Supplementary Information for

Snowball Earths, population bottlenecks, and the evolution of marine photosynthetic bacteria

Hao Zhang, Ying Sun, Qinglu Zeng, Sean A. Crowe, Haiwei Luo

Haiwei Luo

Email: hluo2006@gmail.com

This PDF file includes:

Supplemental Methods

References

Figures S1 to S11

Tables S1 to S5

Supplemental Methods

Table of Contents

1. Taxon sampling and gene annotation of Cyanobacteria genomes

By the time of this study (Dec 2018), a total of 309 oxygenic cyanobacterial genomes were available in the NCBI RefSeq database[1], among which 126 were marked as high-quality reference or representative genomes (Table S2). For refined phylogenomic and relaxed molecular clock analyses, the 126 reference or representative genomes in RefSeq and the *Prochlorococcus* and *Synechococcus* collection included in a previous study[2] were used here, with the total number of 159 genomes (Table S2). The latter genomes were included here because they were used to demonstrate an evolutionary mechanism underlying genome reduction of *Prochlorococcus*[2], which forms the basis of the present study. Completeness of all the genomes were assessed using CheckM v1.0.11[3] (Table S2). Clusters of orthologous group (COG) assignments for protein sequences were performed using RPSBLAST against the NCBI COG database (Dec. 2014 release)[4]. Only the top COG hit for each protein was retained, which satisfied the domain-specific score threshold compiled from NCBI-curated domains and an e-value cutoff of $1e^{-3}$. Additional functional annotations were carried out using the KEGG database (2017 release) by BLASTP v2.2.6 and subsystem annotations at the RAST Server platform[5].

2. Timing the evolution of *Prochlorococcus*

2.1 Relaxed molecular clock method implemented in MCMCTree

The molecular clock hypothesis provides a powerful way to estimate species divergence time based on molecular sequences[6]. Based on this theory, the genetic distance of two homologous sequences increases linearly with the length of time since their separation[6]. However, evolutionary rates are often not constant over time and among lineages, which renders the strict clock hypothesis problematic in deep lineages like the Cyanobacteria studied here, and only occasionally useful for trees with shallow roots[7,8]. For this reason, we employed the software MCMCTree[9] to perform relaxed molecular clock analysis, which is known to be intrinsically associated with the use of phylogenomic tree, fossil calibrations, clock model and input sequence data. All these factors have been discussed below.

2.2 Phylogenomic tree construction of Cyanobacteria

The prerequisite for a reliable estimate of divergence time is to have a resolved phylogeny[10]. In the case of Cyanobacteria, the mainly unresolved part resides at the LPP lineage, which contains *Leptolyngbya*, *Plectonema*, *Phormidium*, and *Synechococcus* sp. PCC7335[11,12]. In published phylogenies, LPP was a monophyletic group either located at the basal of the Microcyanobacteria group which contains *Synechococcus* and *Prochlorococcus*[13,14], or at the basal of the Macrocyanobacteria group which contains the N_2-fixing Pleurocapsales and Nostocales[15,16]. We intended to solve this phylogenetic discrepancy before performing the time estimation.

Ortholog identification among oxygenic Cyanobacteria genomes

Using these 159 oxygenic Cyanobacteria genomes, we identified 381 single-copy orthologous gene families present in at least 155 genomes by implementing the orthology matrix algorithm (OMA v2.1.1)[17]. We further examined whether all the members of each family shared the same COG functional category, and screened for potential inter-phylum horizontal gene transfer (HGT) using a BLASTP-based protocol similar to the one used in a previous study[18]. A potential inter-phylum HGT event was defined as a cyanobacterial query with a non-cyanobacterial top hit (excluding the query itself; with an e-value ≤ 1e-10 and a percent of identity ≥ 35%) from the NCBI nr database[19]. Finally, a total of 214 (out of 381) single-copy orthologous gene families that met all the requirements were retained for downstream analyses (Table S4).

Phylogenomic tree construction of oxygenic Cyanobacteria based on the complete set of orthologs

The orthologous protein sequences were aligned using the E-INS-I refinement method of MAFFT v7.271[20], and gaps were removed. The concatenation of the 214 single-copy orthologous gene families resulted in an alignment of 65,818 amino acid sites. PartitionFinder v2.1.1[21] was used to determine the optimal partitioning schemes and best-fitting models using a greedy search with Bayesian information criterion (BIC). Phylogenetic analyses were performed using RAxML v8.2.10 (100 bootstrap replicates with GAMMA model of rate heterogeneity applied to each partition)[22] (Fig. S4A) and MrBayes v3.2.6[23] (Fig. S4C). Each Bayesian execution computed two independent runs with four chains, running for 4,000,000 generations with a burn-in fraction of 25% and a sampling frequency of 2,000. Convergence between runs and posterior probabilities of the estimates was determined using Tracer v1.6[24].

Phylogenomic tree construction of oxygenic Cyanobacteria based on the compositionally homogeneous subset of orthologs

An evident variation in G+C content among lineages (Table S2) suggested putative compositional heterogeneity across taxa[25]. To assess whether each orthologous gene family significantly departs from the assumption of homogeneity, we carried out the simulation-based test implemented in the P4 phylogenetic toolkit[26], following a previous procedure in the analysis of Alphaproteobacteria phylogeny[27]. For each individual ortholog alignment, we inferred the optimized parameters for the best-fitting substitution model based on ProtTest analysis[28], and used the resulting maximum likelihood (ML) tree as the phylogram on which 1,000 replicates were simulated. The distribution of amino acid compositions in the simulated data was subsequently compared with that of the empirical data under the χ^2 statistic. Eventually, a set of 90 (out of 214) single-copy orthologous gene families confirmed compositional homogeneity at the 0.05 significance level (Table S4). Phylogenetic analyses were performed again using these 90 families in the same way as elucidated above using both ML (Fig. S4B) and Bayesian (Fig. S4D) approaches, except that the MrBayes runs were ensured with convergence at the 3,000,000th generation (instead of 4,000,000th) when the average standard deviation of the split frequency reached as low as 0.002 (< 0.01).

Phylogenomic tree construction of the Cyanobacteria phylum based on the subset of orthologs showing compositionally homogeneity

To incorporate non-oxygenic Cyanobacteria as outgroups in our molecular clock analyses, we obtained eight metagenome-assembled genomes (MAGs) of Melainabacteria and Sericytochromatia from GenBank. All these MAGs are known to be closely related to oxygenic Cyanobacteria, and have been used as outgroups in a previous study[29]. Completeness of these MAGs were assessed using CheckM v1.0.11[3] (Table S2). We predicted protein sequences of these MAGs using the software Prokka v1.12[30], which were then combined into the protein sequence dataset of oxygenic Cyanobacteria for another round of ortholog identification using OMA v2.1.1[17]. To simplify the process of phylogenomic tree construction, we extracted the previously identified set of compositionally homogeneous orthologs without additional simulation tests. They were used to build the phylogenomic tree of Cyanobacteria phylum using the software IQ-Tree v2.0 with automatically assigned amino acid substitution model under 1,000 ultrafast bootstraps (Fig. S4E).

Resolved phylogeny of Cyanobacteria

Using a concatenation of the protein sequences of the complete set (n=214) of single-copy orthologous gene families shared by 159 high quality oxygenic cyanobacterial genomes which contain more LPP members, the LPP lineage forms a polyphyletic group separately located at the basal of both Microcyanobacteria and Macrocyanobacteria in the ML (Fig. S4A) and Bayesian trees (Fig. S4C). Interestingly, using a concatenation of protein sequences of the remaining composition-homogeneous gene families (n=90), the phylogenies with the ML (Fig. S4B) and Bayesian (Fig. S4D) methods became fully congruent in which the LPP lineage became a monophyletic group and located at the basal of the Microcyanobacteria group. This phylogenetic structure has been commonly used in recent studies of time estimates[12-14], and remains stable when non-oxygenic Cyanobacteria outgroups were incorporated (Fig. S4E). We therefore employed the phylogeny shown in Fig. S4D and S4E for molecular dating analyses.

2.3 Justification of calibrations used for the molecular dating analyses

Molecular dating analyses are proposed to be intrinsically tied to calibration points[10]. In the case of Cyanobacteria, there are two major ways to calibrate their evolution depending on whether the non-oxygenic Cyanobacteria lineages are used or not. In both ways, three time constraints are commonly used to calibrate the evolution of Cyanobacteria, which target the origin of oxygenic Cyanobacteria, the origin of Nostocales, and the origin of Pleurocapsales[12,14,31]. However, when non-oxygenic Cyanobacteria lineages are included, additional time constraints on the root of Cyanobacteria phylum are required.

Despite the rigorous considerations of Cyanobacteria time constraints in previous studies, we notice that the way how fossil calibrations were applied in some of those studies was not appropriate (C1-C6 in Table S1). Thus, we modified the commonly used calibration sets in the present study (C7-C14 in Table S1) and also proposed a new strategy to calibrate the evolution of Cyanobacteria when non-oxygenic Cyanobacteria lineages are included (C15-C38 in Table S1). Details were provided below.

Calibration of the Nostocales group

The time constraints for the crown group of Nostocales have been heavily debated. The maximum boundary of Nostocales was set at different ages in previous studies. First, it was inferred based on heterocysts, which are specialized cells for nitrogen fixation under oxic conditions[32]. As heterocysts were proposed to originate at the time when the atmospheric oxygen became increasingly available at 2,450 Mya[33,34], this age was once set as the maximum boundary of Nostocales. Second, it was inferred based on akinetes, which is another type of differentiated cell of Nostocales for survival under extreme environmental conditions[34]. Since Nostocales is not the only group in Cyanobacteria that produce akinetes, the age (2,100 Ma) of the earliest known akinetes fossil discovered in West Africa was used as the maximum boundary of Nostocales[14,34]. Third, the Nostocales cells are featured with morphological characters including the presence of sheath (condensed part of the akinete coat) and large cell diameter[15]. As ancestral state reconstruction indicates that these characters occurred before the presence of Nostocales, the maximum age of Nostocales was set to 1,900 Ma when microfossils with both sheath and large cell diameter first appeared[15,35]. In terms of the minimum boundary, since the previously mentioned akinete fossil identified at 2,100 Ma was later inferred to be affiliated with Nostocales, the minimum boundary of Nostocales was set to 2,100 Ma in previous study[34,36]. An alternative minimum age of this lineage was set to 1,600 Ma due to the discovery of the nostocalean akinetes fossil in McArthur Group, northern Australia[37]. We noticed that the akinete fossil identified to 2,100 Ma was used as either the maximum boundary or the minimum boundary of Nostocales in different Cyanobacteria dating analyses[12,14]. Although being self-contradictory, we still employed this boundary in different calibration sets (Table S1) for the purpose of comparison.

We note that morphological fossils such as akinetes and heterocysts have been used as the maximum bound to calibrate the crown group of Nostocales in previous studies[14,31]. However, given the potentially large gap between the initial appearance of an apomorphic character and its first fossilization time[38], the placement of these fossils on crown group of Nostocales may overly constrain the age prior and lead to false precisions in time estimates. Given the fact that apomorphic characters must have evolved earlier than the divergence of the crown group of assigned lineage, a more secure way to use these morphological fossils is to constrain the minimum age on total groups[38]. From this perspective, the use of the nostocalean akinete fossils as the minimum constraints in previous studies are also inappropriate, as they were placed on the crown group of Nostocales[12,31]. Given these considerations, in the present study, we employed these morphological fossils to calibrate the lower bounds of the Nostocales total group regardless of whether the non-oxygenic Cyanobacteria lineages were used or not, and left the upper limit of Nostocales group open to avoid overly precise age estimates (C9-C38 in Table S1).

Calibration of the Pleurocapsales group

The time constraints for the crown group of Pleurocapsales are also contentious. Members of Pleurocapsales have large cell diameters[15]. Since this character has been proposed to evolve earlier than the ancestor of Pleurocapsales, the maximum age of Pleurocapsales was once set to 2,450 Ma when the large cell diameter appeared in microfossil[15]. Alternatively, since Pleurocapsales evolved later than filamentous and coccoid Cyanobacteria[14], which were proposed to occur at 1,900 Ma based on the microfossil identified in Gunflint chert[39], the maximum boundary of Pleurocapsales was set to 1,900 Ma in previous Cyanobacteria dating analyses[14,31]. The minimum age of Pleurocapsales was set to 1,700 Ma because of the microfossil identified in Hebei, China[40,41].

We argue that the use of morphological fossils such as filamentous and coccoid cells as the maximum bound of Pleurocapsales in previous studies was not appropriate for the same reason we provided in the last section 'Calibration of the Nostocales group'. Thus, we modified the use of microfossils at 1,900 Ma as the minimum bound of total Pleurocapsales group. Moreover, since the maximum bound is hard to be established using fossil records[38], we left the upper limit of Pleurocapsales group open (C9-C38 in Table S1).

Calibrations of the root of oxygenic Cyanobacteria

For the root of oxygenic Cyanobacteria (i.e., the root of the phylogenomic tree when non-oxygenic Cyanobacteria lineages are not included; Fig. S4D), the minimum age at 2,320 Mya is commonly applied because of the convincing geochemical evidence for the rise of atmospheric oxygen at that time known as the Great Oxidation Event (GOE)[42], though recent studies showed that GOE may antedate the crown group of oxygenic Cyanobacteria[43,44]. The upper limit calibration of this root has been even more contentious. It was initially reported that 2-methylhopane can be used as a biomarker for Cyanobacteria[45], and the oldest record of this biomarker is dated back to 2,700 Mya[46], but the taxonomic link of 2-methylhopane to Cyanobacteria was challenged by the discoveries that 2-methylhopane is produced by the anoxygenic phototroph *Rhodopseudomonas palustris* under anaerobic conditions[47], and that the key gene for the methylation at the C-2 position of hopanoids was also found in α-Proteobacteria and Acidobacteria[48]. The use of 2,700 Mya as the maximum age of the emergence of oxygenic Cyanobacteria was further weakened by a recent finding that the previously studied samples contained contaminants[49]. On the other hand, ample geochemical evidence based on various sensitive redox proxies indicates the appreciable levels of the atmospheric oxygen at 3,000 Mya[50-52], which has been used as the upper bound of crown oxygenic Cyanobacteria in recent molecular clock analyses[14,31]. Consequently, in the cases when non-oxygenic Cyanobacteria lineages were not included, we calibrated the lower and upper limit of the crown oxygenic Cyanobacteria at 2,320 Mya and 3,000 Mya, respectively (C9-C14 in Table S1; Fig. S4D).

Recently, two lineages have been identified as the outgroups of oxygenic Cyanobacteria: Melainabacteria and Sericytochromatia[29,53]. Members of these outgroup lineages are proposed to lack essential genes for photosynthesis and carbon fixation, suggesting that the last common ancestor of Cyanobacteria was non-phototrophic[29]. If this is the case, the oxygenic photosynthesis could be an evolutionary synapomorphy, which likely evolved at the stem lineage of oxygenic Cyanobacteria. Thus, when non-oxygenic Cyanobacteria lineages are incorporated, it is more appropriate to constrain the lower bound of total oxygenic Cyanobacteria instead of the upper bound of crown oxygenic Cyanobacteria using the geochemical evidence that atmospheric oxygen became available at 3,000 Mya[50-52] (C15-C38 in Table S1; Fig. S4E).

<u>Calibrations of the root of Cyanobacteria phylum</u>

In the cases when non-oxygenic Cyanobacteria lineages were included, we have to calibrate the root of phylogeny (i.e., the root of the Cyanobacteria phylum; Fig. S4E). To avoid overly precise age estimates, we constrained the upper limit of the Cyanobacteria root as ancient as possible. Given the potentially great influence of root prior on time estimates[54], we attempted different maximum prior ages for comparison. For example, we used 4,200 Mya, 4,000 Mya and 3,800 Mya by considering the time when the planet Earth became habitable and fostered the earliest life[55,56] (C15-C32; Table S1). Additionally, a more conservative age at 4,500 Mya was used, since it was the time when the planet Earth formed[55] (C33-C38; Table S1).

2.4 Selection of molecular clock model

Molecular clock model is known to have a strong impact on posterior age estimates[57]. The software MCMCTree implements different relaxed molecular clock models for time estimation, including auto-correlated rates (AR) model and independent rates (IR) model. The former assumes that the evolutionary rates in daughter species are statistically distributed around the parental rates, whereas the latter assumes a fully independent rate among evolutionary branches[58].

To assess the fitness of each clock model in our data, we compared the Bayes factors (BF) of these models using the thermodynamic integration method in the package "mcmc3r"[58]. While the method is powerful, it is very computationally intensive. Thus, we used the calibration set C9 as the representative for Bayesian model selection. Our results indicate that the IR model is superior to the AR model, as the BF value of the former is much higher than that of the latter (0.999 vs 0.001). We therefore employed the IR model in the following molecular clock analyses.

2.5 Input sequence data for molecular clock analysis

As an enlarged sequence dataset is able to improve the precision of time estimation based on the infinite-site theory[59], we employed as many as 25 core protein-coding genes[60] and two rRNA genes (16S, 23S) in the present study. Since substitutions at the third codon positions are largely silent and thus reach saturation rapidly, only the first and second codon positions of the 25 protein-coding genes were used. These 25 conserved protein-coding genes were previously identified from a genomic dataset spanning multiple bacterial and archaeal phyla and used to infer the evolutionary timeline of those groups[60]. For each gene, we selected the best-fitting nucleotide substitution model by jModelTest[61], and calculated a rough substitution rate using BASEML[9] under a strict molecular clock. Further, the mean substitution rate was calculated based on the substitution rates of all input gene sequences, and then was used to inform the Dirichlet-gamma prior (rgene_gamma) in MCMCTree.

2.6 Assessing the precisions of molecular clock analyses

Evaluation of molecular clock analyses is important, since using different calibration set leads to a difference up to over 320 Ma in the estimates of the SBE-LCA when the non-oxygenic Cyanobacteria were not included (i.e., the last common ancestor of *Prochlorococcus* HL, LLI and LLII/III) (655 Mya under calibration set C6 vs. 981 Mya under calibration set C3; Fig. S2). Although the variation reduces to less than 10 Ma when the ages were estimated with the modified calibration sets (C7-C14 in Table S1), statistical evaluations of these analyses are valuable. The Bayesian inference approach that implemented in MCMCTree integrates the information from both calibrations and genetic data for posterior age estimation[62]. Once the use of a calibration set is settled, according to the infinite-site theory, increased number of sites are recommended for molecular clock analysis, as they reduce the uncertainty in genetic distance estimate and increase the precision of the posterior time estimates[59]. Theoretically, if sequences of infinite sites are used, the uncertainties in posterior time estimates are solely imposed by the uncertainties of the calibrations[59]. By plotting the widths of 95% HPD interval against the posterior mean ages, we are able to assess the precision of the molecular clock analyses by comparing the slopes of the regression lines. A greater slope represents a lower precision of the time estimates[62,63].

It has been repeatedly proposed that using multiple and more calibrations often lead to more reliable estimation than using less or even a single calibration[59,64]. Consistently, we showed that the time estimates based on calibration set C7 and C8 with a single calibration node has a high slope of 0.19 and 0.29, respectively (Fig. S3). It means that every 100 Ma divergence adds 19 and 29 Ma uncertainty in the posterior time estimates, respectively. According to this rule, the time estimates based on calibration set C14 has the lowest slope (0.149; Fig. S3), suggesting that the posterior time estimates of the SBE-LCA derived from this set are most precise. We did not further consider the analyses based on the calibration sets C1-C6 because the calibrations were not appropriately placed on the phylogeny.

We note that including Melainabacteria and Sericytochromatia consistently lead to less precise age estimates, as shown by higher slopes of the regression line between HPD width and the posterior age estimates (C15-C38 versus C1-C14 in Fig. S3). Given that genomes of these non-oxygenic Cyanobacteria lineages are fully represented by metagenome-assembled genomes (MAGs) but genomes of oxygenic Cyanobacteria used in our analyses are all derived from pure cultures, we hypothesize that the use of MAGs in molecular dating analysis, particularly those of lineages occupying important phylogenetic positions, may increase the uncertainties of the poterior age estimates. The quality of MAGs is questionable. While the CheckM[3] predicted that all of the MAGs used here show high level of completeness and low level of contamination (Table S2), these assessments may not be reliable, as shown in a recent benchmarking study[65]. For example, MAGs with estimated completeness as high as 95% may capture only three-fourths of the population core genes and a half of the variable genes, suggesting a greater amount of DNA is missing in the assemblies than estimated[65]. Moreover, MAGs with estimated contamination as low as 1.5% may incorporate up to 5% of their genes with other taxonomic origins, suggesting a potentially higher contamination rate in the MAGs[65].

3. Reconstruction of gene gain and loss processes

3.1 Using AnGST

Genome content evolution via gene gains and losses was inferred using the gene tree vs. species tree reconciliation approach implemented in AnGST[66]. The 62 genomes comprising the *Synechococcus-Prochlorococcus* monophyletic group were retrieved from the ultrametric cladograms yielded by the molecular dating analyses as described above (Fig. S11). Gene trees were constructed using the following procedure. Firstly, homology relationships among proteins of the 62 *Synechococcus* and *Prochlorococcus* genomes were determined using OrthoFinder v2.2.1[67] with DIAMOND as the alignment program[68]. We identified 4,689 orthogroups (out of 5,615 orthogroups in total) each with at least three sequences. Next, multi-sequence alignments were constructed for each orthogroup using-E-INS-I method implemented in the software MAFFT v7.222[20], and trimmed with trimAl v1.4 ('-gappyout' option) to remove poorly aligned and excessively gapped regions[69]. Lastly, gene trees were built using IQ-TREE v1.6.2[70] under the ModelFinder feature (-m MFP) with ultrafast bootstrapping (1,000 replicates).

The reconciliation was inferred for each orthogroup under a generalized parsimony framework to achieve a minimum number of evolutionary events (gene loss, gene duplication, horizontal gene transfer [HGT], gene birth and speciation) along the species tree, using event penalties determined by the genome flux analysis[66]. The genome flux analysis requires a minimal average difference in genome size between the ancestor and the descendant across the branches of the species tree, resulting in a set of optimized event penalties. We implemented the genome flux analysis with the speciation penalty fixed at 0.0 and the loss penalty at 1.0 as recommended in a previous study[66]. The minimal genome flux was achieved when the HGT and duplication penalties are equal to 3.0 and 4.0, respectively (Fig. S8). The HGT penalty inferred here agreed with the value achieved in a previous study based on a wide range of taxa across all three domains[66], and also confirmed HGT as the strongest effect on the genome flux as suggested in previous studies[66,71,72].

For all reconciliations performed, we enforced the time consistency (ultrametric = True) and restricted transfers to occur only between contemporaneous lineages. All 1,000 bootstrap replicates of each gene tree were provided to AnGST to resolve the gene tree phylogenetic uncertainties through amalgamation[66]. AnGST incorporates the gene tree refinement procedure into the reconciliation process, and yields a chimeric gene tree (from the bootstrap replicates) which results in the lowest reconciliation cost, satisfying a generalized parsimony criterion[66]. The numbers of gain, loss, and transfer events were summarized based on the AnGST output for each orthogroup across all branches along the species tree.

3.2 Using BadiRate

Gene gains and losses were also inferred with the likelihood-based method equipped in BadiRate v1.35[73], which uses a full ML approach to determine the gene family turnover rates that maximize the probability of observing the gene count patterns provided in the family size table. A table of gene counts, consisting of all the aforementioned 5,615 orthogroups inferred by OrthoFinder v2.2.1[67], and the same ultrametric time tree used in the AnGST analysis were used as the inputs. We fit nine different combinations of turnover rates (e.g., Birth-Death-Innovation model [BDI], Gain-Death model [GD], Lambda model [L] and Lambda-Innovation model [LI]) and branch models (e.g., Global-Rates model [GR], Branch-Specific-Rates model [BR], or Free-Rates model [FR]), including BDI/GD/L/LI+GR+ML, BDI/GD/L/LI+BR+ML and GD+FR+ML. Due to the computational intensity of the FR branch model, it was only implemented with the GD model under the ML framework. In the BR model, the four branches leading to the last common ancestor (LCA) of all *Prochlorococcus*, of the HL, LLI and LLII/III clades, of the HL and LLI clades, and of the HL clade, were allowed to have branch-specific turnover rates, whereas other branches were assumed to share the same rate. To avoid local optima, we ran 100 replicates for each ML analysis using different starting values (-start_val 1 accompanied with distinct seeds [-seed] provided by a random number generator). The likelihood of different runs among distinct models were compared (Fig. S9). The presented estimates were based on the run with the maximum likelihood in each selected model.

3.3 Gene gain and loss data integration

For both methods, the corresponding results were compared and summarized to determine the common patterns shared by all analyses. AnGST categorizes the variation of genome contents into born, loss, duplication and horizontal gene transfer, whereas BadiRate only reports gene gain and loss through copy number changes. To smooth the comparison of all attempts, we standardized a "gain" event as the increase in the copy number of a gene family (including born, duplication and HGT), and accordingly a "loss" event as the decrease in the copy number of a gene family (including complete and partial loss) (Fig. S12). Since the two methods gave a similar pattern of genome size reduction, we presented the number of gene gains and losses derived from the AnGST in the main text.

4. Calculating the rate of nonsynonymous nucleotide substitutions leading to radical and conservative amino acid changes, respectively

Previous study identified an excess of radical changes in *Prochlorococcus* HL and LLI/II/III lineages in comparison to their LLIV relatives[2]. Here, radical changes are defined as nonsynonymous nucleotide substitutions leading to the replacements between amino acids with distinct physicochemical properties (charge, volume and polarity; Table S5), while conservative changes are among similar amino acids. The Radical and

Conservative change Calculator (RCCalculator http://www.geneorder.org/RCCalculator/) was developed to compute the radical and conservative substitution rates (d_R and d_C) which takes into account the GC biases of the DNA sequences[2].

In the present study, a total of 543 single-copy orthologous gene families, shared by all the 61 genomes of *Prochlorococcus* and *Synechococcus* clade 5.1/5.2, were retrieved from the aforementioned results of OrthoFinder v2.2.1[67]. Genes were aligned at the amino acid level using MAFFT v7.271[20], and DNA sequences were imposed on the alignments. Gaps and codons with ambiguous nucleotides were removed. The ratio of nonsynonoymous to synonymous substitution rates (d_N/d_S) was calculated using KaKs_Calculator under YN model for each of the orthologous gene pairs[74], and the median value of each gene family was used for RCCalculator. The transition/transversion ratio (t_S/t_V) of each gene family, also required by RCCalculator, was estimated using MEGA-CC v7.0.26[75]. By incorporating the uncultivated lineages of *Prochlorococcus,* a total of 751 single-copy orthologous gene families shared by 62 out of 65 genomes were retrieved and subject to the same procedures as described.

A total of six cases were considered for the calculation of d_R and d_C, including two ways of categorizing amino acids (by charge and by volume and polarity) and three approaches of GC-bias correction (uncorrected, on codon frequency correction, and on amino acid composition correction). Under each case, given a gene family, RCCalculator estimates the number of radical and conservative sites for each sequence (R_i and C_i, where $i \in [1,61]$), as well as the numbers of radical and conservative differences of each sequence pair (ΔR_{ij} and ΔC_{ij}, where $i \in [1,61]$, $j \in [1,61]$, and $i \neq j$). Then, the pairwise d_R/d_C ratio was defined as $[\frac{\Delta R_{ij}}{mean(R_i,R_j)}]\Big/[\frac{\Delta C_{ij}}{mean(C_i,C_j)}]$, where $i \in [1,61]$, $j \in$

$[1, 61]$, and $i \neq j$. In our study, each gene family had approximately 240 d_R/d_C ratios resulted from the comparisons between sequences of the target group and the reference group (40 genomes in the target group vs. six genomes in the reference group), and 90 d_R/d_C ratios from the control vs. reference comparisons (15 genomes in the control group vs. six reference ones). The mean values of these two categories were then used to represent the "target" and "control" d_R/d_C ratios of the gene family. Further, after pooling all the 543 pairs of d_R/d_C ratios together, sign test and paired t-test were used to determine significant differences between the d_R/d_C ratios from the "target" and "control" groups.

References

1 Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-65 (2007).

2 Luo, H., Huang, Y., Stepanauskas, R. & Tang, J. Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. *Nat Microbiol* **2**, 17091 (2017).

3 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043-1055 (2015).

4 Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33-36 (2000).

5 Aziz, R. K. *et al.* The RAST server: Rapid annotations using subsystems technology. *Bmc Genomics* **9** (2008).

6 Zuckerkandl, E. P., Linus. Evolutionary divergence and convergence in proteins. In Evolving genes and proteins 97-166 (Elsevier, 1965).

7 Kumar, S. Molecular clocks: four decades of evolution. *Nature Reviews Genetics* **6**, 654 (2005).

8 Brown, R. P. & Yang, Z. Rate variation and estimation of divergence times using strict and relaxed clocks. *BMC Evol Biol* **11**, 271 (2011).

9 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591 (2007).

10 Schirrmeister, B. E., Sanchez-Baracaldo, P. & Wacey, D. Cyanobacterial evolution during the Precambrian. *Int J Astrobiol* **15**, 187-204 (2016).

11 Sarma T A. Handbook of cyanobacteria. (CRC Press, 2012).

12 Sánchez-Baracaldo, P., Ridgwell, A. & Raven, J. A. A neoproterozoic transition in the marine nitrogen cycle. *Current Biology* **24**, 652-657 (2014).

13 Shih, P. M. *et al.* Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci USA* **110**, 1053-1058 (2013).

14 Sánchez-Baracaldo, P. Origin of marine planktonic cyanobacteria. *Sci Rep* **5**, 17418 (2015).

15 Blank, C. & Sanchez-Baracaldo, P. Timing of morphological and ecological innovations in the cyanobacteria–a key to understanding the rise in atmospheric oxygen. *Geobiology* **8**, 1-23 (2010).

16 Uyeda, J. C., Harmon, L. J. & Blank, C. E. A comprehensive study of cyanobacterial morphological and ecological evolutionary dynamics through deep geologic time. *PLoS One* **11**, e0162539 (2016).

17 Train, C. M., Glover, N. M., Gonnet, G. H., Altenhoff, A. M. & Dessimoz, C. Orthologous Matrix (OMA) algorithm 2.0: more robust to asymmetric evolutionary rates and more scalable hierarchical orthologous group inference. *Bioinformatics* **33**, i75-i82 (2017).

18 Williams, T. A. *et al.* Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proc Natl Acad Sci U S A* **114**, E4602-E4611 (2017).

19 Coordinators, N. R. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **46**, D8-D13 (2018).

20 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013).

21 Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T. & Calcott, B. PartitionFinder 2: New Methods for Selecting Partitioned Models of Evolution for Molecular and Morphological Phylogenetic Analyses. *Mol Biol Evol* **34**, 772-773 (2017).

22 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of

large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

23 Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* **61**, 539-542 (2012).

24 Drummond, A. J., Rambaut, A., Shapiro, B. & Pybus, O. G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* **22**, 1185-1192 (2005).

25 Foster, P. G. & Hickey, D. A. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol* **48**, 284-290 (1999).

26 Foster, P. G. Modeling compositional heterogeneity. *Syst Biol* **53**, 485-495 (2004).

27 Luo, H. Evolutionary origin of a streamlined marine bacterioplankton lineage. *ISME J* **9**, 1423 (2015).

28 Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164-1165 (2011).

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

30 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069 (2014).

31 Sánchez-Baracaldo, P., Raven, J. A., Pisani, D. & Knoll, A. H. Early photosynthetic eukaryotes inhabited low-salinity habitats. *Proc Natl Acad Sci USA*, 201620089 (2017).

32 Wolk, C. P., Ernst, A. & Elhai, J. in *The molecular biology of cyanobacteria* 769-823 (Springer, 1994).

33 Farquhar, J., Bao, H. & Thiemens, M. Atmospheric influence of Earth's earliest sulfur cycle. *Science* **289**, 756-758 (2000).

34 Tomitani, A., Knoll, A. H., Cavanaugh, C. M. & Ohno, T. The evolutionary diversification of cyanobacteria: molecular–phylogenetic and paleontological perspectives. *Proc Natl Acad Sci USA* **103**, 5442-5447 (2006).

35 Schirrmeister, B. E., Gugger, M. & Donoghue, P. C. Cyanobacteria and the Great Oxidation Event: evidence from genes and fossils. *Palaeontology* **58**, 769-785 (2015).

36 Knoll, A., Golubic, S., Green, J. & Swett, K. Organically preserved microbial endoliths from the late Proterozoic of East Greenland. *Nature* **321**, 856 (1986).

37 Golubic, S., Sergeev, V. N. & Knoll, A. H. Mesoproterozoic Archaeoellipsoides: akinetes of heterocystous cyanobacteria. *Lethaia* **28**, 285-298 (1995).

38 Marshall, C. R. Using the Fossil Record to Evaluate Timetree Timescales. *Front Genet* **10**, 1049 (2019).

39 Sergeev, V., Gerasimenko, L. & Zavarzin, G. The proterozoic history and present state of cyanobacteria. *Microbiology* **71**, 623-637 (2002).

40 Zhang, Y. & Golubic, S. Endolithic microfossils (cyanophyta) from early Proterozoic stromatolites, Hebei, China. *Acta Micropaleontol. Sin* **4**, 1-3 (1987).

41 Golubic, S. & Seong-Joo, L. Early cyanobacterial fossil record: preservation, palaeoenvironments and identification. *Eur J Phycol* **34**, 339-348 (1999).

42 Bekker, A. *et al.* Dating the rise of atmospheric oxygen. *Nature* **427**, 117 (2004).

43 Shih, P. M., Hemp, J., Ward, L. M., Matzke, N. J. & Fischer, W. W. Crown group Oxyphotobacteria postdate the rise of oxygen. *Geobiology* **15**, 19-29 (2017).

44 Betts, H. C. *et al.* Integrated genomic and fossil evidence illuminates life’s early evolution and eukaryote origin. *Nat Ecol Evol* **2**, 1556-1562 (2018).

45 Summons, R. E., Jahnke, L. L., Hope, J. M. & Logan, G. A. 2-Methylhopanoids as biomarkers for cyanobacterial oxygenic photosynthesis. *Nature* **400**, 554 (1999).

46 Brocks, J. J., Buick, R., Summons, R. E. & Logan, G. A. A reconstruction of Archean biological diversity based on molecular fossils from the 2.78 to 2.45 billion-year-old Mount Bruce Supergroup, Hamersley Basin, Western Australia. *Geochim Cosmochim Acta* **67**,

4321-4335 (2003).

47 Rashby, S. E., Sessions, A. L., Summons, R. E. & Newman, D. K. Biosynthesis of 2-methylbacteriohopanepolyols by an anoxygenic phototroph. *Proc Natl Acad Sci USA* **104**, 15099-15104 (2007).

48 Welander, P. V., Coleman, M. L., Sessions, A. L., Summons, R. E. & Newman, D. K. Identification of a methylase required for 2-methylhopanoid production and implications for the interpretation of sedimentary hopanes. *Proc Natl Acad Sci USA* **107**, 8537-8542 (2010).

49 French, K. L. *et al.* Reappraisal of hydrocarbon biomarkers in Archean rocks. *Proc Natl Acad Sci USA* **112**, 5915-5920 (2015).

50 Crowe, S. A. *et al.* Atmospheric oxygenation three billion years ago. *Nature* **501**, 535 (2013).

51 Lalonde, S. V. & Konhauser, K. O. Benthic perspective on Earth's oldest evidence for oxygenic photosynthesis. *Proc Natl Acad Sci USA* **112**, 995-1000 (2015).

52 Planavsky, N. J. *et al.* Evidence for oxygenic photosynthesis half a billion years before the Great Oxidation Event. *Nat Geosci* **7**, 283 (2014).

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

54 Battistuzzi, F. U., Billing-Ross, P., Murillo, O., Filipski, A. & Kumar, S. A Protocol for Diagnosing the Effect of Calibration Priors on Posterior Time Estimates: A Case Study for the Cambrian Explosion of Animal Phyla. *Mol Biol Evol* **32**, 1907-1912 (2015).

55 Nisbet, E. G. & Sleep, N. H. The habitat and nature of early life. *Nature* **409**, 1083-1091 (2001).

56 Sleep, N. H., Zahnle, K. J., Kasting, J. F. & Morowitz, H. J. Annihilation of ecosystems by large asteroid impacts on the early Earth. *Nature* **342**, 139-142 (1989).

57 Dos Reis, M. *et al.* Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. *Current Biology* **25**, 2939-2950 (2015).

58 Reis, M. D. *et al.* Using phylogenomic data to explore the effects of relaxed clocks and calibration strategies on divergence time estimation: primates as a test case. *Syst Biol* **67**, 594-615 (2018).

59 Yang, Z. & Rannala, B. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol* **23**, 212-226 (2005).

60 Battistuzzi, F. U. & Hedges, S. B. A major clade of prokaryotes with ancient adaptations to life on land. *Mol Biol Evol* **26**, 335-343 (2008).

61 Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* **9**, 772 (2012).

62 Inoue, J., Donoghue, P. C. & Yang, Z. The impact of the representation of fossil calibrations on Bayesian estimation of species divergence times. *Syst Biol* **59**, 74-89 (2009).

63 Dos Reis, M. & Yang, Z. The unbearable uncertainty of Bayesian divergence time estimation. *J Syst Evol* **51**, 30-43 (2013).

64 Sauquet, H. *et al.* Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of Nothofagus (Fagales). *Syst Biol* **61**, 289-313 (2011).

65 Meziti, A. *et al.* How reliably do metagenome-assembled genomes (MAGs) represent natural populations? Insights from comparing MAGs against isolate genomes derived from the same fecal sample. *Appl Environ Microbiol* (2021).

66 David, L. A. & Alm, E. J. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* **469**, 93 (2010).

67 Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**, 157 (2015).

68 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using

DIAMOND. *Nature Methods* **12**, 59-60 (2015).

69 Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).

70 Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274 (2015).

71 Kamneva, O. K., Knight, S. J., Liberles, D. A. & Ward, N. L. Analysis of genome content evolution in pvc bacterial super-phylum: assessment of candidate genes associated with cellular organization and lifestyle. *Genome Biol Evol* **4**, 1375-1390 (2012).

72 Richards, V. P. *et al.* Phylogenomics and the dynamic genome evolution of the genus Streptococcus. *Genome Biol Evol* **6**, 741-753 (2014).

73 Librado, P., Vieira, F. G. & Rozas, J. BadiRate: estimating family turnover rates by likelihood-based methods. *Bioinformatics* **28**, 279-281 (2012).

74 Wang, D., Zhang, Y., Zhang, Z., Zhu, J. & Yu, J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, proteomics & bioinformatics* **8**, 77-80 (2010).

75 Kumar, S., Stecher, G., Peterson, D. & Tamura, K. MEGA-CC: computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* **28**, 2685-2686 (2012).

Fig. S1

🟦 Gain (Gene birth, Gene duplication, HGT)

🟥 Loss (Gene family loss, Gene family reduction)

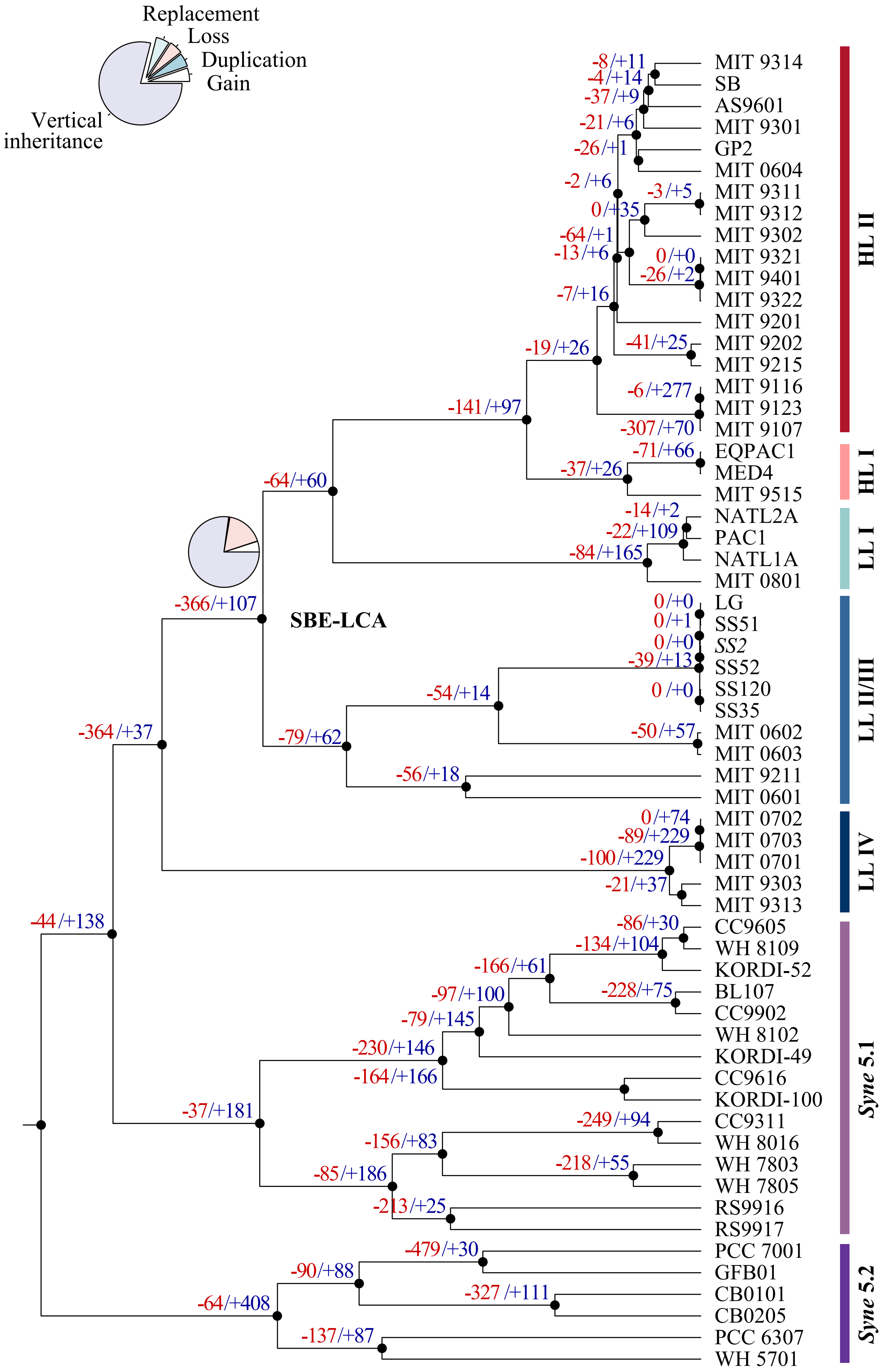


Fig. S1 The number of gene gain and loss events along the genome tree of *Prochlorococcus* and *Synechococcus* reconstructed by AnGST. Gene gain events include gene birth, duplication and HGT, while gene loss events comprise gene family size reduction and loss of entire gene families. The pie chart on the ancestral branches leading to SBE-LCA provides the detailed proportion of each type of genomic event in these key evolutionary stages.

Fig. S2

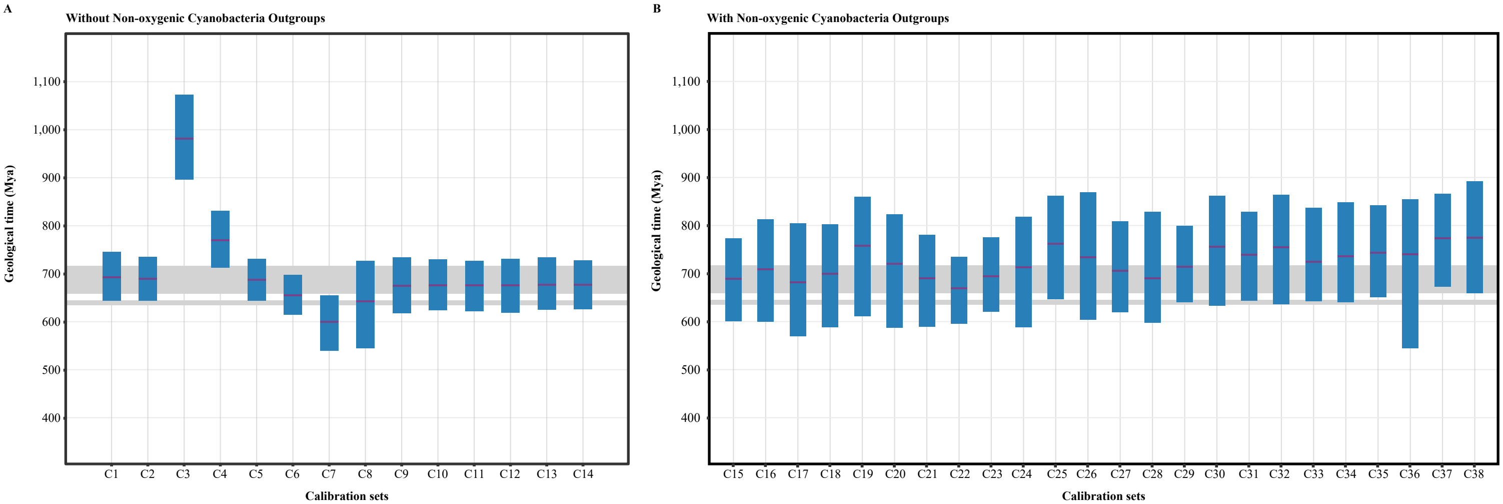


Fig. S2 Divergence time estimates of the ancestral node ‘SBE-LCA’ based on different calibration sets. (A) Calibration sets used for the phylgeony of oxygenic Cyanobacteria group, including some adapted from previous studies (C1-C6) and others modified in the current study (C7-C14). (B) Calibration sets used for the phylogeny of Cyanobacteria phylum including both oxygenic and non-oxygenic groups (C15-C38). The purple lines and blue vertical bars represent the posterior age estimates and the 95% highest probability density (HPD) intervals, respectively. The upper and the lower horizontal grey bars represent the time of Sturtian glaciation and Marinoan glaciation, respectively.

Fig. S3

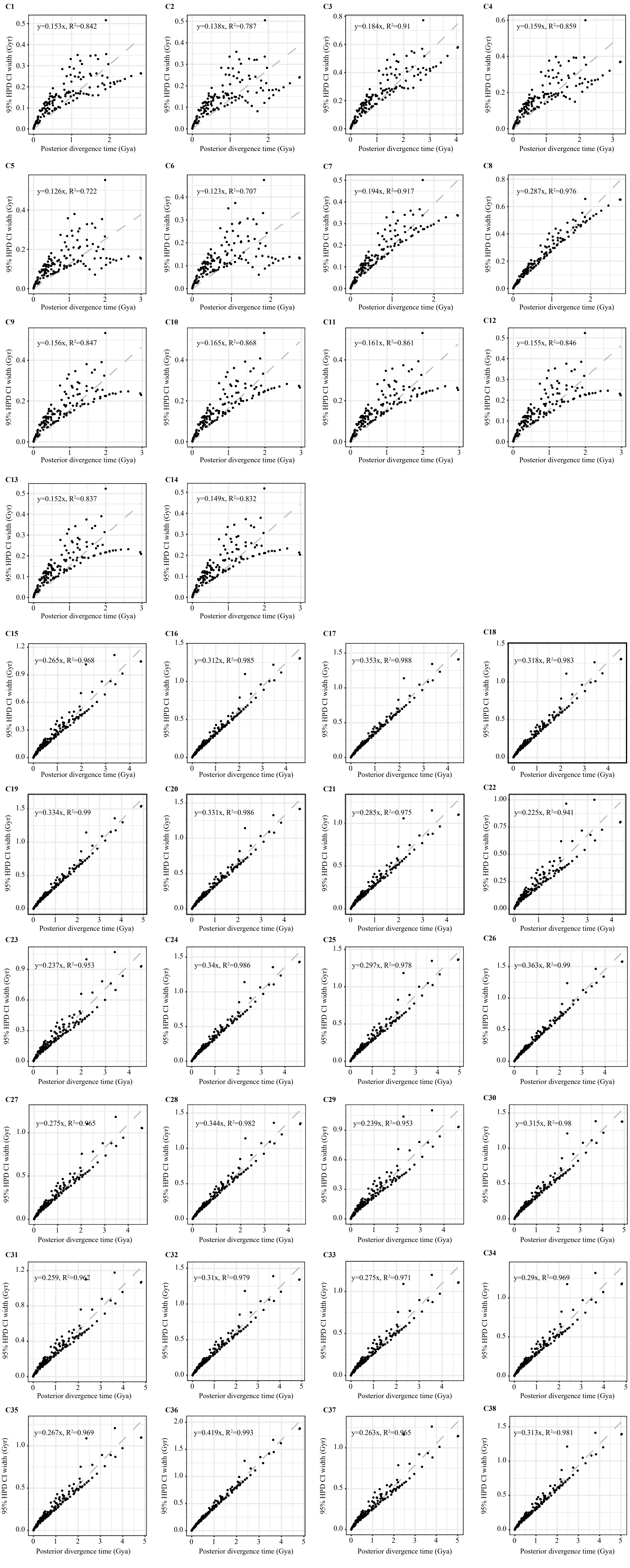


Fig. S3 The infinite-site plots of time estimates based on different calibration sets (C1-C38; see Table S1). The width of the 95% highest probability density (HPD) interval was plotted against the posterior means of the divergence time. A lower slope of the regression line suggests a higher precision of the molecular clock analysis.

Fig. S4 Phylogenomic trees of cyanobacteria based on concatenation of single-copy orthologous gene families at the amino acid sequence level. (A) Maximum likelihood phylogeny of 159 oxygenic Cyanobacteria genomes (Table S2) based on the 214 single-copy gene families shared by these genomes (Table S4). (B) Maximum likelihood phylogeny of 159 oxygenic Cyanobacteria genomes based on the 90 gene families (TableS4) with evidence of compositional homogeneity in the protein sequences. (C) Bayesian inference phylogeny of 159 oxygenic Cyanobacteria genomes based on the 214 gene families. (D) Bayesian inference phylogeny of 159 oxygenic Cyanobacteria genomes based on the 90 gene families with evidence of compositional homogeneity in the protein sequences. The taxonomic classification is adapted from Sanchez-Baracaldo et al. (2015). (E) Maximum likelihood phylogeny of 159 oxygenic Cyanobacteria genomes as well as eight non-oxygenic Cyanobacteria outgroups (Table S2) based on the 90 gene families with evidence of compositional homogeneity in the protein sequences. Trees shown in (D) and (E) are used for molecular dating analyses, and calibrated ancestor nodes are marked with solid orange circle. Solid and open circles at ancestral nodes indicate the percentage of posterior probability or the frequency of the group defined by that node in 100 bootstrapped replicates is at least 95 and 85, respectively.

Macrocyanobacteria

Microcyanobacteria

Basal lineage

Fig. S4 A

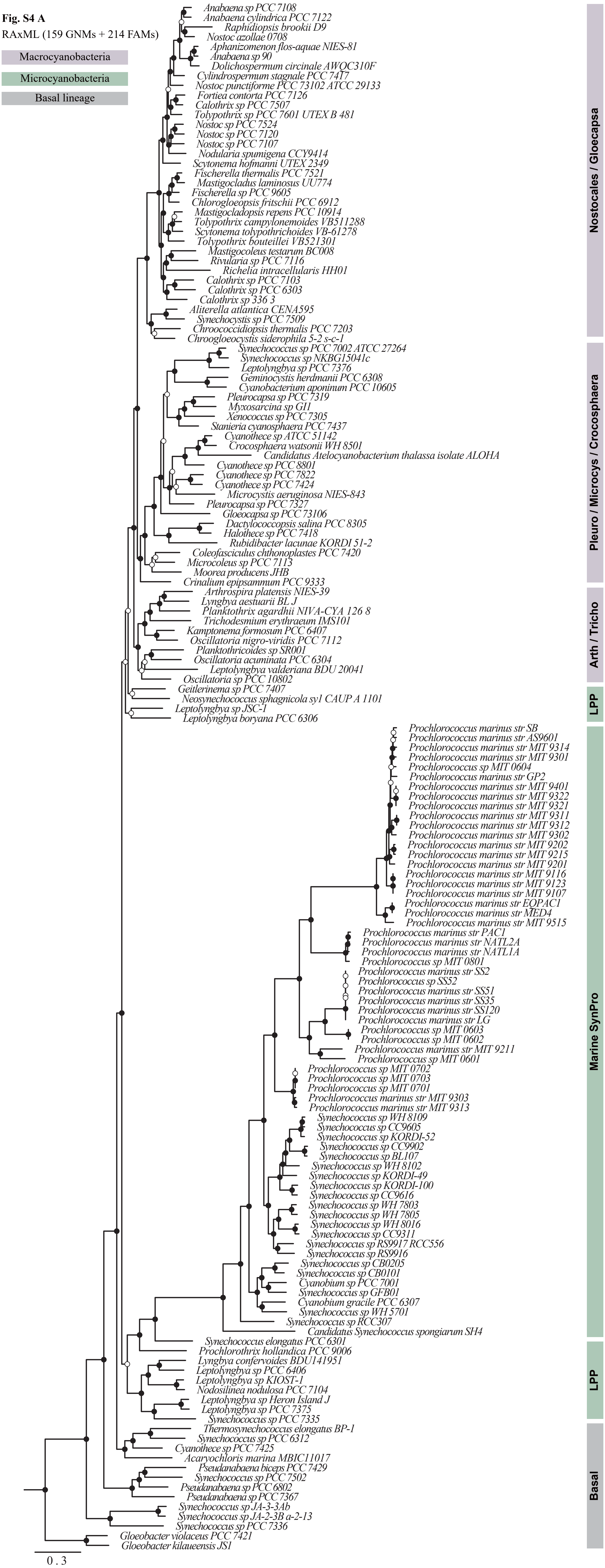

RAxML (159 GNMs + 214 FAMs)
Macrocyanobacteria
Microcyanobacteria
Basal lineage
Anabaena sp PCC 7108
Anabaena cylindrica PCC 7122
Raphidiopsis brookii D9
Nostoc azollae 0708
Aphanizomenon flos-aquae NIES-81
Anabaena sp 90
Dolichospermum circinale AWQC310F
Cylindrospermum stagnale PCC 7417
Nostoc punctiforme PCC 73102 ATCC 29133
Fortiea contorta PCC 7126
Calothrix sp PCC 7507
Tolypothrix sp PCC 7601 UTEX B 481
Nostoc sp PCC 7524
Nostoc sp PCC 7120
Nostoc sp PCC 7107
Nodularia spumigena CCY9414
Scytonema hofmanni UTEX 2349
Fischerella thermalis PCC 7521
Mastigocladus laminosus UU774
Fischerella sp PCC 9605
Chlorogloeopsis fritschii PCC 6912
Mastigocladopsis repens PCC 10914
Tolypothrix campylonemoides VB511288
Scytonema tolypothrichoides VB-61278
Tolypothrix bouteillei VB521301
Mastigocoleus testarum BC008
Rivularia sp PCC 7116
Richelia intracellularis HH01
Calothrix sp PCC 7103
Calothrix sp PCC 6303
Calothrix sp 336 3
Aliterella atlantica CENA595
Synechocystis sp PCC 7509
Chroococcidiopsis thermalis PCC 7203
Chroogloeocystis siderophila 5-2 s-c-1
Synechococcus sp PCC 7002 ATCC 27264
Synechococcus sp NKBG15041c
Leptolyngbya sp PCC 7376
Geminocystis herdmanii PCC 6308
Cyanobacterium aponinum PCC 10605
Pleurocapsa sp PCC 7319
Myxosarcina sp GI1
Xenococcus sp PCC 7305
Stanieria cyanosphaera PCC 7437
Cyanothece sp ATCC 51142
Crocosphaera watsonii WH 8501
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp PCC 8801
Cyanothece sp PCC 7822
Cyanothece sp PCC 7424
Microcystis aeruginosa NIES-843
Pleurocapsa sp PCC 7327
Gloeocapsa sp PCC 73106
Dactylococcopsis salina PCC 8305
Halothece sp PCC 7418
Rubidibacter lacunae KORDI 51-2
Coleofasciculus chthonoplastes PCC 7420
Microcoleus sp PCC 7113
Moorea producens JHB
Crinalium epipsammum PCC 9333
Arthrospira platensis NIES-39
Lyngbya aestuarii BL J
Planktothrix agardhii NIVA-CYA 126 8
Trichodesmium erythraeum IMS101
Kamptonema formosum PCC 6407
Oscillatoria nigro-viridis PCC 7112
Planktothricoides sp SR001
Oscillatoria acuminata PCC 6304
Leptolyngbya valderiana BDU 20041
Oscillatoria sp PCC 10802
Geitlerinema sp PCC 7407
Neosynechococcus sphagnicola sy1 CAUP A 1101
Leptolyngbya sp JSC-1
Leptolyngbya boryana PCC 6306
Prochlorococcus marinus str SB
Prochlorococcus marinus str AS9601
Prochlorococcus marinus str MIT 9314
Prochlorococcus marinus str MIT 9301
Prochlorococcus sp MIT 0604
Prochlorococcus marinus str GP2
Prochlorococcus marinus str MIT 9401
Prochlorococcus marinus str MIT 9322
Prochlorococcus marinus str MIT 9321
Prochlorococcus marinus str MIT 9311
Prochlorococcus marinus str MIT 9312
Prochlorococcus marinus str MIT 9302
Prochlorococcus marinus str MIT 9202
Prochlorococcus marinus str MIT 9215
Prochlorococcus marinus str MIT 9201
Prochlorococcus marinus str MIT 9116
Prochlorococcus marinus str MIT 9123
Prochlorococcus marinus str MIT 9107
Prochlorococcus marinus str EQPAC1
Prochlorococcus marinus str MED4
Prochlorococcus marinus str MIT 9515
Prochlorococcus marinus str PAC1
Prochlorococcus marinus str NATL2A
Prochlorococcus marinus str NATL1A
Prochlorococcus sp MIT 0801
Prochlorococcus marinus str SS2
Prochlorococcus sp SS52
Prochlorococcus marinus str SS51
Prochlorococcus marinus str SS35
Prochlorococcus marinus str SS120
Prochlorococcus marinus str LG
Prochlorococcus sp MIT 0603
Prochlorococcus sp MIT 0602
Prochlorococcus marinus str MIT 9211
Prochlorococcus sp MIT 0601
Prochlorococcus sp MIT 0702
Prochlorococcus sp MIT 0703
Prochlorococcus sp MIT 0701
Prochlorococcus marinus str MIT 9303
Prochlorococcus marinus str MIT 9313
Synechococcus sp WH 8109
Synechococcus sp CC9605
Synechococcus sp KORDI-52
Synechococcus sp CC9902
Synechococcus sp BL107
Synechococcus sp WH 8102
Synechococcus sp KORDI-49
Synechococcus sp KORDI-100
Synechococcus sp CC9616
Synechococcus sp WH 7803
Synechococcus sp WH 7805
Synechococcus sp WH 8016
Synechococcus sp CC9311
Synechococcus sp RS9917 RCC556
Synechococcus sp RS9916
Synechococcus sp CB0205
Synechococcus sp CB0101
Cyanobium sp PCC 7001
Synechococcus sp GFB01
Cyanobium gracile PCC 6307
Synechococcus sp WH 5701
Synechococcus sp RCC307
Candidatus Synechococcus spongiarum SH4
Synechococcus elongatus PCC 6301
Prochlorothrix hollandica PCC 9006
Lyngbya confervoides BDU141951
Leptolyngbya sp PCC 6406
Leptolyngbya sp KIOST-1
Nodosilinea nodulosa PCC 7104
Leptolyngbya sp Heron Island J
Leptolyngbya sp PCC 7375
Synechococcus sp PCC 7335
Thermosynechococcus elongatus BP-1
Synechococcus sp PCC 6312
Cyanothece sp PCC 7425
Acaryochloris marina MBIC11017
Pseudanabaena biceps PCC 7429
Synechococcus sp PCC 7502
Pseudanabaena sp PCC 6802
Pseudanabaena sp PCC 7367
Synechococcus sp JA-3-3Ab
Synechococcus sp JA-2-3B a-2-13
Synechococcus sp PCC 7336
Gloeobacter violaceus PCC 7421
Gloeobacter kilaueensis JS1
0 . 3
Nostocales / Gloecapsa
Pleuro / Microcys / Crocosphaera
Arth / Tricho
LPP
Marine SynPro
LPP
Basal

Fig. S4 B

RAxML (159 GNMs + 90 FAMs)

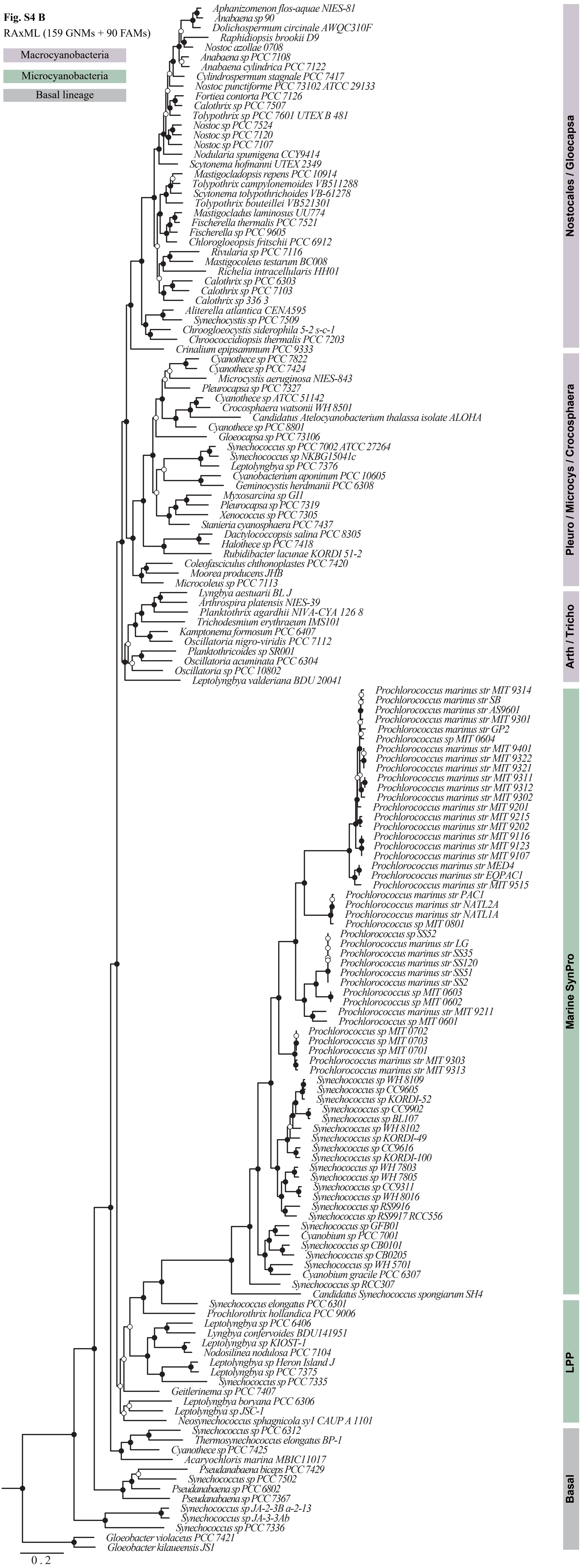

Macrocyanobacteria
Microcyanobacteria
Basal lineage
Aphanizomenon flos-aquae NIES-81
Anabaena sp 90
Dolichospermum circinale AWQC310F
Raphidiopsis brookii D9
Nostoc azollae 0708
Anabaena sp PCC 7108
Anabaena cylindrica PCC 7122
Cylindrospermum stagnale PCC 7417
Nostoc punctiforme PCC 73102 ATCC 29133
Fortiea contorta PCC 7126
Calothrix sp PCC 7507
Tolypothrix sp PCC 7601 UTEX B 481
Nostoc sp PCC 7524
Nostoc sp PCC 7120
Nostoc sp PCC 7107
Nodularia spumigena CCY9414
Scytonema hofmanni UTEX 2349
Mastigocladopsis repens PCC 10914
Tolypothrix campylonemoides VB511288
Scytonema tolypothrichoides VB-61278
Tolypothrix bouteillei VB521301
Mastigocladus laminosus UU774
Fischerella thermalis PCC 7521
Fischerella sp PCC 9605
Chlorogloeopsis fritschii PCC 6912
Rivularia sp PCC 7116
Mastigocoleus testarum BC008
Richelia intracellularis HH01
Calothrix sp PCC 6303
Calothrix sp PCC 7103
Calothrix sp 336 3
Aliterella atlantica CENA595
Synechocystis sp PCC 7509
Chroogloeocystis siderophila 5-2 s-c-1
Chroococcidiopsis thermalis PCC 7203
Crinalium epipsammum PCC 9333
Cyanothece sp PCC 7822
Cyanothece sp PCC 7424
Microcystis aeruginosa NIES-843
Pleurocapsa sp PCC 7327
Cyanothece sp ATCC 51142
Crocosphaera watsonii WH 8501
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp PCC 8801
Gloeocapsa sp PCC 73106
Synechococcus sp PCC 7002 ATCC 27264
Synechococcus sp NKBG15041c
Leptolyngbya sp PCC 7376
Cyanobacterium aponinum PCC 10605
Geminocystis herdmanii PCC 6308
Myxosarcina sp GI1
Pleurocapsa sp PCC 7319
Xenococcus sp PCC 7305
Stanieria cyanosphaera PCC 7437
Dactylococcopsis salina PCC 8305
Halothece sp PCC 7418
Rubidibacter lacunae KORDI 51-2
Coleofasciculus chthonoplastes PCC 7420
Moorea producens JHB
Microcoleus sp PCC 7113
Lyngbya aestuarii BL J
Arthrospira platensis NIES-39
Planktothrix agardhii NIVA-CYA 126 8
Trichodesmium erythraeum IMS101
Kamptonema formosum PCC 6407
Oscillatoria nigro-viridis PCC 7112
Planktothricoides sp SR001
Oscillatoria acuminata PCC 6304
Oscillatoria sp PCC 10802
Leptolyngbya valderiana BDU 20041
Prochlorococcus marinus str MIT 9314
Prochlorococcus marinus str SB
Prochlorococcus marinus str AS9601
Prochlorococcus marinus str MIT 9301
Prochlorococcus marinus str GP2
Prochlorococcus sp MIT 0604
Prochlorococcus marinus str MIT 9401
Prochlorococcus marinus str MIT 9322
Prochlorococcus marinus str MIT 9321
Prochlorococcus marinus str MIT 9311
Prochlorococcus marinus str MIT 9312
Prochlorococcus marinus str MIT 9302
Prochlorococcus marinus str MIT 9201
Prochlorococcus marinus str MIT 9215
Prochlorococcus marinus str MIT 9202
Prochlorococcus marinus str MIT 9116
Prochlorococcus marinus str MIT 9123
Prochlorococcus marinus str MIT 9107
Prochlorococcus marinus str MED4
Prochlorococcus marinus str EQPAC1
Prochlorococcus marinus str MIT 9515
Prochlorococcus marinus str PAC1
Prochlorococcus marinus str NATL2A
Prochlorococcus marinus str NATL1A
Prochlorococcus sp MIT 0801
Prochlorococcus sp SS52
Prochlorococcus marinus str LG
Prochlorococcus marinus str SS35
Prochlorococcus marinus str SS120
Prochlorococcus marinus str SS51
Prochlorococcus marinus str SS2
Prochlorococcus sp MIT 0603
Prochlorococcus sp MIT 0602
Prochlorococcus marinus str MIT 9211
Prochlorococcus sp MIT 0601
Prochlorococcus sp MIT 0702
Prochlorococcus sp MIT 0703
Prochlorococcus sp MIT 0701
Prochlorococcus marinus str MIT 9303
Prochlorococcus marinus str MIT 9313
Synechococcus sp WH 8109
Synechococcus sp CC9605
Synechococcus sp KORDI-52
Synechococcus sp CC9902
Synechococcus sp BL107
Synechococcus sp WH 8102
Synechococcus sp KORDI-49
Synechococcus sp CC9616
Synechococcus sp KORDI-100
Synechococcus sp WH 7803
Synechococcus sp WH 7805
Synechococcus sp CC9311
Synechococcus sp WH 8016
Synechococcus sp RS9916
Synechococcus sp RS9917 RCC556
Synechococcus sp GFB01
Cyanobium sp PCC 7001
Synechococcus sp CB0101
Synechococcus sp CB0205
Synechococcus sp WH 5701
Cyanobium gracile PCC 6307
Synechococcus sp RCC307
Candidatus Synechococcus spongiarum SH4
Synechococcus elongatus PCC 6301
Prochlorothrix hollandica PCC 9006
Leptolyngbya sp PCC 6406
Lyngbya confervoides BDU141951
Leptolyngbya sp KIOST-1
Nodosilinea nodulosa PCC 7104
Leptolyngbya sp Heron Island J
Leptolyngbya sp PCC 7375
Synechococcus sp PCC 7335
Geitlerinema sp PCC 7407
Leptolyngbya boryana PCC 6306
Leptolyngbya sp JSC-1
Neosynechococcus sphagnicola sy1 CAUP A 1101
Synechococcus sp PCC 6312
Thermosynechococcus elongatus BP-1
Cyanothece sp PCC 7425
Acaryochloris marina MBIC11017
Pseudanabaena biceps PCC 7429
Synechococcus sp PCC 7502
Pseudanabaena sp PCC 6802
Pseudanabaena sp PCC 7367
Synechococcus sp JA-2-3B a-2-13
Synechococcus sp JA-3-3Ab
Synechococcus sp PCC 7336
Gloeobacter violaceus PCC 7421
Gloeobacter kilaueensis JS1
0 . 2
Nostocales / Gloecapsa
Pleuro / Microcys / Crocosphaera
Arth / Tricho
Marine SynPro
LPP
Basal

Fig. S4 C

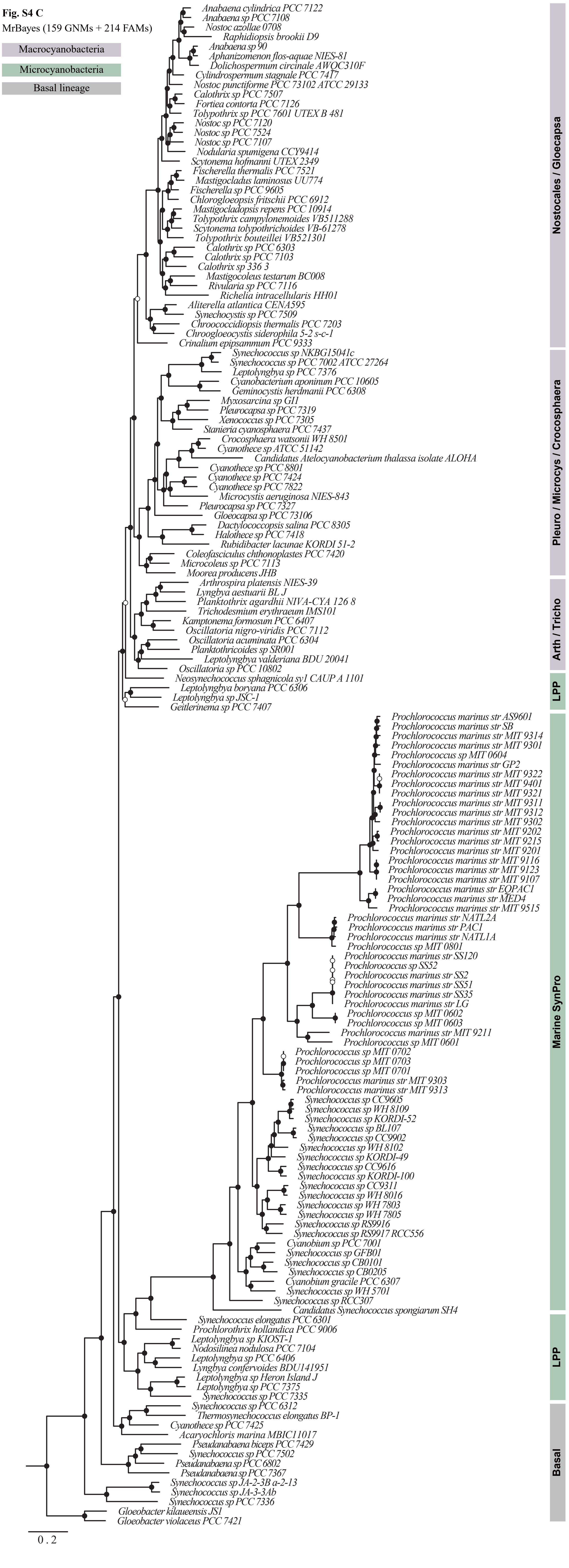

MrBayes (159 GNMs + 214 FAMs)
Macrocyanobacteria
Microcyanobacteria
Basal lineage
Anabaena cylindrica PCC 7122
Anabaena sp PCC 7108
Nostoc azollae 0708
Raphidiopsis brookii D9
Anabaena sp 90
Aphanizomenon flos-aquae NIES-81
Dolichospermum circinale AWQC310F
Cylindrospermum stagnale PCC 7417
Nostoc punctiforme PCC 73102 ATCC 29133
Calothrix sp PCC 7507
Fortiea contorta PCC 7126
Tolypothrix sp PCC 7601 UTEX B 481
Nostoc sp PCC 7120
Nostoc sp PCC 7524
Nostoc sp PCC 7107
Nodularia spumigena CCY9414
Scytonema hofmanni UTEX 2349
Fischerella thermalis PCC 7521
Mastigocladus laminosus UU774
Fischerella sp PCC 9605
Chlorogloeopsis fritschii PCC 6912
Mastigocladopsis repens PCC 10914
Tolypothrix campylonemoides VB511288
Scytonema tolypothrichoides VB-61278
Tolypothrix bouteillei VB521301
Calothrix sp PCC 6303
Calothrix sp PCC 7103
Calothrix sp 336 3
Mastigocoleus testarum BC008
Rivularia sp PCC 7116
Richelia intracellularis HH01
Aliterella atlantica CENA595
Synechocystis sp PCC 7509
Chroococcidiopsis thermalis PCC 7203
Chroogloeocystis siderophila 5-2 s-c-1
Crinalium epipsammum PCC 9333
Synechococcus sp NKBG15041c
Synechococcus sp PCC 7002 ATCC 27264
Leptolyngbya sp PCC 7376
Cyanobacterium aponinum PCC 10605
Geminocystis herdmanii PCC 6308
Myxosarcina sp GI1
Pleurocapsa sp PCC 7319
Xenococcus sp PCC 7305
Stanieria cyanosphaera PCC 7437
Crocosphaera watsonii WH 8501
Cyanothece sp ATCC 51142
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp PCC 8801
Cyanothece sp PCC 7424
Cyanothece sp PCC 7822
Microcystis aeruginosa NIES-843
Pleurocapsa sp PCC 7327
Gloeocapsa sp PCC 73106
Dactylococcopsis salina PCC 8305
Halothece sp PCC 7418
Rubidibacter lacunae KORDI 51-2
Coleofasciculus chthonoplastes PCC 7420
Microcoleus sp PCC 7113
Moorea producens JHB
Arthrospira platensis NIES-39
Lyngbya aestuarii BL J
Planktothrix agardhii NIVA-CYA 126 8
Trichodesmium erythraeum IMS101
Kamptonema formosum PCC 6407
Oscillatoria nigro-viridis PCC 7112
Oscillatoria acuminata PCC 6304
Planktothricoides sp SR001
Leptolyngbya valderiana BDU 20041
Oscillatoria sp PCC 10802
Neosynechococcus sphagnicola sy1 CAUP A 1101
Leptolyngbya boryana PCC 6306
Leptolyngbya sp JSC-1
Geitlerinema sp PCC 7407
Prochlorococcus marinus str AS9601
Prochlorococcus marinus str SB
Prochlorococcus marinus str MIT 9314
Prochlorococcus marinus str MIT 9301
Prochlorococcus sp MIT 0604
Prochlorococcus marinus str GP2
Prochlorococcus marinus str MIT 9322
Prochlorococcus marinus str MIT 9401
Prochlorococcus marinus str MIT 9321
Prochlorococcus marinus str MIT 9311
Prochlorococcus marinus str MIT 9312
Prochlorococcus marinus str MIT 9302
Prochlorococcus marinus str MIT 9202
Prochlorococcus marinus str MIT 9215
Prochlorococcus marinus str MIT 9201
Prochlorococcus marinus str MIT 9116
Prochlorococcus marinus str MIT 9123
Prochlorococcus marinus str MIT 9107
Prochlorococcus marinus str EQPAC1
Prochlorococcus marinus str MED4
Prochlorococcus marinus str MIT 9515
Prochlorococcus marinus str NATL2A
Prochlorococcus marinus str PAC1
Prochlorococcus marinus str NATL1A
Prochlorococcus sp MIT 0801
Prochlorococcus marinus str SS120
Prochlorococcus sp SS52
Prochlorococcus marinus str SS2
Prochlorococcus marinus str SS51
Prochlorococcus marinus str SS35
Prochlorococcus marinus str LG
Prochlorococcus sp MIT 0602
Prochlorococcus sp MIT 0603
Prochlorococcus marinus str MIT 9211
Prochlorococcus sp MIT 0601
Prochlorococcus sp MIT 0702
Prochlorococcus sp MIT 0703
Prochlorococcus sp MIT 0701
Prochlorococcus marinus str MIT 9303
Prochlorococcus marinus str MIT 9313
Synechococcus sp CC9605
Synechococcus sp WH 8109
Synechococcus sp KORDI-52
Synechococcus sp BL107
Synechococcus sp CC9902
Synechococcus sp WH 8102
Synechococcus sp KORDI-49
Synechococcus sp CC9616
Synechococcus sp KORDI-100
Synechococcus sp CC9311
Synechococcus sp WH 8016
Synechococcus sp WH 7803
Synechococcus sp WH 7805
Synechococcus sp RS9916
Synechococcus sp RS9917 RCC556
Cyanobium sp PCC 7001
Synechococcus sp GFB01
Synechococcus sp CB0101
Synechococcus sp CB0205
Cyanobium gracile PCC 6307
Synechococcus sp WH 5701
Synechococcus sp RCC307
Candidatus Synechococcus spongiarum SH4
Synechococcus elongatus PCC 6301
Prochlorothrix hollandica PCC 9006
Leptolyngbya sp KIOST-1
Nodosilinea nodulosa PCC 7104
Leptolyngbya sp PCC 6406
Lyngbya confervoides BDU141951
Leptolyngbya sp Heron Island J
Leptolyngbya sp PCC 7375
Synechococcus sp PCC 7335
Synechococcus sp PCC 6312
Thermosynechococcus elongatus BP-1
Cyanothece sp PCC 7425
Acaryochloris marina MBIC11017
Pseudanabaena biceps PCC 7429
Synechococcus sp PCC 7502
Pseudanabaena sp PCC 6802
Pseudanabaena sp PCC 7367
Synechococcus sp JA-2-3B a-2-13
Synechococcus sp JA-3-3Ab
Synechococcus sp PCC 7336
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
0 . 2
Nostocales / Gloecapsa
Pleuro / Microcys / Crocosphaera
Arth / Tricho
LPP
Marine SynPro
LPP
Basal

Fig. S4 D

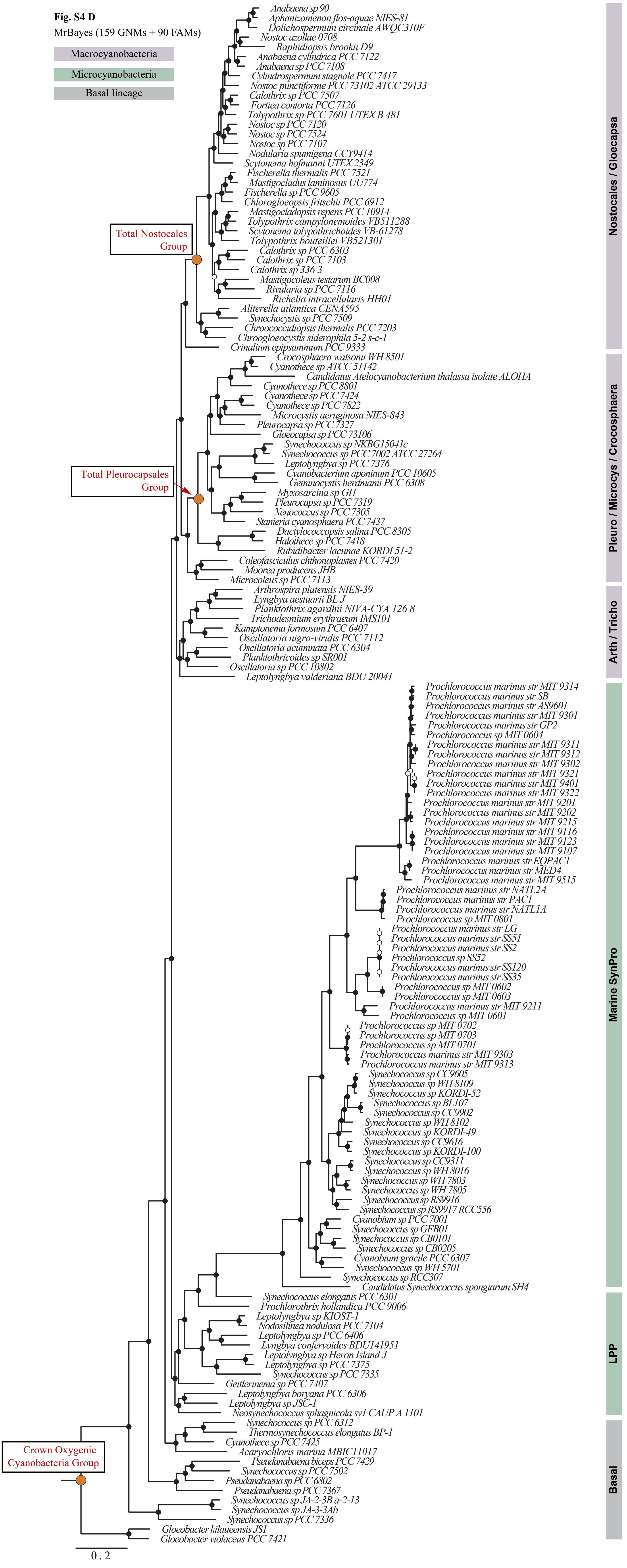

MrBayes (159 GNMs + 90 FAMs)
Macrocyanobacteria
Microcyanobacteria
Basal lineage
Anabaena sp 90
Aphanizomenon flos-aquae NIES-81
Dolichospermum circinale AWQC310F
Nostoc azollae 0708
Raphidiopsis brookii D9
Anabaena cylindrica PCC 7122
Anabaena sp PCC 7108
Cylindrospermum stagnale PCC 7417
Nostoc punctiforme PCC 73102 ATCC 29133
Calothrix sp PCC 7507
Fortiea contorta PCC 7126
Tolypothrix sp PCC 7601 UTEX B 481
Nostoc sp PCC 7120
Nostoc sp PCC 7524
Nostoc sp PCC 7107
Nodularia spumigena CCY9414
Scytonema hofmanni UTEX 2349
Fischerella thermalis PCC 7521
Mastigocladus laminosus UU774
Fischerella sp PCC 9605
Chlorogloeopsis fritschii PCC 6912
Mastigocladopsis repens PCC 10914
Tolypothrix campylonemoides VB511288
Scytonema tolypothrichoides VB-61278
Tolypothrix bouteillei VB521301
Calothrix sp PCC 6303
Calothrix sp PCC 7103
Calothrix sp 336 3
Mastigocoleus testarum BC008
Rivularia sp PCC 7116
Richelia intracellularis HH01
Aliterella atlantica CENA595
Synechocystis sp PCC 7509
Chroococcidiopsis thermalis PCC 7203
Chroogloeocystis siderophila 5-2 s-c-1
Crinalium epipsammum PCC 9333
Total Nostocales Group
Nostocales / Gloecapsa
Crocosphaera watsonii WH 8501
Cyanothece sp ATCC 51142
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp PCC 8801
Cyanothece sp PCC 7424
Cyanothece sp PCC 7822
Microcystis aeruginosa NIES-843
Pleurocapsa sp PCC 7327
Gloeocapsa sp PCC 73106
Synechococcus sp NKBG15041c
Synechococcus sp PCC 7002 ATCC 27264
Leptolyngbya sp PCC 7376
Cyanobacterium aponinum PCC 10605
Geminocystis herdmanii PCC 6308
Myxosarcina sp GI1
Pleurocapsa sp PCC 7319
Xenococcus sp PCC 7305
Stanieria cyanosphaera PCC 7437
Dactylococcopsis salina PCC 8305
Halothece sp PCC 7418
Rubidibacter lacunae KORDI 51-2
Coleofasciculus chthonoplastes PCC 7420
Moorea producens JHB
Microcoleus sp PCC 7113
Total Pleurocapsales Group
Pleuro / Microcys / Crocosphaera
Arthrospira platensis NIES-39
Lyngbya aestuarii BL J
Planktothrix agardhii NIVA-CYA 126 8
Trichodesmium erythraeum IMS101
Kamptonema formosum PCC 6407
Oscillatoria nigro-viridis PCC 7112
Oscillatoria acuminata PCC 6304
Planktothricoides sp SR001
Oscillatoria sp PCC 10802
Leptolyngbya valderiana BDU 20041
Arth / Tricho
Prochlorococcus marinus str MIT 9314
Prochlorococcus marinus str SB
Prochlorococcus marinus str AS9601
Prochlorococcus marinus str MIT 9301
Prochlorococcus marinus str GP2
Prochlorococcus sp MIT 0604
Prochlorococcus marinus str MIT 9311
Prochlorococcus marinus str MIT 9312
Prochlorococcus marinus str MIT 9302
Prochlorococcus marinus str MIT 9321
Prochlorococcus marinus str MIT 9401
Prochlorococcus marinus str MIT 9322
Prochlorococcus marinus str MIT 9201
Prochlorococcus marinus str MIT 9202
Prochlorococcus marinus str MIT 9215
Prochlorococcus marinus str MIT 9116
Prochlorococcus marinus str MIT 9123
Prochlorococcus marinus str MIT 9107
Prochlorococcus marinus str EQPAC1
Prochlorococcus marinus str MED4
Prochlorococcus marinus str MIT 9515
Prochlorococcus marinus str NATL2A
Prochlorococcus marinus str PAC1
Prochlorococcus marinus str NATL1A
Prochlorococcus sp MIT 0801
Prochlorococcus marinus str LG
Prochlorococcus marinus str SS51
Prochlorococcus marinus str SS2
Prochlorococcus sp SS52
Prochlorococcus marinus str SS120
Prochlorococcus marinus str SS35
Prochlorococcus sp MIT 0602
Prochlorococcus sp MIT 0603
Prochlorococcus marinus str MIT 9211
Prochlorococcus sp MIT 0601
Prochlorococcus sp MIT 0702
Prochlorococcus sp MIT 0703
Prochlorococcus sp MIT 0701
Prochlorococcus marinus str MIT 9303
Prochlorococcus marinus str MIT 9313
Synechococcus sp CC9605
Synechococcus sp WH 8109
Synechococcus sp KORDI-52
Synechococcus sp BL107
Synechococcus sp CC9902
Synechococcus sp WH 8102
Synechococcus sp KORDI-49
Synechococcus sp CC9616
Synechococcus sp KORDI-100
Synechococcus sp CC9311
Synechococcus sp WH 8016
Synechococcus sp WH 7803
Synechococcus sp WH 7805
Synechococcus sp RS9916
Synechococcus sp RS9917 RCC556
Cyanobium sp PCC 7001
Synechococcus sp GFB01
Synechococcus sp CB0101
Synechococcus sp CB0205
Cyanobium gracile PCC 6307
Synechococcus sp WH 5701
Synechococcus sp RCC307
Candidatus Synechococcus spongiarum SH4
Marine SynPro
Synechococcus elongatus PCC 6301
Prochlorothrix hollandica PCC 9006
Leptolyngbya sp KIOST-1
Nodosilinea nodulosa PCC 7104
Leptolyngbya sp PCC 6406
Lyngbya confervoides BDU141951
Leptolyngbya sp Heron Island J
Leptolyngbya sp PCC 7375
Synechococcus sp PCC 7335
Geitlerinema sp PCC 7407
Leptolyngbya boryana PCC 6306
Leptolyngbya sp JSC-1
Neosynechococcus sphagnicola sy1 CAUP A 1101
LPP
Synechococcus sp PCC 6312
Thermosynechococcus elongatus BP-1
Cyanothece sp PCC 7425
Acaryochloris marina MBIC11017
Pseudanabaena biceps PCC 7429
Synechococcus sp PCC 7502
Pseudanabaena sp PCC 6802
Pseudanabaena sp PCC 7367
Synechococcus sp JA-2-3B a-2-13
Synechococcus sp JA-3-3Ab
Synechococcus sp PCC 7336
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Crown Oxygenic Cyanobacteria Group
Basal
0 . 2

Fig. S4 E

IQ-Tree (167 GNMs + 90 FAMs)

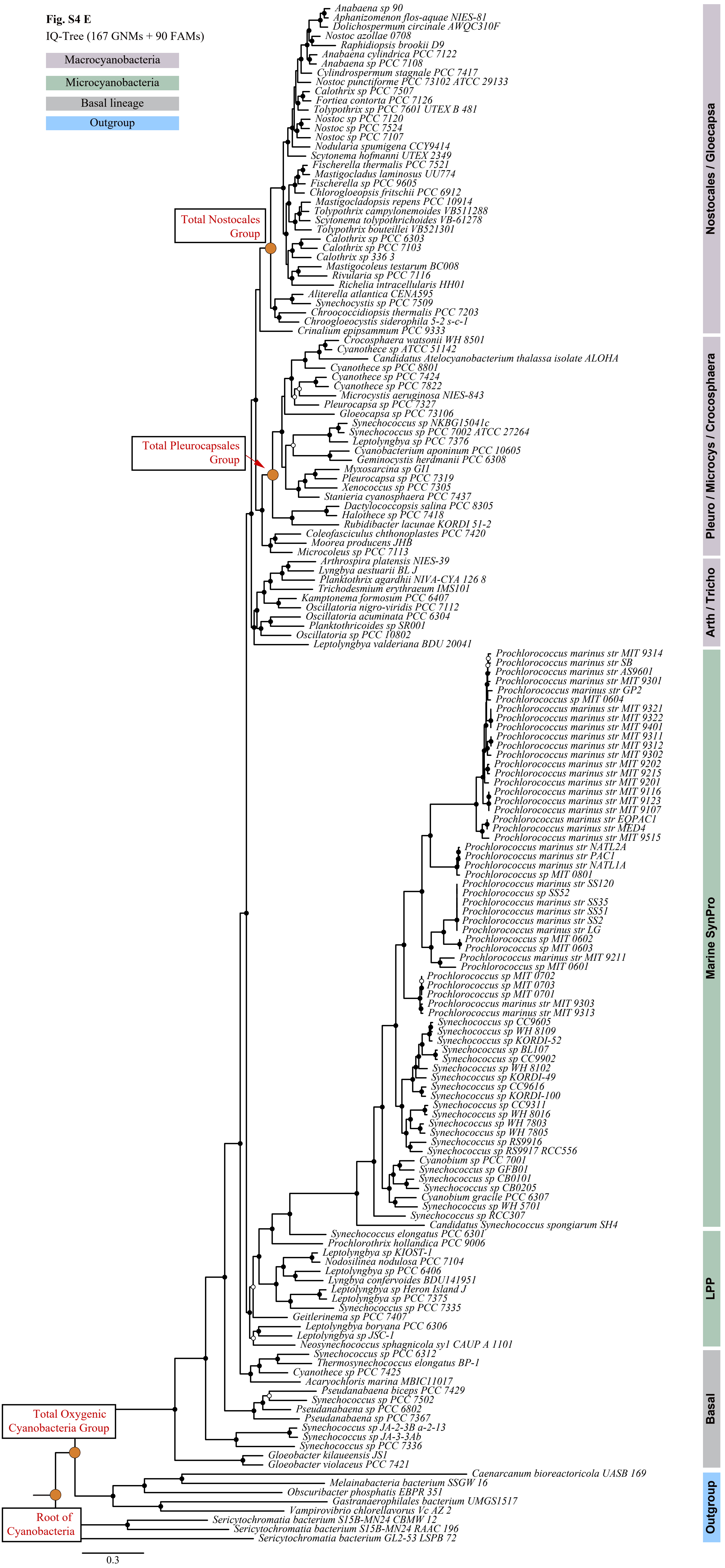

Macrocyanobacteria
Microcyanobacteria
Basal lineage
Outgroup
Anabaena sp 90
Aphanizomenon flos-aquae NIES-81
Dolichospermum circinale AWQC310F
Nostoc azollae 0708
Raphidiopsis brookii D9
Anabaena cylindrica PCC 7122
Anabaena sp PCC 7108
Cylindrospermum stagnale PCC 7417
Nostoc punctiforme PCC 73102 ATCC 29133
Calothrix sp PCC 7507
Fortiea contorta PCC 7126
Tolypothrix sp PCC 7601 UTEX B 481
Nostoc sp PCC 7120
Nostoc sp PCC 7524
Nostoc sp PCC 7107
Nodularia spumigena CCY9414
Scytonema hofmanni UTEX 2349
Fischerella thermalis PCC 7521
Mastigocladus laminosus UU774
Fischerella sp PCC 9605
Chlorogloeopsis fritschii PCC 6912
Mastigocladopsis repens PCC 10914
Tolypothrix campylonemoides VB511288
Scytonema tolypothrichoides VB-61278
Tolypothrix bouteillei VB521301
Calothrix sp PCC 6303
Calothrix sp PCC 7103
Calothrix sp 336 3
Mastigocoleus testarum BC008
Rivularia sp PCC 7116
Richelia intracellularis HH01
Aliterella atlantica CENA595
Synechocystis sp PCC 7509
Chroococcidiopsis thermalis PCC 7203
Chroogloeocystis siderophila 5-2 s-c-1
Crinalium epipsammum PCC 9333
Total Nostocales Group
Crocosphaera watsonii WH 8501
Cyanothece sp ATCC 51142
Candidatus Atelocyanobacterium thalassa isolate ALOHA
Cyanothece sp PCC 8801
Cyanothece sp PCC 7424
Cyanothece sp PCC 7822
Microcystis aeruginosa NIES-843
Pleurocapsa sp PCC 7327
Gloeocapsa sp PCC 73106
Synechococcus sp NKBG15041c
Synechococcus sp PCC 7002 ATCC 27264
Leptolyngbya sp PCC 7376
Cyanobacterium aponinum PCC 10605
Geminocystis herdmanii PCC 6308
Myxosarcina sp GI1
Pleurocapsa sp PCC 7319
Xenococcus sp PCC 7305
Stanieria cyanosphaera PCC 7437
Dactylococcopsis salina PCC 8305
Halothece sp PCC 7418
Rubidibacter lacunae KORDI 51-2
Coleofasciculus chthonoplastes PCC 7420
Moorea producens JHB
Microcoleus sp PCC 7113
Total Pleurocapsales Group
Arthrospira platensis NIES-39
Lyngbya aestuarii BL J
Planktothrix agardhii NIVA-CYA 126 8
Trichodesmium erythraeum IMS101
Kamptonema formosum PCC 6407
Oscillatoria nigro-viridis PCC 7112
Oscillatoria acuminata PCC 6304
Planktothricoides sp SR001
Oscillatoria sp PCC 10802
Leptolyngbya valderiana BDU 20041
Prochlorococcus marinus str MIT 9314
Prochlorococcus marinus str SB
Prochlorococcus marinus str AS9601
Prochlorococcus marinus str MIT 9301
Prochlorococcus marinus str GP2
Prochlorococcus sp MIT 0604
Prochlorococcus marinus str MIT 9321
Prochlorococcus marinus str MIT 9322
Prochlorococcus marinus str MIT 9401
Prochlorococcus marinus str MIT 9311
Prochlorococcus marinus str MIT 9312
Prochlorococcus marinus str MIT 9302
Prochlorococcus marinus str MIT 9202
Prochlorococcus marinus str MIT 9215
Prochlorococcus marinus str MIT 9201
Prochlorococcus marinus str MIT 9116
Prochlorococcus marinus str MIT 9123
Prochlorococcus marinus str MIT 9107
Prochlorococcus marinus str EQPAC1
Prochlorococcus marinus str MED4
Prochlorococcus marinus str MIT 9515
Prochlorococcus marinus str NATL2A
Prochlorococcus marinus str PAC1
Prochlorococcus marinus str NATL1A
Prochlorococcus sp MIT 0801
Prochlorococcus marinus str SS120
Prochlorococcus sp SS52
Prochlorococcus marinus str SS35
Prochlorococcus marinus str SS51
Prochlorococcus marinus str SS2
Prochlorococcus marinus str LG
Prochlorococcus sp MIT 0602
Prochlorococcus sp MIT 0603
Prochlorococcus marinus str MIT 9211
Prochlorococcus sp MIT 0601
Prochlorococcus sp MIT 0702
Prochlorococcus sp MIT 0703
Prochlorococcus sp MIT 0701
Prochlorococcus marinus str MIT 9303
Prochlorococcus marinus str MIT 9313
Synechococcus sp CC9605
Synechococcus sp WH 8109
Synechococcus sp KORDI-52
Synechococcus sp BL107
Synechococcus sp CC9902
Synechococcus sp WH 8102
Synechococcus sp KORDI-49
Synechococcus sp CC9616
Synechococcus sp KORDI-100
Synechococcus sp CC9311
Synechococcus sp WH 8016
Synechococcus sp WH 7803
Synechococcus sp WH 7805
Synechococcus sp RS9916
Synechococcus sp RS9917 RCC556
Cyanobium sp PCC 7001
Synechococcus sp GFB01
Synechococcus sp CB0101
Synechococcus sp CB0205
Cyanobium gracile PCC 6307
Synechococcus sp WH 5701
Synechococcus sp RCC307
Candidatus Synechococcus spongiarum SH4
Synechococcus elongatus PCC 6301
Prochlorothrix hollandica PCC 9006
Leptolyngbya sp KIOST-1
Nodosilinea nodulosa PCC 7104
Leptolyngbya sp PCC 6406
Lyngbya confervoides BDU141951
Leptolyngbya sp Heron Island J
Leptolyngbya sp PCC 7375
Synechococcus sp PCC 7335
Geitlerinema sp PCC 7407
Leptolyngbya boryana PCC 6306
Leptolyngbya sp JSC-1
Neosynechococcus sphagnicola sy1 CAUP A 1101
Synechococcus sp PCC 6312
Thermosynechococcus elongatus BP-1
Cyanothece sp PCC 7425
Acaryochloris marina MBIC11017
Pseudanabaena biceps PCC 7429
Synechococcus sp PCC 7502
Pseudanabaena sp PCC 6802
Pseudanabaena sp PCC 7367
Synechococcus sp JA-2-3B a-2-13
Synechococcus sp JA-3-3Ab
Synechococcus sp PCC 7336
Gloeobacter kilaueensis JS1
Gloeobacter violaceus PCC 7421
Caenarcanum bioreactoricola UASB 169
Melainabacteria bacterium SSGW 16
Obscuribacter phosphatis EBPR 351
Gastranaerophilales bacterium UMGS1517
Vampirovibrio chlorellavorus Vc AZ 2
Sericytochromatia bacterium S15B-MN24 CBMW 12
Sericytochromatia bacterium S15B-MN24 RAAC 196
Sericytochromatia bacterium GL2-53 LSPB 72
Total Oxygenic Cyanobacteria Group
Root of Cyanobacteria
0.3
Nostocales / Gloecapsa
Pleuro / Microcys / Crocosphaera
Arth / Tricho
Marine SynPro
LPP
Basal
Outgroup

Fig. S5

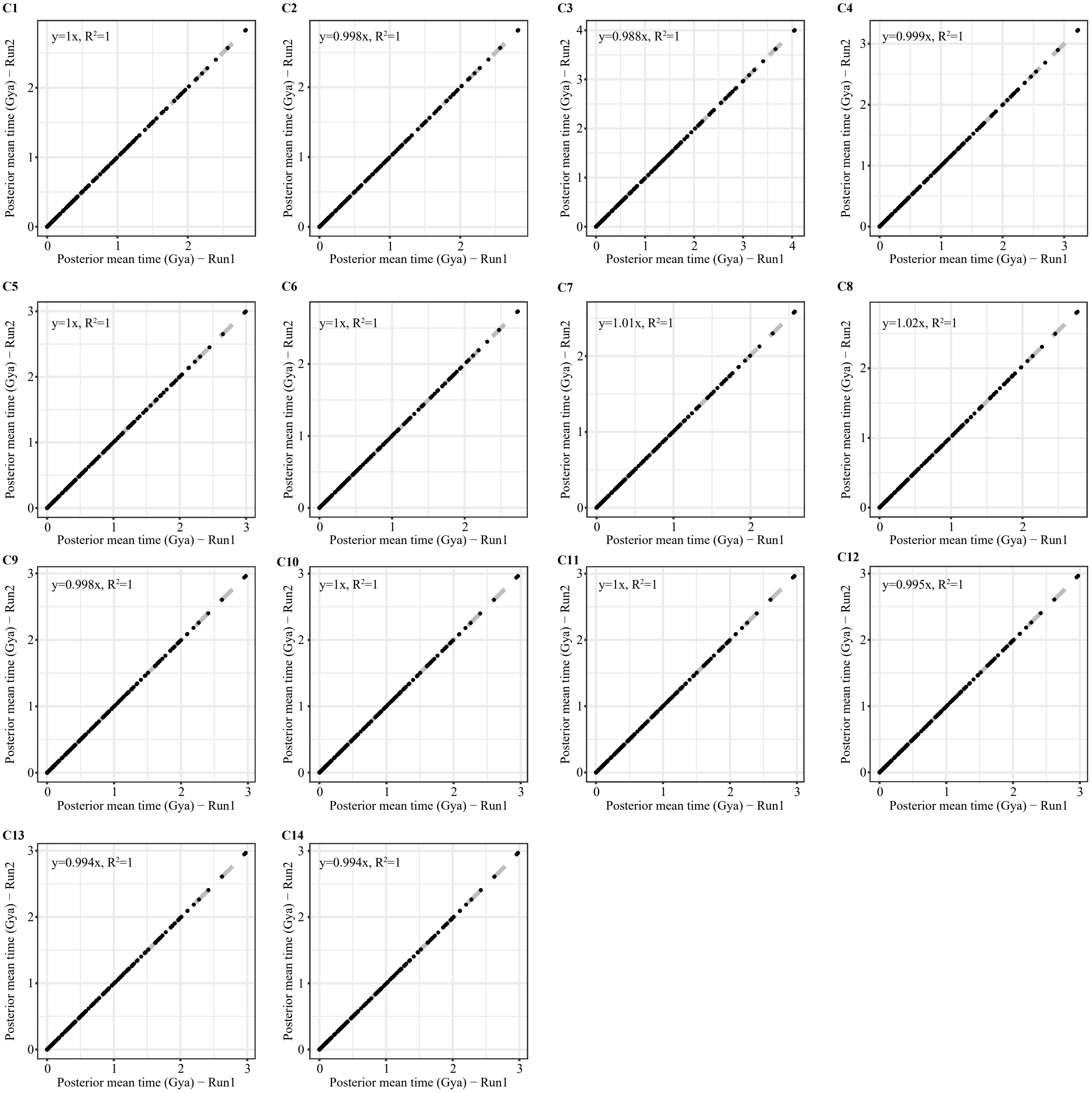


Fig. S5 Correlation of the posterior mean of estimated ages on ancestral nodes in replicated MCMC runs based on calibration sets C1-C14. Convergence of independent runs is achieved if points fall almost perfectly on the y=x line.

Fig. S6

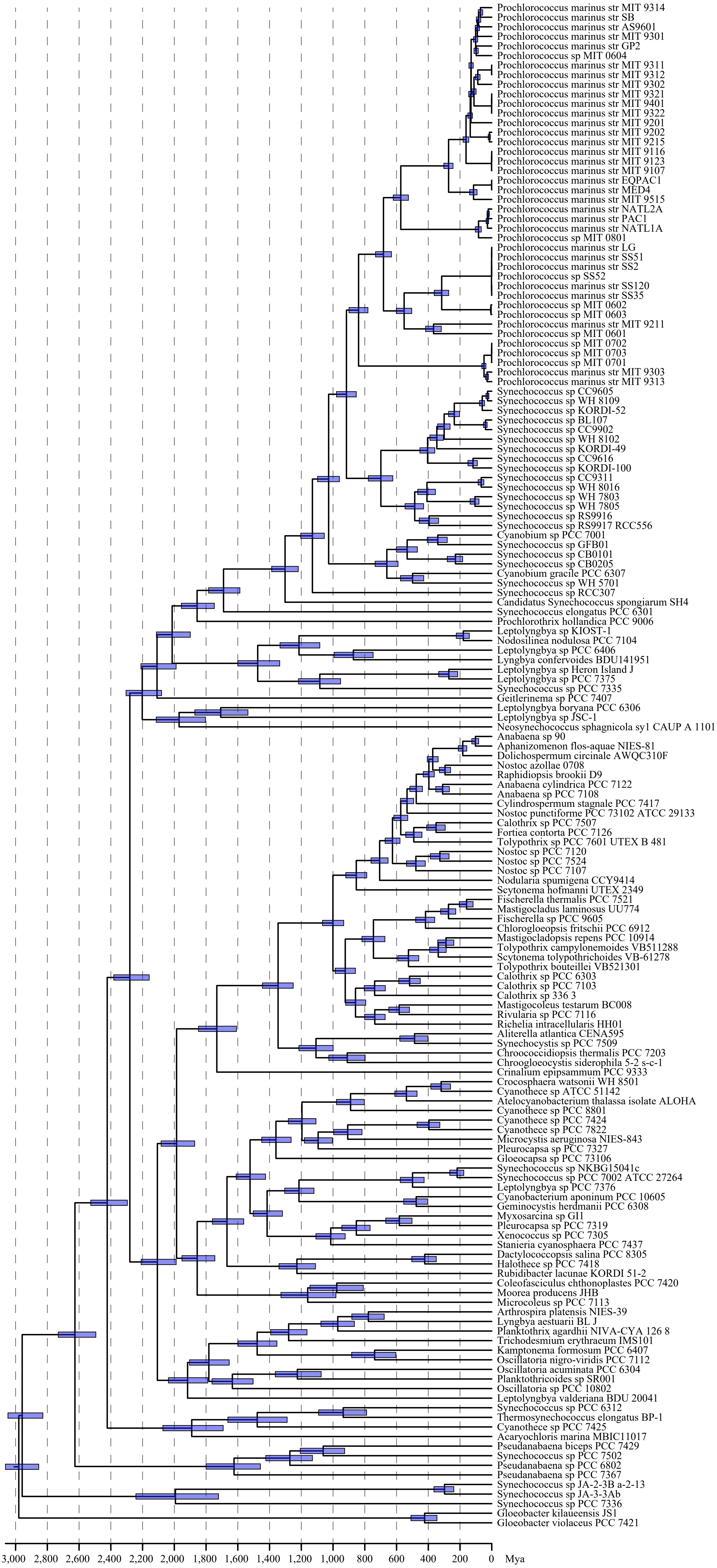


Fig. S6 A chronogram of cyanobacteria reconstructed with a relaxed molecular analysis implemented in MCMCTree. The molecular dating analysis uses 27 genes, a Bayesian phylogenomic tree of 159 genomes constructed with protein sequences of 90 gene families under the calibration set C14.

Fig. S7

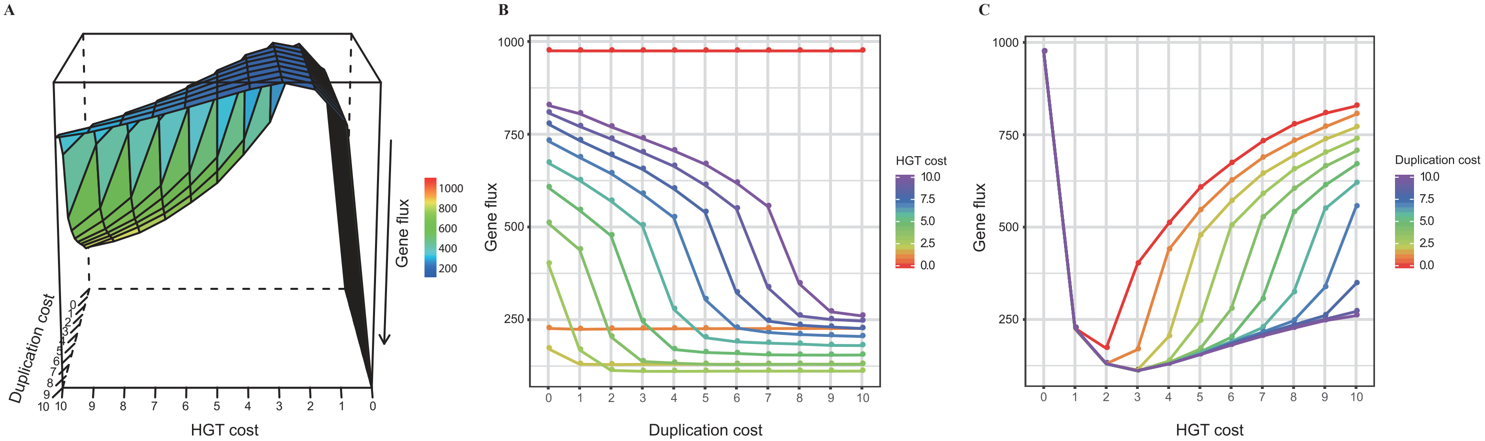


Fig. S7 The number of gene gains and losses (i.e., gene flux) reconstructed by AnGST depends on the penalty set for horizontal gene transfer (HGT) and gene duplication events relative to the penalty of gene loss events which was fixed to 1. (A) A 3-D plot showing gene flux with increased penalty of duplication and HGT events which ranges from 0 to 10. The color scheme represents the amount of gene flux. (B) The change of gene flux along with the increased penalty of duplication under different settings of HGT penalty (color). (C) The change of gene flux along with the increased penalty of HGT under different settings of duplication penalty.

Fig. S8

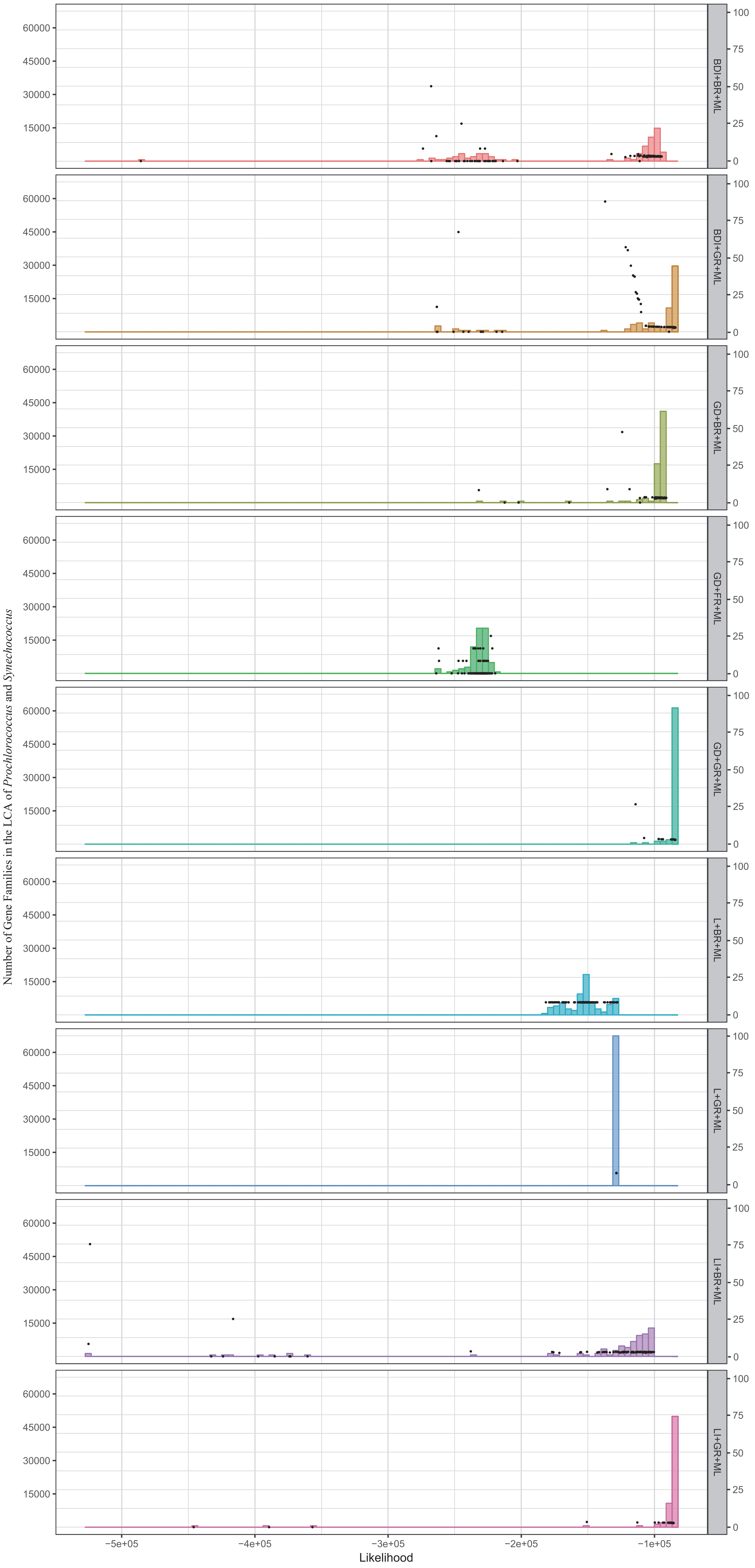
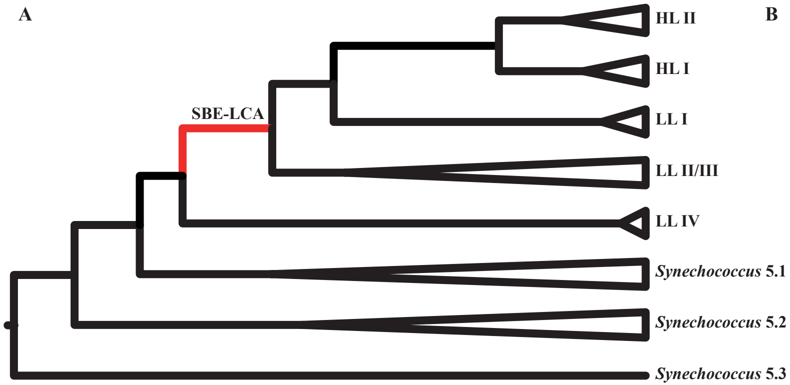


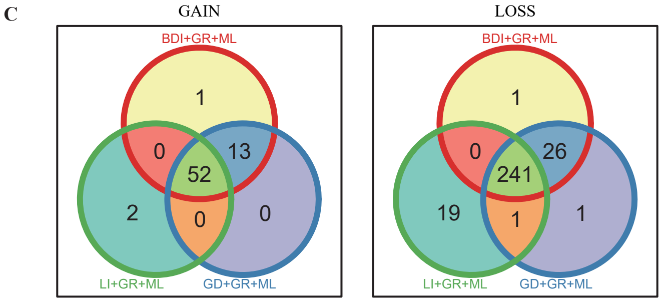
Fig. S8 The histogram distribution of likelihoods derived from 100 replicated analyses under each BadiRate model. Vertical bars represent the frequency of replicates within the same likelihood range (right y-axis). Black dots represent the estimated number of gene families at the root node (i.e., the last common ancestor of *Prochlorococcus* and *Synechococcus*) at a given likelihood (left y-axis). The ancestral reconstruction with the largest likelihood under each BadiRate model is considered for further analyses.

Fig. S9



B Best-fitting Models were chosen based on the likelihood.

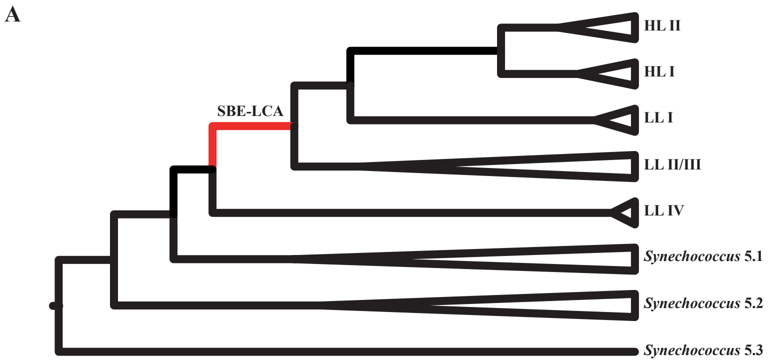
NAME	MEAN	MEDIAN	MAX
BDI+GR+ML	-111264.43	-88323.13	-84508.82
GD+GR+ML	-85685.10	-84551.35	-84551.35
LI+GR+ML	-97116.23	-85956.71	-85956.71
GD+BR+ML	-100939.50	-94705.62	-91345.30
BDI+BR+ML	-155458.90	-107205.09	-94802.30
LI+BR+ML	-149392.61	-115007.38	-100625.35
L+BR+ML	-152360.18	-151987.12	-128020.33
L+GR+ML	-128738.83	-128738.83	-128738.83
GD+FR+ML	-232431.80	-230572.17	-219759.66



	BDI+GR+ML	GD+GR+ML	LI+GR+ML	GAIN
BDI+GR+ML	-	65	52	66
GD+GR+ML	267	-	52	65
LI+GR+ML	241	242	-	54
LOSS	268	269	261	-

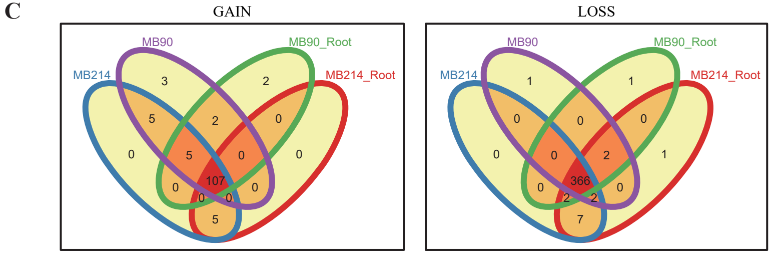
Fig. S9 Comparison of the gene gain and loss events reconstructed by BadiRate with different models during the evolution of *Prochlorococcus*. (A) The diagram highlights the evolutionary stages that led to the ancestral nodes 'SBE-LCA' (Fig. 1) of *Prochlorococcus*. (B) Multiple models are implemented in BadiRate for ancestral reconstruction, with the mean, median, and maximum likelihood values of each in 100 replicated analyses are shown. Models with their maximum likelihood values ranking at the top three (shaded in grey) are subject to further analyses. (C) Venn diagrams show the number of gain and loss events, respectively, reconstructed with the three models shown in (B). The detailed statistics are provided in the table on the right, in which the number of gain and loss events that are consistently inferred by distinct models are shaded with blue and pink, respectively. For example, 267 and 65 gene families are consistently inferred to be lost and gained following the model GD+GR+ML and BDI+GR+ML, respectively. The numbers of gain and loss events following each individual model are shown in the rightmost column (blue bold) and in the bottom row (red bold), respectively.

Fig. S10



B

Full Label	Description
MB214	(1) HGT Time Consistency (2) Ultrametric Tree (MCMCTREE + MrBayes Tree built on the concatenation of 214 gene familes + 3 Calibrations Points)
MB214-Root	(1) HGT Time Consistency (2) Ultrametric Tree (MCMCTREE + MrBayes Tree built on the concatenation of 214 gene familes + 1 Root calibration)
MB90	(1) HGT Time Consistency (2) Ultrametric Tree (MCMCTREE + MrBayes Tree built on the concatenation of 90 gene familes + 3 Calibrations Points)
MB90-Root	(1) HGT Time Consistency (2) Ultrametric Tree (MCMCTREE + MrBayes Tree built on the concatenation of 90 gene familes + 1 Root calibration)



	MB214	MB214-Root	MB90	MB90-Root	GAIN
MB214	-	112	117	112	**122**
MB214-Root	377	-	107	107	**112**
MB90	368	370	-	114	**122**
MB90-Root	368	370	368	-	**116**
LOSS	**377**	**380**	**371**	**371**	-

Fig. S10 Comparison of the gene gain and loss events reconstructed by AnGST with different strategies during the evolution of *Prochlorococcus*. (A) The diagram highlights the evolutionary stage that led to the ancestral nodes 'SBE-LCA' (Fig. 1) of *Prochlorococcus*. (B) For ancestral reconstructions with AnGST, a chronogram is used to limit HGT events occurring between contemporaneous lineages (HGT Time Consistency). Chronograms are estimated with MCMCTree based on 214 gene families or 90 gene families under 3 calibration points or single root calibration. (C) Venn diagrams show the number of gain and loss events, respectively, reconstructed based on different strategies shown in (B). The detailed statistics are provided in the table on the right, in which the number of gain and loss events that are consistently inferred by distinct strategies are shaded with blue and pink, respectively. For example, 377 and 112 gene families are consistently to be lost and gained following the strategy MB214 and MB214-Root, respectively. The numbers of gain and loss events following each individual strategy are shown in the rightmost column (blue bold) and in the bottom row (red bold), respectively.

Fig. S11

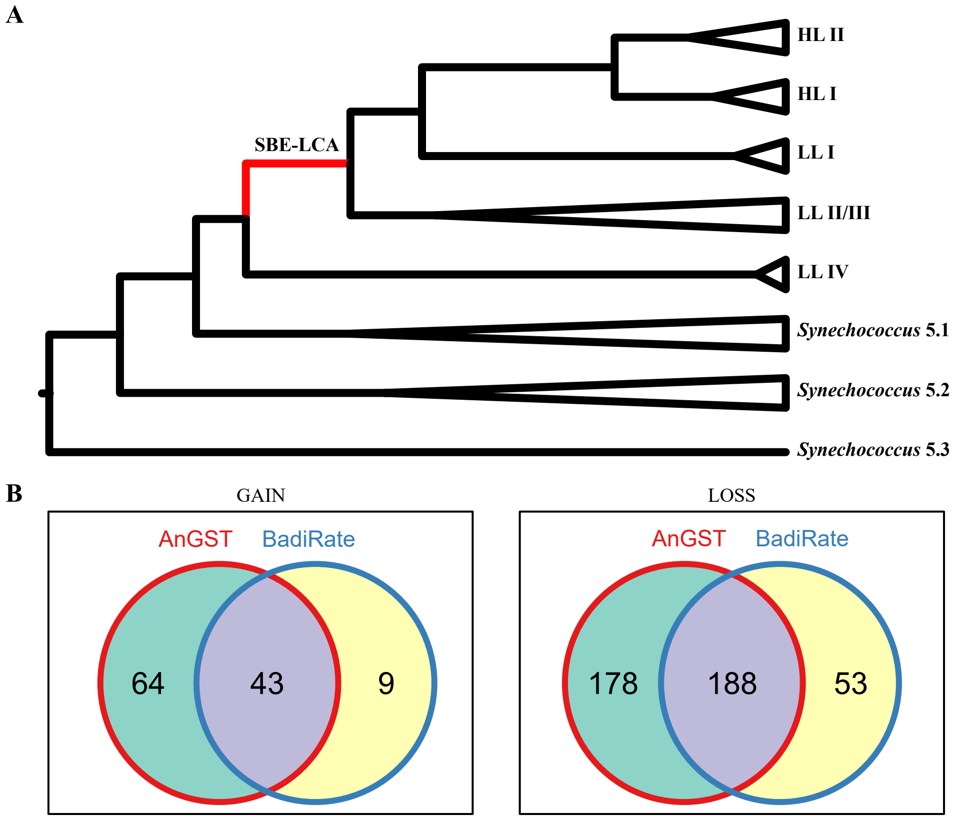


Fig. S11 (A) The diagram highlights the evolutionary stage that led to the ancestral nodes ‘SBE-LCA’ (Fig. 1) of *Prochlorococcus*. (B) Venn diagrams show the number of gene families consistently predicted to be gained or lost by AnGST and BadiRate during the evolutionary.

Table S1 A list of fossil calibration sets employed in the present study. Maximum and minimum time constraints are in the unit of billion years ago (Ga). References for calibrations are provided.

	Calibration Set	**Cyanobacteria Root**	**Total Oxygenic Cyanobacteria**	**Crown Oxygenic Cyanobacteria**	**Total Pleurocapsales**	**Total Nostocales**	**Crown Nostocales**
Without Non-oxygenic Cyanobacteria Outgroup	C1 [15]	-	-	2.32-2.7 [1,2]	1.7-2.45 [3,4]	-	2.1-2.45 [3,5]
	C2 [16]	-	-	2.32-2.7	1.7-1.9 [6,7,9]	-	1.6-1.9 [6,7,8]
	C3 [15]	-	-	2.32-3.0 [2,10,11,12]	1.7-2.45	-	2.1-2.45
	C4 [16]	-	-	2.32-3.0	1.7-1.9	-	1.6-1.9
	C5 [17]	-	-	2.32-3.0	1.7-1.9	-	<2.1 [5]
	C6 [17]	-	-	2.32-2.7	1.7-1.9	-	<2.1
	C7	-	-	2.32-2.7	-	-	-
	C8	-	-	2.32-3.0	-	-	-
	C9	-	-	2.32-3.0	>1.7	>1.6	-
	C10	-	-	2.32-3.0	>1.7	>1.9	-
	C11	-	-	2.32-3.0	>1.7	>2.1	-
	C12	-	-	2.32-3.0	>1.9	>1.6	-
	C13	-	-	2.32-3.0	>1.9	>1.9	-
	C14	-	-	2.32-3.0	>1.9	>2.1	-
With Non-oxygenic Cyanobacteria Outgroup	C15	<3.8 [13,14]	>3.0 [10,11,12]	-	>1.7	>1.6	-
	C16	<3.8	>3.0	-	>1.7	>1.9	-
	C17	<3.8	>3.0	-	>1.7	>2.1	-
	C18	<3.8	>3.0	-	>1.9	>1.6	-
	C19	<3.8	>3.0	-	>1.9	>1.9	-
	C20	<3.8	>3.0	-	>1.9	>2.1	-
	C21	<4.0 [13,14]	>3.0	-	>1.7	>1.6	-
	C22	<4.0	>3.0	-	>1.7	>1.9	-
	C23	<4.0	>3.0	-	>1.7	>2.1	-
	C24	<4.0	>3.0	-	>1.9	>1.6	-
	C25	<4.0	>3.0	-	>1.9	>1.9	-
	C26	<4.0	>3.0	-	>1.9	>2.1	-
	C27	<4.2 [13,14]	>3.0	-	>1.7	>1.6	-
	C28	<4.2	>3.0	-	>1.7	>1.9	-
	C29	<4.2	>3.0	-	>1.7	>2.1	-
	C30	<4.2	>3.0	-	>1.9	>1.6	-
	C31	<4.2	>3.0	-	>1.9	>1.9	-
	C32	<4.2	>3.0	-	>1.9	>2.1	-
	C33	<4.5 [13,14]	>3.0	-	>1.7	>1.6	-
	C34	<4.5	>3.0	-	>1.7	>1.9	-
	C35	<4.5	>3.0	-	>1.7	>2.1	-
	C36	<4.5	>3.0	-	>1.9	>1.6	-
	C37	<4.5	>3.0	-	>1.9	>1.9	-
	C38	<4.5	>3.0	-	>1.9	>2.1	-

Reference

1 J. J. Brocks, R. Buick, R. E. Summons, G. A. Logan, A reconstruction of Archean biological diversity based on molecular fossils from the 2.78 to 2.45 billion-year-old Mount Bruce Supergroup, Hamersley Basin, Western Australia. Geochim Cosmochim Acta 67, 4321-4335 (2003).

2 H. D. Holland, Volcanic gases, black smokers, and the Great Oxidation Event. Geochim Cosmochim Acta 66, 3811-3826 (2002).

3 C. Blank, P. Sanchez‐Baracaldo, Timing of morphological and ecological innovations in the cyanobacteria ‐a key to understanding the rise in atmospheric oxygen. Geobiology 8, 1-23 (2010).

4 A. Knoll, S. Golubic, J. Green, K. Swett, Organically preserved microbial endoliths from the late Proterozoic of East Greenland. Nature 321, 856 (1986).

5 A. Tomitani, A. H. Knoll, C. M. Cavanaugh, T. Ohno, The evolutionary diversification of cyanobacteria: molecular ‐phylogenetic and paleontological perspectives. Proc Natl Acad Sci USA 103, 5442-5447 (2006).

6 H. Hofmann, Precambrian microflora, Belcher Islands, Canada: significance and systematics. J Paleontol, 1040-1073 (1976).

7 S. Golubic, L. Seong-Joo, Early cyanobacterial fossil record: preservation, palaeoenvironments and identification. Eur J Phycol 34, 339-348 (1999).

8 S. Golubic, V. N. Sergeev, A. H. Knoll, Mesoproterozoic Archaeoellipsoides: akinetes of heterocystous cyanobacteria. Lethaia 28, 285-298 (1995).

9 Y. Zhang, S. Golubic, Endolithic microfossils (cyanophyta) from early Proterozoic stromatolites, Hebei, China. Acta Micropaleontol. Sin 4, 1-3 (1987).

10 N. J. Planavsky et al., Evidence for oxygenic photosynthesis half a billion years before the Great Oxidation Event. Nat Geosci 7, 283 (2014).

11 S. A. Crowe et al., Atmospheric oxygenation three billion years ago. Nature 501, 535 (2013).

12 A. M. Satkoski, N. J. Beukes, W. Li, B. L. Beard, C. M. Johnson, A redox-stratified ocean 3.2 billion years ago. Earth Planet Sci Lett 430, 43-53 (2015).

13 Nisbet, E. G. & Sleep, N. H. The habitat and nature of early life. *Nature* **409**, 1083-1091, doi:10.1038/35059210 (2001).

14 Sleep, N. H., Zahnle, K. J., Kasting, J. F. & Morowitz, H. J. Annihilation of ecosystems by large asteroid impacts on the early Earth. *Nature* **342**, 139-142, doi:10.1038/342139a0 (1989).

15 P. Sánchez-Baracaldo, A. Ridgwell, J. A. Raven, A neoproterozoic transition in the marine nitrogen cycle. Current Biology 24, 652-657 (2014).

16 P. Sánchez-Baracaldo, J. A. Raven, D. Pisani, A. H. Knoll, Early photosynthetic eukaryotes inhabited low-salinity habitats. Proc Natl Acad Sci USA, 201620089 (2017).

17 P. Sánchez-Baracaldo, Origin of marine planktonic cyanobacteria. Sci Rep 5, 17418 (2015).

Table S2 Genomic properties of 317 cyanobacteria downloaded from the NCBI public database, among which the 159 high-quality reference or representative genomes are shaded.

Organism	Taxonomy ID	Genome Size	G+C%	Completeness (Contamination)	Refseq Category	Assembly Level	Ecotype
Prochlorococcus_marinus_str_AS9601	146891	1.67	31.32	99.64	representative genome	Complete genome	HLII
Prochlorococcus_marinus_str_MED4	59919	1.66	30.8	99.46	representative genome	Complete genome	HLI
Prochlorococcus_marinus_str_MIT_9107	59921	1.7	31.02	99.46	representative genome	Contig	HLII
Prochlorococcus_marinus_str_MIT_9201	93057	1.67	31.28	100	representative genome	Contig	HLII
Prochlorococcus_marinus_str_MIT_9301	167546	1.64	31.34	99.46	representative genome	Complete genome	HLII
Prochlorococcus_marinus_str_MIT_9303	59922	2.68	50.01	99.73	reference genome	Complete genome	LLIV
Prochlorococcus_marinus_str_MIT_9312	74546	1.71	31.21	99.73	representative genome	Complete genome	HLII
Prochlorococcus_marinus_str_MIT_9313	74547	2.41	50.74	99.18	representative genome	Complete genome	LLIV
Prochlorococcus_marinus_str_MIT_9515	167542	1.7	30.79	100	representative genome	Complete genome	HLI
Prochlorococcus_marinus_str_NATL2A	59920	1.84	35.12	98.64	representative genome	Complete genome	LLI
Prochlorococcus_marinus_str_SS120	167539	1.75	36.44	100	reference genome	Complete genome	LLII/III
Prochlorococcus_sp_MIT_0601	1499498	1.71	37.02	99.73	representative genome	Contig	LLII/III
Prochlorococcus_sp_MIT_0603	1499500	1.75	36.35	100	representative genome	Contig	LLII/III
Synechococcus_sp_CC9311	64471	2.61	52.45	99.73	representative genome	Complete genome	Synechococcus_5.1
Synechococcus_sp_CC9902	316279	2.23	54.16	99.46	representative genome	Complete genome	Synechococcus_5.1
Synechococcus_sp_KORDI-100	1280380	2.79	57.5	99.46	representative genome	Complete genome	Synechococcus_5.1
Synechococcus_sp_KORDI-49	585423	2.59	61.37	99.37	representative genome	Complete genome	Synechococcus_5.1
Synechococcus_sp_KORDI-52	585425	2.57	59.09	100	representative genome	Complete genome	Synechococcus_5.1
Synechococcus_sp_RS9916	221359	2.66	59.8	99.73	representative genome	Scaffold	Synechococcus_5.1
Synechococcus_sp_WH_7803	32051	2.37	60.24	99.18	representative genome	Complete genome	Synechococcus_5.1
Synechococcus_sp_WH_8102	84588	2.43	59.41	99.46	representative genome	Complete genome	Synechococcus_5.1
Prochlorococcus_marinus_str_EQPAC1	190047	1.65	30.79	99.46	na	Contig	HLI
Prochlorococcus_marinus_str_GP2	59925	1.62	31.16	99.46	na	Contig	HLII
Prochlorococcus_marinus_str_LG	167556	1.75	36.43	99.86	na	Contig	LLII/III
Prochlorococcus_marinus_str_MIT_9116	167544	1.69	31.01	99.18	na	Contig	HLII
Prochlorococcus_marinus_str_MIT_9123	167545	1.7	31.02	99.18	na	Contig	HLII
Prochlorococcus_marinus_str_MIT_9202	93058	1.69	31.08	98.78	na	Scaffold	HLII
Prochlorococcus_marinus_str_MIT_9211	93059	1.69	38.01	99.73	na	Chromosome	LLII/III
Prochlorococcus_marinus_str_MIT_9215	93060	1.74	31.15	99.73	na	Complete genome	HLII
Prochlorococcus_marinus_str_MIT_9302	74545	1.75	31.12	99.18	na	Contig	HLII
Prochlorococcus_marinus_str_MIT_9311	167547	1.71	31.21	99.73	na	Contig	HLII
Prochlorococcus_marinus_str_MIT_9314	167548	1.69	31.18	99.73	na	Contig	HLII
Prochlorococcus_marinus_str_MIT_9321	167549	1.66	31.2	99.73	na	Contig	HLII
Prochlorococcus_marinus_str_MIT_9322	167550	1.66	31.21	99.73	na	Contig	HLII
Prochlorococcus_marinus_str_MIT_9401	167551	1.67	31.21	99.73	na	Contig	HLII
Prochlorococcus_marinus_str_NATL1A	167555	1.86	34.98	98.91	na	Complete genome	LLI
Prochlorococcus_marinus_str_PAC1	59924	1.84	35.09	99.18	na	Contig	LLI
Prochlorococcus_marinus_str_SB	59926	1.67	31.5	99.91	na	Contig	HLII
Prochlorococcus_marinus_str_SS2	167552	1.75	36.44	100	na	Contig	LLII/III
Prochlorococcus_marinus_str_SS35	167553	1.75	36.44	99.86	na	Contig	LLII/III
Prochlorococcus_marinus_str_SS51	167554	1.75	36.43	100	na	Contig	LLII/III
Prochlorococcus_sp_MIT_0602	1499499	1.75	36.34	100	na	Contig	LLII/III
Prochlorococcus_sp_MIT_0604	1501268	1.78	31.17	99.73	na	Complete genome	HLII
Prochlorococcus_sp_MIT_0701	1499502	2.59	50.6	99.73	na	Contig	LLIV
Prochlorococcus_sp_MIT_0702	1499503	2.58	50.6	99.73	na	Contig	LLIV
Prochlorococcus_sp_MIT_0703	1499504	2.58	50.61	99.59	na	Contig	LLIV
Prochlorococcus_sp_MIT_0801	1501269	1.93	34.91	99.18	na	Complete genome	LLI
Prochlorococcus_sp_SS52	1499501	1.75	36.44	99.86	na	Contig	LLII/III
Synechococcus_sp_BL107	313625	2.29	54.2	99.46	na	Scaffold	Synechococcus_5.1
Synechococcus_sp_CC9605	110662	2.51	59.22	99.73	na	Complete genome	Synechococcus_5.1
Synechococcus_sp_CC9616	110663	2.65	56.52	99.46	na	Scaffold	Synechococcus_5.1
Synechococcus_sp_WH_7805	59931	2.63	57.49	99.73	na	Scaffold	Synechococcus_5.1
Synechococcus_sp_WH_8016	166318	2.69	54.09	99.18	na	Contig	Synechococcus_5.1
Synechococcus_sp_WH_8109	166314	2.11	60.09	99.46	na	Complete genome	Synechococcus_5.1
Acaryochloris_marina_MBIC11017	329726	8.36	46.96	99.53	representative genome	Complete genome	-
Aliterella_atlantica_CENA595	1618023	5.27	42.6	94.22	representative genome	Contig	-
Anabaena_cylindrica_PCC_7122	272123	7.06	38.79	99.44	representative genome	Complete genome	-
Anabaena_sp_90	46234	5.31	38.1	99.67	representative genome	Complete genome	-
Anabaena_sp_PCC_7108	163908	5.89	38.77	99.63	representative genome	Scaffold	-
Aphanizomenon_flos-aquae_NIES-81	284502	5.85	37.37	99.44	representative genome	Scaffold	-
Arthrospira_platensis_NIES-39	696747	6.79	43.65	99.13	representative genome	Chromosome	-
Calothrix_sp_336_3	1337936	6.42	41.1	100	representative genome	Complete genome	-
Calothrix_sp_PCC_6303	1170562	6.96	39.8	99.76	representative genome	Complete genome	-
Calothrix_sp_PCC_7103	32057	11.58	38.53	99.39	representative genome	Scaffold	-
Calothrix_sp_PCC_7507	99598	7.02	42.25	99.11	representative genome	Complete genome	-
Candidatus_Atelocyanobacterium_thalassa_isolate_ALOHA	1453429	1.44	31.12	73.92	representative genome	Complete genome	-
Candidatus_Synechococcus_spongiarum_SH4	1451353	1.66	63.05	83.83	representative genome	Contig	-
Chlorogloeopsis_fritschii_PCC_6912	211165	7.75	41.48	99.64	representative genome	Contig	-
Chroococcidiopsis_thermalis_PCC_7203	251229	6.69	44.47	99.63	representative genome	Complete genome	-
Chroogloeocystis_siderophila_5-2_s-c-1	247279	5.01	42.88	99.78	representative genome	Contig	-
Coleofasciculus_chthonoplastes_PCC_7420	118168	8.68	45.29	98.93	representative genome	Scaffold	-
Crinalium_epipsammum_PCC_9333	1173022	5.62	40.16	99.48	representative genome	Complete genome	-
Crocosphaera_watsonii_WH_8501	165597	6.24	37.11	99.74	representative genome	Scaffold	-
Cyanobacterium_aponinum_PCC_10605	755178	4.18	34.93	99.45	representative genome	Complete genome	-
Cyanobium_gracile_PCC_6307	292564	3.34	68.71	99.73	representative genome	Complete genome	-
Cyanobium_sp_PCC_7001	180281	2.83	68.7	99.46	representative genome	Scaffold	-
Cyanothece_sp_ATCC_51142	43989	5.46	37.94	99.96	representative genome	Complete genome	-
Cyanothece_sp_PCC_7424	65393	6.55	38.51	99.71	representative genome	Complete genome	-
Cyanothece_sp_PCC_7425	395961	5.79	50.65	99.29	representative genome	Complete genome	-
Cyanothece_sp_PCC_7822	497965	7.84	39.9	99.82	representative genome	Complete genome	-
Cyanothece_sp_PCC_8801	41431	4.79	39.76	99.56	representative genome	Complete genome	-
Cylindrospermum_stagnale_PCC_7417	56107	7.61	42.2	99.78	representative genome	Chromosome	-
Dactylococcopsis_salina_PCC_8305	13035	3.78	42.44	99.55	representative genome	Complete genome	-
Dolichospermum_circinale_AWQC310F	553470	4.41	37.33	99.56	representative genome	Scaffold	-
Fischerella_sp_PCC_9605	1173024	8.08	42.6	100	representative genome	Scaffold	-
Fischerella_thermalis_PCC_7521	98439	5.44	41.02	99.76	representative genome	Contig	-
Fortiea_contorta_PCC_7126	643473	5.74	42.2	99.56	representative genome	Scaffold	-
Geitlerinema_sp_PCC_7407	1173025	4.68	58.46	99.87	representative genome	Complete genome	-
Geminocystis_herdmanii_PCC_6308	113355	4.26	34.28	99.78	representative genome	Chromosome	-
Gloeobacter_kilaueensis_JS1	1183438	4.72	60.54	98.29	representative genome	Complete genome	-
Gloeobacter_violaceus_PCC_7421	251221	4.66	62	99.15	reference genome	Complete genome	-
Gloeocapsa_sp_PCC_73106	102232	4.03	41.11	98.84	representative genome	Contig	-
Halothece_sp_PCC_7418	65093	4.18	42.92	99.48	representative genome	Complete genome	-
Kamptonema_formosum_PCC_6407	402777	6.89	43.37	99.56	representative genome	Scaffold	-
Leptolyngbya_boryana_PCC_6306	272134	7.26	47.01	99.41	representative genome	Scaffold	-
Leptolyngbya_sp_Heron_Island_J	1385935	8.06	48.05	98.64	representative genome	Contig	-
Leptolyngbya_sp_JSC-1	1487953	7.87	48.86	99.53	representative genome	Scaffold	-
Leptolyngbya_sp_KIOST-1	1229172	6.32	59.44	99.18	representative genome	Contig	-
Leptolyngbya_sp_PCC_6406	1173264	5.78	55.11	98.64	representative genome	Scaffold	-
Leptolyngbya_sp_PCC_7375	102129	9.42	47.62	99.73	representative genome	Scaffold	-
Leptolyngbya_sp_PCC_7376	111781	5.13	43.87	99.42	representative genome	Complete genome	-
Leptolyngbya_valderiana_BDU_20041	322866	6.99	59.77	87.93	representative genome	Contig	-
Lyngbya_aestuarii_BL_J	1348334	6.87	41.16	99.74	representative genome	Contig	-
Lyngbya_confervoides_BDU141951	1574623	8.8	49.67	99.34	representative genome	Scaffold	-
Mastigocladopsis_repens_PCC_10914	221288	6.47	43.52	99.76	representative genome	Scaffold	-
Mastigocladus_laminosus_UU774	1594576	8.56	37.59	74.6	representative genome	Scaffold	-
Mastigocoleus_testarum_BC008	371196	12.7	37.28	99.04	representative genome	Contig	-
Microcoleus_sp_PCC_7113	1173027	7.97	46.21	99.56	representative genome	Complete genome	-
Microcystis_aeruginosa_NIES-843	449447	5.84	42.33	99.89	representative genome	Complete genome	-
Moorea_producens_JHB	1454205	9.38	43.55	99	representative genome	Chromosome	-
Myxosarcina_sp_GI1	1541065	7.07	40.1	99.56	representative genome	Contig	-
Neosynechococcus_sphagnicola_sy1_CAUP_A_1101	1497020	4.33	51.59	96.7	representative genome	Contig	-
Nodosilinea_nodulosa_PCC_7104	118166	6.89	57.64	99.18	representative genome	Scaffold	-
Nodularia_spumigena_CCY9414	313624	5.46	41.19	99.76	representative genome	Chromosome	-
Nostoc_azollae_0708	551115	5.49	38.37	98.89	representative genome	Complete genome	-
Nostoc_punctiforme_PCC_73102_ATCC_29133	63737	9.06	41.35	99.56	representative genome	Complete genome	-
Nostoc_sp_PCC_7107	317936	6.33	40.36	99.26	representative genome	Complete genome	-
Nostoc_sp_PCC_7120	103690	7.21	41.27	99.19	representative genome	Complete genome	-
Nostoc_sp_PCC_7524	28072	6.72	41.53	99.33	representative genome	Complete genome	-
Oscillatoria_acuminata_PCC_6304	56110	7.8	47.61	99.71	representative genome	Complete genome	-
Oscillatoria_nigro-viridis_PCC_7112	179408	8.27	45.78	99.78	representative genome	Complete genome	-
Oscillatoria_sp_PCC_10802	1173028	8.59	53.94	100	representative genome	Scaffold	-
Planktothricoides_sp_SR001	1705388	7.07	43.4	100	representative genome	Scaffold	-
Planktothrix_agardhii_NIVA-CYA_126_8	388467	5.05	39.57	100	representative genome	Chromosome	-
Pleurocapsa_sp_PCC_7319	118161	7.39	38.73	99.56	representative genome	Scaffold	-
Pleurocapsa_sp_PCC_7327	118163	4.99	45.19	99.49	representative genome	Complete genome	-
Prochlorothrix_hollandica_PCC_9006	317619	5.65	54.19	99.18	representative genome	Scaffold	-
Pseudanabaena_biceps_PCC_7429	927668	5.48	43.18	99.29	representative genome	Contig	-
Pseudanabaena_sp_PCC_6802	118173	5.62	47.81	99.76	representative genome	Scaffold	-
Pseudanabaena_sp_PCC_7367	82654	4.89	46.22	98.23	representative genome	Complete genome	-
Raphidiopsis_brookii_D9	533247	3.19	40.06	99.37	representative genome	Contig	-
Richelia_intracellularis_HH01	1165094	3.24	33.71	93.44	representative genome	Contig	-
Rivularia_sp_PCC_7116	373994	8.73	37.53	99.78	representative genome	Complete genome	-
Rubidibacter_lacunae_KORDI_51-2	582515	4.15	56.22	99.7	representative genome	Contig	-
Scytonema_hofmanni_UTEX_2349	1469607	8.13	41.14	99.76	representative genome	Scaffold	-
Scytonema_tolypothrichoides_VB-61278	1233231	7.89	42.33	100	representative genome	Scaffold	-
Stanieria_cyanosphaera_PCC_7437	111780	5.54	36.22	99.56	representative genome	Complete genome	-
Synechococcus_elongatus_PCC_6301	269084	2.7	55.48	99.73	representative genome	Complete genome	-
Synechococcus_sp_CB0101	232348	2.69	64.21	99.73	representative genome	Contig	-
Synechococcus_sp_CB0205	232363	2.43	62.97	99.18	representative genome	Contig	-
Synechococcus_sp_GFB01	1662190	2.34	67.77	84.81	representative genome	Contig	-
Synechococcus_sp_JA-2-3B_a-2-13	321332	3.05	58.45	100	representative genome	Complete genome	-
Synechococcus_sp_JA-3-3Ab	321327	2.93	60.24	100	representative genome	Complete genome	-
Synechococcus_sp_NKBG15041c	1407650	3.18	49.25	99.45	representative genome	Contig	-
Synechococcus_sp_PCC_6312	195253	3.72	48.5	99.29	representative genome	Complete genome	-
Synechococcus_sp_PCC_7002_ATCC_27264	32049	3.41	49.19	100	representative genome	Complete genome	-
Synechococcus_sp_PCC_7335	91464	5.97	48.16	98.91	representative genome	Scaffold	-
Synechococcus_sp_PCC_7336	195250	5.07	53.71	100	representative genome	Chromosome	-
Synechococcus_sp_PCC_7502	1173263	3.58	40.62	99.76	representative genome	Complete genome	-
Synechococcus_sp_RCC307	316278	2.22	60.84	99.64	representative genome	Complete genome	-
Synechococcus_sp_RS9917_RCC556	221360	2.58	64.33	99.46	representative genome	Scaffold	-
Synechococcus_sp_WH_5701	69042	3.28	60.68	99.18	representative genome	Scaffold	-
Synechocystis_sp_PCC_7509	927677	4.91	41.67	99.67	representative genome	Scaffold	-
Thermosynechococcus_elongatus_BP-1	197221	2.59	53.92	99.76	reference genome	Complete genome	-
Tolypothrix_bouteillei_VB521301	1479485	11.57	39.57	97.22	representative genome	Scaffold	-
Tolypothrix_campylonemoides_VB511288	1245935	9.47	44.67	99.52	representative genome	Scaffold	-
Tolypothrix_sp_PCC_7601_UTEX_B_481	1188	9.98	40.69	99.11	representative genome	Scaffold	-
Trichodesmium_erythraeum_IMS101	203124	7.75	34.14	99.71	representative genome	Complete genome	-
Xenococcus_sp_PCC_7305	102125	5.93	39.68	99.78	representative genome	Contig	-
Acaryochloris_sp_CCMEE_5410	310037	7.88	47.12	99.29	na	Contig	-
Anabaena_sp_4-3	1811979	5.6	41.36	99.56	na	Scaffold	-
Anabaena_sp_CA	52271	5.56	41.3	98.56	na	Scaffold	-
Anabaena_sp_WA102	1647413	5.78	38.38	99.78	na	Complete genome	-
Aphanizomenon_flos-aquae_2012_KM1_D3	1532906	5.74	38.22	87.52	na	Contig	-
Arthrospira_maxima_CS-328	513049	6	44.76	96.58	na	Contig	-
Arthrospira_platensis_C1	459495	6.09	43.55	99.71	na	Chromosome	-
Arthrospira_platensis_str_Paraca	634502	6.5	44.31	99.34	na	Contig	-
Arthrospira_platensis_YZ	1738638	6.52	44.19	99.71	na	Chromosome	-
Arthrospira_sp_PCC_8005	376219	6.23	44.73	99.42	na	Chromosome	-
Arthrospira_sp_TJSD091	1640536	5.98	44.75	98.51	na	Contig	-
Calothrix_rhizosoleniae_SC01	439482	5.97	39.47	98.78	na	Contig	-
Calothrix_sp_HK-06	1137096	9.99	38.81	99.15	na	Scaffold	-
Candidatus_Synechococcus_spongiarum	431041	1.41	61.86	74.82	na	Contig	-
Candidatus_Synechococcus_spongiarum_LMB_bulk15N	1943583	1.47	59.8	73.91	na	Scaffold	-
Chamaesiphon_minutus_PCC_6605	1173020	6.76	45.67	99.48	na	Chromosome	-
Chlorogloeopsis_fritschii_PCC_9212	184925	7.65	41.5	99.76	na	Contig	-
Crocosphaera_watsonii_WH_0003	423471	5.89	37.68	99.49	na	Contig	-
Crocosphaera_watsonii_WH_0005	423472	5.96	37.66	98.4	na	Contig	-
Crocosphaera_watsonii_WH_0401	555881	4.55	37.66	99.34	na	Contig	-
Crocosphaera_watsonii_WH_0402	1284629	5.88	37.69	98.14	na	Contig	-
Crocosphaera_watsonii_WH_8502	423474	4.68	37.63	99.67	na	Contig	-
Cyanobacterium_sp_IPPAS_B-1200	1562720	3.41	37.73	99.34	na	Contig	-
Cyanobium_sp_NIES-981	1851505	3.02	68.62	99.73	na	Complete genome	-
Cyanothece_sp_ATCC_51472	860575	5.46	37.93	99.96	na	Contig	-
Cyanothece_sp_CCY0110	391612	5.88	36.69	99.78	na	Contig	-
Cyanothece_sp_PCC_8802	395962	4.8	39.82	99.56	na	Complete genome	-
Cylindrospermopsis_raciborskii_CENA302	1170768	3.48	40.08	99.74	na	Contig	-
Cylindrospermopsis_raciborskii_CENA303	1170769	3.4	40.25	99.37	na	Contig	-
Cylindrospermopsis_raciborskii_CS-505	533240	3.88	40.23	99.85	na	Contig	-
Cylindrospermopsis_raciborskii_CS-505_CS505	533240	4.16	40.28	99.85	na	Contig	-
Cylindrospermopsis_raciborskii_CS-508	533243	3.56	40.16	97.41	na	Contig	-
Cylindrospermopsis_raciborskii_ITEP-A1	1810942	3.61	40.15	99.85	na	Contig	-
Cylindrospermopsis_raciborskii_MVCC14	940191	3.59	40.08	97.74	na	Contig	-
Cylindrospermopsis_sp_CR12	1747196	3.72	40.04	99.93	na	Scaffold	-
Desertifilum_sp_IPPAS_B-1220	1781255	6.09	48.68	99.29	na	Contig	-
Dolichospermum_circinale_AWQC131C	398007	4.45	37.01	99.56	na	Scaffold	-
Fischerella_major_NIES-592	210994	5.64	41.03	99.76	na	Scaffold	-
Fischerella_muscicola_PCC_7414	306281	6.9	41.26	99.52	na	Contig	-
Fischerella_muscicola_SAG_1427-1	372781	7.36	40.32	99.52	na	Contig	-
Fischerella_sp_JSC-11	741277	5.38	41.05	99.76	na	Contig	-
Fischerella_sp_NIES-3754	1752063	5.83	40.98	99.76	na	Complete genome	-
Fischerella_sp_PCC_9339	1174528	8.01	40.12	99.76	na	Scaffold	-
Fischerella_sp_PCC_9431	1173023	7.17	40.17	99.76	na	Scaffold	-
Geitlerinema_sp_PCC_7105	102127	6.15	51.38	93.75	na	Scaffold	-
Geitlerinema_sp_PCC_9228	111611	4.8	48.8	99.37	na	Scaffold	-
Geminocystis_sp_NIES-3708	1615909	4.04	32.29	99.78	na	Complete genome	-
Geminocystis_sp_NIES-3709	1617448	4.43	33.36	99.78	na	Complete genome	-
Gloeocapsa_sp_PCC_7428	1173026	5.88	43.36	99.78	na	Complete genome	-
Gloeomargarita_lithophora_Alchichica-D10	1188229	3.05	51.34	98.29	na	Chromosome	-
Halomicronema_hongdechloris_C2206	1641165	5.57	54.62	98.82	na	Complete genome	-
Hapalosiphon_sp_MRB220	1704290	7.43	40.19	99.52	na	Scaffold	-
Hydrococcus_rivularis_NIES-593	1921803	5.15	44.98	99.49	na	Scaffold	-
Leptolyngbya_boryana_dg5	1260963	6.8	47.01	99.41	na	Complete genome	-
Leptolyngbya_boryana_IAM_M-101	411966	6.8	47.01	99.41	na	Complete genome	-
Leptolyngbya_sp_NIES-2104	1552121	6.39	47.43	99.29	na	Contig	-
Leptolyngbya_sp_NIES-3755	1752064	6.76	46.7	99.41	na	Complete genome	-
Leptolyngbya_sp_O-77	1080068	5.48	55.93	98.7	na	Complete genome	-
Limnoraphis_robusta_CS-951	1637645	7.31	41.62	84.72	na	Scaffold	-
Limnothrix_rosea_IAM_M-220	454133	4.07	45.42	100	na	Scaffold	-
Lyngbya_sp_PCC_8106	313612	7.04	41.11	99.3	na	Contig	-
Mastigocladus_laminosus_74	1913078	7.33	32.9	73.45	na	Scaffold	-
Microcoleus_vaginatus_FGP-2	756067	6.7	46.04	99.67	na	Contig	-
Microcystis_aeruginosa_CHAOHU_1326	1914535	5.27	42.54	99.67	na	Scaffold	-
Microcystis_aeruginosa_DIANCHI905	1235808	4.86	42.45	99.01	na	Contig	-
Microcystis_aeruginosa_KW	1960155	5.89	42.81	97.92	na	Contig	-
Microcystis_aeruginosa_NaRes975	1914537	5.12	42.41	99.89	na	Contig	-
Microcystis_aeruginosa_NIES-2481	1698524	4.29	42.91	99.82	na	Complete genome	-
Microcystis_aeruginosa_NIES-2549	1641812	4.29	42.92	99.89	na	Complete genome	-
Microcystis_aeruginosa_NIES-44	449439	4.57	43.19	99.89	na	Contig	-
Microcystis_aeruginosa_NIES-88	449441	5.26	41.37	99.45	na	Scaffold	-
Microcystis_aeruginosa_NIES-98	267869	4.98	42.41	99.67	na	Scaffold	-
Microcystis_aeruginosa_PCC_7005	267870	4.91	42.51	99.89	na	Contig	-
Microcystis_aeruginosa_PCC_7806SL	1903187	5.14	42.09	99.67	na	Complete genome	-
Microcystis_aeruginosa_PCC_7941	213618	4.8	42.25	98.57	na	Scaffold	-
Microcystis_aeruginosa_PCC_9432	1160280	4.99	42.16	99.67	na	Scaffold	-
Microcystis_aeruginosa_PCC_9443	1160281	5.19	42.14	98.36	na	Scaffold	-
Microcystis_aeruginosa_PCC_9701	721123	4.76	42.42	99.12	na	Scaffold	-
Microcystis_aeruginosa_PCC_9717	1160286	5.3	42.11	98.57	na	Scaffold	-
Microcystis_aeruginosa_PCC_9806	1160282	4.26	42.79	99.01	na	Scaffold	-
Microcystis_aeruginosa_PCC_9807	1160283	5.16	42.01	99.01	na	Scaffold	-
Microcystis_aeruginosa_PCC_9808	1160284	5.05	42.04	99.01	na	Scaffold	-
Microcystis_aeruginosa_PCC_9809	1160285	5.01	42.15	98.36	na	Scaffold	-
Microcystis_aeruginosa_SPC777	482300	5.45	42.65	94.85	na	Contig	-
Microcystis_aeruginosa_TAIHU98	1134457	4.85	42.45	99.89	na	Contig	-
Microcystis_sp_T1-4	1160279	4.69	42.78	99.67	na	Contig	-
Moorea_bouillonii_PNG_PNG5-198	568701	8.32	43.46	98.11	na	Scaffold	-
Moorea_producens_3L	489825	8.48	43.22	98.11	na	Scaffold	-
Moorea_producens_PAL_NAK12DEC93-3La	1905009	8.37	43.57	97.74	na	Scaffold	-
Moorea_producens_PAL-8-15-08-1	1458985	9.71	43.52	99.26	na	Complete genome	-
Nodularia_spumigena_CENA596	1819295	5.19	41.2	100	na	Contig	-
Nostoc_calcicola_FACHB-389	1357508	9.05	41.47	99.22	na	Scaffold	-
Nostoc_piscinale_CENA21	224013	7.09	40.54	96.99	na	Chromosome	-
Nostoc_sp_KVJ20	457944	9.18	41.69	99.67	na	Scaffold	-
Nostoc_sp_NIES-3756	1751286	6.99	40.58	99.3	na	Complete genome	-
Oscillatoria_sp_PCC_6506	272129	6.68	43.4	99.12	na	Contig	-
Phormidesmis_priestleyi_BC1401	1850361	5.55	49.16	99.29	na	Scaffold	-
Phormidesmis_priestleyi_ULC007	1920490	5.68	48.56	95.99	na	Scaffold	-
Phormidium_ambiguum_IAM_M-71	454136	7.41	39.53	99.26	na	Scaffold	-
Phormidium_tenue_NIES-30	549789	5.82	55.99	99.15	na	Scaffold	-
Phormidium_willei_BDU_130791	399101	4.6	53.36	88.92	na	Contig	-
Planktothrix_agardhii_NIVA-CYA_15	329553	5.49	38.81	99.56	na	Scaffold	-
Planktothrix_agardhii_NIVA-CYA_34	213631	5.52	39.48	99.56	na	Scaffold	-
Planktothrix_agardhii_NIVA-CYA_56_3	329545	5.51	39.32	99.56	na	Scaffold	-
Planktothrix_mougeotii_NIVA-CYA_405	545630	5.5	39.29	99.34	na	Scaffold	-
Planktothrix_paucivesiculata_PCC_9631	671071	68	3.58	98.9	na	Scaffold	-
Planktothrix_prolifica_NIVA-CYA_406	545626	5.76	38.62	100	na	Scaffold	-
Planktothrix_prolifica_NIVA-CYA_540	1255359	5.57	39.02	99.78	na	Scaffold	-
Planktothrix_prolifica_NIVA-CYA_98	329562	5.74	38.76	99.56	na	Scaffold	-
Planktothrix_rubescens_7821	59512	11.39	19.4	100	na	Scaffold	-
Planktothrix_rubescens_NIVA-CYA_407	1255374	5.51	38.69	99.89	na	Scaffold	-
Planktothrix_serta_PCC_8927	671068	55.18	4.48	98.77	na	Scaffold	-
Planktothrix_tepida_PCC_9214	671072	6.78	39.14	99.78	na	Scaffold	-
Prochlorococcus_marinus_SCGC_AAA795-F05	1219	1.42	31.36	87.38	na	Contig	-
Prochlorococcus_marinus_SCGC_AAA795-I06	1219	1.39	31.12	82.32	na	Contig	-
Prochlorococcus_marinus_SCGC_AAA795-I15	1219	1.28	31.34	82.38	na	Contig	-
Prochlorococcus_marinus_SCGC_AAA795-J16	1219	1.46	31	86.05	na	Contig	-
Prochlorococcus_marinus_SCGC_AAA795-M23	1219	1.44	31.19	86.46	na	Contig	-
Prochlorococcus_marinus_str_MIT_1320	1801623	2.5	50.35	99.46	na	Contig	-
Prochlorococcus_marinus_str_MIT_1323	1801625	2.44	50.56	99.46	na	Contig	-
Prochlorococcus_marinus_str_MIT_1342	1801627	2.55	50.26	99.73	na	Contig	-
Prochlorococcus_sp_MIT_1303	1723647	2.56	50.47	99.46	na	Contig	-
Prochlorococcus_sp_MIT_1306	1799667	2.5	50.5	99.73	na	Contig	-
Pseudanabaena_sp_Roaring_Creek_	1681830	5.48	43.33	99.06	na	Contig	-
Richelia_intracellularis_HM01	1165095	2.21	33.76	64.75	na	Contig	-
Scytonema_hofmannii_PCC_7110	128403	12.29	41.49	99.76	na	Scaffold	-
Scytonema_millei_VB511283	1245923	7.25	39.85	53.79	na	Scaffold	-
Scytonema_sp_HK-05	1137095	9.56	43.71	99.76	na	Scaffold	-
Spirulina_major_PCC_6313	129961	5.05	53.46	99.67	na	Scaffold	-
Spirulina_subsalsa_PCC_9445	1173029	5.32	47.39	99.56	na	Scaffold	-
Synechococcus_elongatus_PCC_7942	1140	2.74	55.43	100	na	Complete genome	-
Synechococcus_sp_MIT_S9504	1801628	3.09	55.43	99.73	na	Scaffold	-
Synechococcus_sp_MIT_S9508	1801629	2.5	56.05	100	na	Scaffold	-
Synechococcus_sp_MIT_S9509	1801630	3.09	55.42	99.18	na	Scaffold	-
Synechococcus_sp_NKBG042902	490193	3.32	49.35	99.34	na	Contig	-
Synechococcus_sp_PCC_7003	374981	3.35	49.26	100	na	Complete genome	-
Synechococcus_sp_PCC_7117	195498	3.43	49.13	100	na	Complete genome	-
Synechococcus_sp_PCC_73109	374982	3.3	49.34	99.78	na	Complete genome	-
Synechococcus_sp_PCC_8807	195248	3.3	49.37	100	na	Complete genome	-
Synechococcus_sp_UTEX_2973	1350461	2.74	55.44	99.73	na	Complete genome	-
Synechococcus_sp_WH_8020	32052	2.66	53.12	99.73	na	Chromosome	-
Synechococcus_sp_WH_8103	29410	2.43	59.47	99.73	na	Complete genome	-
Synechocystis_sp_PCC_6714	1147	3.74	47.55	99.56	na	Complete genome	-
Synechocystis_sp_PCC_6803	1148	3.95	47.37	99.78	na	Complete genome	-
Synechocystis_sp_PCC_6803_GT-S	1148	3.57	47.73	99.78	na	Complete genome	-
Synechocystis_sp_PCC_6803_substr_GT-I	1080228	3.57	47.74	99.78	na	Complete genome	-
Synechocystis_sp_PCC_6803_substr_PCC-N	1080229	3.57	47.74	99.78	na	Complete genome	-
Synechocystis_sp_PCC_6803_substr_PCC-P	1080230	3.57	47.74	99.78	na	Complete genome	-
Thermosynechococcus_sp_NK55a	1394889	2.52	53.81	99.76	na	Chromosome	-
Trichormus_sp_NMC-1	1853259	5.94	38.92	99	na	Scaffold	-
Candidatus Caenarcanum bioreactoricola_UASB_169	1906166	0.51	38.37	84.62 (0)	na	Scaffold	Non-oxygenic outgroup
Candidatus_Gastranaerophilales_bacterium_UMGS1517	2137880	0.68	37.2	87.61 (3.53)	na	Contig	Non-oxygenic outgroup
Candidatus_Melainabacteria_bacterium_SSGW_16	2052166	0.67	40	94.87 (0.85)	na	Contig	Non-oxygenic outgroup
Candidatus_Obscuribacter_phosphatis__EBPR_351	1906157	1.62	36.29	97.44 (2.28)	na	Scaffold	Non-oxygenic outgroup
Candidatus_Sericytochromatia_bacterium_GL2-53_LSPB_72	2762020	1.25	34.23	75.25 (3.85)	na	Contig	Non-oxygenic outgroup
Candidatus_Sericytochromatia_bacterium_S15B-MN24_CBMW_12	2762020	1.16	35.25	89.74 (2.56)	na	Contig	Non-oxygenic outgroup
Candidatus_Sericytochromatia_bacterium_S15B-MN24_RAAC_196	2762020	0.97	40.04	80.10 (1.71)	na	Contig	Non-oxygenic outgroup
Vampirovibrio_chlorellavorus_Vc_AZ_2	758823	0.84	35	95.73 (3.97)	na	Contig	Non-oxygenic outgroup

Table S3 A list of the 27 genes used for the relaxed molecular clock analyses.

ID*	**Name**	**Description**
23S	LSU	23S ribosomal RNA
16S	SSU	16S ribosomal RNA
COG0049	RpsG	Ribosomal protein S7 [Translation, ribosomal structure and biogenesis].
COG0050	TufB	Translation elongation factor EF-Tu, a GTPase [Translation, ribosomal structure and biogenesis].
COG0052	RpsB	Ribosomal protein S2 [Translation, ribosomal structure and biogenesis].
COG0080	RplK	Ribosomal protein L11 [Translation, ribosomal structure and biogenesis].
COG0081	RplA	Ribosomal protein L1 [Translation, ribosomal structure and biogenesis].
COG0085	RpoB	DNA-directed RNA polymerase, beta subunit/140 kD subunit [Transcription].
-	RpoC1	DNA-directed RNA polymerase, gamma subunit/160 kD subunit [Transcription].
COG0086	RpoC2	DNA-directed RNA polymerase, beta' subunit/160 kD subunit [Transcription].
COG0087	RplC	Ribosomal protein L3 [Translation, ribosomal structure and biogenesis].
COG0090	RplB	Ribosomal protein L2 [Translation, ribosomal structure and biogenesis].
COG0092	RpsC	Ribosomal protein S3 [Translation, ribosomal structure and biogenesis].
COG0094	RplE	Ribosomal protein L5 [Translation, ribosomal structure and biogenesis].
COG0097	RplF	Ribosomal protein L6P/L9E [Translation, ribosomal structure and biogenesis].
COG0098	RpsE	Ribosomal protein S5 [Translation, ribosomal structure and biogenesis].
COG0100	RpsK	Ribosomal protein S11 [Translation, ribosomal structure and biogenesis].
COG0102	RplM	Ribosomal protein L13 [Translation, ribosomal structure and biogenesis].
COG0103	RpsI	Ribosomal protein S9 [Translation, ribosomal structure and biogenesis].
COG0197	RplP	Ribosomal protein L16/L10AE [Translation, ribosomal structure and biogenesis].
COG0201	SecY	Preprotein translocase subunit SecY [Intracellular trafficking, secretion, and vesicular transport].
COG0202	RpoA	DNA-directed RNA polymerase, alpha subunit/40 kD subunit [Transcription].
COG0250	NusG	Transcription antitermination factor NusG [Transcription].
COG0480	FusA	Translation elongation factor EF-G, a GTPase [Translation, ribosomal structure and biogenesis].
COG0522	RpsD	Ribosomal protein S4 or related protein [Translation, ribosomal structure and biogenesis].
COG0533	TsaD	tRNA A37 threonylcarbamoyltransferase TsaD [Translation, ribosomal structure and biogenesis].
COG0592	DnaN	DNA polymerase III sliding clamp (beta) subunit, PCNA homolog [Replication, recombination and repair].

* The 25 core gene families were identified in (Battistuzzi and Hedges 2008). The translation initiation factor IF-2 family (PRK05306) is not recorded in COG and was not included in our analysis. Instead, the DNA-directed RNA polymerase, gamma subunit (*RpoC1*) was added. *RpoC1* was used as one of the highly-conserved core genes in (Sánchez-Baracaldo et al. 2014) to date the rise of marine picocyanobacteria and planktonic N_2-fixers.

Reference

Battistuzzi FU, Hedges SB (2008). A major clade of prokaryotes with ancient adaptations to life on land. *Mol Biol Evol* **26:** 335-343.

Sánchez-Baracaldo P, Ridgwell A, Raven JA (2014). A neoproterozoic transition in the marine nitrogen cycle. *Current Biology* **24:** 652-657.

Table S4 A list of the single-copy orthologous gene families used for phylogenomic construction. Among the 214 families, 90 (marked with asterisks) each show composition homogeneity in the protein sequences. The Clusters of Orthologous Groups (COGs) annotation is also provided.

Family_ID	COG_ID	Gene	Description
OG3691	COG0772	FtsW	Bacterial cell division protein FtsW, lipid II flippase [Cell cycle control, cell division, chromosome partitioning].
OG4449*	COG0206	FtsZ	Cell division GTPase FtsZ [Cell cycle control, cell division, chromosome partitioning].
OG2545	COG0771	MurD	UDP-N-acetylmuramoylalanine-D-glutamate ligase [Cell wall/membrane/envelope biogenesis].
OG2602	COG1207	GlmU	Bifunctional protein GlmU, N-acetylglucosamine-1-phosphate-uridyltransferase/glucosamine-1-phosphate-acetyltransferase [Cell wall/membrane/envelope biogenesis].
OG3799	COG0438	RfaB	Glycosyltransferase involved in cell wall bisynthesis [Cell wall/membrane/envelope biogenesis].
OG4486	COG0438	RfaB	Glycosyltransferase involved in cell wall bisynthesis [Cell wall/membrane/envelope biogenesis].
OG5750	COG0451	WcaG	Nucleoside-diphosphate-sugar epimerase [Cell wall/membrane/envelope biogenesis].
OG7226	COG0812	MurB	UDP-N-acetylenolpyruvoylglucosamine reductase [Cell wall/membrane/envelope biogenesis].
OG8607	COG0084	TatD	Tat protein secretion system quality control protein TatD (DNase activity) [Cell motility].
OG425*	COG0542	ClpA	ATP-dependent Clp protease ATP-binding subunit ClpA [Posttranslational modification, protein turnover, chaperones].
OG1182*	COG0443	DnaK	Molecular chaperone DnaK (HSP70) [Posttranslational modification, protein turnover, chaperones].
OG1291*	COG0465	HflB	ATP-dependent Zn proteases [Posttranslational modification, protein turnover, chaperones].
OG1903*	COG0459	GroEL	Chaperonin GroEL (HSP60 family) [Posttranslational modification, protein turnover, chaperones].
OG2491	COG0719	SufB	Fe-S cluster assembly scaffold protein SufB [Posttranslational modification, protein turnover, chaperones].
OG2661	COG0544	Tig	FKBP-type peptidyl-prolyl cis-trans isomerase (trigger factor) [Posttranslational modification, protein turnover, chaperones].
OG2975*	COG1219	ClpX	ATP-dependent protease Clp, ATPase subunit [Posttranslational modification, protein turnover, chaperones].
OG4543	COG0484	DnaJ	DnaJ-class molecular chaperone with C-terminal Zn finger domain [Posttranslational modification, protein turnover, chaperones].
OG6664	COG0755	CcmC	ABC-type transport system involved in cytochrome c biogenesis, permease component [Posttranslational modification, protein turnover, chaperones].
OG9714	COG0396	SufC	Fe-S cluster assembly ATPase SufC [Posttranslational modification, protein turnover, chaperones].
OG12234*	COG0740	ClpP	ATP-dependent protease ClpP, protease subunit [Posttranslational modification, protein turnover, chaperones].
OG17759*	COG0691	SmpB	tmRNA-binding protein [Posttranslational modification, protein turnover, chaperones].
OG25376*	COG0278	GrxD	Glutaredoxin-related protein [Posttranslational modification, protein turnover, chaperones].
OG25557*	COG0526	TrxA	Thiol-disulfide isomerase or thioredoxin [Posttranslational modification, protein turnover, chaperones].
OG907	COG0642	BaeS	Signal transduction histidine kinase [Signal transduction mechanisms].
OG1352*	COG1217	TypA	Predicted membrane GTPase involved in stress response [Signal transduction mechanisms].
OG2082*	COG0467	RAD55	RecA-superfamily ATPase, KaiC/GvpD/RAD55 family [Signal transduction mechanisms].
OG4587	COG0642	BaeS	Signal transduction histidine kinase [Signal transduction mechanisms].
OG10988*	COG0664	Crp	cAMP-binding domain of CRP or a regulatory subunit of cAMP-dependent protein kinases [Signal transduction mechanisms].
OG18366	COG0394	Wzb	Protein-tyrosine-phosphatase [Signal transduction mechanisms].
OG265	COG0653	SecA	Preprotein translocase subunit SecA (ATPase, RNA helicase) [Intracellular trafficking, secretion, and vesicular transport].
OG2397	COG0541	Ffh	Signal recognition particle GTPase [Intracellular trafficking, secretion, and vesicular transport].
OG2511	COG0342	SecD	Preprotein translocase subunit SecD [Intracellular trafficking, secretion, and vesicular transport].
OG3395*	COG0201	SecY	Preprotein translocase subunit SecY [Intracellular trafficking, secretion, and vesicular transport].
OG1420	COG1132	MdlB	ABC-type multidrug transport system, ATPase and permease component [Defense mechanisms].
OG1560	COG1132	MdlB	ABC-type multidrug transport system, ATPase and permease component [Defense mechanisms].
OG8574	COG0842	YadH	ABC-type multidrug transport system, permease component [Defense mechanisms].
OG14048*	COG0450	AhpC	Alkyl hydroperoxide reductase subunit AhpC (peroxiredoxin) [Defense mechanisms].
OG15254*	COG1403	McrA	5-methylcytosine-specific restriction endonuclease McrA [Defense mechanisms].
OG1373	COG0768	FtsI	Cell division protein FtsI/penicillin-binding protein 2 [Cell cycle control, cell division, chromosome partitioning, Cell wall/membrane/envelope biogenesis].
OG783*	COG1185	Pnp	Polyribonucleotide nucleotidyltransferase (polynucleotide phosphorylase) [Translation, ribosomal structure and biogenesis].
OG891*	COG0480	FusA	Translation elongation factor EF-G, a GTPase [Translation, ribosomal structure and biogenesis].
OG935*	COG0595	RnjA	mRNA degradation ribonuclease J1/J2 [Translation, ribosomal structure and biogenesis].
OG972	COG0445	MnmG	tRNA U34 5-carboxymethylaminomethyl modifying enzyme MnmG/GidA [Translation, ribosomal structure and biogenesis].
OG2488	COG0154	GatA	Asp-tRNAAsn/Glu-tRNAGln amidotransferase A subunit or related amidase [Translation, ribosomal structure and biogenesis].
OG2515	COG0621	MiaB	tRNA A37 methylthiotransferase MiaB [Translation, ribosomal structure and biogenesis].
OG3424	COG0172	SerS	Seryl-tRNA synthetase [Translation, ribosomal structure and biogenesis].
OG3654	COG0162	TyrS	Tyrosyl-tRNA synthetase [Translation, ribosomal structure and biogenesis].
OG4878*	COG0539	RpsA	Ribosomal protein S1 [Translation, ribosomal structure and biogenesis].
OG5039	COG0216	PrfA	Protein chain release factor A [Translation, ribosomal structure and biogenesis].
OG5194	COG0012	GTP1	Ribosome-binding ATPase YchF, GTP1/OBG family [Translation, ribosomal structure and biogenesis].
OG5390	COG0533	TsaD	tRNA A37 threonylcarbamoyltransferase TsaD [Translation, ribosomal structure and biogenesis].
OG5757	COG0223	Fmt	Methionyl-tRNA formyltransferase [Translation, ribosomal structure and biogenesis].
OG5935	COG0016	PheS	Phenylalanyl-tRNA synthetase alpha subunit [Translation, ribosomal structure and biogenesis].
OG6058	COG1600	QueG	Epoxyqueuosine reductase QueG (queuosine biosynthesis) [Translation, ribosomal structure and biogenesis].
OG6399	COG0564	RluA	Pseudouridylate synthase, 23S rRNA- or tRNA-specific [Translation, ribosomal structure and biogenesis].
OG6588	COG1234	ElaC	Ribonuclease BN, tRNA processing enzyme [Translation, ribosomal structure and biogenesis].
OG6946	COG1159	Era	GTPase Era, involved in 16S rRNA processing [Translation, ribosomal structure and biogenesis].
OG7548	COG0101	TruA	tRNA U38,U39,U40 pseudouridine synthase TruA [Translation, ribosomal structure and biogenesis].
OG7880	COG1161	RbgA	Ribosome biogenesis GTPase A [Translation, ribosomal structure and biogenesis].
OG8167	COG0566	SpoU	tRNA G18 (ribose-2'-O)-methylase SpoU [Translation, ribosomal structure and biogenesis].
OG8506	COG0024	Map	Methionine aminopeptidase [Translation, ribosomal structure and biogenesis].
OG9130	COG1189	YqxC	Predicted rRNA methylase YqxC, contains S4 and FtsJ domains [Translation, ribosomal structure and biogenesis].
OG9541*	COG0052	RpsB	Ribosomal protein S2 [Translation, ribosomal structure and biogenesis].
OG11488*	COG0081	RplA	Ribosomal protein L1 [Translation, ribosomal structure and biogenesis].
OG13244*	COG0098	RpsE	Ribosomal protein S5 [Translation, ribosomal structure and biogenesis].
OG13767	COG0193	Pth	Peptidyl-tRNA hydrolase [Translation, ribosomal structure and biogenesis].
OG14137*	COG0522	RpsD	Ribosomal protein S4 or related protein [Translation, ribosomal structure and biogenesis].
OG15443*	COG0231	Efp	Translation elongation factor P (EF-P)/translation initiation factor 5A (eIF-5A) [Translation, ribosomal structure and biogenesis].
OG15930*	COG0233	Frr	Ribosome recycling factor [Translation, ribosomal structure and biogenesis].
OG16246	COG0097	RplF	Ribosomal protein L6P/L9E [Translation, ribosomal structure and biogenesis].
OG16323*	COG0094	RplE	Ribosomal protein L5 [Translation, ribosomal structure and biogenesis].
OG16585	COG0590	TadA	tRNA(Arg) A34 adenosine deaminase TadA [Translation, ribosomal structure and biogenesis].
OG16960*	COG0244	RplJ	Ribosomal protein L10 [Translation, ribosomal structure and biogenesis].
OG18594*	COG0049	RpsG	Ribosomal protein S7 [Translation, ribosomal structure and biogenesis].
OG19291	COG0200	RplO	Ribosomal protein L15 [Translation, ribosomal structure and biogenesis].
OG19392*	COG0102	RplM	Ribosomal protein L13 [Translation, ribosomal structure and biogenesis].
OG20601	COG0858	RbfA	Ribosome-binding factor A [Translation, ribosomal structure and biogenesis].
OG20776*	COG0080	RplK	Ribosomal protein L11 [Translation, ribosomal structure and biogenesis].
OG21459*	COG0103	RpsI	Ribosomal protein S9 [Translation, ribosomal structure and biogenesis].
OG21598*	COG0096	RpsH	Ribosomal protein S8 [Translation, ribosomal structure and biogenesis].
OG22505*	COG0048	RpsL	Ribosomal protein S12 [Translation, ribosomal structure and biogenesis].
OG23198*	COG0099	RpsM	Ribosomal protein S13 [Translation, ribosomal structure and biogenesis].
OG23416*	COG0256	RplR	Ribosomal protein L18 [Translation, ribosomal structure and biogenesis].
OG24026*	COG0198	RplX	Ribosomal protein L24 [Translation, ribosomal structure and biogenesis].
OG26629*	COG0089	RplW	Ribosomal protein L23 [Translation, ribosomal structure and biogenesis].
OG26634*	COG0199	RpsN	Ribosomal protein S14 [Translation, ribosomal structure and biogenesis].
OG27183*	COG0254	RpmE	Ribosomal protein L31 [Translation, ribosomal structure and biogenesis].
OG27508	COG0721	GatC	Asp-tRNAAsn/Glu-tRNAGln amidotransferase C subunit [Translation, ribosomal structure and biogenesis].
OG28588*	COG0361	InfA	Translation initiation factor IF-1 [Translation, ribosomal structure and biogenesis].
OG29079*	COG0184	RpsO	Ribosomal protein S15P/S13E [Translation, ribosomal structure and biogenesis].
OG29422*	COG0211	RpmA	Ribosomal protein L27 [Translation, ribosomal structure and biogenesis].
OG31261*	COG0227	RpmB	Ribosomal protein L28 [Translation, ribosomal structure and biogenesis].
OG32944*	COG0238	RpsR	Ribosomal protein S18 [Translation, ribosomal structure and biogenesis].
OG2477*	COG0195	NusA	Transcription antitermination factor NusA, contains S1 and KH domains [Transcription].
OG3271*	COG0568	RpoD	DNA-directed RNA polymerase, sigma subunit (sigma70/sigma32) [Transcription].
OG6679*	COG0583	LysR	DNA-binding transcriptional regulator, LysR family [Transcription].
OG7113*	COG0202	RpoA	DNA-directed RNA polymerase, alpha subunit/40 kD subunit [Transcription].
OG12597*	COG0250	NusG	Transcription antitermination factor NusG [Transcription].
OG16737	COG1386	ScpB	Chromosome segregation and condensation protein ScpB [Transcription].
OG229	COG0178	UvrA	Excinuclease UvrABC ATPase subunit [Replication, recombination and repair].
OG390	COG0188	GyrA	DNA gyrase/topoisomerase IV, subunit A [Replication, recombination and repair].
OG443	COG1200	RecG	RecG-like helicase [Replication, recombination and repair].
OG495	COG0210	UvrD	Superfamily I DNA or RNA helicase [Replication, recombination and repair].
OG929	COG0556	UvrB	Excinuclease UvrABC helicase subunit UvrB [Replication, recombination and repair].
OG944	COG0322	UvrC	Excinuclease UvrABC, nuclease subunit [Replication, recombination and repair].
OG4412	COG0592	DnaN	DNA polymerase III sliding clamp (beta) subunit, PCNA homolog [Replication, recombination and repair].
OG4823	COG2255	RuvB	Holliday junction resolvasome RuvABC, ATP-dependent DNA helicase subunit [Replication, recombination and repair].
OG4880	COG1195	RecF	Recombinational DNA repair ATPase RecF [Replication, recombination and repair].
OG7676	COG0266	Nei	Formamidopyrimidine-DNA glycosylase [Replication, recombination and repair].
OG9041	COG0496	SurE	Broad specificity polyphosphatase and 5'/3'-nucleotidase SurE [Replication, recombination and repair].
OG14257*	COG0353	RecR	Recombinational DNA repair protein RecR [Replication, recombination and repair].
OG22108*	COG0629	Ssb	Single-stranded DNA-binding protein [Replication, recombination and repair].
OG5326	COG0750	RseP	Membrane-associated protease RseP, regulator of RpoE activity [Posttranslational modification, protein turnover, chaperones, Transcription].
OG593	COG0317	SpoT	(p)ppGpp synthase/hydrolase, HD superfamily [Signal transduction mechanisms, Transcription].
OG9248*	COG0745	OmpR	DNA-binding response regulator, OmpR family, contains REC and winged-helix (wHTH) domain [Signal transduction mechanisms, Transcription].
OG10630*	COG0745	OmpR	DNA-binding response regulator, OmpR family, contains REC and winged-helix (wHTH) domain [Signal transduction mechanisms, Transcription].
OG10989*	COG2197	CitB	DNA-binding response regulator, NarL/FixJ family, contains REC and HTH domains [Signal transduction mechanisms, Transcription].
OG1699*	COG1008	NuoM	NADH:ubiquinone oxidoreductase subunit 4 (chain M) [Energy production and conversion].
OG2323*	COG0056	AtpA	FoF1-type ATP synthase, alpha subunit [Energy production and conversion].
OG2575	COG1249	Lpd	Pyruvate/2-oxoglutarate dehydrogenase complex, dihydrolipoamide dehydrogenase (E3) component or related enzyme [Energy production and conversion].
OG2577*	COG0055	AtpD	FoF1-type ATP synthase, beta subunit [Energy production and conversion].
OG2710*	COG0644	FixC	Dehydrogenase (flavoprotein) [Energy production and conversion].
OG4536*	COG1005	NuoH	NADH:ubiquinone oxidoreductase subunit 1 (chain H) [Energy production and conversion].
OG4848*	COG1071	AcoA	TPP-dependent pyruvate or acetoin dehydrogenase subunit alpha [Energy production and conversion].
OG6952	COG0224	AtpG	FoF1-type ATP synthase, gamma subunit [Energy production and conversion].
OG10155*	COG0377	NuoB	NADH:ubiquinone oxidoreductase 20 kD subunit (chhain B) or related Fe-S oxidoreductase [Energy production and conversion].
OG14604*	COG0839	NuoJ	NADH:ubiquinone oxidoreductase subunit 6 (chain J) [Energy production and conversion].
OG15982*	COG0723	QcrA	Rieske Fe-S protein [Energy production and conversion].
OG16389*	COG0712	AtpH	FoF1-type ATP synthase, delta subunit [Energy production and conversion].
OG21795*	COG0355	AtpC	FoF1-type ATP synthase, epsilon subunit [Energy production and conversion].
OG23609	COG0838	NuoA	NADH:ubiquinone oxidoreductase subunit 3 (chain A) [Energy production and conversion].
OG25542*	COG0713	NuoK	NADH:ubiquinone oxidoreductase subunit 11 or 4L (chain K) [Energy production and conversion].
OG1837	COG0119	LeuA	Isopropylmalate/homocitrate/citramalate synthases [Amino acid transport and metabolism].
OG2870	COG0019	LysA	Diaminopimelate decarboxylase [Amino acid transport and metabolism].
OG3131	COG0141	HisD	Histidinol dehydrogenase [Amino acid transport and metabolism].
OG3348	COG0460	ThrA	Homoserine dehydrogenase [Amino acid transport and metabolism].
OG3426	COG0112	GlyA	Glycine/serine hydroxymethyltransferase [Amino acid transport and metabolism].
OG4135	COG4992	ArgD	Acetylornithine/succinyldiaminopimelate/putrescine aminotransferase [Amino acid transport and metabolism].
OG4171	COG0436	AspB	Aspartate/methionine/tyrosine aminotransferase [Amino acid transport and metabolism].
OG4947	COG0337	AroB	3-dehydroquinate synthetase [Amino acid transport and metabolism].
OG5079	COG0263	ProB	Glutamate 5-kinase [Amino acid transport and metabolism].
OG5810	COG0136	Asd	Aspartate-semialdehyde dehydrogenase [Amino acid transport and metabolism].
OG6960	COG0083	ThrB	Homoserine kinase [Amino acid transport and metabolism].
OG7112	COG0287	TyrA	Prephenate dehydrogenase [Amino acid transport and metabolism].
OG7727	COG0685	MetF	5,10-methylenetetrahydrofolate reductase [Amino acid transport and metabolism].
OG10138	COG0106	HisA	Phosphoribosylformimino-5-aminoimidazole carboxamide ribonucleotide (ProFAR) isomerase [Amino acid transport and metabolism].
OG11912	COG0135	TrpF	Phosphoribosylanthranilate isomerase [Amino acid transport and metabolism].
OG12701	COG0040	HisG	ATP phosphoribosyltransferase [Amino acid transport and metabolism].
OG17107*	COG0440	IlvH	Acetolactate synthase, small subunit [Amino acid transport and metabolism].
OG22052*	COG0509	GcvH	Glycine cleavage system H protein (lipoate-binding) [Amino acid transport and metabolism].
OG2375	COG0034	PurF	Glutamine phosphoribosylpyrophosphate amidotransferase [Nucleotide transport and metabolism].
OG3322*	COG0015	PurB	Adenylosuccinate lyase [Nucleotide transport and metabolism].
OG7387*	COG0061	NadF	NAD kinase [Nucleotide transport and metabolism].
OG11355	COG0528	PyrH	Uridylate kinase [Nucleotide transport and metabolism].
OG15375	COG0194	Gmk	Guanylate kinase [Nucleotide transport and metabolism].
OG15941	COG0563	Adk	Adenylate kinase or related kinase [Nucleotide transport and metabolism].
OG1418*	COG0469	PykF	Pyruvate kinase [Carbohydrate transport and metabolism].
OG1775	COG0696	GpmI	Phosphoglycerate mutase (BPG-independent, AlkP superfamily) [Carbohydrate transport and metabolism].
OG1885	COG1543	COG1543	Predicted glycosyl hydrolase, contains GH57 and DUF1957 domains [Carbohydrate transport and metabolism].
OG2085*	COG0364	Zwf	Glucose-6-phosphate 1-dehydrogenase [Carbohydrate transport and metabolism].
OG2129*	COG0297	GlgA	Glycogen synthase [Carbohydrate transport and metabolism].
OG2430	COG1109	ManB	Phosphomannomutase [Carbohydrate transport and metabolism].
OG3103	COG0406	PhoE	Broad specificity phosphatase PhoE [Carbohydrate transport and metabolism].
OG3332*	COG0448	GlgC	ADP-glucose pyrophosphorylase [Carbohydrate transport and metabolism].
OG4126*	COG0126	Pgk	3-phosphoglycerate kinase [Carbohydrate transport and metabolism].
OG5742*	COG0057	GapA	Glyceraldehyde-3-phosphate dehydrogenase/erythrose-4-phosphate dehydrogenase [Carbohydrate transport and metabolism].
OG6085*	COG1494	GlpX	Fructose-1,6-bisphosphatase/sedoheptulose 1,7-bisphosphatase or related protein [Carbohydrate transport and metabolism].
OG7852	COG0483	SuhB	Archaeal fructose-1,6-bisphosphatase or related enzyme of inositol monophosphatase family [Carbohydrate transport and metabolism].
OG10721	COG0149	TpiA	Triosephosphate isomerase [Carbohydrate transport and metabolism].
OG11531	COG0120	RpiA	Ribose 5-phosphate isomerase [Carbohydrate transport and metabolism].
OG64	COG1429	CobN	Cobalamin biosynthesis protein CobN, Mg-chelatase [Coenzyme transport and metabolism].
OG1524*	COG0108	RibB	3,4-dihydroxy-2-butanone 4-phosphate synthase [Coenzyme transport and metabolism].
OG1634	COG0171	NadE	NH3-dependent NAD+ synthetase [Coenzyme transport and metabolism].
OG2655*	COG0422	ThiC	Thiamine biosynthesis protein ThiC [Coenzyme transport and metabolism].
OG3338*	COG0001	HemL	Glutamate-1-semialdehyde aminotransferase [Coenzyme transport and metabolism].
OG3341	COG0373	HemA	Glutamyl-tRNA reductase [Coenzyme transport and metabolism].
OG5256*	COG1239	ChlI	Mg-chelatase subunit ChlI [Coenzyme transport and metabolism].
OG5429	COG0352	ThiE	Thiamine monophosphate synthase [Coenzyme transport and metabolism].
OG5535	COG0379	NadA	Quinolinate synthase [Coenzyme transport and metabolism].
OG6659*	COG0142	IspA	Geranylgeranyl pyrophosphate synthase [Coenzyme transport and metabolism].
OG6798*	COG0382	UbiA	4-hydroxybenzoate polyprenyltransferase [Coenzyme transport and metabolism].
OG6889*	COG0181	HemC	Porphobilinogen deaminase [Coenzyme transport and metabolism].
OG6983	COG0320	LipA	Lipoate synthase [Coenzyme transport and metabolism].
OG7197	COG0196	RibF	FAD synthase [Coenzyme transport and metabolism].
OG7620	COG0190	FolD	5,10-methylene-tetrahydrofolate dehydrogenase/Methenyl tetrahydrofolate cyclohydrolase [Coenzyme transport and metabolism].
OG9475	COG0007	CysG	Uroporphyrinogen-III methylase (siroheme synthase) [Coenzyme transport and metabolism].
OG16135	COG0054	RibE	6,7-dimethyl-8-ribityllumazine synthase (Riboflavin synthase beta chain) [Coenzyme transport and metabolism].
OG17443*	COG0669	CoaD	Phosphopantetheine adenylyltransferase [Coenzyme transport and metabolism].
OG5924	COG0332	FabH	3-oxoacyl-[acyl-carrier-protein] synthase III [Lipid transport and metabolism].
OG6668	COG1562	ERG9	Phytoene/squalene synthetase [Lipid transport and metabolism].
OG7447	COG0331	FabD	Malonyl CoA-acyl carrier protein transacylase [Lipid transport and metabolism].
OG7697	COG0575	CdsA	CDP-diglyceride synthetase [Lipid transport and metabolism].
OG7764*	COG0777	AccD	Acetyl-CoA carboxylase beta subunit [Lipid transport and metabolism].
OG11745	COG0204	PlsC	1-acyl-sn-glycerol-3-phosphate acyltransferase [Lipid transport and metabolism].
OG18593*	COG0764	FabA	3-hydroxymyristoyl/3-hydroxydecanoyl-(acyl carrier protein) dehydratase [Lipid transport and metabolism].
OG7445	COG0329	DapA	Dihydrodipicolinate synthase/N-acetylneuraminate lyase [Amino acid transport and metabolism, Cell wall/membrane/envelope biogenesis].
OG1656	COG0661	AarF	Predicted unusual protein kinase regulating ubiquinone biosynthesis, AarF/ABC1/UbiB family [Coenzyme transport and metabolism, Signal transduction mechanisms].
OG6915	COG0189	RimK	Glutathione synthase/RimK-type ligase, ATP-grasp superfamily [Coenzyme transport and metabolism, Translation, ribosomal structure and biogenesis].
OG5150	COG0473	LeuB	Isocitrate/isopropylmalate dehydrogenase [Energy production and conversion, Amino acid transport and metabolism].
OG148	COG0458	CarB	Carbamoylphosphate synthase large subunit [Amino acid transport and metabolism, Nucleotide transport and metabolism].
OG6196	COG0059	IlvC	Ketol-acid reductoisomerase [Amino acid transport and metabolism, Coenzyme transport and metabolism].
OG2943	COG0391	CofD	Archaeal 2-phospho-L-lactate transferase/Bacterial gluconeogenesis factor, CofD/UPF0052 family [Coenzyme transport and metabolism, Carbohydrate transport and metabolism].
OG1087	COG1154	Dxs	Deoxyxylulose-5-phosphate synthase [Coenzyme transport and metabolism, Lipid transport and metabolism].
OG24322	COG0537	Hit	Diadenosine tetraphosphate (Ap4A) hydrolase or other HIT family hydrolase [Nucleotide transport and metabolism, Carbohydrate transport and metabolism, General function prediction only].
OG10633	COG1028	FabG	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family [Lipid transport and metabolism, Secondary metabolites biosynthesis, transport and catabolism, General function prediction only].
OG2042	COG0111	SerA	Phosphoglycerate dehydrogenase or related dehydrogenase [Coenzyme transport and metabolism, General function prediction only].
OG11575	COG1122	EcfA2	Energy-coupling factor transporter ATP-binding protein EcfA2 [Inorganic ion transport and metabolism, General function prediction only].
OG12926	COG1122	EcfA2	Energy-coupling factor transporter ATP-binding protein EcfA2 [Inorganic ion transport and metabolism, General function prediction only].
OG2000	COG1100	Gem1	GTPase SAR1 family domain [General function prediction only].
OG2927	COG1160	Der	Predicted GTPases [General function prediction only].
OG7813	COG0457	TPR	Tetratricopeptide (TPR) repeat [General function prediction only].
OG8800*	COG0388	YafV	Predicted amidohydrolase [General function prediction only].
OG9628	COG0491	GloB	Glyoxylase or a related metal-dependent hydrolase, beta-lactamase superfamily II [General function prediction only].
OG26026*	COG0762	Ycf19	Uncharacterized conserved protein YggT, Ycf19 family [Function unknown].

Table S5 Classification of amino acids by two independent schemes based on physiochemical properties of the amino acids.

Classification by charge (Hughes et al., 1990)
Positive R, H, K
Negative D, E
Neutral A, N, C, Q, G, I, L, M, F, P, S, T, W, Y, V
Classification by volume and polarity (Miyata et al., 1979)
Special C
Neutral and small A, G, P, S, T
Polar and relative small N, Q, D, E
Polar and relative large R, H, K
Nonpolar and relatively small I, L, M, V
Nonpolar and relatively large F, W, Y

Hughes, A.L., Ota, T., and Nei, M. (1990) Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Molecular biology and evolution 7: 515-524.

Miyata, T., Miyazawa, S., and Yasunaga, T. (1979) Two types of amino acid substitutions in protein evolution. Journal of Molecular Evolution 12: 219-236.