecting particular *psbA* combinations associated with cyanobacteria found in different microhabitats.

Methods

Phylogenetic analysis

A total of 206 cyanobacterial genomes and the G0, 16S and 23S rRNA sequences for *Gloeobacter kilaueensis* JS1 were retrieved from JGI (Grigoriev et al. 2012; Nordberg et al. 2014) and NCBI (Benson et al. 2017) from the 3rd – 7th of January, 2017 and 796 *psbA* gene sequences were extracted from these genomes. The minimum length criteria for inclusion in analyses was approximately two-thirds of the entire sequence (600 bp minimum sequence length). The 16S-23S rRNA (ribosomal RNA) gene sequences were retrieved from the same database as the *psbA* genes with the exception of *Leptolyngbya* sp. JSC-1 for which these data were unavailable. In this case, a partial 16S rRNA gene copy was retrieved from the SILVA ribosomal RNA database (Quast et al. 2013).

Phylogenetic analyses of D1 sequences were performed using the same approach as Cardona et al. (2015) using the atypical sequence from *Gloeobacter kilaueensis* JS1 (G0), described by Saw et al. (2013) as the outgroup. Briefly, the D1 phylogeny was constructed in

PhyML using the LG model of amino acid substitution, four gamma rate categories and the nearest neighbour interchange method for tree improvement. All other parameters were left as default, with the software allowed to estimate the equilibrium frequencies, proportion of invariant sites and the gamma shape parameter. Branch supports were calculated using the SH-like approximate likelihood ratio test option (Shimodaira and Hasegawa 1999) with branch supports above 0.85 (85%) being used as the cutoff threshold. The creation of multiple sequence alignments was aided by generating PDB files for a representative D1 sequence from each D1 protein group using the SWISS-MODEL online service from ExPasy (Guex et al. 2009; Bertoni et al. 2017; Bienert et al. 2017; Waterhouse et al. 2018). The PDB file creation utilised the crystal structure from *Thermosynechococcus vulcanus* (4UB6) as reference (Suga et al. 2015). The resulting PDB files were then aligned using the CEalign function (Shindyalov and Bourne 1998) in PyMOL (DeLano, 2002, 2009) and used in the creation of PyMOL figures. Pairwise alignments of all G3 sequences, as well as the $D1^{INT}$ and $D1^{FR}$ found in this analysis, were also conducted.

A species tree of the 206 cyanobacterial strains, along with the outgroup, was created based on rRNA gene sequences. Briefly, the 16S and 23S rRNA gene sequences were concatenated and aligned using the default parameters of ClustalW (Larkin et al. 2007) and manually checked. As rRNA gene sequences cannot always definitively discriminate between two closely related species (Jaspers and Overmann, 2004), SNPs within multiple copies of the 16S or 23S rRNA gene sequence were utilised to assist in discrimination (Hakovirta et al. 2016). This was achieved by taking the consensus sequence to build the 16S-23S rRNA species tree. In accordance with Hilton et al. (2016) only those alignment sites which had at least 90% coverage were used in the subsequent phylogenetic analysis (Felsenstein 1985). The best-fit model of nucleotide substitution was determined using the JmodelTest 2.1 to generate both the maximum likelihood RAxML and maximum parsimony (PAUP) trees,

168 respectively (Swofford 1985; Stamatakis 2006; Darriba et al. 2012). The data were analysed
169 in both cases using generalised time reversible (GTR) + Γ + I. The most parsimonious trees
170 were found following 1000 replicate heuristic searches with 100 trees saved per replicate to
171 produce a maximum of 10,000 trees. The branch support was then calculated using bootstrap
172 of 1000 replicates. The bootstrap values from the maximum parsimony analysis were
173 transferred to the corresponding branches of the maximum likelihood tree. The maximum
174 likelihood tree was found using 1000 bootstrap iterations. Bootstrap support over 0.95 was
175 used as the threshold cutoff.
176
177 **Identification of genes under purifying selective pressure**
178
179 Pairwise comparison estimates of rates of synonymous (dS) and non-synonymous
180 substitutions (dN) were calculated using codeML in the graphical interface for PAML,
181 PAMLX (Yang 2007; Xu and Yang 2013). Estimates of the ratio of non-synonymous to
182 synonymous mutations, ω (dN/dS), was used to investigate whether each subgroup of *psbA*
183 homologs encoding mature D1 protein sequences were undergoing patterns of neutral drift (ω
184 = 1), purifying selection (ω < 1) or positive selection (ω > 1). The nucleotide multiple
185 sequence alignment of the *psbA* genes encoding each group of D1 proteins was built using
186 the protein alignment for reference. In accordance with Fletcher and Yang (2010), gaps and
187 uncertainties within the multiple sequence alignment were stripped from the alignment to
188 avoid false positives. Additionally, identical nucleotide sequences present in single
189 cyanobacterial strains were also removed to avoid spurious replication of data (Hongo et al.
190 2015). The *rbcL* gene from all strains was included as a reference in this analysis, this gene
191 encodes the large subunit of ribulose-l,5-bisphosphate carboxylase/oxygenase (Rubisco).
192

193

194 Results and discussion

195

196 Diversity of the D1 protein family

197 The analysis of the D1 protein family used in this study employed the LG model of amino
198 acid substitution (Le and Gascuel, 2008). This accounts for among-site rate variation and
199 provides replacement rate estimates using rescaling of amino acid changes observed in the
200 data depending on whether they occur in slow or fast sites. It should be noted that this model
201 is based on a large, diverse data set to estimate a general replacement matrix rather than a
202 more specific matrix. The maximum likelihood phylogeny of D1 proteins (Fig. 1 and Fig. S1)
203 generated using D1 sequences from 206 cyanobacterial strains and the G0 sequence from
204 *Gloeobacter kilaueensis* JS1 showed a similar structure to the previously reported work of
205 Cardona and colleagues with the grouping of D1 proteins not following cyanobacterial
206 phylogenies (Cardona et al. 2015; Grim and Dick 2016). The G0-D1 sequence from
207 *Gloeobacter kilaueensis* JS1 currently has no identified function, and has been suggested to
208 represent the most ancestral D1 sequence based its position in in the type II reaction centre
209 phylogeny of Cardona et al. (2018). Both the amino acid and nucleotide sequences for this
210 purported ancestral D1 has been used as the outgroup in previous phylogenetic studies
211 (Cardona et al. 2015; Grim and Dick 2016). The largest D1 group corresponded to the G4 of
212 Cardona et al. (2015) and this contained 612 of the 796 sequences; including the well-
213 characterised proteins from *Synechocystis* sp. PCC 6803 (D1:1) and *Thermosynechococcus*
214 *elongatus* BP-1 (both D1:1 and D1:2 proteins) (Fig. 1 and Fig. S1, shown in green).

215 Sister to G4 was a group that contained two D1 subgroups, one with moderate support
216 and one well-supported (Fig. 1). This group represents an expansion of the intermediate
217 group of Cardona et al. (2015) from 9 to 47 sequences (Fig. 1 and Fig. S1; subgroups shown

in pink and brown). One subgroup contained 27 D1 sequences, this group will be referred to as $D1^{INT}$ (INT for 'intermediate' as no current function has been ascribed to this group and on the phylogenetic tree these sequences are intermediate between G3 and G4). The second subgroup contained 20 D1 sequences and strains containing these sequences have been shown to contain the FaRLiP gene cluster (Gan et al. 2015). This group will be hereafter referred to as $D1^{FR}$ (FR for far-red).

The next group corresponding to the G3 category from Cardona et al. (2015), which contained the micro-aerobically induced D1′, had increased from 39 to 64 sequences, with almost a third of the analysed cyanobacteria having a *psbA* gene encoding G3-D1. The G3 sequences formed three well-supported subgroups (Fig. 1 and Fig. S1, shown in yellow-orange). Groups corresponding to G2 and G1 of Cardona et al. (2015) were also resolved. The G2 category was increased from 36 to 52 sequences (Fig. 1 and Fig. S1; shown in red), with genes encoding G2-D1 in approximately a quarter of cyanobacteria analysed. The G1 category was increased from 8 to 20 D1 proteins (Fig. 1 and Fig. S1; shown in purple). An alignment of consensus sequences for each D1 type is shown in Fig. 2 at 95% consensus and Fig. S2 for 50% consensus.

A phylogenetically distinct group of D1 protein sequences, $D1^{INT}$

All 27 $D1^{INT}$ sequences have two conserved amino acid changes compared to the G4 proteins: Tyr126 to Trp and Phe260 to Trp. In addition, there are four conserved residues in at least 85% of the $D1^{INT}$ sequences that occur in less than 5% of G4 sequences: Ala68 to Ser, Ser79 to Thr, Ser85 to Thr, Ala156 to Ser (Fig. 3a; and for full length alignment, see Fig. S3). The residues Ser68, Thr79 and Thr85 are located in the lumenal ab-loop. The Tyr126 to Trp substitution is in helix B (Fig. 3b, c) and may directly affect active branch pheophytin ($Pheo_{D1}$) through the loss of the hydrogen bond to the 13^3-ester C=O of $Pheo_{D1}$ (Zabelin et al.

2014). On the other side of $Pheo_{D1}$, in helix C, the Ala156 to Ser substitution may alter hydrogen bonding to both Ala152 and the Tyr161. The alanine at position 152 is thought to interact with the Phe435 of CP43, potentially modulating interactions between D1 and CP43 in the vicinity of $Pheo_{D1}$ (Fig. 3d, e) which Vinyard et al. (2014) suggest may alter the midpoint potential of this pheophytin. The alteration of the Phe260 to Trp is predicted, using in silico modelling, to open a hydrogen bond to the nearby phosphatidylglycerol (PG), a constitutive lipid within the PS II structure (Fig. 3f, g; and see Wada and Murata (2007) and Endo et al. (2019)) and studies by Narusaka et al. (1996, 1999) have suggested that this residue may be involved in phototolerance.

The majority of $D1^{INT}$ encoding genes were in diazotrophic cyanobacteria (25/27) and most of these cyanobacteria were heterocystous (24/27), this represented approximately one third of the heterocystous cyanobacteria analysed in this study (24/71 heterocystous strains). To date specific conditions inducing the up-regulation of $D1^{INT}$ have not been identified.

The D1 proteins associated with the Far-Red Light Photoacclimation (FaRLiP) cluster

The 20 sequences belonging to the $D1^{FR}$ group in Fig. 1 are encoded by *psbA* genes in the far-red-inducible gene cluster described by Gan et al. (2014, 2015). This gene cluster has been identified in multiple cyanobacterial strains including: *Calothrix* sp. PCC 7507, *Chlorogloeopsis fritschii* PCC 9212, *Chroococcidiopsis thermalis* PCC 7203, *Fischerella thermalis* PCC 7521, *Halomicronema hongdechloris* C2206 and *Synechococcus* sp. PCC 7335 (Nürnberg et al. 2018; Partensky et al. 2018; Ho and Bryant 2019; Ho et al. 2019; Chen et al. 2012, 2019). The gene cluster was shown to contain several genes encoding isoforms of PS II, PS I, and phycobilisome proteins as well as regulatory genes. The far-red-inducible PS II genes include two annotated as *psbA* – one encoding chlorophyll *f* synthase and the other encoding $D1^{FR}$ (Gan et al. 2014, 2015). Our analysis supports the conclusion that all genes

encoding $D1^{FR}$ are in a putative FaRLiP cluster (Fig. 4a; for gene context of the 20 *psbA* genes encoding $D1^{FR}$ in far-red-inducible gene clusters, see Fig. S4).

The $D1^{FR}$ proteins retain the essential ligands for binding the OEC. There were 16 conserved changes in the $D1^{FR}$ sequences compared to the 95% consensus of the G4 proteins, as well as three additional changes in which the $D1^{FR}$ proteins had one of two residues that differed from the G4-D1 residues at those positions. The majority of the altered residues are in the first three helices (for consensus, see Fig. 2 and full alignment, see Fig. S5). Within helix A, these proteins share deletion of a frequently observed Thr at position 40, and an insertion of Val before a conserved Phe and a characteristic Gly-Val-Ser motif between residues 43 and 45 (Fig. 4b). These residues occur in the vicinity of the bound ß-carotene and the accessory chlorophyll, Chl_{zD1}, that might serve as side-path electron donors in PS II under specific conditions (Cardona et al. 2012). Between helices A and B there is a Ser79 to Thr change also found in the $D1^{INT}$ sequences.

In the $D1^{FR}$ protein helix B, the His118 ligand of Chl_{zD1} and the putative Tyr126 ligand of $Pheo_{D1}$, are unaltered; however, several residues are altered between Leu114 and Val/Ile/Cys123 which may modify the properties of these cofactors (Fig. 4c, d). The $D1^{FR}$ sequences usually contain the substitution of Gln to Glu at position 130 which is characteristic of the G4 high-light form, D1:2. In addition, the $D1^{FR}$ sequences have the Ala156 to Ser change observed in the $D1^{INT}$ but Ala154 is changed to a Thr in this group which may further modify the efficiency of charge recombination (Fig. 4e, f) (Vinyard et al. 2014). It has been suggested that Thr154 and Tyr119 (instead of Phe) of $D1^{FR}$ may also have a hydrogen bond to the formyl group of chlorophyll *f* (Nürnberg et al. 2018). Between helices C and D, the $D1^{FR}$ Met172 to Leu and Leu174 to Met changes are found; these are located in a region separating the Mn_4CaO_5 cluster from Chl_{zD1} and P_{D1} of P680 (Kern et al. 2007). A

292 Phe184 change is also found in this region in $D1^{FR}$ sequences while in helix D there is a
293 Ser212 to Cys change (Fig. 4b).
294 Ho et al. (2016) and Shen et al. (2019) showed that the G1-D1 is required for the
295 production of chlorophyll *f*. A G1-*psbA* null mutant abolished chlorophyll *f* production in
296 both *Chlorogloeopsis fritschii* PCC 9212 and *Synechococcus* sp. PCC 7335, while
297 chlorophyll *f* could be produced in far-red light in the non FaRLiP strain, *Synechococcus* sp.
298 PCC 7002, when this strain contained a G1-encoding *psbA* gene. Chlorophyll *f* is present in
299 the reaction centres of both PS II and PS I (Ho et al. 2016; Nürnberg et al. 2018; Shen et al.
300 2019). In studies using isolated PS II centres of *Chroococcidiopsis thermalis* PCC 7202, the
301 isolated PS II appeared to contain the $D1^{FR}$ protein when subjected to far-red light (Nürnberg
302 et al. 2018).
303
304 **The G3 D1 group contains multiple subgroups**
305 The D1 phylogeny divided the G3 proteins into three well-supported subgroups (SH-like
306 aLRT > 0.9). Each subgroup contained proteins encoded by *psbA* genes that are up-regulated
307 under micro-aerobic conditions (Summerfield et al. 2008; Sicora et al. 2009) (Fig. 1 and Fig.
308 S1); these were *Nostoc* sp. PCC 7120 and *Cyanothece* sp. ATCC 51142 in subgroup I,
309 *Thermosynechococcus elongatus* BP-1 in subgroup II, and *Synechocystis* sp. PCC 6803 in
310 subgroup III. The separation of G3-D1 into these subgroups was also observed when these 64
311 sequences were analysed using the original outgroup or a representative sequence from each
312 of the other D1 groups to root the tree (Fig. S6 – S9). The G3 subgroups contain 33, 2 and 29
313 sequences, respectively (Fig. 1). The two main subgroups have alterations in the amino acids
314 that frequently contribute to the secondary ligand sphere of the OEC (highlighted in Fig. 5;
315 for full alignment of G3 protein sequences see Fig. S10 and Table S1).

316 The three characteristic amino acid changes of low-oxygen-induced *psbA* encoded
317 proteins identified by Sicora et al. (2009) (Gly80 to Ala, Phe158 to Leu and Thr286 to Ala
318 were in 61 of the 64 protein sequences in G3. However, the Gly80 to Ala substitution was not
319 in the G3 protein sequence from *Oscillatoria* sp. PCC 6506 or *Kamptonema formosum* PCC
320 6407 in subgroup I. All G3 sequences contained the Phe158 to Leu change, but the
321 *Geitlerinema* sp. PCC 7105 subgroup I sequence did not have the Thr286 to Ala change (Fig.
322 5 and Fig. S10).
323 In subgroup I, residues that differed to the 95% G4 consensus sequence in at least
324 90% of the sequences included: Leu41 to Ala (rarely Ile or Gly), Cys47 to Val (rarely Ala or
325 Thr) both in helix A, and in the a-b loop, both Ala81 to Thr and Ser85 to Thr. The Asn87
326 residue is replaced with an Ala in almost 80% of the subgroup I sequences; this Asn has been
327 reported to interact with a chloride-binding site associated with a proton exit channel for the
328 OEC (Banerjee et al. 2018, 2019). In addition, Asn87 may also interact with CP43-Glu354
329 and CP43-Arg357 through hydrogen bonding but these interactions would in all likelihood be
330 lost when the residue is Ala (Fig. 5b, c). Also in subgroup I (and subgroup II) a Pro to Met
331 change is observed at position 173 in the c-d loop; this substitution in *T. elongatus* has been
332 shown to affect oxidation of the redox active Tyr161 (Y_Z) and weaken the hydrogen bond
333 between Y_Z and His190 (Sugiura et al. 2014).
334 In subgroup III, residues that differed to the 95% G4 protein consensus sequence are
335 more frequently found between helix C and the C-terminus. Residues changed with respect to
336 the G4 sequence that are characteristic of this G3 subgroup include: Pro162 to Ser (rarely
337 Ala, in helix C), Phe186 to Leu in helix CD in the c-d loop, Ile192 to Val (also found in the
338 c-d loop of 8 out of 33 subgroup I sequences), as well as, Thr292 to Cys or Ser and Met293
339 to Phe in helix E and Ala336 to Val (Fig. 2).

340 Introduction of the Pro162 to Ser change found in D1′ in *Synechocystis* sp. PCC 6803
341 did not alter oxygen evolution; however, the F186L and F186L:P162S mutants exhibited
342 perturbed oxygen evolution and Q_A to Q_B electron transfer (Funk et al. 2001; Wiklund et al.
343 2001; Sicora et al. 2004). Phe186 is hydrogen bonded to His190 and Phe182 as part of a
344 putative hydrogen bond network involving several bound waters in the vicinity of Y_Z (Fig.
345 5d, e). Both Phe186 and Phe182 along with Met293 contribute to a hydrophobic pocket, as
346 previously noted, separating the OEC from P680 (Kern et al. 2007). The Met to Phe
347 substitution at position 293 likely disrupts hydrogen bonding involving Asn296 and
348 potentially Gln165. Asn296 and Gln165 of G4 are hydrogen bonded to oxygen atoms which
349 interact with the OEC.
350 The Ile at position 192 in G4 that becomes a Val in G3 subgroup III is located on the
351 lumenal side of the D1 protein, while no specific role for this residue could be ascertained in
352 silico, a I192F:N267I double mutant in *Synechocystis* sp. PCC 6803 prevented
353 photoautotrophic growth (Yamasato et al. 2002). The G4 Ala336 position that is a Val in
354 subgroup III is likely to interact with the OEC ligand, His337, and may interact with Asp61,
355 which binds the OEC through a water molecule (W567 in PDB 4UB6) (Fig. 5f, g).
356
357 **The G1 and G2 D1 proteins**
358 The G2 of Cardona et al. (2015) included the rogue and sentinel D1s described by Murray
359 (2012) and Wegener et al. (2015), respectively, these lack a number of key amino acids
360 required to support normal PS II function. In our extended analysis, 52 G2 D1 proteins were
361 identified: the additional sequences had the same donor and acceptor side changes reported
362 previously (Cardona et al. 2015) (Fig. 2) but three residues were no longer conserved across
363 all the G2 members (Glu65, His252 and Gly256).

In agreement with previous reports, none of the G2 members have the 341-344 Leu-Asp-Leu-Ala motif that is conserved in G4, $D1^{INT}$, $D1^{FR}$ and G3 (except one G3-D1 with a Leu341to Met change) on the N-terminal side of the CtpA cleavage site. The C-terminus was altered in four G2 D1 sequences from unicellular strains (*Cyanobacterium aponinum* strains and *Stanieria* spp.), these ended at position 343, in addition, 23 G2 sequences had an Ala344 to Ser change. The remainder of the strains (25) had Ala at position 344 with the number of amino acids following this residue varying from zero to 27 amino acids. This sequence variation in G2-D1 would be consistent with no processing of the C-terminus suggested by Wegener et al. (2015).

The G1 group of Cardona et al. (2015) contained eight protein sequences of the far-red-inducible chlorophyll *f* synthase, which catalyses the production of chlorophyll *f* (Chen et al. 2010) and was first identified by Murray (2012). The ligands necessary to bind the OEC, which are provided by Asp170, Glu189, His332, Glu333, Asp342 and Ala344 were absent or not conserved in the G1 category of proteins as previously reported by Cardona et al. (2015). The G1 sequences did retain other ligands necessary to bind PS II cofactors, e.g., His118 which provides the axial ligand to the accessory chlorophyll *a* (Chl_{zD1}), the residues binding pheophytin ($Pheo_{D1}$) at positions Thr126 and Glu130 and the axial ligand at His198 for the reaction centre chlorophyll P_{D1}, as well as the key Tyr161 (Yz) and His190 pairing on the donor side. However, the G1 sequences contain substitutions around the Chl_{zD1} binding site with all sequences having changes: Ile116 to Val, Phe117 to Leu, Leu121 to Ile and Ala123 to Ile. In the vicinity of the $Pheo_{D1}$ binding site, the G1 sequences included the changes Met127 to Gln and Gly128 to Asp (Fig. 2).

Purifying selection pressure within the *psbA* genes encoding the D1 protein family

388 The *psbA* genes encoding all the different D1 protein sequences are subject to similar,
389 relatively strong, purifying selection; this was similar to that observed for the gene *rbcL* that
390 encodes the Rubisco large subunit (Fig. 6). Of the seven groups, the G1 sequences exhibited
391 slightly more relaxed selection (mean $\omega = 0.071 \pm 0.045$). Genes encoding the $D1^{FR}$ and
392 $D1^{INT}$ proteins were found to be undergoing the highest amount of purifying selection (mean
393 $\omega = 0.020 \pm 0.013$ and mean $\omega = 0.026 \pm 0.003$, respectively). This may indicate that amino
394 acid changes in the mature D1 protein of all the D1 isoforms can either impair or retard the
395 performance of PS II, suggesting that this protein family is retaining amino acids critical to
396 their function: indicating that all of these proteins are likely to be physiologically relevant
397 (Fig. 6).
398
399 **Distribution of the *psbA* genes encoding the D1 protein family in cyanobacteria**
400 The 16S-23S rRNA gene phylogeny shows the relationship of the 206 cyanobacterial strains
401 used in this study. This phylogeny has been annotated with the type of D1 proteins found in
402 each strain along with the number of genes encoding each type of D1 (Fig. 7). The
403 cyanobacterial clades recovered in this analysis were compared to the previous analysis of
404 Shih et al. (2013) (Fig. 7). While the two analyses differed in that the analysis of Shih et al.
405 (2013) used 31 concatenated protein sequences to generate the species tree, both approaches
406 produced similar cyanobacterial groupings and therefore the clade annotation used in Fig. 7 is
407 the same as that used in Shih et al. (2013). All cyanobacterial genomes examined contain at
408 least one copy of a *psbA* gene encoding G4-D1 (either a D1:1 or D1:2 or both). It should be
409 noted that this analysis includes draft genomes and in some cases updated genomes may vary
410 (for example, in the contig assembly of *Fischerella* sp. PCC 9605, ALVT00000000, the
411 $D1^{INT}$ was not identified, but it was present in the scaffold assembly of these contigs
412 (KI912148-KI91254)).

413 The heterocystous cyanobacteria (subsection IV, Nostocales and subsection V,
414 Stigonematales) form group B1 (Fig. 7). The majority of the Stigonematales formed a
415 moderately supported subgroup within B1, these included cyanobacteria with *psbA* genes
416 encoding the largest number of D1 types. Genes encoding $D1^{INT}$ and G2-D1 were very
417 common in these strains and more than half had genes encoding G1-D1, G2-D1, G4-D1,
418 $D1^{FR}$ and $D1^{INT}$ whereas only two strains had genes encoding G3-D1. The rest of the B1
419 subgroup were predominately Nostocales strains and these had greater variation in *psbA* gene
420 diversity. Only four of the 54 Nostocales strains in this analysis contained genes coding for
421 G1-D1 and $D1^{FR}$ but genes encoding G2-D1, $D1^{INT}$ and G3-D1 where in 9, 10 and 16 strains,
422 respectively. The B1 strains had between 1 and 11 *psbA* copies encoding G4-D1, this
423 included the draft genomes of *Cylindrospermopis raciborskii* strains CENA302 and ITEP-A1
424 which each had 11 copies. In addition, the draft genome of *Fischerella* sp. PCC 9605 had
425 nine copies and the draft genomes of *C. raciborskii* MVCC14, *Leptolyngbya* Heron Island J
426 and *Nostoc* NIES-403 each had eight copies of *psbA* encoding G4-D1. Several Nostocales
427 strains (26 strains) had only genes coding for G4-D1, this included the obligate symbionts
428 *Nostoc azollae* 0708, *Richelia intracellularis* HM01 and *Richelia intracellularis* HH01 and
429 also free-living strains from marine, freshwater and terrestrial environments. Some of these
430 strains contained only genes for D1:1 or D1:2, although most strains contained both.
431 There is a striking decrease in diversity of *psbA* genes in the filamentous non-
432 heterocystous cyanobacteria in A1 and B2a groups compared to the heterocystous
433 cyanobacteria. More than half of the A1 and B2a strains (14/24) contain only genes encoding
434 G4-D1; in addition, genes encoding G3-D1 and G2-D1 were found in ten strains and one
435 strain, respectively. Sister to these is a moderately supported group, B2b, that contains
436 unicellular and filamentous cyanobacteria, the majority (>70%) of these strains have genes
437 coding for at least two D1 types. Similar to groups A1 and B2a, the gene encoding G3-D1 is

438 common, being present in half these strains; in contrast, many more of the strains (~40%)
439 have genes encoding G2-D1 but only two strains have the FaRLiP gene cluster.

440 The well-supported group C1 includes members of the *Prochlorococcus* genus, these
441 strains have contracted genomes relative to other cyanobacteria and inhabit the nutrient poor,
442 oligotrophic oceans (Scanlan et al. 2009). This genus utilises a range of light-inducible
443 proteins for photoprotection (Rocap et al. 2003), which may result in a reduced reliance on
444 D1:2 to reduce the rates of photoinhibition, consistent with these strains having one to three
445 copies of *psbA* encoding the G4 D1:1 protein (Mella-Flores et al. 2012). Sister to the
446 *Prochlorococcus* subgroup is a well-supported subgroup of marine *Synechococcus* strains;
447 these have genes encoding both G4 isoforms and sister to this is a smaller group containing
448 four unicellular strains that each contain *psbA* genes encoding G3-D1. The C2 subgroup
449 contains three *Synechococcus* strains with genes for both the G4 D1 proteins. In contrast, the
450 well-supported subgroup C3 contains both unicellular and filamentous cyanobacteria and
451 these exhibit variation in their *psbA* diversity, all containing genes for G4-D1 (both D1:1 and
452 D1:2) and for up to three other D1 types, including two strains containing the FaRLiP cluster:
453 *Synechococcus* sp. PCC 7335 and *Halomicronema hongdechloris* C2206. The subgroups E,
454 D and F contain cyanobacteria with genes encoding G4-D1 alone or in combination with G3-
455 D1 (14 strains), with the exception of two strains with the FaRLiP gene cluster
456 (*Oscillatoriales* sp. JSC-12 and *Leptolyngbya* sp. JSC-1), one of which also has the gene
457 encoding $D1^{INT}$ (*Leptolyngbya* sp. JSC-1). The hot-spring-inhabiting *Synechococcus* spp. JA-
458 3-3Ab and JA-2-3B'a (2-13) (subgroup G, Fig. 7) are amongst the most deeply branching
459 cyanobacteria identified (Shih et al. 2013; Li et al. 2014; Sánchez-Baracaldo et al. 2017;
460 Moore et al. 2019) and have genes encoding G4 D1:2 and G2-D1.

461

462 **Potential roles for the *psbA* gene family in cyanobacteria**

463 The grouping of D1 proteins did not follow the topology of the 16S-23S rRNA gene
464 phylogeny (Fig. 1 and Fig. 7). The D1 phylogeny showed six groups of D1 proteins, and the
465 16S-23S rRNA phylogeny annotated with the distribution of the D1 protein types indicates
466 the presence of the different D1 types in strains across the phylogeny. Well-supported groups
467 of closely related strains tend to have similar D1 protein complements, suggesting different
468 cyanobacterial lineages have retained and lost specific D1 types. More than half the strains
469 (106/206) had at least one D1 type in addition to G4-D1, with ~30% and ~14% of all strains
470 having one or two additional D1 types, respectively. Furthermore, ~8% of strains have three
471 or more D1 types in addition to G4-D1. Out of the 100 strains with only G4-D1 proteins, 43
472 strains have genes encoding both D1:1 and D1:2 proteins, 31 have only D1:1 proteins and 26
473 have only D1:2 proteins. Only ~10% of the cyanobacterial strains had a single copy of *psbA*
474 but it should be noted that many of these are draft genomes. We interpret the prevalence of
475 different D1 types and multiple copies of the same D1 type in most cyanobacterial strains to
476 be indicative of a selective advantage to maintaining these copies, although the function of
477 some D1 types is not clear.
478 Microenvironments occupied by the cyanobacteria may have led to the retention of
479 different D1 types: for example, Gan and Bryant (2015) suggested the far-red-inducible gene
480 cluster may confer an advantage when green light is either scattered or absorbed by the
481 environment or competing photoautotrophic organisms are present. In this our analysis, a
482 phylogenetically diverse collection of cyanobacteria had the FaRLiP cluster and these were
483 isolated from environments that had the potential to be competitive for light. For example, a
484 niche for chlorophyll *f*-containing cyanobacteria was identified below the surface of a hot
485 spring microbial mat where only wavelengths of light > 700 nm remained (Ohkubo and
486 Miyashita 2017) and eleven strains with the FaRLiP cluster were isolated from hot springs.
487 Two strains were isolated from associations with other phototrophs: one as an endophyte of a

488 red alga and one from a stromatolite. In addition, two strains were from soil and one from a
489 sphagnum bog and all of these environments have potential to be far-red light enriched (Gan
490 and Bryant 2015).

491 The $D1^{INT}$ protein was found predominately in heterocystous cyanobacteria, but only
492 a third of heterocystous strains had this protein. Both Nostocales and Stigonmatales strains
493 contained the gene encoding $D1^{INT}$ along with three additional non-heterocystous strains. The
494 three strains were: the unicellular *Gloeocapsa* sp. PCC 7428 which also has four genes
495 encoding D1:2 copies and was isolated from a hot spring; and the filamentous strains
496 *Leptolyngbya* JSC-1 isolated from a hot spring and *Halomicronema hongdechloris* C2206
497 isolated from a stromatolite, both of these filamentous strains also have the FaRLiP cluster.
498 The $D1^{INT}$ was found in a similar number of strains as the FaRLiP gene cluster. There was no
499 clear pattern of co-occurrence with other *psbA* genes; however, 24 of the 27 strains had at
500 least three D1 types.

501 The G3-D1 protein was in a phylogenetically broad range of cyanobacteria that
502 represented about ~30% of the strains in this analysis. The gene encoding G3-D1 is up-
503 regulated under low oxygen in several cyanobacterial strains (Summerfield et al. 2008; Sicora
504 et al. 2009). Cardona et al. (2018) estimate the G3-D1 to have evolved around the time of the
505 Great Oxidation Event (GOE) branching slightly before G4-D1, raising the possibility that
506 these genes evolved under low-oxygen conditions and were down-regulated in the presence
507 of oxygen. This regulation has been demonstrated in a *Synechocystis* sp. PCC 6803 strain
508 containing only the low oxygen expressed *psbA* gene (Summerfield et al. 2008; Crawford et
509 al. 2016). The *psbA* genes encoding G3-D1 are under relatively strong purifying selection in
510 both diazotrophic and non-diazotrophic strains indicating a current physiological function.
511 Low-oxygen conditions are also associated with the up-regulation of genes encoding other
512 components of the photosynthetic electron transport chain (Summerfield et al. 2008) In

addition, under low oxygen G3-D1 PS II centres were less susceptible to photoinhibition than G4-D1 PS II centres in *Synechocystis* sp. PCC 6803 (Crawford et al. 2016).

The G2-D1 protein has been suggested to be involved in protecting nitrogenase in unicellular diazotrophs (Wegener et al. 2015). Of the strains analysed from the unicellular diazotrophs *Crocosphaera watsonii* and *Cyanothece* spp., most have genes encoding G2-D1 except *Cyanothece* sp. CCY 0110, for which only a draft genome was available and therefore data may be missing, and *Cyanothece* sp. PCC 7425 for which a complete genome was available. Unlike the other *Cyanothece* strains, *Cyanothece* sp. PCC 7425 is not an aerobic diazotroph and has been identified as belonging to the Synechococcales based on thylakoid structure and molecular phylogenetic analysis (Mares et al. 2019). The presence of G2-D1 in these strains is consistent with subjective dark detection of low levels of G2-D1-containing PS II centres in *Crocosphaera watsonii* WH8501 (Masuda et al. 2018) and the suggestion G2-D1-containing PS II centres have a role in the temporal regulation of diazotrophy and photosynthesis (Wegener et al. 2015; Masuda et al. 2018; Sicora et al. 2019).

Genes encoding G2-D1 were identified in genomes of heterocystous, filamentous non-heterocystous and unicellular strains: most of which have been demonstrated to be nitrogen fixing or have the *nif* gene cluster but a further seven strains had the G2-D1-encoding gene but did not have genes encoding nitrogenase. Strains containing *psbA* encoding G2-D1 were members of the orders: Chroococcales, Pleurocapsales, Chroococcidiopsidales, Synechococcales, Oscillatoriales, Nostocales and Stigonematales. The wide distribution of *psbA* encoding G2-D1 in strains that employ different strategies for separating photosynthesis and nitrogen fixation appears to indicate additional roles for G2-D1-containing PS II centres. In our analysis, 22 of the 71 heterocystous strains contained a gene coding for G2-D1, these strains would not require G2-D1 PS II centres to protect nitrogenase as PS II and nitrogenase would be spatially separated. Several unicellular and

538 filamentous diazotrophs lack the gene, including: *Xenococcus* sp. PCC 9228, *Pseudanabaena*
539 sp. PCC 6802, *Microcoleus* sp. PCC 7113, *Trichodesmium erythraeum* IMS101, and *Lyngbya*
540 sp. PCC 8106. The distribution of the gene encoding G2-D1 included absence from some
541 non-heterocystous diazotrophs, and presence in some heterocystous strains and a small
542 number of non-diazotrophic strains indicate additional or alternative roles of G2-D1. In total,
543 a quarter of strains in our analysis had the *psbA* that encoded G2-D1 and it has been shown to
544 be upregulated in *Anabaena variabilis* ATCC 29413 in heterotrophically grown filaments
545 (Park et al. 2013) consistent with a physiological role for this isoform.
546 We propose all six different copies of D1 may confer selective advantages in specific
547 microhabitats. Furthermore, carrying a large suite of D1 proteins might impart a competitive
548 advantage in a fluctuating environment and may explain the diversity of D1 proteins in some
549 cyanobacterial strains.
550
551 **Conclusions**
552 Our analysis of the D1 family members and their distribution in cyanobacteria has identified
553 a phylogenetically distinct D1 group; this contains two subgroups: $D1^{FR}$ and $D1^{INT}$. The
554 genes encoding these proteins were under similar selective pressure to the genes encoding
555 other types of D1. The $D1^{INT}$ protein has the ligands necessary to bind the OEC and was
556 found in a phylogenetically diverse range of cyanobacteria but predominantly in
557 heterocystous cyanobacteria and this was in about one third of the heterocystous strains. The
558 gene encoding the $D1^{FR}$ protein was part of the FaRLiP cluster, which also contains a gene
559 encoding the enzymatic form of D1 — the G1, chlorophyll *f* synthase. The $D1^{FR}$ protein has
560 the ligands necessary to bind the OEC and several amino acid changes that might be
561 associated with binding of chlorophyll *f*, rather than chlorophyll *a*, consistent with its
562 involvement in the far-red light acclimation process. Furthermore, the previously identified

G3-D1 group was shown to contain three subgroups. Subgroup I had changes predominately towards the N-terminus of the D1 protein whereas subgroup III had most variation from the G4 consensus towards the C-terminus. In this analysis ~30% of cyanobacteria contained a gene encoding one of these two G3-D1 subgroups.

The gene encoding G2-D1 was found in 25% of cyanobacteria, many of which, but not all, are diazotrophic strains. However, many diazotrophic strains (both unicellular and filamentous) do not contain genes encoding G2-D1. Each group of D1 proteins was found in a phylogenetically diverse range of cyanobacteria consistent with ancestral cyanobacteria having multiple copies of D1. The filamentous heterocystous cyanobacteria tended to have more D1 types, perhaps reflecting an enhanced capacity to adapt to changing environmental conditions. These analyses support that distinct D1 types confer a selective advantage under specific conditions that has led to their retention in a phylogenetically diverse range of cyanobacteria.

Acknowledgements

The authors would like to acknowledge Andy Nilsen for the valuable discussions while creating the phylogenetic trees and Bronwyn Carlisle for helping to finalise the figures for publication. KJS is supported by a University of Otago Division of Sciences PhD Scholarship. Additional funding was provided by a University of Otago research grant to TCS.

Compliance with ethical standards

Conflict of interest

588 The authors declare that they have no conflict of interest.

589

590

591 **Additional information**

592 The data reported in this paper have come from genomes deposited in both the Genbank and
593 JGI databases (accession nos. CP000117, CP000393, CP003614, CP003620, CP003642,
594 CP006269, CP006270, CP006271, CP006471, CP006882, CP007203, CP007542, CP007753,
595 CP007754, CP011304, CP011382, CP011456, CP011941, CP012036, CP012375, CP013008,
596 CP013998, CP016474, CP016483, CP017599, CP017675, CP017708, CP018091, CP018344,
597 CP018345, CP018346, CP019636, CP020771, CP021983, FO818640, Ga0010025,
598 Ga0012361, Ga0012362, Ga0014323, Ga0025054, Ga0025357, Ga0025386, Ga0025408,
599 Ga0026686, Ga0064116, Ga0064117, Ga0078583, Ga0079976, Ga0166459, NC_003272,
600 NC_004113, NC_005042, NC_005070, NC_005071, NC_005072, NC_006576, NC_007335,
601 NC_007513, NC_007516, NC_007577, NC_007604, NC_007775, NC_007776, NC_008319,
602 NC_008816, NC_008817, NC_008819, NC_009091, NC_009481, NC_009482, NC_009840,
603 NC_009976, NC_010296, NC_010475, NC_010546, NC_010628, NC_011726, NC_011729,
604 NC_011884, NC_013161, NC_014248, NC_014501, NC_019427, NC_019675, NC_019676,
605 NC_019678, NC_019680, NC_019682, NC_019683, NC_019684, NC_019689, NC_019693,
606 NC_019695, NC_019697, NC_019701, NC_019702, NC_019703, NC_019738, NC_019745,
607 NC_019748, NC_019751, NC_019771, NC_019776, NC_019779, NC_019780, NC_020286,
608 NC_022600, NC_023033, AAVU00000000, AAXW00000000, ABRV00000000,
609 ABRS00000000, ABSE00000000, ABYK00000000, ACYA00000000, AGCR00000000,
610 AGIZ00000000, AJLJ00000000, AJLK00000000, AJLL00000000, AJLM00000000,
611 AJLN00000000, AJWF00000000, ALVI00000000, ALVJ00000000, ALVK00000000,
612 ALVL00000000, ALVP00000000, ALVQ00000000, ALVR00000000, ALVS00000000,

ALVT00000000, ALVW00000000, ALVX00000000, ALVY00000000, ALVZ00000000, ALWB00000000, ALWD00000000, ANFJ00000000, ANFQ00000000, ANNX00000000, AP014638, AP014642, AP014815, AP014821, AP017295, AP017308, AP017367, AP017375, AP017959, AP018172, AP018174, AP018178, AP018180, AP018184, AP018194, AP018203, AP018207, AP018222, AP018227, AP018233, AP018248, AP018254, AP018255, AP018268, AP018280, AP018281, AP018288, AP018298, AP018307, AP018316, AP017305, AUZM00000000, AVFS00000000, AWNH00000000, BDUC00000000, CACA00000000, CAIS00000000, CAIY00000000, CM001632, CM001775, CM001776, CZCT00000000, CZCU00000000, CZDF00000000, JMKF00000000, JQFA00000000, JTHE00000000, JXCB00000000, JYON00000000, LIRN00000000, LMTZ00000000, LNDC00000000, LT578417, LUBZ00000000, LUHI00000000, MBQX00000000, MBQY00000000, MKZR00000000, MKZS00000000, MQTZ00000000, MRBY00000000, MRCA00000000, MRCB00000000, MTPU00000000, NXIB00000000, PEBC00000000).

628 **References**

629

630 Ago H, Adachi H, Umena Y Tashiro T, Kawakami K, Kamiya N, Tian L, Han G, Kuang T,
631 Liu Z, Wang F, Zou H, Enami I, Miyano M, Shen JR (2016) Novel features of eukaryotic
632 Photosystem II revealed by its crystal structure analysis from a red alga. J Biol Chem 291:
633 5676–5687. https://doi.org/10.1074/jbcM115.711689

634

635 Banerjee G, Ghosh I, Kim CJ, Debus RJ, Brudvig GW (2018) Substitution of the D1-Asn87
636 site in Photosystem II of cyanobacteria mimics the chloride-binding characteristics of spinach
637 Photosystem II. J Biol Chem 293: 2487–2497.
638 https://doi.org /10.1074/jbc.M117.813170

639

640 Banerjee G, Ghosh I, Kim CJ, Debus RJ, Brudvig GW (2019) Bicarbonate rescues damaged
641 proton-transfer pathway in Photosystem II. BBA – Bioenergetics 1860: 611–617.
642 https://doi.org /10.1016/j.bbabio.2019.06.014

643

644 Bertoni M, Kiefer F, Biasini M, Bordoli L, Schwede T (2017) Modeling protein quaternary
645 structure of homo- and hetero-oligomers beyond binary interactions by homology. Sci Rep 7**:**
646 10480.
647 https://doi.org/10.1038/s41598-017-09654-8

648

649 Benson DA, Cavanaug M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW
650 (2017) Genbank. Nucleic Acids Res 45: 37–42.
651 https://doi.org/10.1093/nar/gks1195

652

653

654 Bienert S, Waterhouse A, de Beer TAP, Tauriello G, Studer G, Bordoli L, Schwede T (2017)
655 The SWISS-MODEL Repository - new features and functionality. Nucleic Acids Res 45:
656 313–319.
657 https://doi.org/10.1093/nar/gkw1132
658
659 Cardona T, Sedoud A, Cox N, Rutherford AW (2012) Charge separation in Photosystem II:
660 A comparative and evolutionary overview. BBA – Bioenergetics 1817: 26–43.
661 https://doi:10.1016/j.bbabio.2011.07.012
662
663 Cardona T, Murray JW, Rutherford A.W (2015) Origin and evolution of water oxidation
664 before the last common ancestor of the cyanobacteria. Mol Biol Evol 32: 1310–1328.
665 https://doi.org/10.1093/molbev/msv024
666
667 Chen M, Schliep M, Willows RD, Cai ZL, Neilan BA, Scheer H (2010) A red-shifted
668 chlorophyll. Science 329: 1318–1319.
669 https://doi.org/10.1126/science.1191127
670
671 Chen M, Li Y, Birch D, Willows RD (2012) A cyanobacterium that contains chlorophyll *f* – a
672 red-absorbing photopigment. FEBS Lett 586: 3249–3254.
673 https://doi.org/10.1016/j.febslet.2012.06.045
674
675 Chen M, Hernandez-Prieto MA, Loughlin PC, Li Y, Willows RD (2019): Genome and
676 proteome of the chlorophyll *f*-producing cyanobacterium *Halomicronema hongdechloris*:
677 adaptative proteomic shifts under different light conditions. BMC Genomics 20: 207.

https://doi.org/10.1186/s12864-019-5587-3

Crawford TS, Hanning KR, Chua JP, Eaton-Rye JJ, Summerfield TC (2016) Comparison of D1′-and D1-containing PS II reaction centre complexes under different environmental conditions in *Synechocystis* sp. PCC 6803. Plant Cell Environ 39: 1715–1726. https://doi.org/10.1111/pce.12738

Darriba D, Taboada G.L, Doallo R, Posada D (2012) jModelTest 2 more models new heuristics and parallel computing. Nat Methods 9: 772. https://doi.org/10.1038/nmeth.2109

DeLano WL (2002) Pymol: An open-source molecular graphics tool. CCP4 Newsletter On Protein Crystallography 40: 82–92. https://www.ccp4.ac.uk/newsletters/newsletter40/11_pymol.pdf. Accessed 1 June 2018

DeLano WL (2009) PyMOL molecular viewer Updates and refinements. In: Abstracts of Papers of the American Chemical Society (Vol. 238). American Chemical Society, Washington DC.

El Bissati K, Kirilovsky D (2001) Regulation of *psbA* and *psaE* expression by light quality in *Synechocystis* species PCC 6803. A redox control mechanism. Plant Physiol 125: 1988–2000. https://doi.org/10.1104/pp.125.4.1988

Endo K, Kobayashi K, Wang H-T, Chu H-A, Shen J-R, Wada H (2019) Site-directed mutagenesis of two amino acid residues in cytochrome b_{559} α subunit that interact with a

703 phosphatidylglycerol molecule (PG772) induces quinone-dependent inhibition of
704 Photosystem II activity. Photosynth Res 139: 267–279.
705 https://doi.org/10.1007/s11120-018-0555-3
706
707 Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap.
708 Evolution 39: 783–791.
709 https://doi.org/10.1111/j.1558-5646.1985.tb00420.x.
710
711 Ferreira KN, Iverson TM, Maghlaoui K, Barber J, Iwata S (2004) Architecture of the
712 photosynthetic oxygen-evolving center. Science 303: 1831–1838.
713 https://doi.org/10.1126/science.1093087
714
715 Fletcher W, Yang Z (2010) The effect of insertions deletions and alignment errors on the
716 branch-site test of positive selection. Mol Biol Evol 27: 2257–2267.
717 https://doi.org/10.1093/molbev/msq115
718
719 Gan F, Zhang S, Rockwell NC, Martin SS, Lagarias JC, Bryant DA (2014) Extensive
720 remodeling of a cyanobacterial photosynthetic apparatus in far-red light. Science 345: 1312–
721 1317.
722 https://doi.org/10.1126/science.1256963
723
724 Gan F, Bryant DA. 2015. Adaptive and acclimative responses of cyanobacteria to far-red
725 light. Env Microbiol 17: 3450–3465.
726 https://doi.org/10.1111/1462-2920.12992
727

728 Gan F, Shen G, Bryant DA (2015) Occurrence of far-red light photoacclimation (FaRLiP) in
729 diverse cyanobacteria. Life 5: 4–24.
730 https://doi.org/10.3390/life5010004
731
732 Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky
733 S, Nikitin, R, Ohm, RA, Otillar R, Poliakov A, Ratnere I, Riley R, Smirnova T, Rokhsar D,
734 Dubchak I (2012) The genome portal of the department of energy joint genome institute.
735 Nucleic Acids Res 40: 26–32.
736 https://doi.org/10.1093/nar/gkt1069
737
738 Grim SL, Dick GJ (2016) Photosynthetic versatilityin the genome of *Geitlerinema* sp. PCC
739 9228 (Formerly *Oscillatoria limnetica* 'Solar Lake'), a model anoxygenic photosynthetic
740 cyanobacterium. Front Microbiol 7: 26–32.
741 https://doi.org/10.3389/fmicb.2016.01546
742
743 Guex N, Peitsch MC, Schwede T (2009) Automated comparative protein structure modeling
744 with SWISS-MODEL and Swiss-PdbViewer A historical perspective. Electrophoresis 30:
745 162–173.
746 https://doi.org/10.1002/elps.200900140.
747
748 Hakovirta JR, Prezioso S, Hodge D, Pillai SP, Weigel LM (2016) Identification and analysis
749 of informative single nucleotide polymorphisms in 16S rRNA gene sequences of the *Bacillus*
750 *cereus* group. J Clin Microbiol 54: 2749–2756.
751 https://doi.org/10.1128/JCM.01267-16
752

753 Hilton JA, Meeks JC, Zehr JP (2016) Surveying DNA Elements within functional genes of
754 heterocyst-forming cyanobacteria. PLOS ONE 11: e0156034.
755 https://doi.org/10.1371/journal.pone.0156034
756
757 Ho MY, Shen G, Canniffe DP, Zhao C, Bryant DA (2016) Light-dependent chlorophyll *f*
758 synthase is a highly divergent paralog of PsbA of Photosystem II. Science 353: aaf9178.
759 https://doi.org/10.1126/science.aaf9178
760
761 Ho MY, Bryant DA (2019) Global transcriptional profiling of the cyanobacterium
762 *Chlorogloeopsis fritschii* PCC 9212 in far-red light insights into the regulation of chlorophyll
763 *d* synthesis. Front Microbiol 10: 465.
764 https://doi.org/10.3389/fmicb.2019.00465
765
766 Ho MY, Niedzwiedzki DM, MacGregor-Chatwin C, Gerstenecker G, Hunter CN,
767 Blankenship RE, Bryant DA (2019) Extensive remodeling of the photosynthetic apparatus
768 alters energy transfer among photosynthetic complexes when cyanobacteria acclimate to far-
769 red light. BBA–Bioenergetics: 148064.
770 https://doi.org/10.1016/j.bbabio.2019.148064
771
772 Hongo JA, Castro GM, Cintra LC, Zerlotini A, Lobo FP (2015) POTION an end-to-end
773 pipeline for positive Darwinian selection detection in genome-scale data through
774 phylogenetic comparison of protein-coding genes. BMC Genomics 16: 567.
775 https://doi.org/10.1186/s12864-015-1765-0
776

Jaspers E, Overmann J (2004) Ecological significance of microdiversity: identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. Appl Environ Microbiol 70: 4831–4839. https://doi.org/10.1128/AEM.70.8.4831-4839.2004

Kern J, Biesiadka J, Loll B, Saenger W, Zouni A (2007) Structure of the Mn_4-Ca cluster as derived from X-ray diffraction. Photosynth Res 92: 389–405. https://doi.org/10.1007/s11120-007-9173-1

Kern J, Chatterjee R, Young ID et al. (2018) Structures of the intermediates of Kok's photosynthetic water oxidation clock. Nature 563: 421–425. https://doi.org/10.1038/s41586-018-0681-2

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948. https://doi.org/10.1093/bioinformatics/btm404

Le SQ, Gascuel O. (2008) An Improved General Amino Acid Replacement Matrix. Mol Biol Evol 25: 1307–1320. https://doi.org/10.1093/molbev/msn067

Li B, Lopes JS, Foster PG, Embley TM, Cox CJ (2014) Compositional biases among synonymous substitutions cause conflict between gene and protein trees for plastid origins. Mol Biol Evol 31: 1697–1709.
https://doi.org/10.1093/molbev/msu105

Mares J, Johansen JR, Hauer T, Zima J Jr, Ventura S, Cuzman O, Tiribilli B, Kastovsky J (2019). Taxonomic resolution of the genus *Cyanothece* (Chroococcales, Cyanobacteria), with a treatment on *Gloeothece* and three new genera, *Crocosphaera, Rippkaea*, and *Zehria*. J Phycol 55: 578–610.
https://doi.org/10.1111/jpy.12853

Masuda T, Bernát G, Bečková M, Kotabová E, Lawrenz E, Lukeš M, Komenda J, Prášil O (2018) Diel regulation of photosynthetic activity in the oceanic unicellular diazotrophic cyanobacterium *Crocosphaera watsonii* WH8501. Environ Microbiol 20: 546–560.
http://doi.org/10.1111/1462-2920.13963

Mella-Flores D, Six C, Ratin M, Partensky F, Boutte C, Le Corguillé G, Blot N, Gourvil P, Kolowrat C, Garczarek L, Marie D (2012) *Prochlorococcus* and *Synechococcus* have evolved different adaptive mechanisms to cope with light and UV stress. Front Microbiol 3: 285.
https://doi.org/10.3389/fmicb.2012.00285

Moore KR, Magnabosco C, Momper L, Gold DA, Bosak T, Fournier GP (2019) An expanded ribosomal phylogeny of cyanobacteria supports a deep placement of plastids. Front Microbiol 10: 1612.
https://doi.org/10.3389/fmicb.2019.01612

826

827 Mulo P, Sicora C, Aro EM (2009) Cyanobacterial *psbA* gene family optimization of oxygenic
828 photosynthesis. Cell Mol Life Sci 66: 3697.
829 https://doi.org/10.1007/s00018-009-0103-6

830

831 Mulo P, Sakurai I, Aro EM (2012) Strategies for *psbA* gene expression in cyanobacteria
832 green algae and higher plants from transcription to PSII repair. BBA–Bioenergetics 1817:
833 247–257.
834 https://doi.org/10.1016/j.bbabio.2011.04.011

835

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

839

840 Narusaka Y, Murakami A, Saeki M, Kobayashi H, Satoh K (1996) Preliminary
841 characterization of a photo-tolerant mutant of *Synechocystis* sp. PCC 6803 obtained by in
842 vitro random mutagenesis of *psbA2*. Plant Sci 115: 261–266.
843 https://doi.org/10.1016/0168-9452(96)04393-2

844

845 Narusaka Y, Narusaka M, Satoh K, Kobayashi H (1999) In vitro random mutagenesis of the
846 D1 protein of the Photosystem II reaction center confers phototolerance on the
847 cyanobacterium *Synechocystis* sp. PCC 6803. J Biol Chem 274: 23270–23275.
848 http://doi.org/10.1074/jbc.274.33.23270

849

850

851 Nordberg H, Cantor M, Dusheyko S, Hua S, Polakov A, Shabalov I, Smirnova T, Grigorie
852 IV, Dubchak I (2014) The genome portal of the department of energy joint genome institute
853 2014 updates. Nucleic Acids Res 42: 26–31.
854 https://doi.org/10.1093/nar/gkt1069
855
856 Nürnberg DJ, Morton J, Santabarbara S, Telfer A, Joliot P, Antonaru LA, Ruban AV,
857 Cardona T, Krausz E, Boussac A, Fantuzzi A, Rutherford AW (2018) Photochemistry
858 beyond the red limit in chlorophyll *f*–containing photosystems. Science 360: 1210–1213.
859 https://doi.org/10.1126/science.aar8313
860
861 Ohkubo S, Miyashita H (2017) A niche for cyanobacteria producing chlorophyll f within a
862 microbial mat. The ISME Journal 11: 2368–2378.
863 https://doi.org/10.1038/ismej.2017.98
864
865 Park J-J, Lechno-Yossef S, Wolk CP, Vieille C (2013) Cell-specific gene expression in
866 *Anabaena variabilis* grown phototrophically, mixotrophically, and heterotrophically. BMC
867 Genomics 14: 759.
868 https://doi.org/10.1186/1471-2164-14-759
869
870 Partensky F, Six C, Ratin M, Garczarek L, Vaulot D, Probert I, Calteau A, Gourvil P, Marie
871 D, Grébert T, Bouchier C (2018) A novel species of the marine cyanobacterium
872 *Acaryochloris* with a unique pigment content and lifestyle. Sci Rep 8: 9142.
873 https://doi.org/10.1038/s41598-018-27542-7
874

875 Ponce-Toledo RI, Deschamps P, López-García P, Zivanovic Y, Benzerara K, David Moreira
876 D. 2017. An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids. Curr Biol
877 27: 386-391.
878 http://dx.doi.org/10.1016/j.cub.2016.11.056
879
880 Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO (2013)
881 The SILVA ribosomal RNA gene database project improved data processing and web-based
882 tools. Nucleic Acids Res 41: 590–596.
883 https://doi.org/10.1093/nar/gks1219
884
885 Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, Ahlgren NA, Arellano A, Coleman
886 M, Hauser L, Hess WR, Johnson ZI (2003) Genome divergence in two *Prochlorococcus*
887 ecotypes reflects oceanic niche differentiation. Nature 424: 1042.
888 https://doi.org/10.1038/nature01947
889
890 Sánchez-Baracaldo P, Raven JA, Pisani D, Knoll AH (2017) Early photosynthetic
891 eukaryotes inhabited low-salinity habitats. Proc Natl Acad Sci USA 114: 7737–7745.
892 https://doi.org/10.1073/pnas.1620089114
893
894 Saw JH, Schatz M, Brown MV, Kunkel DD, Foster JS, Shick H, Christensen S, Hou S, Wan
895 X, Donachie SP (2013) Cultivation and complete genome sequencing of *Gloeobacter*
896 *kilaueensis* sp. nov, from a lava cave in Kīlauea Caldera Hawai'i. PLOS ONE 8: e76376.
897 https://doi.org/10.1371/journal.pone.0076376
898

899 Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR, Post AF,
900 Hagemann M, Paulsen I., Partensky F (2009) Ecological genomics of marine
901 picocyanobacteria. Microbiol Mol Biol R 73: 249–299.
902 https://doi.org/10.1128/MMBR.00035-08
903
904 Shen J-R (2015) The structure of Photosystem II and the mechanism of water oxidation in
905 photosynthesis. Annu Rev Plant Biol 66: 23–48.
906 https://doi.org/10.1146/annurev-arplant-050312-120129
907
908 Shen G, Canniffe DP, Ho MY, Kurashov V, van der Est A, Golbeck JH, Bryant DA (2019)
909 Characterization of chlorophyll *f* synthase heterologously produced in *Synechococcus* sp.
910 PCC 7002. Photosynth Res 140: 1–16.
911 https://doi.org/10.1007/s11120-018-00610-9
912
913 Shih PM, Wu D, Latifi A, Axen SD, Fewer DP, Talla E, Calteau A, Cai F, De Marsac NT,
914 Rippka R, Herdman M (2013) Improving the coverage of the cyanobacterial phylum using
915 diversity-driven genome sequencing. Proc Natl Acad Sci USA 110: 1053–1058.
916 https://doi.org/10.1073/pnas.1217107110
917
918 Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with
919 applications to phylogenetic inference. Mol Biol Evol 16: 1114–1116.
920 https://doi.org/10.1093/oxfordjournals.molbev.a026201
921
922 Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial
923 extension (CE) of the optimal path. Prot Eng 11: 739–47.

https://doi.org/10.1093/protein/11.9.739

Sicora C., Ho FM, Salminen T, Styring S, Aro EM (2009) Transcription of a "silent" cyanobacterial *psbA* gene is induced by microaerobic conditions. BBA–Bioenergetics 1787: 105–112.
https://doi.org/10.1016/j.bbabio.2008.12.002

Sicora CI, Chiș I, Chiș C, Sicora O (2019) Regulation of PSII function in *Cyanothece* sp. ATCC 51142 during a light–dark cycle. Photosynth Res 139: 461–473.
https://doi.org/10.1007/s11120-018-0598-5

Stamatakis A (2006) RAxML-VI-HPC maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–2690.
https://doi.org/10.1093/bioinformatics/btl446

Suga M, Akita F, Hirata K, Ueno G, Murakami H, Nakajima Y, Shimizu T, Yamashita K, Yamamoto M, Ago H, Shen JR (2015) Native structure of Photosystem II at 1.95 Å resolution viewed by femtosecond X-ray pulses. Nature 517: 99–103.
https://doi.org/10.1038/nature13991

Suga M, Akita F, Sugahara M et al. (2017) Light-induced structural changes and the site of O=O bond formation in PS II caught by XFEL. Nature 543: 131–135.
https://doi.org/10.1038/nature21400

948 Sugiura M, Ozaki Y, Nakamura M, Cox N, Rappaport F, Boussac A (2014) The D1-173
949 amino acid is a structural determinant of the critical interaction between D1-Tyr161 (Tyr_Z)
950 and D1-His190 in Photosystem II. BBA – Bioenergetics 1837: 1922–1931.
951 https://doi.org/10.1016/j.bbabio.2014.08.008
952
953 Summerfield TC, Toepel J, Sherman LA (2008) Low-oxygen induction of normally cryptic
954 *psbA* genes in cyanobacteria. Biochemistry 47: 12939–12941.
955 https://doi.org/10.1021/bi8018916
956
957 Swofford DL (2001) Paup*: Phylogenetic analysis using parsimony (and other methods) 4.0.
958 B5.
959
960 Toepel J, Welsh E, Summerfield TC, Pakrasi HB, Sherman LA (2008) Differential
961 transcriptional analysis of the cyanobacterium *Cyanothece* sp. strain ATCC 51142 during
962 light-dark and continuous-light growth. J Bacteriol 190: 3904–3913.
963 https://doi.org/10.1128/JB.00206-08
964
965 Umena Y, Kawakami K, Shen JR, Kamiya N (2011) Crystal structure of oxygen-evolving
966 Photosystem II at a resolution of 1.9 Å. Nature 473: 55–60.
967 https://doi.org/10.1038/nature09913
968
969 Vinyard DJ, Gimpel J, Ananyev GM, Mayfield SP, Dismukes GC (2014) Engineered
970 Photosystem II reaction centers optimize photochemistry versus photoprotection at different
971 solar intensities. J Am Chem Soc 136: 4048–4055.
972 https://doi.org/10.1021/ja5002967

973

974 Vinyard DJ and Brudvig GW (2018) Progress toward a molecular mechanism of water
975 oxidation in Photosystem II. Annu Rev Phys Chem 68: 101–116.
976 https://doi.org/10.1146/annurev-physchem-052516-044820

977

978 Wada H, Murata N (2007) The essential role of phosphatidylglycerol in photosynthesis.
979 Photosynth Res 92: 205–215.
980 https://doi.org/10.1007/s11120-007-9203-z

981

982 Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer
983 TAP, Rempfer C, Bordoli L, Lepore R, Schwede T (2018) SWISS-MODEL: homology
984 modelling of protein structures and complexes. Nucleic Acids Res 46: 296–303.
985 https://doi.org/10.1093/nar/gky427

986

987 Wegener KM, Nagarajan A, Pakrasi HB (2015) An atypical *psbA* gene encodes a sentinel D1
988 protein to form a physiologically relevant inactive Photosystem II complex in cyanobacteria.
989 J Biol Chem 290: 3764–3774.
990 https://doi.org/10.1074/jbc.M114.604124

991

992 Wei X, Su X, Cao P et al. (2016) Structure of spinach Photosystem II – LHCII supercomplex
993 at 3.2 Å resolution. Nature 534: 69–74

994

995 Wiklund R, Salih GF, Mäenpää, P, Jansson C (2001) Engineering of the protein environment
996 around the redox-active TyrZ in Photosystem II. The role of F186 and P162 in the D1 protein
997 of *Synechocystis* 6803. Eur J Biochem 268: 5356–5364.

https://doi.org/10.1046/j.0014-2956.2001.02466.x

Xu B, Yang Z (2013) PAMLX a graphical user interface for PAML. Mol Biol Evol 30: 2723–2724.
https://doi.org/10.1093/molbev/mst179

Yamasato A, Kamada T, Satoh K (2002) Random mutagenesis targeted to the *psbA*II gene of *Synechocystis* sp. PCC 6803 to identify functionally important residues in the D1 protein of the Photosystem II reaction center. Plant Cell Physiol 43: 540–548.
https://doi.org/10.1093/pcp/pcf066

Yang Z (2007) PAML 4 phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591.
https://doi.org/10.1093/molbev/msm088

Zabelin AA, Shkuropatova VA, Makhneva ZK, Moskalenko AA, Shuvalov VA, Shkuropatov AY (2014) Chemically modified reaction centers of Photosystem II: Exchange of pheophytin a with 7-deformyl-7-hydroxymethyl-pheophytin b. BBA–Bioenergetics 1837: 1870–1881.
https://doi.org/10.1016/j.bbabio.2014.08.004

Figure Legends

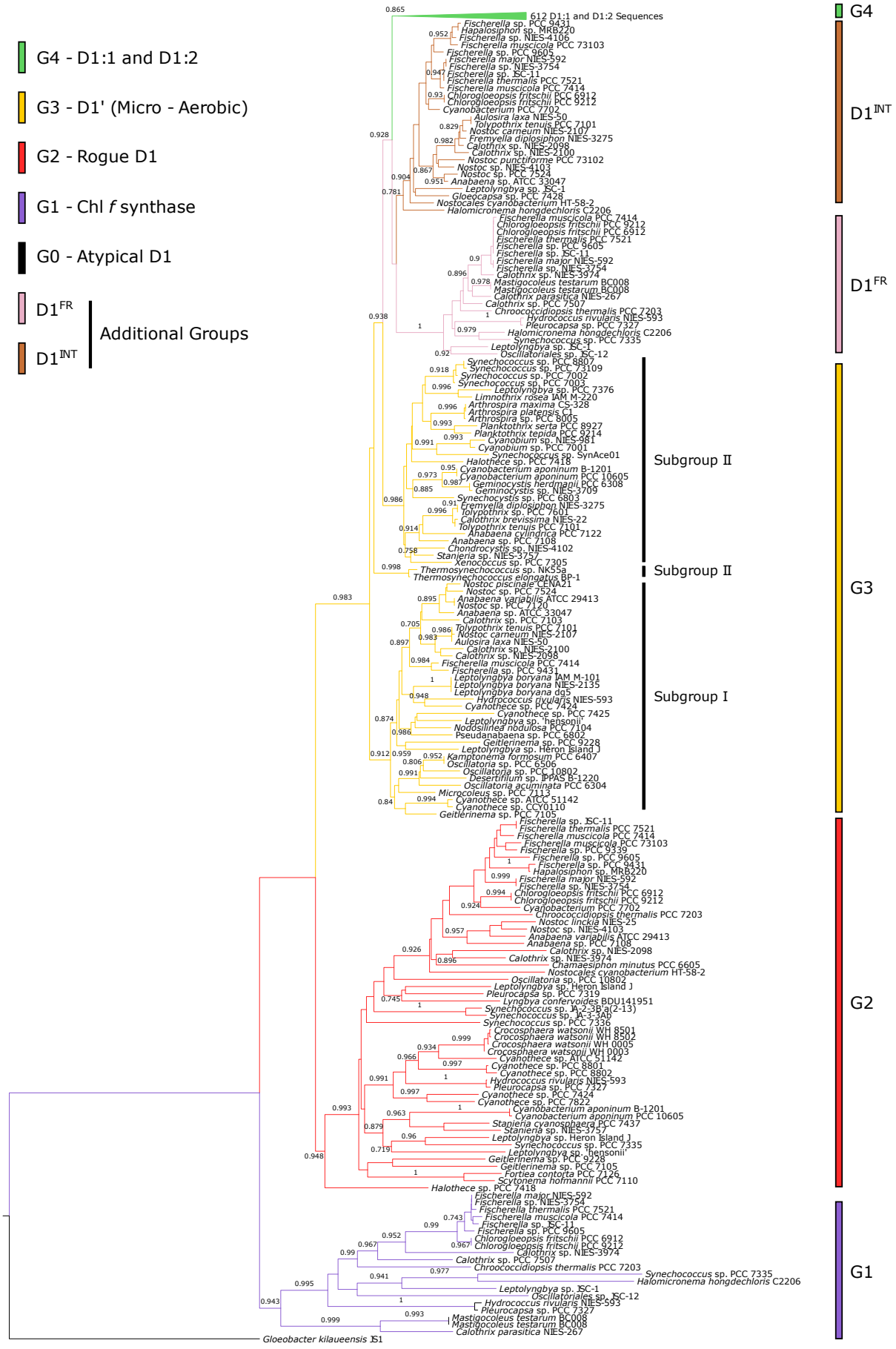
G4 - D1:1 and D1:2
G3 - D1' (Micro - Aerobic)
G2 - Rogue D1
G1 - Chl f synthase
G0 - Atypical D1
D1FR
D1INT
Additional Groups
612 D1:1 and D1:2 Sequences
Subgroup II
Subgroup II
Subgroup I
G4
D1INT
D1FR
G3
G2
G1
0.08

Fig. 1 Rooted maximum likelihood phylogeny of D1 proteins using the atypical D1 from *Gloeobacter kilaueensis* JS1 as the outgroup. Branch supports are expressed as SH-like aLRT probabilities. The G0 sequence from *Gloeobacter kilaueensis* JS1 is coloured in black, with G1, G2, G3 and G4 D1 proteins shown in purple, red, yellow and green, respectively. The two D1 protein groups: $D1^{FR}$ and $D1^{INT}$ are indicated in pink and brown, respectively.

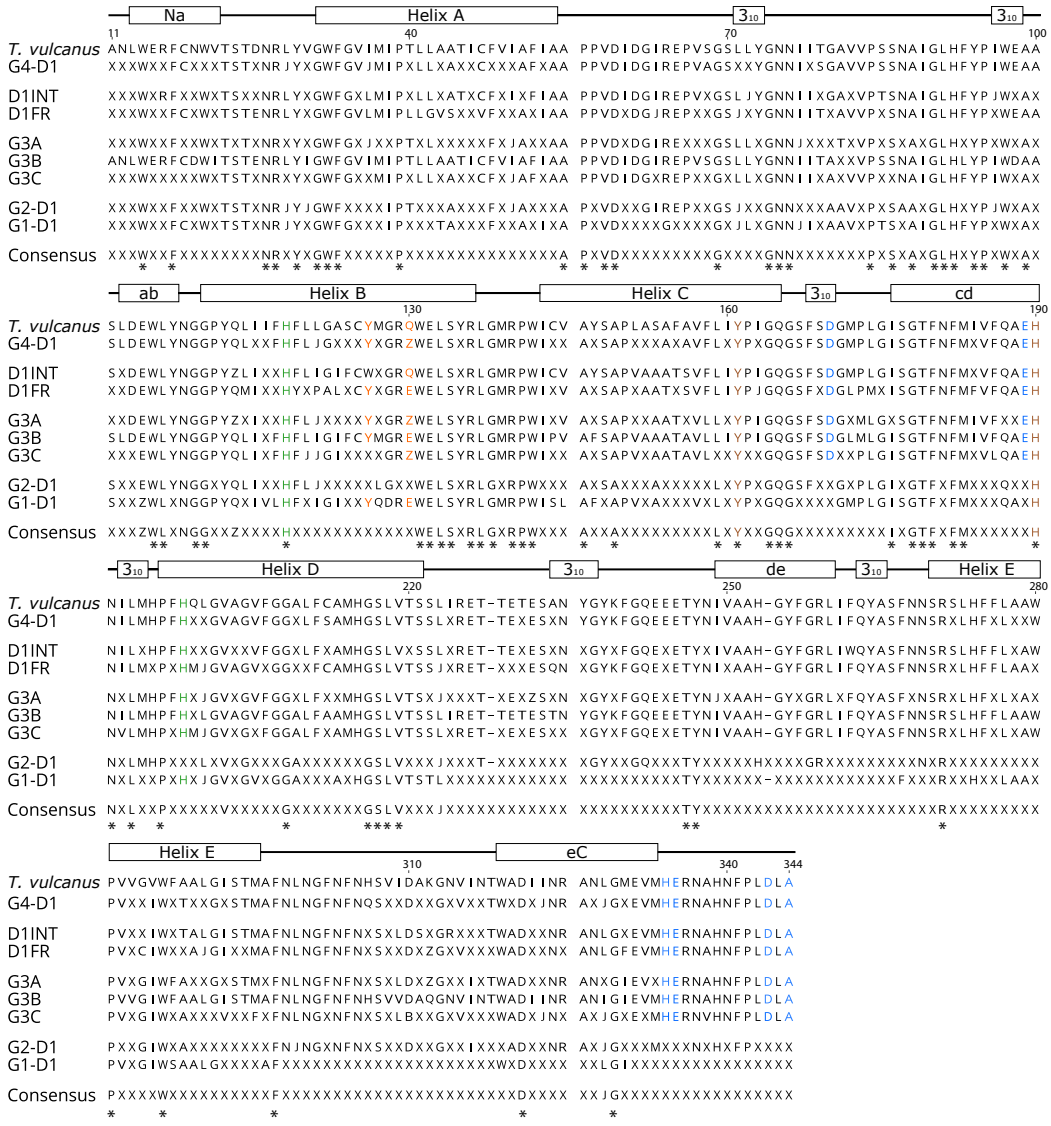

Fig. 2 Alignment of the 95% consensus for each group of D1 in the phylogenetic tree in Fig. 1 with a sequence representing the consensus for all eight D1 groups and the G4 sequence from *Thermosynechococcus vulcanus*. Positions highlighted with an asterisk indicate residues which are fully conserved across all types of D1. Ligands to the OEC, chlorophyll, Y_z, and pheophytin are highlighted in blue, green, purple and orange, respectively.

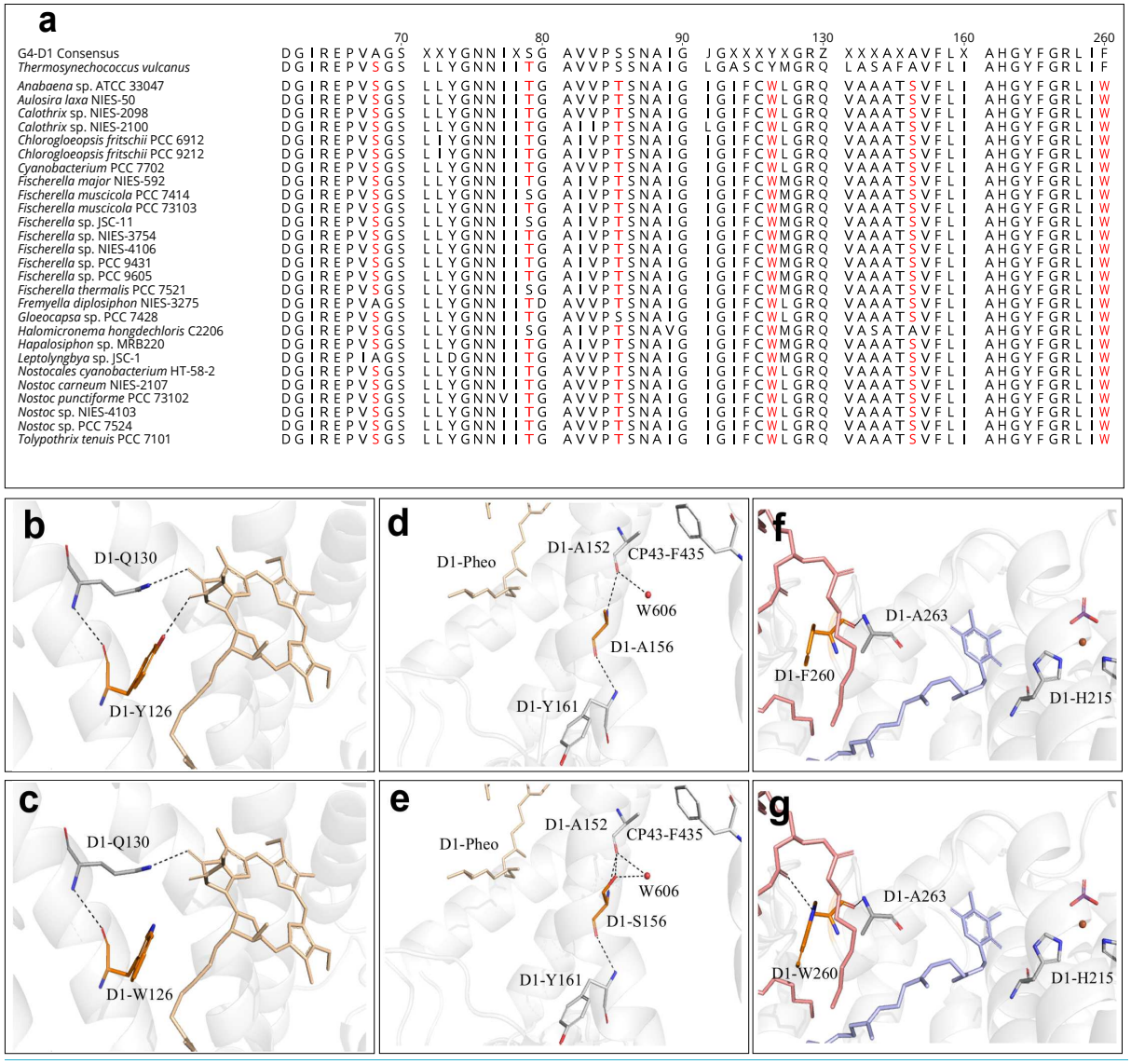

Fig. 3 Alignment of $D1^{INT}$ sequences with conserved residues highlighted. **a** Alignment of $D1^{INT}$ protein sequences compared to the G4 sequence from *Thermosynechococcus vulcanus* and the consensus sequence for G4-D1s, with conserved changes to the protein sequences highlighted in red. **b, d** and **f** show the structure of *Thermosynechococcus vulcanus* at

Tyr126, 156 and Phe260, while **c**, **e** and **g** show the same residues as modelled for the $D1^{INT}$ protein sequence from *Nostoc punctiforme* ATCC 29133. Distances within 3.6 Å, indicating potential hydrogen bonds are shown in dashed, black lines. The pheophytin present in the D1 protein is shown in tan. Q_B is shown in blue. The phosphatidylglycerol adjacent to Phe260 is shown in salmon pink in **f** and **g**.

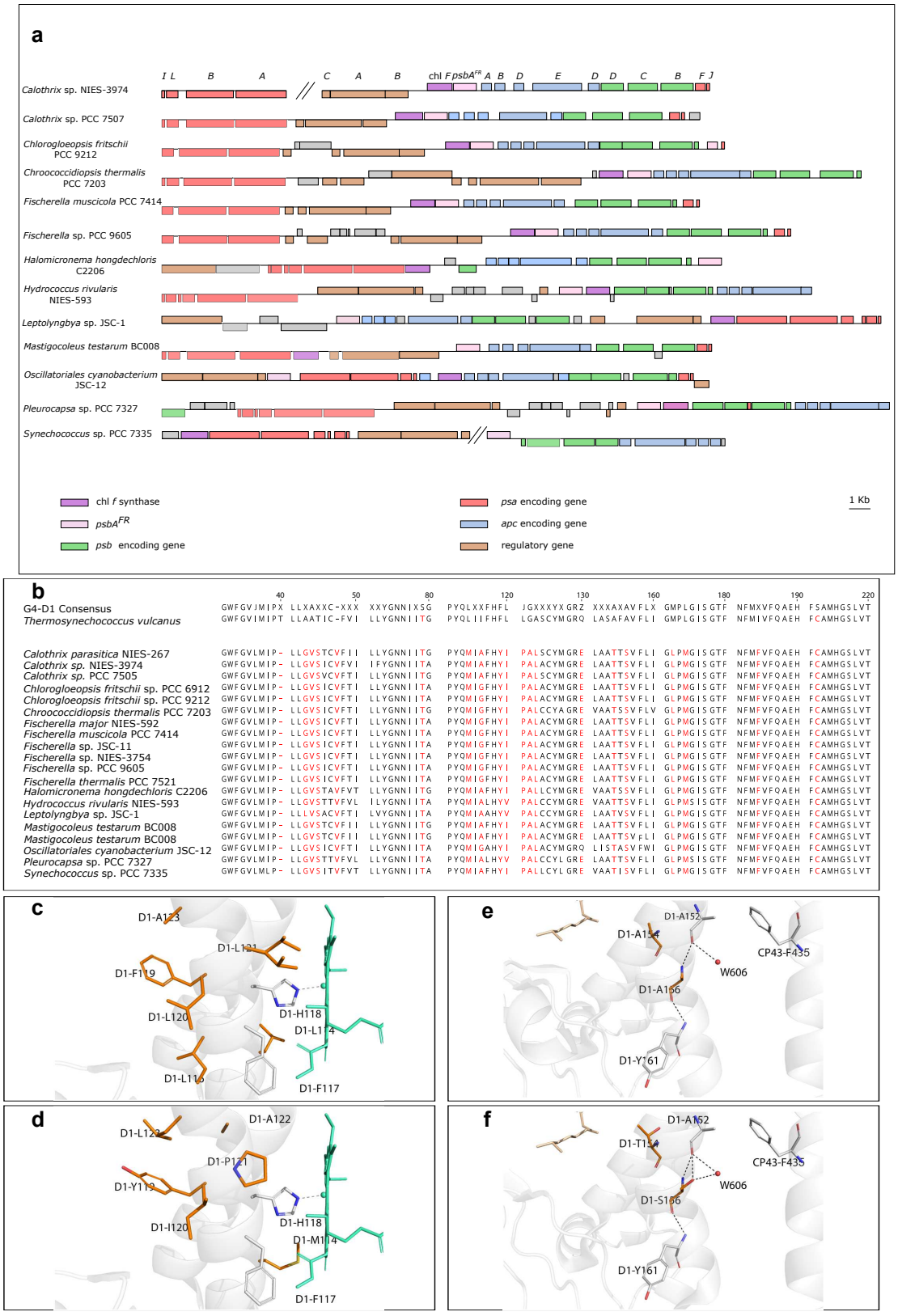

Fig. 4 Gene context, sequence alignment and highlighted residues of interest for D1^{FR} sequences. **a** Gene context of the chlorophyll *f* synthase and D1^{FR} in the far-red-inducible

gene cluster. Identity of the genes present in the *Calothrix* sp. NIES-3974 are given for reference. **b** Alignment of the 20 $D1^{FR}$ sequences with the reference G4 sequence from *Thermosynechococcus vulcanus* and the consensus sequence for all G4-D1s; conserved modified residues in $D1^{FR}$ highlighted in red. **c** and **d** The D1 helix B residues present in *Thermosynechococcus vulcanus* PS II crystal structure and the same region present in the $D1^{FR}$ sequence from *Chlorogloeopsis fritschii* PCC 9212, respectively. **e** and **f** D1 helix C from the PS II structure from *Thermosynechococcus vulcanus* and the corresponding region for the modelled D1 from *C. fritschii* PCC 9212, respectively. In **c** and **d** the accessory chlorophyll in PS II is shown in cyan, while the pheophytin in **e** and **f** is shown in tan.

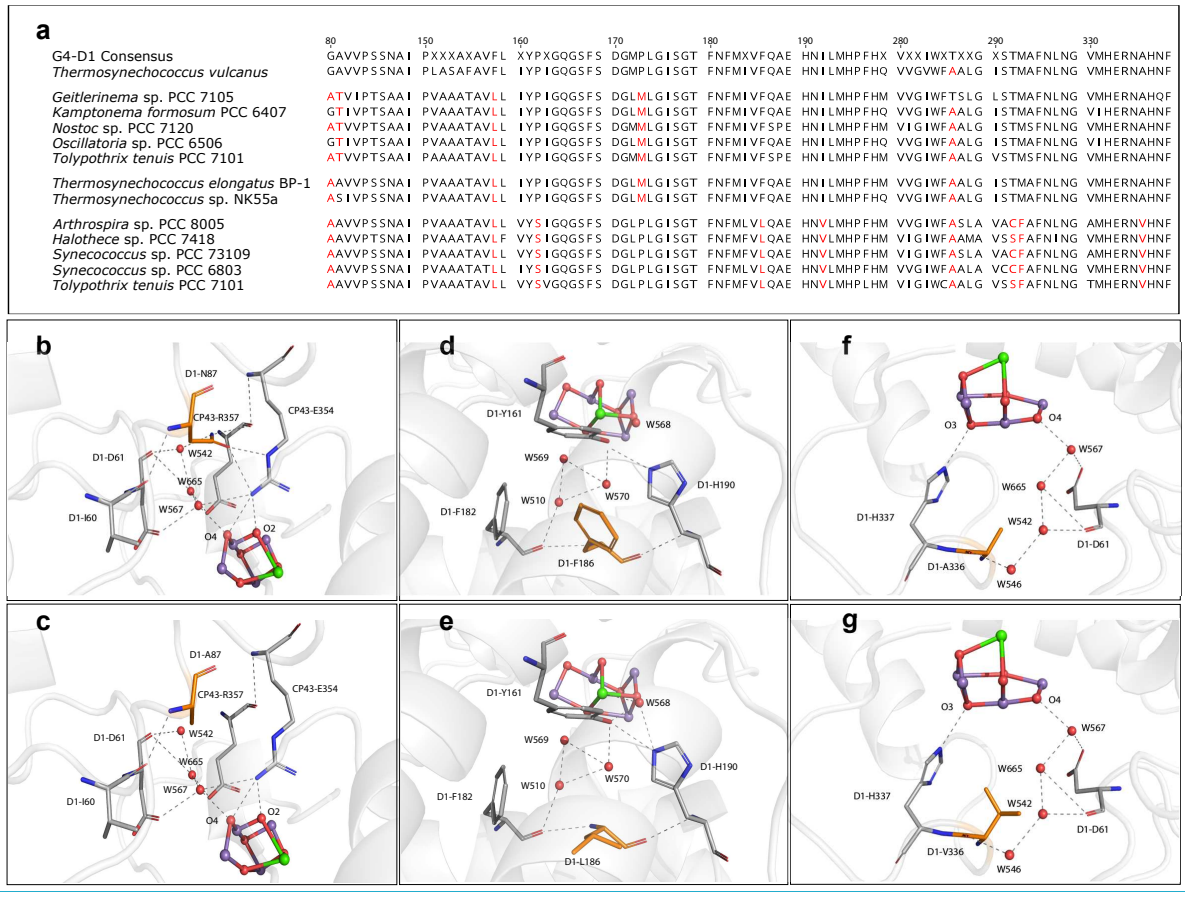

Fig. 5 Alignment of all sequences and highlighted residues of interest for G3-D1 sequences. **a** Alignment of five subgroup I D1′ sequences (*Geitlerinema* sp. PCC 7105 - *Tolypothrix tenuis* PCC 7101, two subgroup II sequences (*Thermosynechococcus elongatus* BP-1 and

Thermosynechococcus sp. NK55a) and five subgroup III sequences (*Arthrospira* sp. PCC 8005 - *Tolypothrix tenuis* PCC 7101) against the G4 reference sequence from *Thermosynechococcus vulcanus* and the consensus sequence for all G4 D1 sequences with subgroup-specific alterations to the D1 protein structure highlighted in red. **b** and **c** comparison of the amino acids around Asn87 in the *Thermosynechococcus vulcanus* PS II crystal structure and the modelled Ala87 from the G3 D1 protein of *Nostoc* sp. PCC 7120. **d**,**e**,**f** and **g** show the interactions of Phe186 and Ala336 of the G4-D1 from the *Thermosynechococcus vulcanus* PS II crystal structure and the modelled alterations of these ligands from the G3 D1 protein of *Synechocystis* sp. PCC 6803, respectively. Both G3-D1 sequences from *Nostoc* sp. PCC 7120 and *Synechocystis* sp. PCC 6803 were modelled based on the known crystal structure of D1 from *Thermosynechococcus vulcanus* as described in methods. The potential hydrogen-bonding network surrounding these residues is shown in dashed, black lines and limited to distances within 3.6 Å. The OEC is shown in balls and sticks with the calcium, manganese and oxygen shown in green, purple and red, respectively.

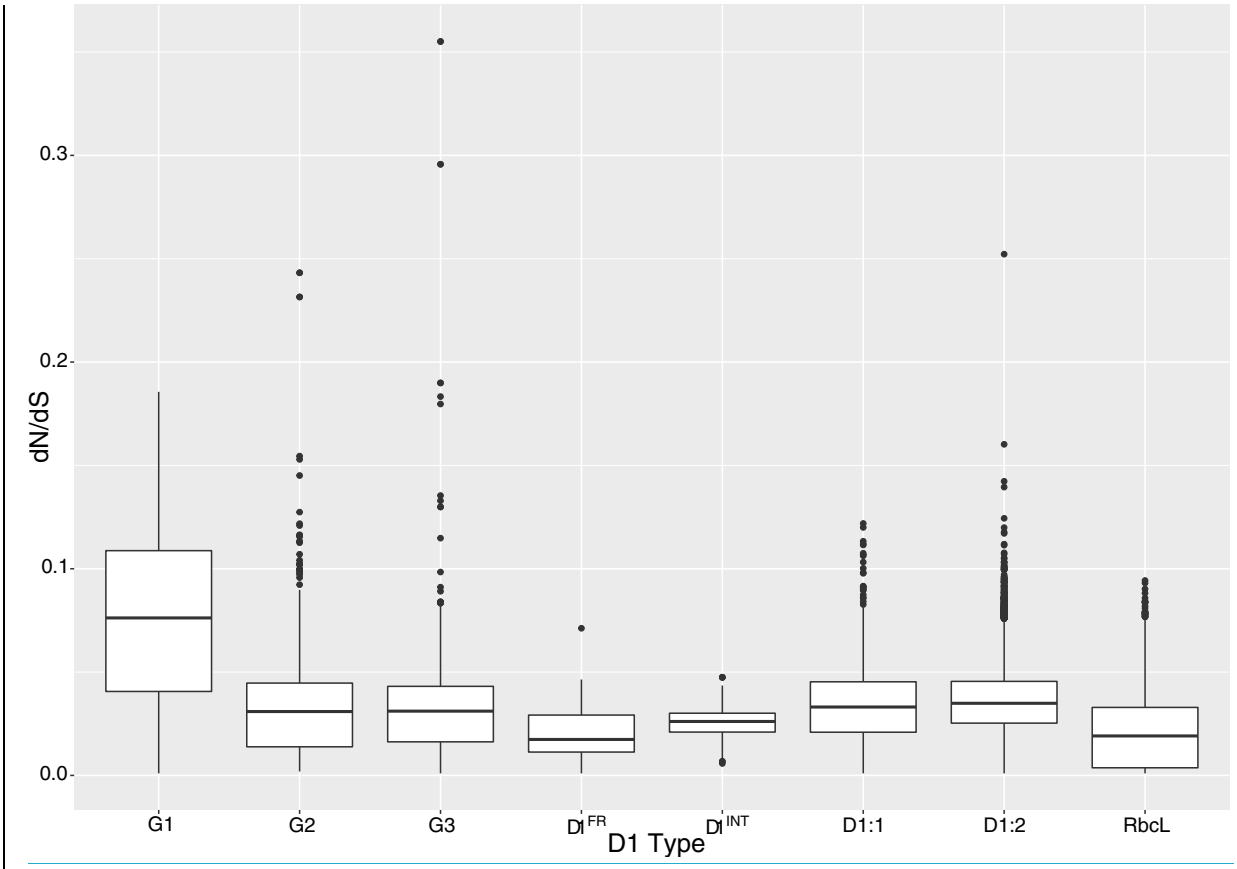

Fig. 6 Boxplot illustrating the range of w (dN / dS) obtained by pairwise comparison of genes encoding for the proteins within each group of D1. Lines indicate the median and boxes delineate first and third quartiles, whiskers illustrate the minimum and maximum values and outliers are shown as individual points.

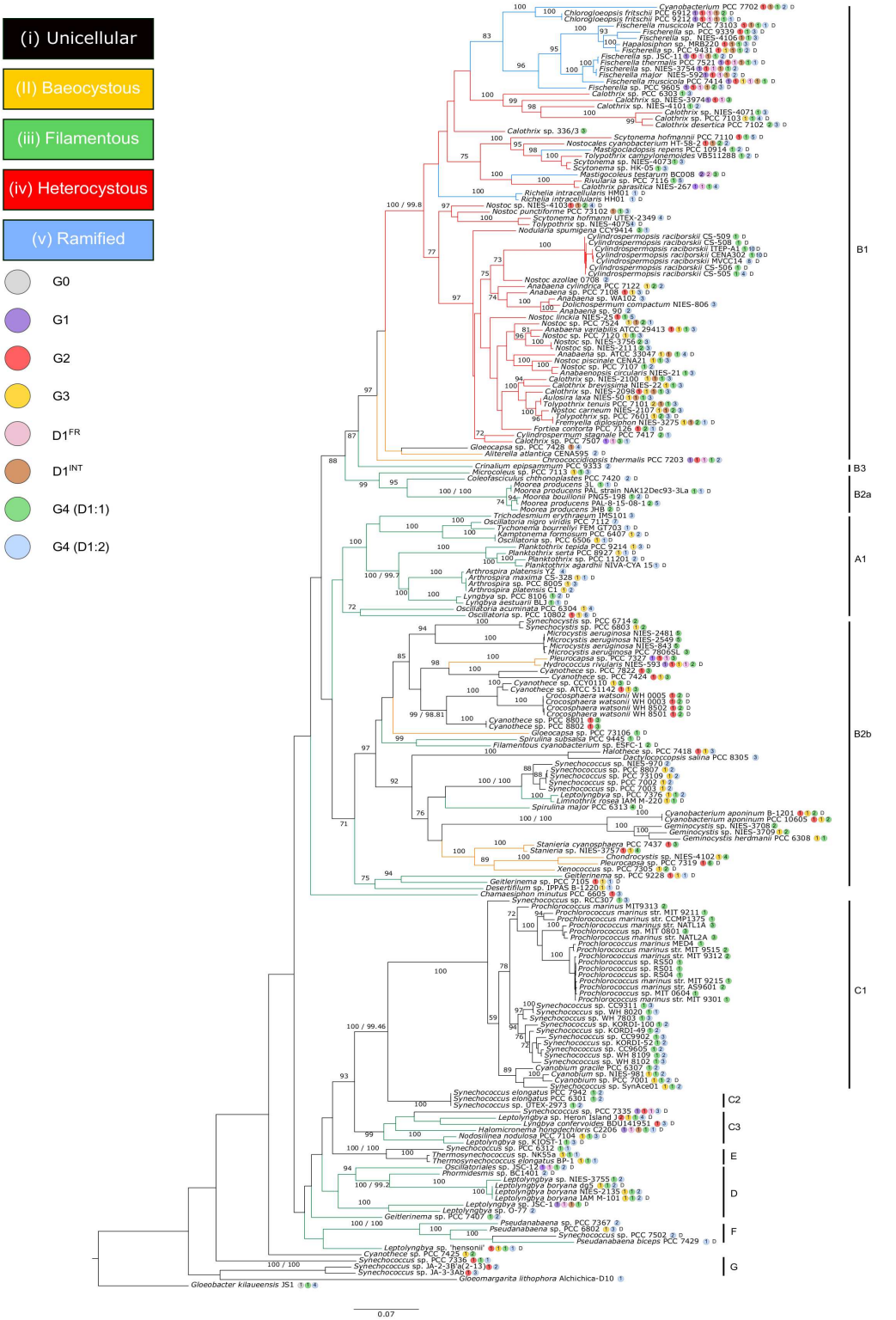
(i) Unicellular
(II) Baeocystous
(iii) Filamentous
(iv) Heterocystous
(v) Ramified
G0
G1
G2
G3
D1FR
D1INT
G4 (D1:1)
G4 (D1:2)
B1
B3
B2a
A1
B2b
C1
C2
C3
E
D
F
G
0.07

Fig. 7 Rooted maximum likelihood phylogeny of 16S-23S rRNA cyanobacterial sequences using *Gloeobacter kilaueensis* JS1 as the outgroup. Branch support over 70% from the maximum likelihood bootstrap are indicated, with branch support over 95% from the maximum parsimony tree also being highlighted (number of iterations = 1000). The D1 type and number of genes encoding each type that are present in each strain are indicated using coloured circles with G1, G2 , G3, G4 D1:1 and G4 D1:2 protein sequences shown in purple, red, yellow, green and blue, respectively and the D1 proteins $D1^{FR}$ and $D1^{INT}$ are indicated in pink and brown, respectively. Phylogenetic subclades recovered in the Shih et al. (2013) analysis are indicated to the right of their corresponding groupings recovered in this analysis. A 'D' next to the D1 types for a strains indicates the data was obtained from a draft genome.

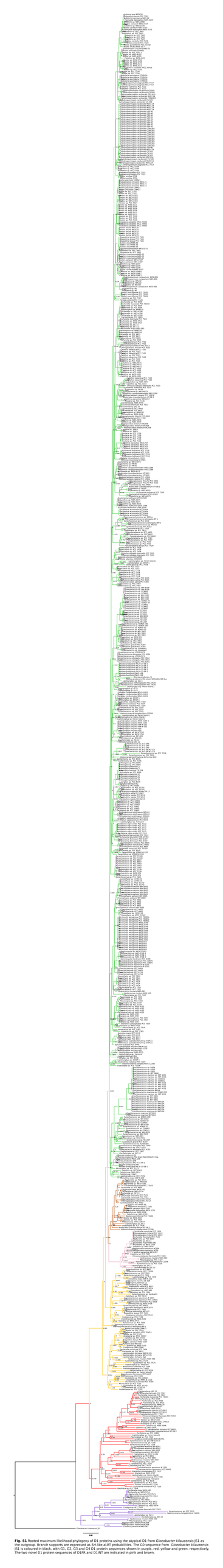

Fig. S1 Rooted maximum likelihood phylogeny of D1 proteins using the atypical D1 from *Gloeobacter kilaueensis* JS1 as the outgroup. Branch supports are expressed as SH-like aLRT probabilities. The G0 sequence from *Gloeobacter kilaueensis* JS1 is coloured in black, with G1, G2, G3 and G4 D1 protein sequences shown in purple, red, yellow and green, respectively. The two novel D1 protein sequences of D1FR and D1NT are indicated in pink and brown.

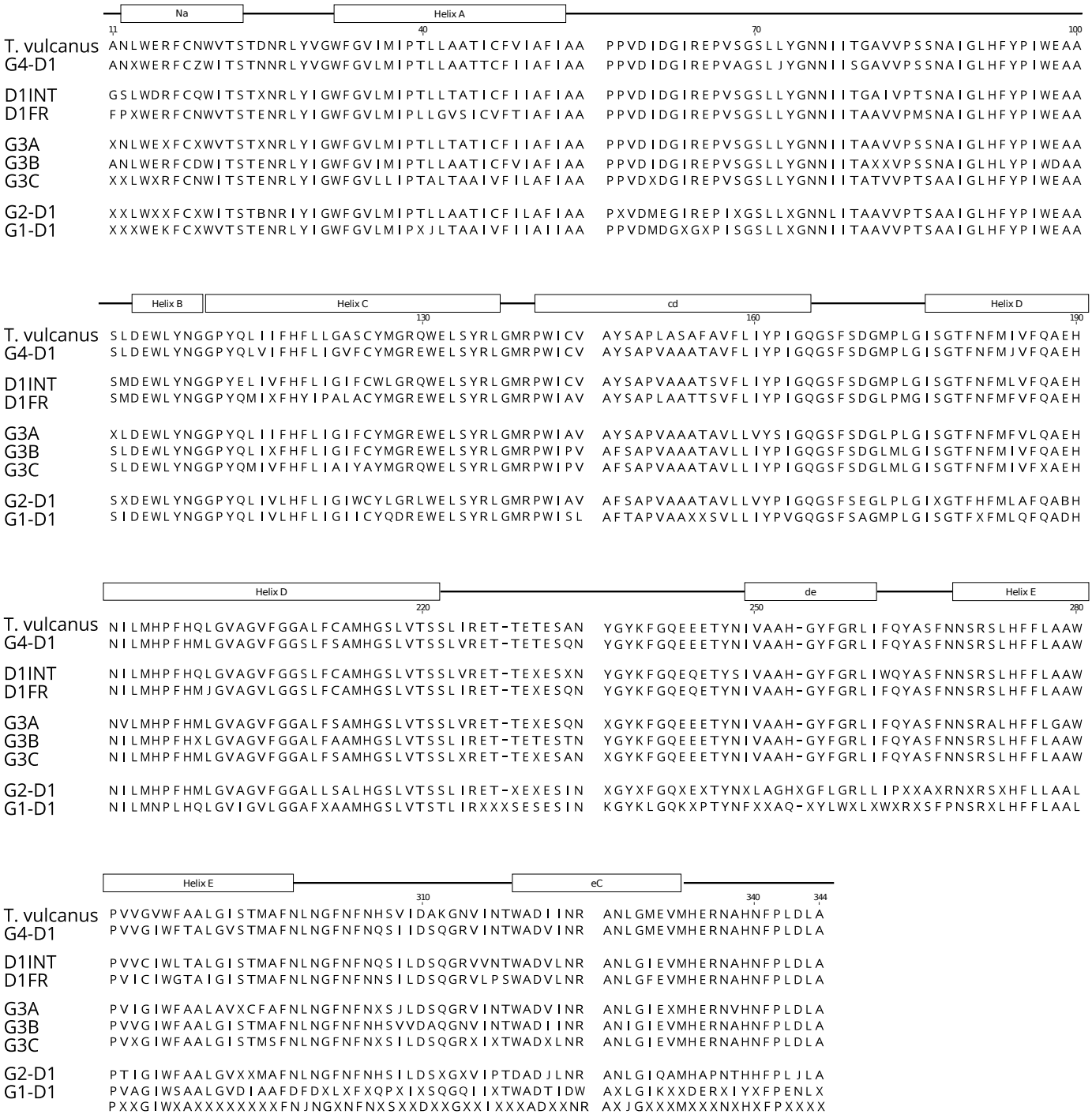

Fig. S2 Alignment of 50% consensus for each group of D1 in the phylogenetic tree shown in Fig. 1 and Fig. S1 with the G4 reference sequence from *Thermosynechococcus vulcanus*.

Fig. S3 Alignment of all D1^{INT} protein sequences recovered in this analysis.

Fig. S4 Gene context of the D1FR protein (shown in pink) and the G1 D1 protein (shown in purple) in the far-red-inducible gene cluster. PS I, PS II, phycobilisome and regulatory encoding genes are coloured in red, green, blue and brown, respectively. Hypothetical and non-photosynthetic encoding genes are coloured in grey. The specific genes present in the far-red cluster of Calothrix sp. NIES-3974 have been included for reference.

```
                                                  1       10        20        30        40        50        60
                                                  |       |         |         |         |         |         |
          Calothrix parasitica NIES-267           MTT-ISQQPSRFPLWDRFCDWVTSTENRLYIGWFGVLMIPLLGVSTCVFIIAFIAAPPVD
                    Calothrix sp. PCC 7507        MTTILQRQKGRSPTWERFCNWVTSTENRLYIGWFGVLMIPLLGVSVCVFTIAFIAAPPVD
     Chlorogloeopsis fritschii PCC 6912           MTTISTRPTSRFPTWDRFCNWVTSTENRLYIGWFGVLMIPLLGVSICVFTIAFIAAPPVD
     Chlorogloeopsis fritschii PCC 9212           MTTISTRPTSRFPTWDRFCNWVTSTENRLYIGWFGVLMIPLLGVSICVFTIAFIAAPPVD
  Chroococcidiopsis thermalis PCC 7203            MTTVLQRQLNRVPAWERFCNWVTSTENRLYIGWFGVLMIPLLGVSICVFTIAFIAAPPVD
                Calothrix sp. NIES-3974           MTTISQPKTTRFPAWERFCDWITSTENRLYIGWFGVLMIPLLGVSICVFVIAFIAAPPVD
            Fischerella major NIES-592            MTTISTRPTSRFPAWDRFCDWVTSTENRLYIGWFGVLMIPLLGVSICVFTIAFIAAPPVD
        Fischerella muscicola PCC 7414            MTTISTRPTSRFPTWDRFCNWVTSTENRLYIGWFGVLMIPLLGVSICVFTIAFIAAPPVD
                  Fischerella sp. JSC-11          MTTISTRPTSRFPTWDRFCNWVTSTENRLYIGWFGVLMIPLLGVSICVFTIAFIAAPPVD
               Fischerella sp. NIES-3754          MTTISTRPTSRFPAWDRFCDWVTSTENRLYIGWFGVLMIPLLGVSICVFTIAFIAAPPVD
                Fischerella sp. PCC 9605          MTTISTRPTSRFPTWDRFCNWVTSTENRLYIGWFGVLMIPLLGVSICVFTIAFIAAPPVD
        Fischerella thermalis PCC 7521            MTTISTRPTSRFPTWDRFCNWVTSTENRLYIGWFGVLMIPLLGVSICVFTIAFIAAPPVD
     Halomicronema hongdechloris C2206            MTTTL-QRPGTIDLWERFCAWVTSTENRLYVGWFGVLMIPLLGVSTAVFVTAFIAAPPVD
        Hydrococcus rivularis NIES-593            MTTTL-QKREIPNLWEQFCQWVTSTENRLYIGWFGVLMIPLLGVSTTVFVLAIIAAPPVD
                 Leptolyngbya sp. JSC-1           MTT-LLQRRGARVQWEQFCNWVTSTENRLYVGWFGVLMIPLLLVSACVFTIAFIAAPPVD
           Mastigocoleus testarum BC008           MTTFSQQQPRPLSLWEKFCNWVTSTENRLYIGWFGVLMIPLLGVSTCVFIIAFIAAPPVD
           Mastigocoleus testarum BC008           MTTFSQQQPRPLSLWEKFCNWVTSTENRLYIGWFGVLMIPLLGVSTCVFIIAFIAAPPVD
  Oscillatoriales cyanobacterium JSC-12           MTTLLQQRKIASDQWERFCNWITSTENRLYIGWFGVLMIPLLGVSICVFTIAFIAAPPVD
              Pleurocapsa sp. PCC 7327            MTTTL-QKREIPNLWEQFCQWVTSTENRLYIGWFGVLMIPLLGVSTTVFVLAIIAAPPVD
            Synechococcus sp. PCC 7335            MTT-L-IQRPNISQWERFCQWITSTENRLYIGWFGVLMLPLLGVSITVFVTAFIAAPPVD

                                                  61      70        80        90        100       110       120
                                                  |       |         |         |         |         |         |
          Calothrix parasitica NIES-267           IDGIREPIAGSLLYGNNIITGAVVPTSNAIGLHFYPLWEAASMDEWLYNGGPYQMIAFHY
                    Calothrix sp. PCC 7507        IDGIREPVAGSLLYGNNIITGAVVPMSNAIGLHFYPIWEAASMDEWLYNGGPYQMIAFHY
     Chlorogloeopsis fritschii PCC 6912           IDGIREPVSGSLLYGNNIITAAVVPMSNAIGLHFYPIWEAASMDEWLYNGGPYQMIGFHY
     Chlorogloeopsis fritschii PCC 9212           IDGIREPVSGSLLYGNNIITAAVVPMSNAIGLHFYPIWEAASMDEWLYNGGPYQMIGFHY
  Chroococcidiopsis thermalis PCC 7203            IDGIREPVSGSLLYGNNIITGAVVPMSNAIGLHFYPIWEAASMDEWLYNGGPYQMIGFHY
                Calothrix sp. NIES-3974           IDGIREPVAGSIFYGNNIITAAVVPMSNAIGLHFYPIWEAASMDEWLYNGGPYQMIAFHY
            Fischerella major NIES-592            IDGIREPVSGSLLYGNNIITAAVVPMSNAIGLHFYPIWEAASMDEWLYNGGPYQMIGFHY
        Fischerella muscicola PCC 7414            IDGIREPVSGSLLYGNNIITAAVVPMSNAIGLHFYPIWEAASMEEWLYNGGPYQMIGFHY
                  Fischerella sp. JSC-11          IDGIREPVSGSLLYGNNIITAAVVPMSNAIGLHFYPIWEAASMDEWLYNGGPYQMIGFHY
               Fischerella sp. NIES-3754          IDGIREPVSGSLLYGNNIITAAVVPMSNAIGLHFYPIWEAASMDEWLYNGGPYQMIGFHY
                Fischerella sp. PCC 9605          IDGIREPVSGSLLYGNNIITAAVVPMSNAIGLHFYPIWEAASMDEWLYNGGPYQMIGFHY
        Fischerella thermalis PCC 7521            IDGIREPVSGSLLYGNNIITAAVVPMSNAIGLHFYPIWEAASMDEWLYNGGPYQMIGFHY
     Halomicronema hongdechloris C2206            IDGLREPLSGSLLYGNNIITGAVVPSSNAIGLHFYPMWEAATLDEWLYNGGPYQMIAFHY
        Hydrococcus rivularis NIES-593            MDGLREPITGSILYGNNIITAAVVPTSNAIGLHFYPIWEAASMDEWLYNGGPYQMIALHY
                 Leptolyngbya sp. JSC-1           IDGIREPIAGSLVYGNNIITAAVVPSSNAIGLHFYPIWEAASMDEWLYNGGPYQMIAAHY
           Mastigocoleus testarum BC008           IDGIREPLSGSLLYGNNIITGAVVPTSNAIGLHFYPIWEAASMDEWLYNGGPYQMIAFHY
           Mastigocoleus testarum BC008           IDGIREPLSGSLLYGNNIITGAVVPTSNAIGLHFYPIWEAASMDEWLYNGGPYQMIAFHY
  Oscillatoriales cyanobacterium JSC-12           IDGIREPVAGSLLYGNNIITAAVVPSSNAIGLHFYPIWEAASIDEWLYNGGPYQMIGAHY
              Pleurocapsa sp. PCC 7327            MDGLREPITGSLLYGNNIITAAVVPTSNAIGLHFYPIWEAASMDEWLYNGGPYQMIALHY
            Synechococcus sp. PCC 7335            IDGIREPLSGSLLYGNNIITAAVVPTSNAIGLHFYPIWEAATLDEWLYNGGPYQMIAFHY

                                                  121     130       140       150       160       170       180
                                                  |       |         |         |         |         |         |
          Calothrix parasitica NIES-267           IPALSCYMGREWELSYRLGMRPWIAVAYSAPLAATTSVFLIYPIGQGSFSDGLPMGISGT
                    Calothrix sp. PCC 7507        IPALSCYMGREWELSYRLGMRPWICVAYSAPLAATTSVFLIYPIGQGSFSDGLPMGISGT
     Chlorogloeopsis fritschii PCC 6912           IPALACYMGREWELSYRLGMRPWIAVAYSAPLAATTSVFLIYPIGQGSFSDGLPMGISGT
     Chlorogloeopsis fritschii PCC 9212           IPALACYMGREWELSYRLGMRPWIAVAYSAPLAATTSVFLIYPIGQGSFSDGLPMGISGT
  Chroococcidiopsis thermalis PCC 7203            IPALCCYAGREWELSYRLGMRPWIAVAYTAPVAATSSVFLVYPIGQGSFSDGLPMGISGT
                Calothrix sp. NIES-3974           IPALACYMGREWELSYRLGMRPWIAVAYSAPLAATTSVFLIYPIGQGSFSDGLPMGISGT
            Fischerella major NIES-592            IPALACYMGREWELSYRLGMRPWIAVAYSAPLAATTSVFLIYPIGQGSFSDGLPMGISGT
        Fischerella muscicola PCC 7414            IPALACYMGREWELSYRLGMRPWIAVAYSAPLAATTSVFLIYPIGQGSFSDGLPMGISGT
                  Fischerella sp. JSC-11          IPALACYMGREWELSYRLGMRPWIAVAYSAPLAATTSVFLIYPIGQGSFSDGLPMGISGT
               Fischerella sp. NIES-3754          IPALACYMGREWELSYRLGMRPWIAVAYSAPLAATTSVFLIYPIGQGSFSDGLPMGISGT
                Fischerella sp. PCC 9605          IPALACYMGREWELSYRLGMRPWIAVAYSAPLAATTSVFLIYPIGQGSFSDGLPMGISGT
        Fischerella thermalis PCC 7521            IPALACYMGREWELSYRLGMRPWIAVAYSAPLAATTSVFLIYPIGQGSFSDGLPMGISGT
     Halomicronema hongdechloris C2206            IPALLCYMGREWELSYRLGMRPWICVAYSAPVAATSSVFLIYPIGQGSFSDGLPMGISGT
        Hydrococcus rivularis NIES-593            VPALCCYLGREWELSYRLGMRPWICVAFSAPVAATTSVFLIYPLGQGSFADGLPMSISGT
                 Leptolyngbya sp. JSC-1           VPALCCYMGREWELSYRLGMRPWIAVAYSAPLAATVSVFLIYPIGQGSFSDGLPMGISGT
           Mastigocoleus testarum BC008           IPALACYMGREWELSYRLGMRPWIAVAYSAPLAATTSVFLIYPIGQGSFSDGLPMGISGT
           Mastigocoleus testarum BC008           IPALACYMGREWELSYRLGMRPWIAVAYSAPLAATTSVFLIYPIGQGSFSDGLPMGISGT
  Oscillatoriales cyanobacterium JSC-12           IPALACYMGRQWELSYRLGMRPWICVAYSAPLISTASVFWIYPIGQGSFSDGLPMGISGT
              Pleurocapsa sp. PCC 7327            VPALCCYLGREWELSYRLGMRPWICVAFSAPLAATTSVFLIYPLGQGSFADGLPMSISGT
            Synechococcus sp. PCC 7335            IPALLCYLGREWELSYRLGMRPWICIAYSAPVAATISVFLIYPIGQGSFSDGLPMGISGT

                                                  181     190       200       210       220       230       240
                                                  |       |         |         |         |         |         |
          Calothrix parasitica NIES-267           FNFMFVFQAEHNILMNPFHMIGVAGVLGGSLFCAMHGSLVTSSLVRETSESESQNYGYKF
                    Calothrix sp. PCC 7507        FNFMFVFQAEHNILMHPFHMLGVAGVLGGSLFCAMHGSLVTSSLIRETTESESQNYGYKF
     Chlorogloeopsis fritschii PCC 6912           FNFMFVFQAEHNILMHPLHMIGVAGVLGGSLFCAMHGSLVTSSLIRETTELESQNYGYKF
     Chlorogloeopsis fritschii PCC 9212           FNFMFVFQAEHNILMHPLHMIGVAGVLGGSLFCAMHGSLVTSSLIRETTELESQNYGYKF
  Chroococcidiopsis thermalis PCC 7203            FNFMFVFQAEHNILMHPLHMIGVAGVLGGSLFCAMHGSLVTSSLIRETTELESQNYGYKF
                Calothrix sp. NIES-3974 3         FNFMFVFQAEHNILMHPLHMIGVAGVLGGSLFCAMHGSLVTSSIIRETTELESQNYGYKF
            Fischerella major NIES-592            FNFMFVFQAEHNILMHPLHMIGVAGVLGGSLFCAMHGSLVTSSLIRETTELESQNYGYKF
        Fischerella muscicola PCC 7414            FNFMFVFQAEHNILMHPLHMIGVAGVLGGSLFCAMHGSLVTSSLIRETTELESQNYGYKF
                  Fischerella sp. JSC-11          FNFMFVFQAEHNILMHPLHMIGVAGVLGGSLFCAMHGSLVTSSLIRETTELESQNYGYKF
               Fischerella sp. NIES-3754          FNFMFVFQAEHNILMHPLHMIGVAGVLGGSLFCAMHGSLVTSSLIRETTELESQNYGYKF
                Fischerella sp. PCC 9605          FNFMFVFQAEHNILMHPLHMIGVAGVLGGSLFCAMHGSLVTSSLIRETTELESQNYGYKF
        Fischerella thermalis PCC 7521            FNFMFVFQAEHNILMHPLHMIGVAGVLGGSLFCAMHGSLVTSSLIRETTELESQNYGYKF
     Halomicronema hongdechloris C2206            FNFMFVFQAEHNILMHPFHMLGVAGVLGGSLFCAMHGSLVTSSLIRETTDSESQNAGYKF
        Hydrococcus rivularis NIES-593            FNFMFVFQAEHNILMHPFHMLGVAGVFGGAFFCAMHGSLVTSSLIRETSENESQNYGYKF
                 Leptolyngbya sp. JSC-1           FNFMFVFQAEHNILMHPFHMLGVAGVFGGSLFCAMHGSLVTSSLVRETTENESQNYGYKF
           Mastigocoleus testarum BC008           FNFMFVFQAEHNILMNPFHMLGVAGVLGGSLFCAMHGSLVTSSIVRETTESESQNYGYKF
           Mastigocoleus testarum BC008           FNFMFVFQAEHNILMNPFHMLGVAGVLGGSLFCAMHGSLVTSSIVRETTESESQNYGYKF
  Oscillatoriales cyanobacterium JSC-12           FNFMFVFQAEHNILMHPFHMLGVIGVFGGSLFCAMHGSLVTSSLVRETTEDESANNGYKF
              Pleurocapsa sp. PCC 7327            FNFMFVFQAEHNILMHPFHMLGVAGVFGGAFFCAMHGSLVTSSLIRETSENESQNYGYKF
            Synechococcus sp. PCC 7335            FNFMFVFQAEHNILMHPFHMLGVAGVLGGSLFCAMHGSLVTSSLVRETSDSQSQNEGYKF

                                                  241     250       260       270       280       290       300
                                                  |       |         |         |         |         |         |
          Calothrix parasitica NIES-267           GQEQETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVICIWGTAIGISTMAFNLNG
                    Calothrix sp. PCC 7507        GQEEETYNIVAAHGYFGRLIFQYASFNNSRTLHFFLAAWPVICIWATALGISTMAFNLNG
     Chlorogloeopsis fritschii PCC 6912           GQEQETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVICIWGTAIGISTMAFNLNG
     Chlorogloeopsis fritschii PCC 9212           GQEQETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVICIWGTAIGISTMAFNLNG
  Chroococcidiopsis thermalis PCC 7203            GQESETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVICIWATALGISTMAFNLNG
                Calothrix sp. NIES-3974           GQEEETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVICIWGTAIGISTMAFNLNG
            Fischerella major NIES-592            GQEQETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVICIWGTAIGISTMAFNLNG
        Fischerella muscicola PCC 7414            GQEQETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVICIWGTAIGISTMAFNLNG
                  Fischerella sp. JSC-11          GQEQETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVICIWGTAIGISTMAFNLNG
               Fischerella sp. NIES-3754          GQEQETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVICIWGTAIGISTMAFNLNG
                Fischerella sp. PCC 9605          GQEQETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVICIWGTAIGISTMAFNLNG
        Fischerella thermalis PCC 7521            GQEQETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVICIWGTAIGISTMAFNLNG
     Halomicronema hongdechloris C2206            GQESETYNILAAHGYFGRLIFQYASFNNSRWLHFFLAAWPVVCIWFVALGISTMAFNLNG
        Hydrococcus rivularis NIES-593            GQEEETYNIVAAHGYFGRLIFQYASFNNSRNLHFFLAAWPVICIWFTALGICVMAFNLNG
                 Leptolyngbya sp. JSC-1           GQEGETYNIVAAHGYFGRLVFQYASFTNSRSLHFFLAAFPVICIWATALGISTMAFNLNG
           Mastigocoleus testarum BC008           GQEQETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVICIWATAIGISTMAFNLNG
           Mastigocoleus testarum BC008           GQEQETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLAAWPVICIWATAIGISTMAFNLNG
  Oscillatoriales cyanobacterium JSC-12           GQEDETYNIVAAHGYFGRLIFQYASFNNSRSLHFFLGAFPVVCIWCTALGICTMGFNLNG
              Pleurocapsa sp. PCC 7327            GQEEETYNIVAAHGYFGRLIFQYASFNNSRNLHFFLAAWPVICIWFTALGICVMAFNLNG
            Synechococcus sp. PCC 7335            GQEEETYNILAAHGYFGRLIFQYASFNNSRQLHFFLAAWPVVCIWFVALGISTMAFNLNG

                                                  301     310       320       330       340       350       360
                                                  |       |         |         |         |         |         |
          Calothrix parasitica NIES-267           FNFNNSILDSQGRVLPSWADVLNRANLGFEVMHERNAHNFPLDLASGKSIPVTVSLKAPA
                    Calothrix sp. PCC 7507        FNFNNSILDSEGHVLPSWADVLNRANLGFEVMHERNAHNFPLDLASSEAYPVPVALQAPA
     Chlorogloeopsis fritschii PCC 6912           FNFNNSILDSQGRVLPSWADVLNRANLGFEVMHERNAHNFPLDLACGEAVPVAI--TAPS
     Chlorogloeopsis fritschii PCC 9212           FNFNNSILDSQGRVLPSWADVLNRANLGFEVMHERNAHNFPLDLACGEAVPVAI--TAPS
  Chroococcidiopsis thermalis PCC 7203            FNFNNSVLDSQGRVLPTWADVLNRANLGFEVMHERNAHNFPLDLACSDAVPV--AFKAPA
                Calothrix sp. NIES-3974           FNFNNSILDSQGRVLPSWADVLNRANLGFEVMHERNAHNFPLDLASGEAIPVPVAMAAPS
            Fischerella major NIES-592            FNFNNSILDSQGRVLPSWADVLNRANLGFEVMHERNAHNFPLDLAGGEAVPVAI--TAPS
        Fischerella muscicola PCC 7414            FNFNNSILDSQGRVLPSWADVLNRANLGFEVMHERNAHNFPLDLACGEAVPVAI--TAPS
                  Fischerella sp. JSC-11          FNFNNSILDSQGRVLPSWADVLNRANLGFEVMHERNAHNFPLDLAGGEAVPVAI--TAPS
               Fischerella sp. NIES-3754          FNFNNSILDSQGRVLPSWADVLNRANLGFEVMHERNAHNFPLDLAGGEAVPVAI--TAPS
                Fischerella sp. PCC 9605          FNFNNSILDSQGRVLPSWADVLNRANLGFEVMHERNAHNFPLDLAGGEAVPVAI--TAPS
        Fischerella thermalis PCC 7521            FNFNNSILDSQGRVLPSWADVLNRANLGFEVMHERNAHNFPLDLAGGEAVPVAI--TAPS
     Halomicronema hongdechloris C2206            FNFNNSVLDSQGRVLPSWADVVNRANIGFEVMHERNAHNFPLDLASGPATPV--AYKAPA
        Hydrococcus rivularis NIES-593            FNFNNSIHDAQGRVLPSWADVINQANLGFEVMHERNAHNFPLDLASGKAIPV--ALRAPA
                 Leptolyngbya sp. JSC-1           FNFNHSVLDSQGRVVNTWADVLNRANLGFEVMHERNAHNFPLDLAGGEAVPIAMMMAAPV
           Mastigocoleus testarum BC008           FNFNNSILDSEGRVLPSWADLLNRANLGFEVMHERNAHNFPLDLACGEAVPVNVSFKAPA
           Mastigocoleus testarum BC008           FNFNNSILDSEGRVLPSWADLLNRANLGFEVMHERNAHNFPLDLACGEAVPVNVSFKAPA
  Oscillatoriales cyanobacterium JSC-12           FNFNQSVLDSQGKVVNTWADVLNRANLGFEVMHERNAHNFPLDLAGADPIPV--ALRAPA
              Pleurocapsa sp. PCC 7327            FNFNNSIHDAQGRVLPSWADVINQANLGFEVMHERNAHNFPLDLASGEAIPV--ALRAPA
            Synechococcus sp. PCC 7335            FNFNHSVLDSQGRVLPSWADVVNRASLGFEVMHERNAHNFPLDLASGESVQV--AMRAPH

                                                  361
                                                  |
          Calothrix parasitica NIES-267           LAG--------
                    Calothrix sp. PCC 7507        IAS--------
     Chlorogloeopsis fritschii PCC 6912           ITA--------
     Chlorogloeopsis fritschii PCC 9212           ITA--------
  Chroococcidiopsis thermalis PCC 7203            IAEVPPQAIAS
                Calothrix sp. NIES-3974           ITM--------
            Fischerella major NIES-592            ITA--------
        Fischerella muscicola PCC 7414            ITA--------
                  Fischerella sp. JSC-11          ITA--------
               Fischerella sp. NIES-3754          ITA--------
                Fischerella sp. PCC 9605          ITA--------
        Fischerella thermalis PCC 7521            ITA--------
     Halomicronema hongdechloris C2206            IHG--------
        Hydrococcus rivularis NIES-593            IAS--------
                 Leptolyngbya sp. JSC-1           LPS--------
           Mastigocoleus testarum BC008           LAAS-------
           Mastigocoleus testarum BC008           LAAS-------
  Oscillatoriales cyanobacterium JSC-12           IAS--------
              Pleurocapsa sp. PCC 7327            IAS--------
            Synechococcus sp. PCC 7335            IGA--------
```

Fig. S5 Alignment of D1FR protein sequences recovered in this analysis.

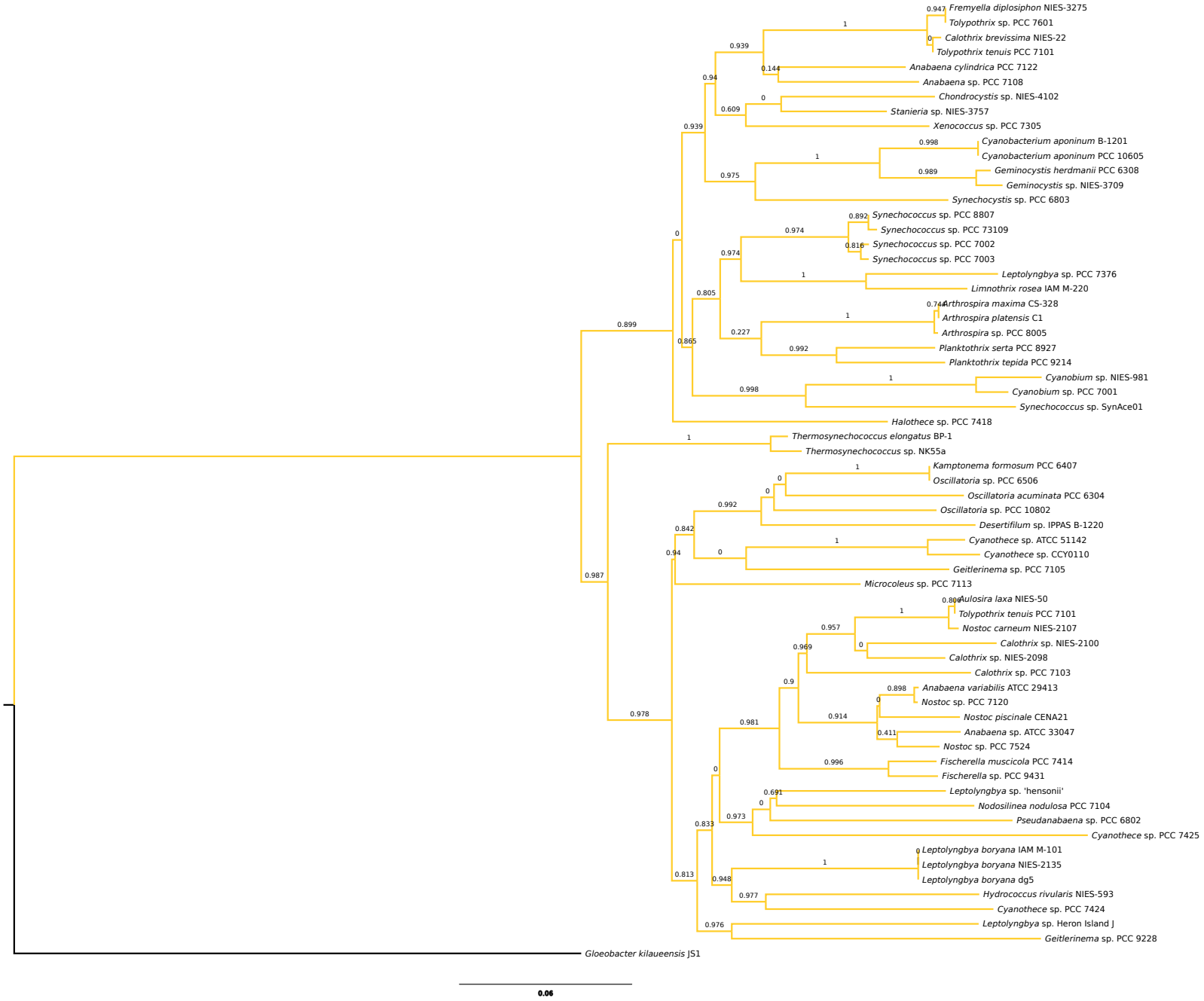

Fig. S6 Rooted maximum likelihood phylogeny of G3 protein sequences using G0 from *Gloeobacter kilaueensis* JS1 as the outgroup. Branch supports are expressed as SH-like aLRT probabilities. Colouring follows that used in Fig. 1 of main text.

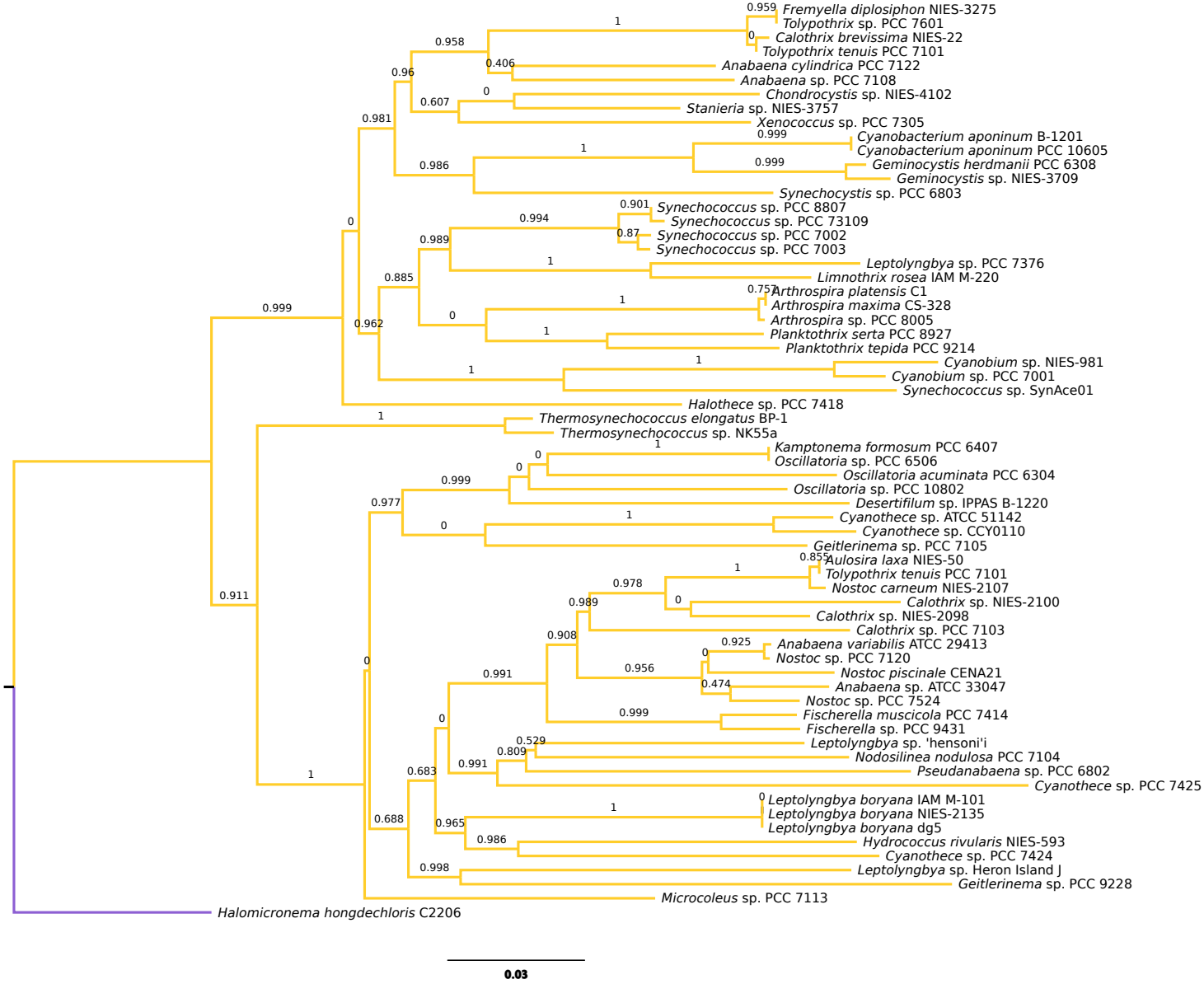

Fig. S7 Rooted maximum likelihood phylogeny of G3 protein sequences using the G1 protein sequence from *Halomicronema hongdechloris* C2206 as the outgroup. Branch supports are expressed as SH-like aLRT probabilities. Colouring follows that used in Fig. 1 of main text.

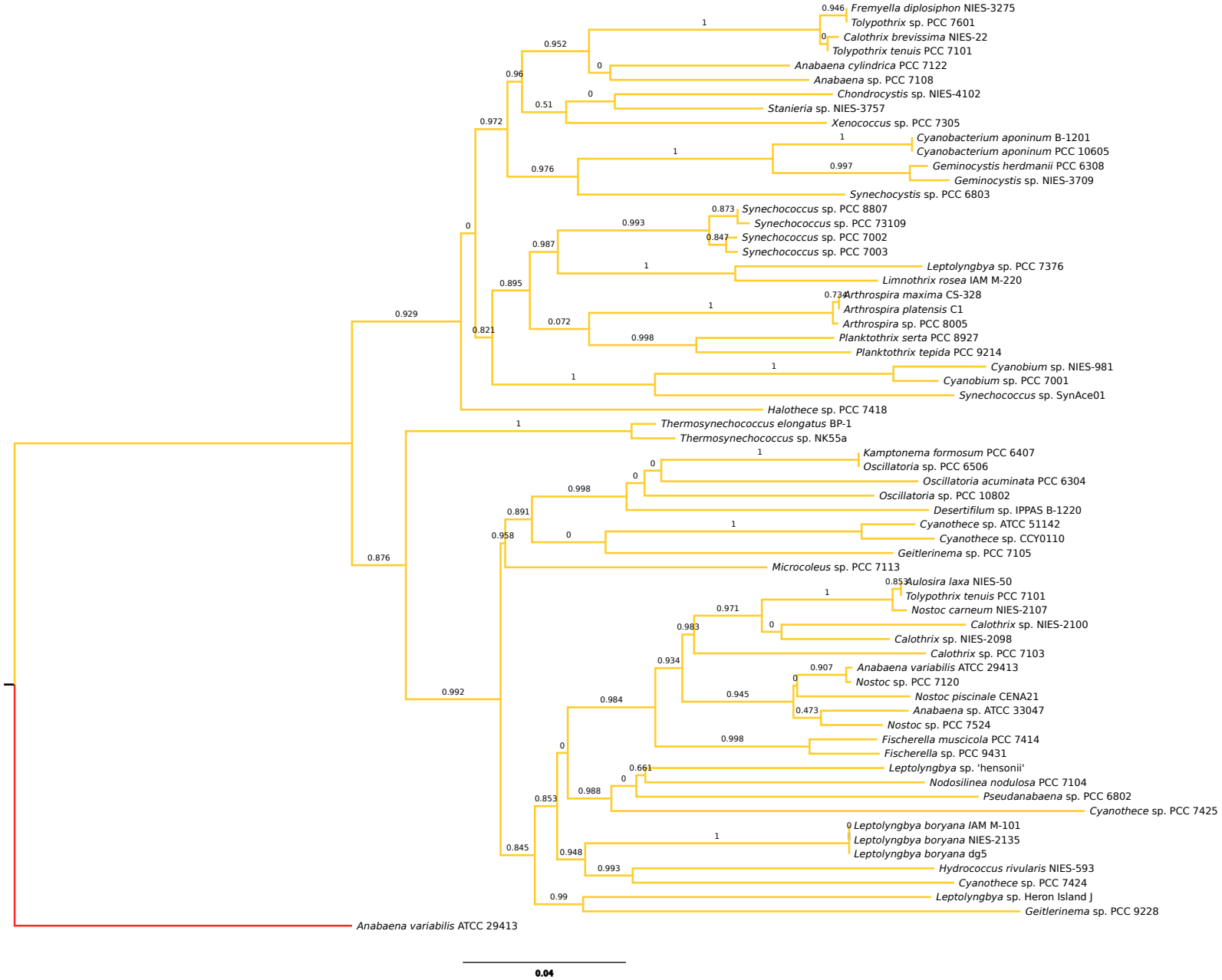

Fig. S8 Rooted maximum likelihood phylogeny of G3 protein sequences using the G2 protein sequence from *Anabaena variabilis* ATCC 29413 as the outgroup. Branch supports are expressed as SH-like aLRT probabilities. Colouring follows that used in Fig. 1 of main text.

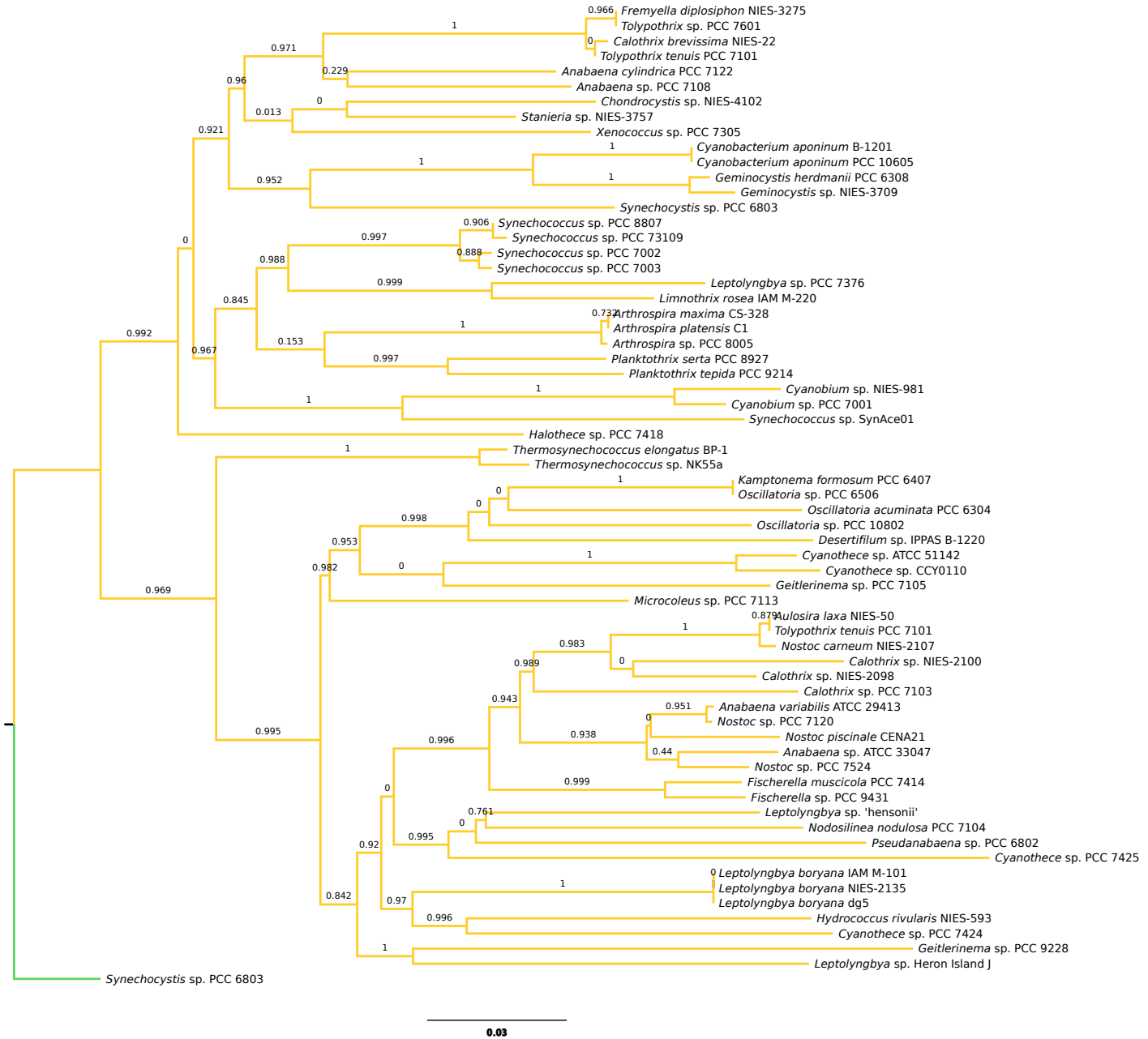

Fig. S9 Rooted maximum likelihood phylogeny of G3 protein sequences using the G4 protein sequence from *Synechocystis* sp. PCC 6803 as the outgroup. Branch supports are expressed as SH-like aLRT probabilities. Colouring follows that used in Fig. 1 of main text.

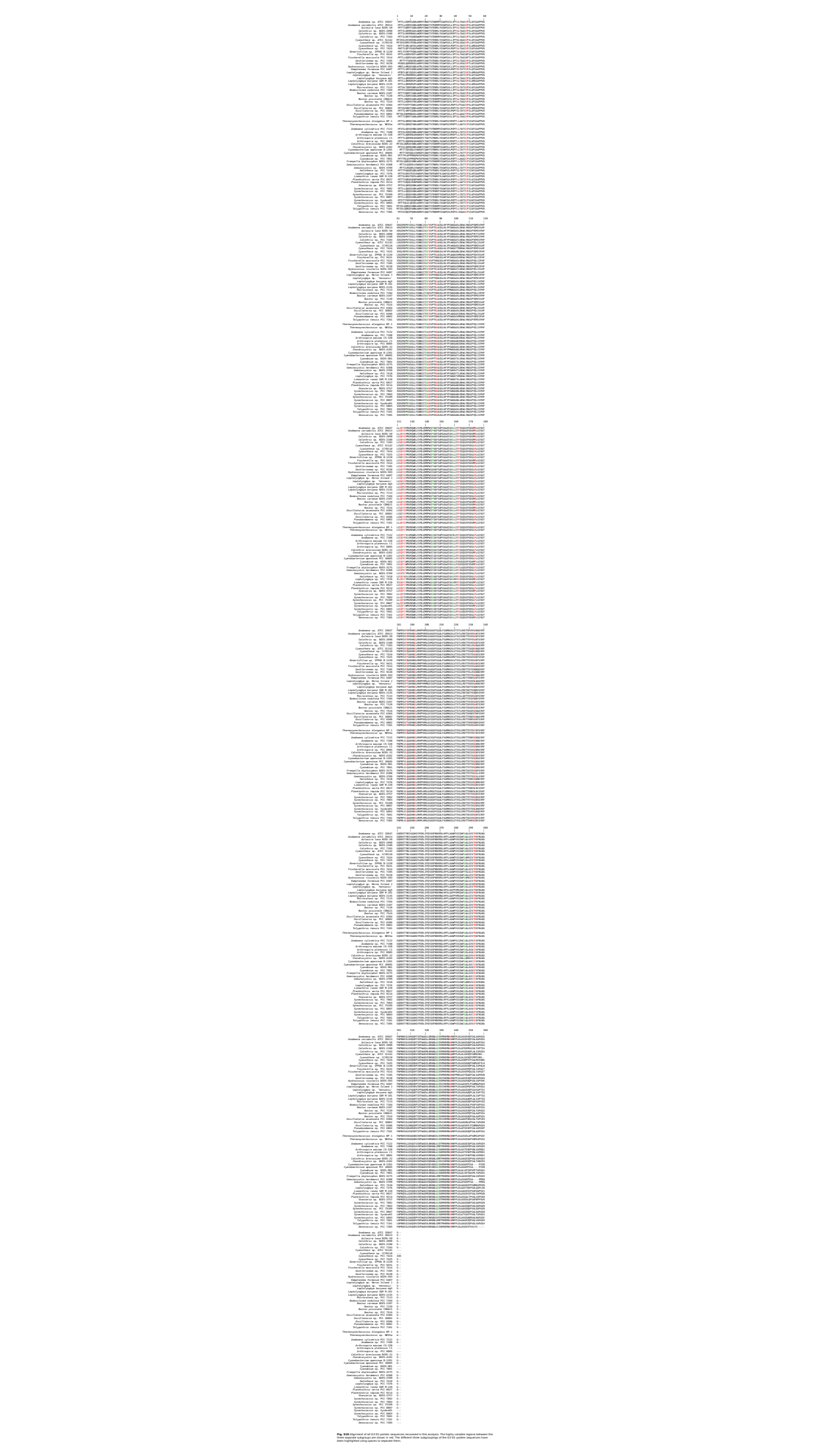

Fig. S10 Alignment of all 0.3 D1 protein sequences recovered in this analysis. The highly variable regions between the three separate subgroups are shown in red. The different three subgroupings of the 0.3 D1 protein sequences have been highlighted using spaces to separate them.