Supporting Information Appendix

Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing

Patrick M. Shih[1,2], Dongying Wu[1,3], Amel Latifi[4], Seth D. Axen[1], David P. Fewer[5], Emmanuel Talla[4], Alexandra Calteau[6], Fei Cai[1], Nicole Tandeau de Marsac[4,7], Rosmarie Rippka[7], Michael Herdman[7], Kaarina Sivonen[5], Therese Coursin[8], Thierry Laurent[8], Lynne Goodwin[9], Matt Nolan[1], Karen W. Davenport[9], Cliff S. Han[9], Edward M. Rubin[1], Jonathan A. Eisen[1,3], Tanja Woyke[1], Muriel Gugger[8*], Cheryl A. Kerfeld[1,2*]

[1]DOE Joint Genome Institute, Walnut Creek, California, 94598, USA.
[2]Department of Plant and Microbial Biology, University of California, Berkeley, California, 94720, USA.
[3]University of California, Davis, Davis, California 95616, USA.
[4]Aix-Marseille University, LCB, CNRS UMR 7283, 13402 Marseille, France.
[5]Food and Environmental Sciences, Division of Microbiology, FIN-00014, University of Helsinki, Finland.
[6]CEA, DSV, IG, Genoscope & CNRS-UMR8030, Laboratoire d'Analyse Bioinformatiques en Génomique et Métabolisme (LABGeM), 91057 Evry, France.
[7]Unité des Cyanobactéries, Institut Pasteur, CNRS URA 2172, 75724 Paris Cedex 15, France.
[8] Collection des Cyanobactéries, Institut Pasteur, 75724 Paris Cedex 15, France.
[9] Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA.

*Corresponding Authors
To whom correspondence should be addressed, email: ckerfeld@lbl.gov or mgugger@pasteur.fr

Table of Contents

Supplementary Materials and Methods:

The 54 strains used for genome sequencing in this study are available at Pasteur Culture collection of Cyanobacteria (http://www.pasteur.fr/pcc_cyanobacteria). The 54 sequenced genomes in this study were compared to 72 publicly available cyanobacterial genomes (Table S1).

A sequence similarity matrix was calculated for alignments of 1,813 16S small subunit rRNA sequences of cyanobacterial isolates from the greengenes database, excluding sequences from environmental samples (December 2008). The cyanobacterial isolates were grouped into 104 clusters by MCL clustering performed on the sequence similarity matrix at similarity cutoff of 95% and inflation value of 2. Type strains, PCC identification numbers and the status of previous sequencing efforts were highlighted for all the isolates in the 104 clusters. This analysis, interest of the strains to the research community and their availability at the Pasteur Culture Collection, was used as guide to choose the strains for genome sequencing. For strains chosen, 1.25 L of liquid cultures in late exponential to linear growth phase were centrifuged at 12,000*g* for 10min at 20°C. After washing twice with sterile distilled water or sterile saline solution (1% NaCl) for marine strains, the pellets were immediately frozen in liquid N2 prior to being lyophilized. DNA of the lyophilized pellets was extracted using Genomic DNA isolation - NucleoBond ® AX (Macherey-Nagel, Hoerdt, France) according manufacturer's instructions for bacterial DNA using the columns Nucleobond AX-G 500.

Genome sequencing and assembly

The 54 CyanoGEBA draft genomes were generated at the DOE Joint Genome Institute (JGI) using either a combination of Illumina (1) and 454 technologies (2) or the Illumina technology (Table S10). The 454 Titanium standard data and the 454 paired end data were assembled using Newbler, versions 2.3 to 2.6, and the resulting consensus sequences were computationally shredded into 2 Kbp overlapping fake reads (shreds). Illumina sequencing data was assembled with Velvet, versions 0.7.55 to 1.105 (3), and the consensus sequence computationally shredded into 1.5 Kbp overlapping fake reads (shreds). The 454 Newbler consensus shreds, the Illumina Velvet consensus shreds and the read pairs in the 454 paired end library were then integrated using parallel Phrap, version SPS - 4.24 (High Performance Software, LLC). The software Consed (4),(5) (6) was used in the following finishing process. Illumina data was used to correct potential base errors and increase consensus quality using the software Polisher developed at JGI. Possible mis-assemblies were corrected using gapResolution, Dupfinisher (7), or sequencing cloned bridging PCR fragments with subcloning. Gaps between contigs were closed by editing in Consed, by PCR and by Bubble PCR primer walks. All general aspects of library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov/. At Los Alamos National Laboratory (LANL), 25 of these genomes underwent manual finishing efforts, while 20 others underwent autofinishing. Gap closure in autofinishing is fully automated and thus less extensive as compared to manually finishing. The 9 remaining CyanoGEBA genomes were not subjected to finishing efforts. For PCC 9605 and PCC 10914, all raw Illumina sequence data was passed through DUK, a filtering program developed at JGI, which removes known Illumina sequencing and library preparation artifacts. Illumina sequence reads were assembled using Allpaths-LG versions 38118 (PCC 9339), 38445 (PCC 9431, PCC 10914, PCC 7702) and 39750 (PCC 9605). For PCC 73106, PCC 7509, and PCC 6406, following steps were performed for genome assembly: 1) filtered Illumina reads were assembled using Velvet (3), 2) 1-3 Kbp simulated paired end reads were created from Velvet contigs using wgsim (https://github.com/lh3/wgsim), 3) Illumina reads were assembled with simulated read pairs using Allpaths-LG (versions 37843 and 38118) (8).

Genome annotation

Genes were identified using Prodigal (9), followed by a round of manual curation using GenePRIMP (10) for finished genomes and draft genomes in fewer than 10 scaffolds. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) nonredundant database, UniProt, TIGRFam, Pfam, KEGG, COG, and InterPro databases. The tRNAScanSE tool (11) was used to find tRNA genes, whereas ribosomal RNA genes were found by searches against models of the ribosomal RNA genes built from SILVA (12). Other non-coding RNAs such as the RNA components of the protein secretion complex and the RNase P were identified by searching the genome for the corresponding Rfam profiles using INFERNAL (http://infernal.janelia.org). Additional gene prediction analysis and manual functional annotation was performed within the Integrated Microbial Genomes (IMG) platform (http://img.jgi.doe.gov) developed by the Joint Genome Institute, Walnut Creek, CA, USA (13).

Species tree

The Species tree was generated by a concatenation of thirty-one conserved proteins as described by Wu et al. (14). Homologs of each ribosomal protein were identified using reciprocal BLAST of the 49 publicly available cyanobacterial genomes in IMG at the end of 2009. These gene families were aligned using MAFFT, using the maxiterative function (15). The subsequent alignment was used to create Hidden Markov Models (HMMs) for the respective ribosomal protein using HMMer v.2.0 (16). Total protein coding sequences for each cyanobacterial genome, and of four outgroups (*Chloroflexus auranticus* J-10, *Rhodobacter sphaeroides* 2.4.1, *Heliobacterium modesticaldum* Ice1, and *Chlorobium tepidum* TLS) were retrieved using IMG (13). Using HMMer, the hmmsearch function was used to identify orthologs and align them using the hmmalign function. The resulting thirty-one alignments were then concatenated. The default setting to omit gappy columns was used with the software Belvu (17). A phylogenetic tree was generated with the alignment using PhyML (18). The LG amino acid substitution model was chosen using ProtTest with gamma-distributed rate variation (four categories) and estimation of a proportion of invariable sites (19).

Tree Imbalance study

Two trees, one with all cyanobacterial genomes (126 species) and one with only the 72 publicly available were generated. The alignments and the phylogenetic trees were generated using the same methods described to construct the Species Tree. *Gloeobacter* *violaceus* PCC 7421 was set as the outgroup in both trees. The tree imbalance of both trees was measured using Colless' Imbalance in the software Mesquite (20, 21). The tree depth was set to 10 and 1000 simulations of both uniform and equiprobable speciations were conducted.

16S rRNA phylogeny

A phylogeny using 16S rRNA sequences retrieved from IMG for all cyanobacteria of this study was generated to compare to the Species tree. Due to incomplete or partial sequences, *Arthrospira* sp. PCC 8005, *Synechococcus* sp. CB 0101, *Synechococcus* sp. CB 0205, and *Crocosphaera watsonii* WH 0003 were omitted from this phylogeny. Sequences were aligned in MAFFT. A maximum likelihood tree was generated using PhyML, using the GTR model with gamma-distributed rate variation (four categories) and an estimation of proportion of invariable sites.

Identification of novel proteins

All 292,935 proteins from the CyanoGEBA genomes were searched against the entire amino acid non-redundant (nr) database downloaded from NCBI, updated April 2^{nd,} 2012, using BLASTP set at an e-value cutoff of 1e-2. The 21,107 proteins with no hits were considered 'novel' as they have no homology to the nr database.

Morphological transitions analysis

Protein families generated from MCL analysis was used. The specific nodes tested for morphological transitions are indicated in Fig. 1. A set of genes involved in the morphological transition were defined by comparison of presence in one genome or a set of genome belonging to a subsection and their absence in another genome or a set of genomes as reported in Table S5. Moreover, a BLASTP search of the 32 proteins from *Prochlorothrix hollandica* PCC 9006 from Event 2 against the 674 proteins from Event 3 was done, yielding 29 out of the 32 hits. We generated a null hypothesis to verify the enrichment in 29 out of the 32 homologous proteins by randomly sampling the *Prochlorothrix hollandica* PCC 9006 genome against the 674 proteins from Event 3 with BLASTP, 10,000 times, which showed that the value (29 out of 32 proteins) was significant (p-value = 0).

Heterocyst, hormogonium, and akinete related gene distribution analysis

Seed proteins (29 and 20 are involved in cell division and cell differentiation, respectively) were downloaded from the cyanobase (http://genome.kazusa.or.jp/cyanobase) and used for BLAST comparison searches. Putative orthology relationships between a seed protein and other cyanobacterial proteins were defined by an alignment threshold of at least 30% sequence identity with an e-value lower than 1e-10.

COG functional categories

Clusters of Orthologous Groups (COG) functional category data was downloaded by Morphological Subsection from the IMG database.

Plastidome tree

The plastidome tree was generated by a concatenation of twenty-five conserved plastid proteins using the same method to generate the species tree. Proteins from fully sequenced plastid genomes were downloaded from the High-quality Automated and Manual Annotation of microbial Proteins (HAMAP) database (22). Plastids downloaded from HAMAP were: *Cyanophora paradoxa*, *Chaetosphaeridium globosum*, *Anthoceros formosae*, *Cycas taitungensis*, *Arabidopsis thaliana*, *Amborella trichopoda*, *Selaginella uncinata*, *Zygnema circumcarinatum*, *Staurastrum punctulatum*, *Chara vulgaris*, *Nephroselmis olivacea*, *Ostreococcus tauri*, *Bigelowiella natans*, *Chlorella vulgaris*, *Pseudendoclonium akinetum*, *Oltmannsiellopsis viridis*, *Scenedesmus obliquus*, *Chlamydomonas reinhardtii*, *Stigeoclonium helveticum*, *Oedogonium cardiacum*, *Euglena gracilis*, *Mesostigma viride*, *Chlorokybus atmophyticus*, *Cyanidioschyzon merolae*, *Cyanidium caldarium*, *Porphyra yezoensis*, *Porphyra purpurea, Gracilaria tenuistipitata* var. liui, *Rhodomonas salina*, *Guillardia theta*, *Emiliania huxleyi*, *Phaeodactylum tricornutum*, *Odontella sinensis*, *Thalassiosira pseudonana*, *Vaucheria litorea*, *Heterosigma akashiwo* CCMP452, *Heterosigma akashiwo* NIES293, and the chromatophore of *Paulinella chromatophora*. A phylogenetic tree was generated with the alignment using PhyML 3.0. The LG amino acid substitution model was chosen by ProtTest and with gamma-distributed rate variation (four categories) and estimation of a proportion of invariable sites. The tree was rooted to *Gloeobacter violaceus* PCC 7421.

Prediction of Endosymbiotic Gene Transfer.

Proteins from the genomes used in this study were divided into four groups: 1) Nuclear genomes from plastid-containing eukaryotes (Table S8), 2) Bacteria not from the phylum *Cyanobacteria* (*Agrobacterium tumefaciens* C58-Cereon, *Aquifex aeolicus* VF5, *Bacillus subtilis subtilis* 168, *Caulobacter crescentus* CB15, *Chlamydia trachomatis* E/150, *Chlorobium limicola* DSM 245, *Chloroflexus aurantiacus* J-10-fl, *Heliobacterium modesticaldum* Ice1, Candidatus *Kuenenia stuttgartiensis*, *Rickettsia peacockii* Rustic, *Thermotoga maritima* MSB8), 3) Archaea (*Archaeoglobus fulgidus* VC-16, DSM 4304, *Cenarchaeum symbiosum* A, *Methanocaldococcus jannaschii* DSM 2661, *Nanoarchaeum equitans* Kin4-M, *Sulfolobus acidocaldarius* DSM 639), 4) Eukaryotes presumably not containing plastids derived from endosymbiosis (*Caenorhabditis elegans* Bristol N2, *Cryptococcus neoformans* var. neoformans JEC 21, *Drosophila melanogaster*, *Monosiga brevicollis* MX1, *Saccharomyces cerevisiae* S288C). The nuclear proteins from Group 1 were used as queries to BLASTP against two databases: 1) all proteins from Groups 2-4 and all cyanobacterial proteins in this study (CyanoGEBA and publicly-available genomes), and 2) all proteins from Groups 2-4 and cyanobacterial proteins from only publicly-available genomes. Those with top-hits to cyanobacterial proteins were considered genes of cyanobacterial descent, and the total counts for each of the nuclear genomes from Group 1 are described in Table S8 and Dataset S3. COGs for all proteins were assigned using the same methods as in the IMG pipeline (13).

Chlorophyll Binding Protein (CBP) studies

Phylogenetic analysis

CBP homologs were collected by performing a BLASTP search on all cyanobacteria in the IMG database using the inner chlorophyll-binding antenna protein CP43 of PSII from *Thermosynechococcus elongatus* BP-1 as the query, setting and e-value threshold of 1e-

10. All homologs were aligned using MAFFT. The alignment was used to build a maximum likelihood phylogenetic tree in PhyML, under the LG model with gamma-distributed rate variation (four categories) and an estimation of a proportion of invariable sites, after choosing the best-suited model in ProtTest.

Alignment and analysis of chlorophyll binding amino acids

An alignment of a subset of CBP proteins was generated in order to investigate the presence of conserved amino acids that are known to ligate chlorophyll to the protein. The amino acid sequences for the N- and C- termini of PsaA and PsaB, PsbB, PsbC, and IsiA from *Thermosynechococcus elongatus* were aligned to various CBP; the sequences were aligned using MAFFT, followed by manual curation of the alignment, using only the alignments of the first six helices (Fig. S6).

Further analysis of the C-terminal PsaL-like domain of the CBPV was carried out by truncating CBPV sequences to examine specifically the ladder domain. PsaL subunits from *Synechococcus elongatus* PCC 7942, *Synechocystis* sp. PCC 6803, and *Thermosynechococcus elongatus* BP1 were aligned with the truncated CPBV sequences using MAFFT (Fig. S7).

Homology model

The CBPV homolog from *Chroococcidiopsis thermalis* PCC 7203 (Chro_2988) was submitted to the SWISS-MODEL web server (http://swissmodel.expasy.org/) for three-dimensional structural homology modeling. Two homology models were made. 1) The N-terminal domain (first six transmembrane helices) was homology modeled off of the template from the Protein Data Bank, 3ARC_C (the PsbC subunit of Photosystem II from *Thermosynechococcus vulcanus* modeling amino acid positions 6-346). The C-terminal domain (last three transmembrane helices) was modeled off of the template, 1JB0_L (the PsaL subunit of Photosystem I from *Synechococcus elongatus* modeling amino acid positions 342-504). The last five amino acids were removed from the N-terminal domain and the C-terminal domain was positioned near it using PyMol (http://www.pymol.org/). A monomeric subunit of the Photosystem I structure, 1JB0, was used to model the CBPV homolog interaction when replacing the PsaL subunit (Fig. S8).

CRISPR analysis

CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeats) loci were predicted using both CRISPRfinder (23) and CRISPR Recognition Tool (24) (CRT, which is integrated into the IMG pipeline). The presence of CRISPR/Cas systems was confirmed by examining the co-existence of CRISPR loci and the ubiquitous CRISPR-associated (*cas*) genes, namely *cas1* and *cas2*, within one genome.

Figures S1-S10

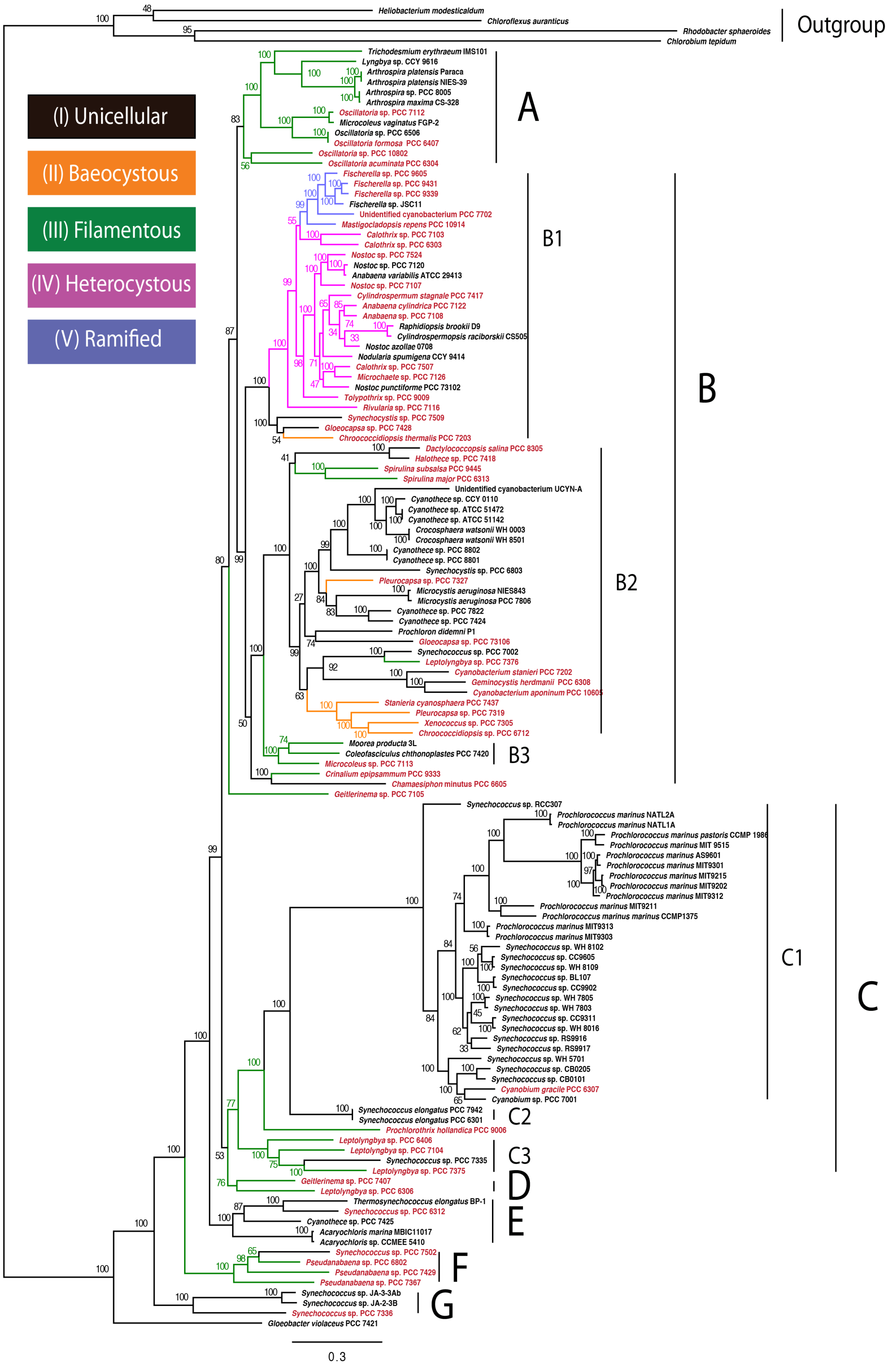

(I) Unicellular
(II) Baeocystous
(III) Filamentous
(IV) Heterocystous
(V) Ramified
Heliobacterium modesticaldum
Chloroflexus auranticus
Rhodobacter sphaeroides
Chlorobium tepidum
Outgroup
Trichodesmium erythraeum IMS101
Lyngbya sp. CCY 9616
Arthrospira platensis Paraca
Arthrospira platensis NIES-39
Arthrospira sp. PCC 8005
Arthrospira maxima CS-328
Oscillatoria sp. PCC 7112
Microcoleus vaginatus FGP-2
Oscillatoria sp. PCC 6506
Oscillatoria formosa PCC 6407
Oscillatoria sp. PCC 10802
Oscillatoria acuminata PCC 6304
A
Fischerella sp. PCC 9605
Fischerella sp. PCC 9431
Fischerella sp. PCC 9339
Fischerella sp. JSC11
Unidentified cyanobacterium PCC 7702
Mastigocladopsis repens PCC 10914
Calothrix sp. PCC 7103
Calothrix sp. PCC 6303
Nostoc sp. PCC 7524
Nostoc sp. PCC 7120
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7107
Cylindrospermum stagnale PCC 7417
Anabaena cylindrica PCC 7122
Anabaena sp. PCC 7108
Raphidiopsis brookii D9
Cylindrospermopsis raciborskii CS505
Nostoc azollae 0708
Nodularia spumigena CCY 9414
Calothrix sp. PCC 7507
Microchaete sp. PCC 7126
Nostoc punctiforme PCC 73102
Tolypothrix sp. PCC 9009
Rivularia sp. PCC 7116
Synechocystis sp. PCC 7509
Gloeocapsa sp. PCC 7428
Chroococcidiopsis thermalis PCC 7203
B1
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
Spirulina subsalsa PCC 9445
Spirulina major PCC 6313
Unidentified cyanobacterium UCYN-A
Cyanothece sp. CCY 0110
Cyanothece sp. ATCC 51472
Cyanothece sp. ATCC 51142
Crocosphaera watsonii WH 0003
Crocosphaera watsonii WH 8501
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Synechocystis sp. PCC 6803
Pleurocapsa sp. PCC 7327
Microcystis aeruginosa NIES843
Microcystis aeruginosa PCC 7806
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Prochloron didemni P1
Gloeocapsa sp. PCC 73106
Synechococcus sp. PCC 7002
Leptolyngbya sp. PCC 7376
Cyanobacterium stanieri PCC 7202
Geminocystis herdmanii PCC 6308
Cyanobacterium aponinum PCC 10605
Stanieria cyanosphaera PCC 7437
Pleurocapsa sp. PCC 7319
Xenococcus sp. PCC 7305
Chroococcidiopsis sp. PCC 6712
B2
Moorea producta 3L
Coleofasciculus chthonoplastes PCC 7420
B3
Microcoleus sp. PCC 7113
Crinalium epipsammum PCC 9333
Chamaesiphon minutus PCC 6605
Geitlerinema sp. PCC 7105
B
Synechococcus sp. RCC307
Prochlorococcus marinus NATL2A
Prochlorococcus marinus NATL1A
Prochlorococcus marinus pastoris CCMP 1986
Prochlorococcus marinus MIT 9515
Prochlorococcus marinus AS9601
Prochlorococcus marinus MIT9301
Prochlorococcus marinus MIT9215
Prochlorococcus marinus MIT9202
Prochlorococcus marinus MIT9312
Prochlorococcus marinus MIT9211
Prochlorococcus marinus marinus CCMP1375
Prochlorococcus marinus MIT9313
Prochlorococcus marinus MIT9303
Synechococcus sp. WH 8102
Synechococcus sp. CC9605
Synechococcus sp. WH 8109
Synechococcus sp. BL107
Synechococcus sp. CC9902
Synechococcus sp. WH 7805
Synechococcus sp. WH 7803
Synechococcus sp. CC9311
Synechococcus sp. WH 8016
Synechococcus sp. RS9916
Synechococcus sp. RS9917
Synechococcus sp. WH 5701
Synechococcus sp. CB0205
Synechococcus sp. CB0101
Cyanobium gracile PCC 6307
Cyanobium sp. PCC 7001
C1
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
C2
Prochlorothrix hollandica PCC 9006
Leptolyngbya sp. PCC 6406
Leptolyngbya sp. PCC 7104
Synechococcus sp. PCC 7335
Leptolyngbya sp. PCC 7375
C3
C
Geitlerinema sp. PCC 7407
Leptolyngbya sp. PCC 6306
D
Thermosynechococcus elongatus BP-1
Synechococcus sp. PCC 6312
Cyanothece sp. PCC 7425
Acaryochloris marina MBIC11017
Acaryochloris sp. CCMEE 5410
E
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 6802
Pseudanabaena sp. PCC 7429
Pseudanabaena sp. PCC 7367
F
Synechococcus sp. JA-3-3Ab
Synechococcus sp. JA-2-3B
Synechococcus sp. PCC 7336
G
Gloeobacter violaceus PCC 7421
0.3

Figure S1. Maximum likelihood tree of *Cyanobacteria* with bootstrap support

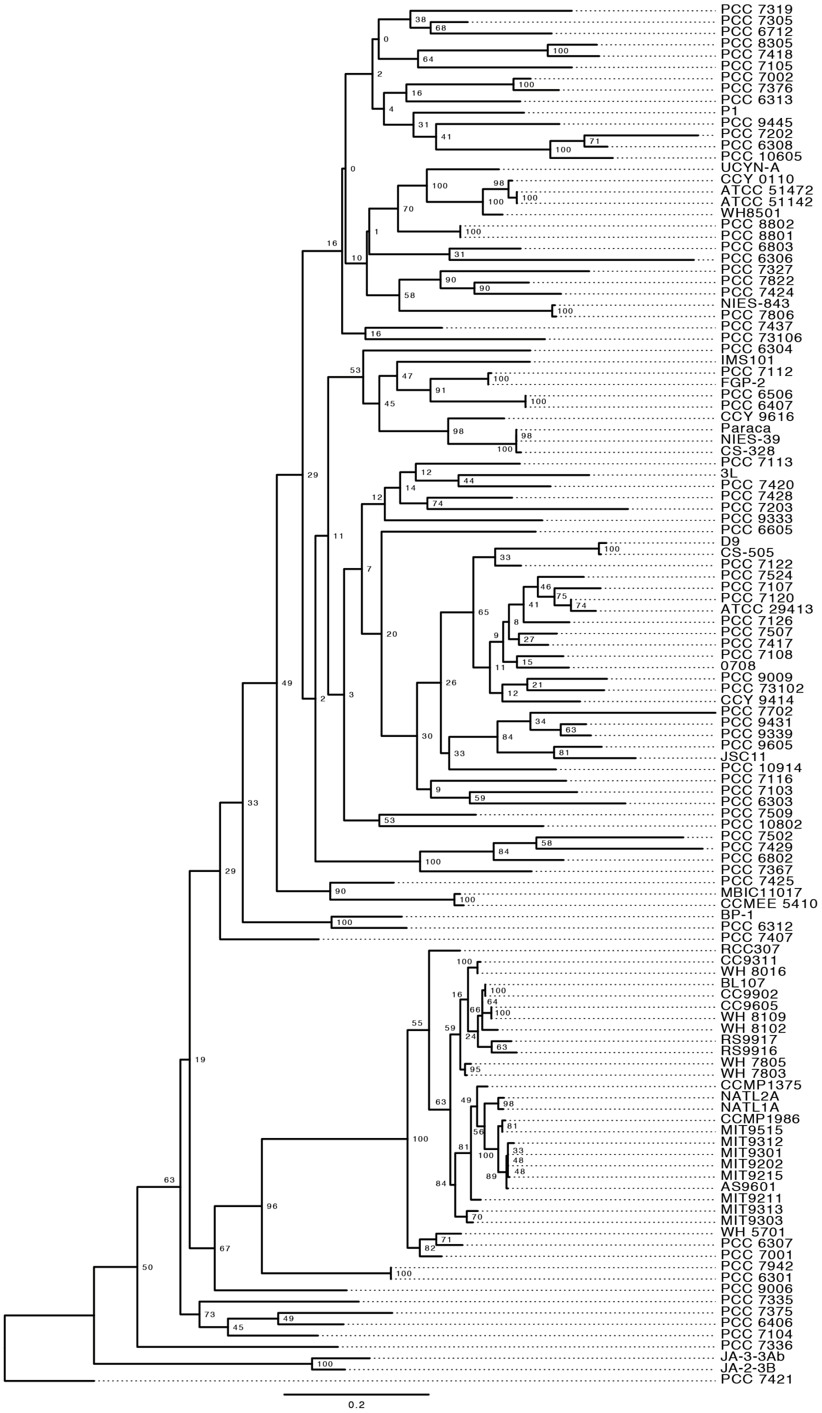


Figure S2. 16S rRNA gene phylogeny of *Cyanobacteria*. Maximum Likelihood tree based on 16S rRNA gene sequences from cyanobacteria included in this study and named accordingly to the Strain_ID in Table S1. Many of the clades defined in Fig. 1 are retrieved in 16S rRNA gene phylogeny. However, given poor bootstrap supports in the latter, there are incongruences between the topologies of the two trees.

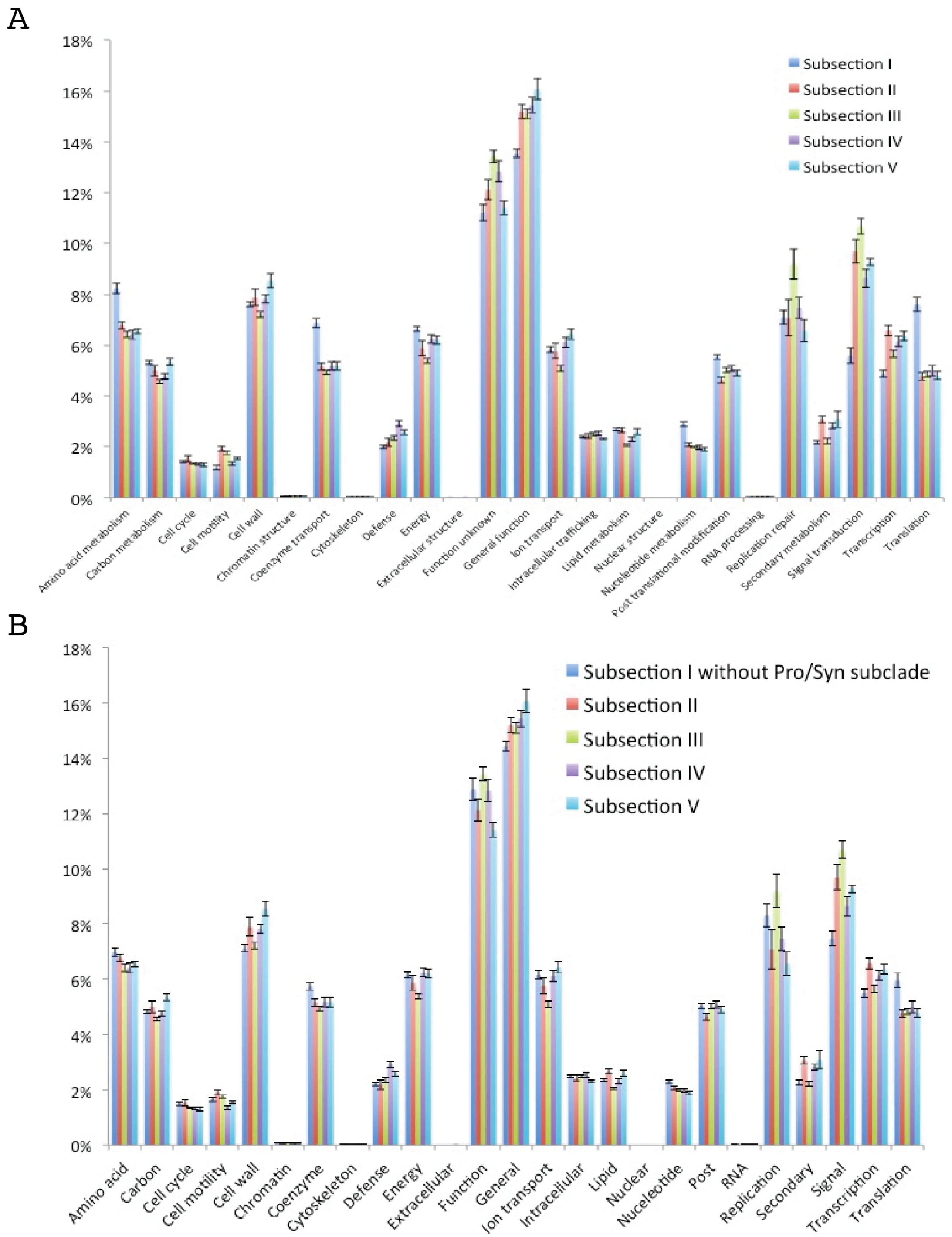


Figure S3. COG functional categories within morphological subsections. Bars represent the standard error given the sampling size of each morphological Subsection. **A**, COG analysis of all cyanobacteria included in this study. **B**, COG analysis of all cyanobacteria, excluding the *Prochlorococcus/Synechococcus* subclade in order to decrease bias within Subsection I.

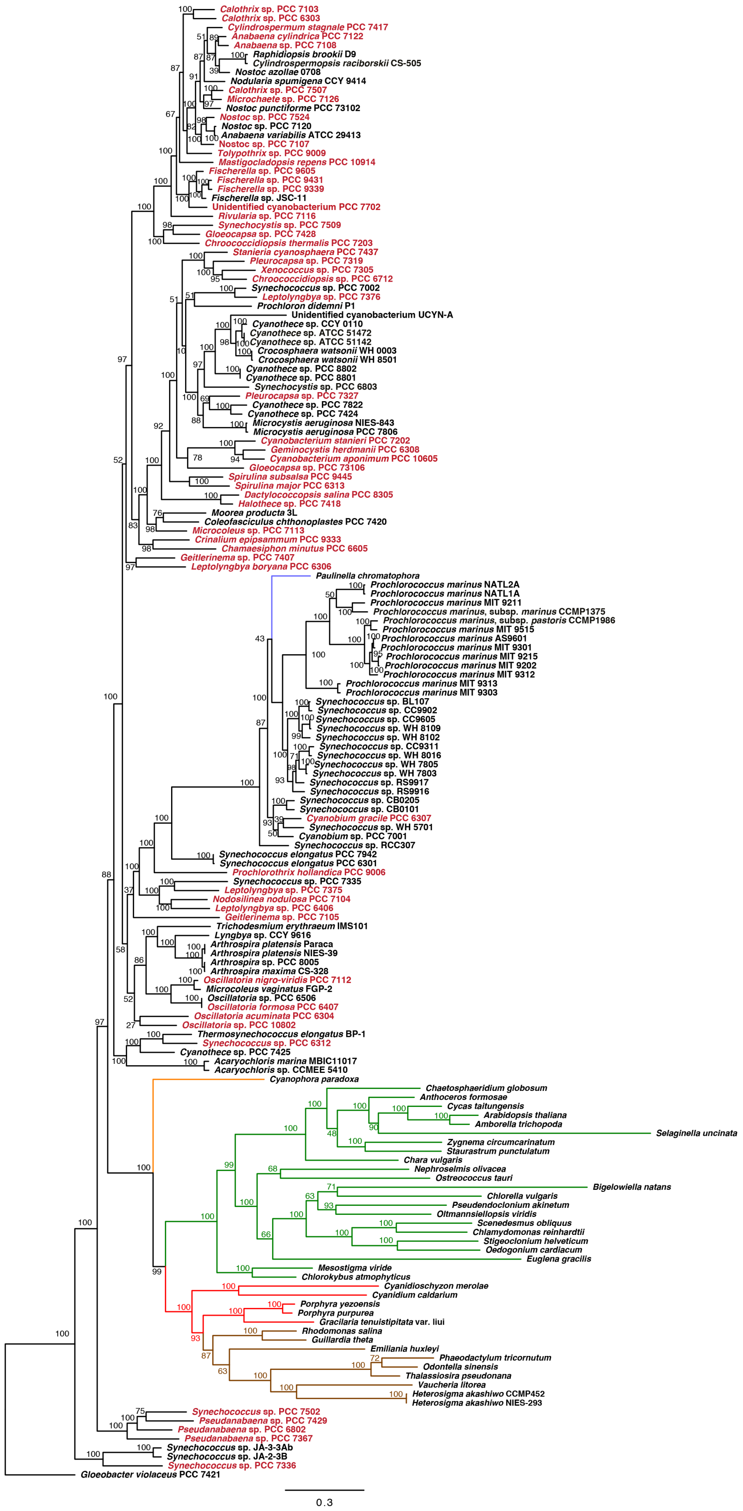

Calothrix sp. PCC 7103
Calothrix sp. PCC 6303
Cylindrospermum stagnale PCC 7417
Anabaena cylindrica PCC 7122
Anabaena sp. PCC 7108
Raphidiopsis brookii D9
Cylindrospermopsis raciborskii CS-505
Nostoc azollae 0708
Nodularia spumigena CCY 9414
Calothrix sp. PCC 7507
Microchaete sp. PCC 7126
Nostoc punctiforme PCC 73102
Nostoc sp. PCC 7524
Nostoc sp. PCC 7120
Anabaena variabilis ATCC 29413
Nostoc sp. PCC 7107
Tolypothrix sp. PCC 9009
Mastigocladopsis repens PCC 10914
Fischerella sp. PCC 9605
Fischerella sp. PCC 9431
Fischerella sp. PCC 9339
Fischerella sp. JSC-11
Unidentified cyanobacterium PCC 7702
Rivularia sp. PCC 7116
Synechocystis sp. PCC 7509
Gloeocapsa sp. PCC 7428
Chroococcidiopsis thermalis PCC 7203
Stanieria cyanosphaera PCC 7437
Pleurocapsa sp. PCC 7319
Xenococcus sp. PCC 7305
Chroococcidiopsis sp. PCC 6712
Synechococcus sp. PCC 7002
Leptolyngbya sp. PCC 7376
Prochloron didemni P1
Unidentified cyanobacterium UCYN-A
Cyanothece sp. CCY 0110
Cyanothece sp. ATCC 51472
Cyanothece sp. ATCC 51142
Crocosphaera watsonii WH 0003
Crocosphaera watsonii WH 8501
Cyanothece sp. PCC 8802
Cyanothece sp. PCC 8801
Synechocystis sp. PCC 6803
Pleurocapsa sp. PCC 7327
Cyanothece sp. PCC 7822
Cyanothece sp. PCC 7424
Microcystis aeruginosa NIES-843
Microcystis aeruginosa PCC 7806
Cyanobacterium stanieri PCC 7202
Geminocystis herdmanii PCC 6308
Cyanobacterium aponimum PCC 10605
Gloeocapsa sp. PCC 73106
Spirulina subsalsa PCC 9445
Spirulina major PCC 6313
Dactylococcopsis salina PCC 8305
Halothece sp. PCC 7418
Moorea producta 3L
Coleofasciculus chthonoplastes PCC 7420
Microcoleus sp. PCC 7113
Crinalium epipsammum PCC 9333
Chamaesiphon minutus PCC 6605
Geitlerinema sp. PCC 7407
Leptolyngbya boryana PCC 6306
Paulinella chromatophora
Prochlorococcus marinus NATL2A
Prochlorococcus marinus NATL1A
Prochlorococcus marinus MIT 9211
Prochlorococcus marinus, subsp. marinus CCMP1375
Prochlorococcus marinus, subsp. pastoris CCMP1986
Prochlorococcus marinus MIT 9515
Prochlorococcus marinus AS9601
Prochlorococcus marinus MIT 9301
Prochlorococcus marinus MIT 9215
Prochlorococcus marinus MIT 9202
Prochlorococcus marinus MIT 9312
Prochlorococcus marinus MIT 9313
Prochlorococcus marinus MIT 9303
Synechococcus sp. BL107
Synechococcus sp. CC9902
Synechococcus sp. CC9605
Synechococcus sp. WH 8109
Synechococcus sp. WH 8102
Synechococcus sp. CC9311
Synechococcus sp. WH 8016
Synechococcus sp. WH 7805
Synechococcus sp. WH 7803
Synechococcus sp. RS9917
Synechococcus sp. RS9916
Synechococcus sp. CB0205
Synechococcus sp. CB0101
Cyanobium gracile PCC 6307
Synechococcus sp. WH 5701
Cyanobium sp. PCC 7001
Synechococcus sp. RCC307
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301
Prochlorothrix hollandica PCC 9006
Synechococcus sp. PCC 7335
Leptolyngbya sp. PCC 7375
Nodosilinea nodulosa PCC 7104
Leptolyngbya sp. PCC 6406
Geitlerinema sp. PCC 7105
Trichodesmium erythraeum IMS101
Lyngbya sp. CCY 9616
Arthrospira platensis Paraca
Arthrospira platensis NIES-39
Arthrospira sp. PCC 8005
Arthrospira maxima CS-328
Oscillatoria nigro-viridis PCC 7112
Microcoleus vaginatus FGP-2
Oscillatoria sp. PCC 6506
Oscillatoria formosa PCC 6407
Oscillatoria acuminata PCC 6304
Oscillatoria sp. PCC 10802
Thermosynechococcus elongatus BP-1
Synechococcus sp. PCC 6312
Cyanothece sp. PCC 7425
Acaryochloris marina MBIC11017
Acaryochloris sp. CCMEE 5410
Cyanophora paradoxa
Chaetosphaeridium globosum
Anthoceros formosae
Cycas taitungensis
Arabidopsis thaliana
Amborella trichopoda
Selaginella uncinata
Zygnema circumcarinatum
Staurastrum punctulatum
Chara vulgaris
Nephroselmis olivacea
Ostreococcus tauri
Bigelowiella natans
Chlorella vulgaris
Pseudendoclonium akinetum
Oltmannsiellopsis viridis
Scenedesmus obliquus
Chlamydomonas reinhardtii
Stigeoclonium helveticum
Oedogonium cardiacum
Euglena gracilis
Mesostigma viride
Chlorokybus atmophyticus
Cyanidioschyzon merolae
Cyanidium caldarium
Porphyra yezoensis
Porphyra purpurea
Gracilaria tenuistipitata var. liui
Rhodomonas salina
Guillardia theta
Emiliania huxleyi
Phaeodactylum tricornutum
Odontella sinensis
Thalassiosira pseudonana
Vaucheria litorea
Heterosigma akashiwo CCMP452
Heterosigma akashiwo NIES-293
Synechococcus sp. PCC 7502
Pseudanabaena sp. PCC 7429
Pseudanabaena sp. PCC 6802
Pseudanabaena sp. PCC 7367
Synechococcus sp. JA-3-3Ab
Synechococcus sp. JA-2-3B
Synechococcus sp. PCC 7336
Gloeobacter violaceus PCC 7421
0.3

Figure S4. Maximum likelihood plastidome tree with full names and bootstrap support. Cyanobacteria are named accordingly to the Strain ID in Table S1.

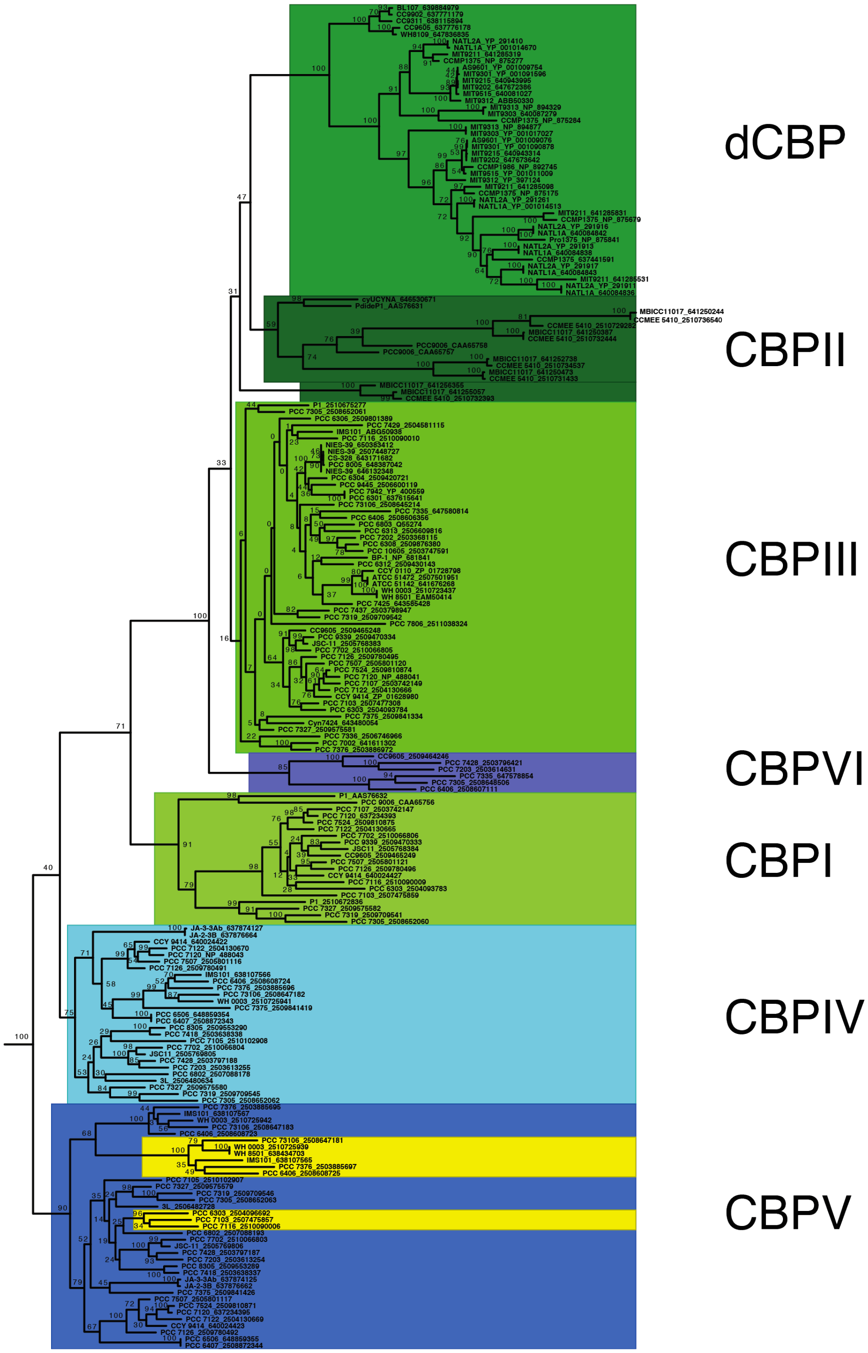
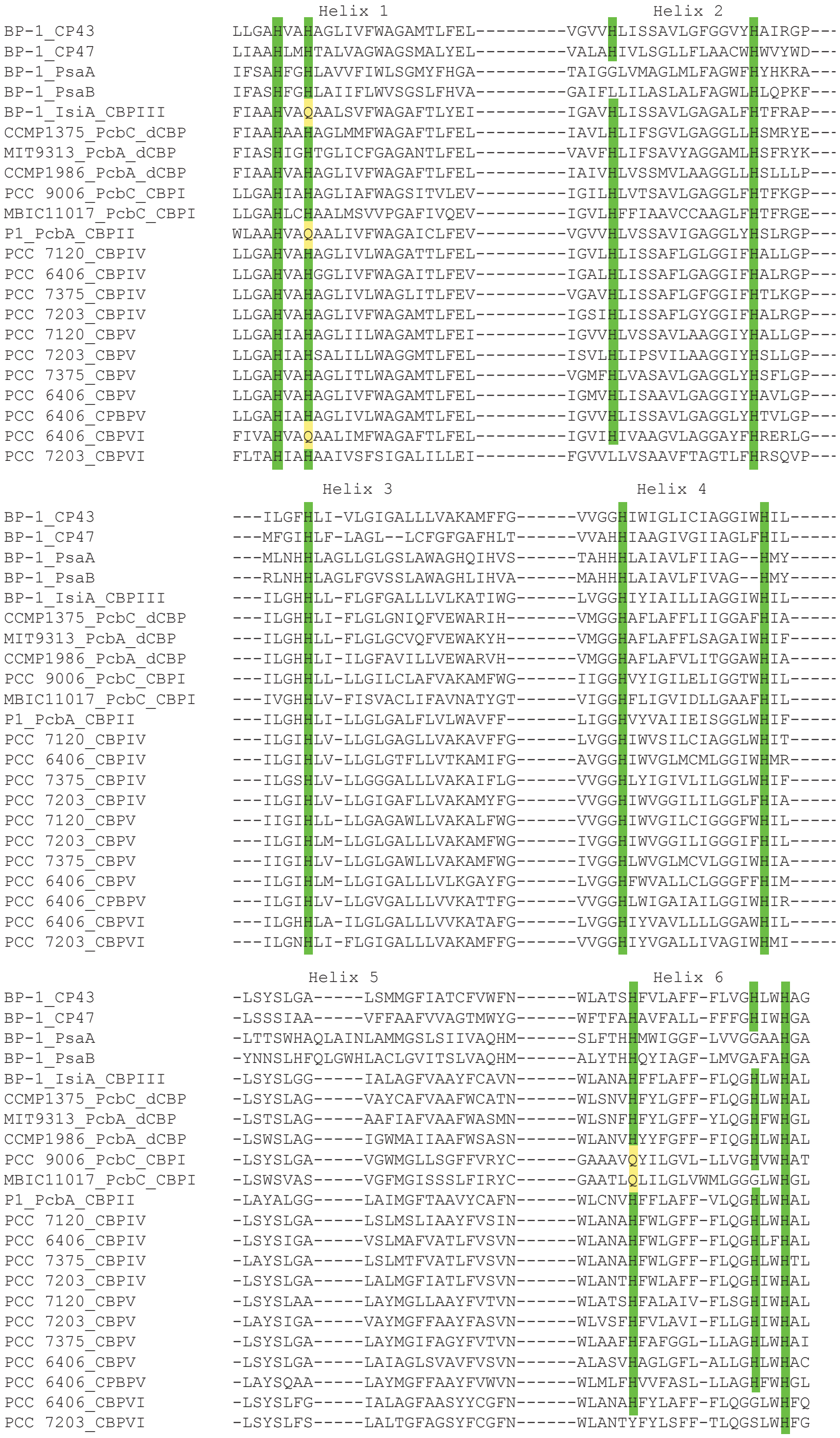


Figure S5. Maximum-likelihood CBP phylogeny reveals a diversity of previously uncharacterized clades. CP43 sequences are used as an outgroup (not shown, Newick file is available upon request), while the major CBP clades are color-coded. Shades of green represent previously characterized CBP clades (divinyl CBP = dCBP for their use of divinyl chlorophyll), whereas shades of blue represent new clades distinctly supported with the addition of CyanoGEBA genomes. Yellow subclades indicate CBPV proteins that lack the C-terminal PsaL-like domain. We find very little support for subclades CBPIII and CBPII. Taxa are named by their strain IDs abbreviation and followed by their IMG Gene Object ID or their GenBank Accessions.



```
                               Helix 1                          Helix 2
BP-1_CP43              LLGAHVAHAGLIVFWAGAMTLFEL---------VGVVHLISSAVLGFGGVYHAIRGP---
BP-1_CP47              LIAAHLMHTALVAGWAGSMALYEL---------VALAHIVLSGLLFLAACWHWVYWD---
BP-1_PsaA              IFSAHFGHLAVVFIWLSGMYFHGA---------TAIGGLVMAGLMLFAGWFHYHKRA---
BP-1_PsaB              IFASHFGHLAIIFLWVSGSLFHVA---------GAIFLLILASLALFAGWLHLQPKF---
BP-1_IsiA_CBPIII       FIAAHVAQAALSVFWAGAFTLYEI---------IGAVHLISSAVLGAGALFHTFRAP---
CCMP1375_PcbC_dCBP     FIAAHAAHAGLMMFWAGAFTLFEL---------IAVLHLIFSGVLGAGGLLHSMRYE---
MIT9313_PcbA_dCBP      FIASHIGHTGLICFGAGANTLFEL---------VAVFHLIFSAVYAGGAMLHSFRYK---
CCMP1986_PcbA_dCBP     FIAAHVAHAGLIVFWAGAFTLFEL---------IAIVHLVSSMVLAAGGLLHSLLLP---
PCC 9006_PcbC_CBPI     LLGAHIAHAGLIAFWAGSITVLEV---------IGILHLVTSAVLGAGGLFHTFKGP---
MBIC11017_PcbC_CBPI    LLGAHLCHAALMSVVPGAFIVQEV---------IGVLHFFIAAVCCAAGLFHTFRGE---
P1_PcbA_CBPII          WLAAHVAQAALIVFWAGAICLFEV---------VGVVHLVSSAVIGAGGLYHSLRGP---
PCC 7120_CBPIV         LLGAHVAHAGLIVLWAGATTLFEL---------IGVLHLISSAFLGLGGIFHALLGP---
PCC 6406_CBPIV         LLGAHVAHGGLIVFWAGAITLFEV---------IGALHLISSAFLGAGGIFHALRGP---
PCC 7375_CBPIV         LLGAHVAHAGLIVLWAGLITLFEV---------VGAVHLISSAFLGFGGIFHTLKGP---
PCC 7203_CBPIV         LLGAHVAHAGLIVFWAGAMTLFEL---------IGSIHLISSAFLGYGGIFHALRGP---
PCC 7120_CBPV          LLGAHIAHAGLIILWAGAMTLFEI---------IGVVHLVSSAVLAAGGIYHALLGP---
PCC 7203_CBPV          LLGAHIAHSALILLWAGGMTLFEL---------ISVLHLIPSVILAAGGIYHSLLGP---
PCC 7375_CBPV          LLGAHVAHAGLITLWAGAMTLFEL---------VGMFHLVASAVLGAGGLYHSFLGP---
PCC 6406_CBPV          LLGAHVAHAGLIVFWAGAMTLFEL---------IGMVHLISAAVLGAGGIYHAVLGP---
PCC 6406_CPBPV         LLGAHIAHAGLIVLWAGAMTLFEL---------IGVVHLISSAVLGAGGLYHTVLGP---
PCC 6406_CBPVI         FIVAHVAQAALIMFWAGAFTLFEL---------IGVIHIVAAGVLAGGAYFHRERLG---
PCC 7203_CBPVI         FLTAHIAHAAIVSFSIGALILLEI---------FGVVLLVSAAVFTAGTLFHRSQVP---

                                Helix 3                       Helix 4

BP-1_CP43              ---ILGFHLI-VLGIGALLLVAKAMFFG------VVGGHIWIGLICIAGGIWHIL-----
BP-1_CP47              ---MFGIHLF-LAGL--LCFGFGAFHLT------VVAHHIAAGIVGIIAGLFHIL-----
BP-1_PsaA              ---MLNHHLAGLLGLGSLAWAGHQIHVS------TAHHHLAIAVLFIIAG--HMY-----
BP-1_PsaB              ---RLNHHLAGLFGVSSLAWAGHLIHVA------MAHHHLAIAVLFIVAG--HMY-----
BP-1_IsiA_CBPIII       ---ILGHHLL-FLGFGALLLVLKATIWG------LVGGHIYIAILLIAGGIWHIL-----
CCMP1375_PcbC_dCBP     ---ILGHHLI-FLGLGNIQFVEWARIH-------VMGGHAFLAFFLIIGGAFHIA-----
MIT9313_PcbA_dCBP      ---ILGHHLL-FLGLGCVQFVEWAKYH-------VMGGHAFLAFFLSAGAIWHIF-----
CCMP1986_PcbA_dCBP     ---ILGHHLI-ILGFAVILLVEWARVH-------VMGGHAFLAFVLITGGAWHIA-----
PCC 9006_PcbC_CBPI     ---ILGHHLL-LLGILCLAFVAKAMFWG------IIGGHVYIGILELIGGTWHIL-----
MBIC11017_PcbC_CBPI    ---IVGHHLV-FISVACLIFAVNATYGT------VIGGHFLIGVIDLLGAAFHIL-----
P1_PcbA_CBPII          ---ILGHHLI-LLGLGALFLVLWAVFF-------LIGGHVYVAIIEISGGLWHIF-----
PCC 7120_CBPIV         ---ILGIHLV-LLGLGAGLLVAKAVFFG------LVGGHIWVSILCIAGGLWHIT-----
PCC 6406_CBPIV         ---ILGIHLV-LLGLGTFLLVTKAMIFG------AVGGHIWVGLMCMLGGIWHMR-----
PCC 7375_CBPIV         ---ILGSHLV-LLGGGALLLVAKAIFLG------VVGGHLYIGIVLILGGLWHIF-----
PCC 7203_CBPIV         ---ILGIHLV-LLGIGAFLLVAKAMYFG------VVGGHIWVGGILILGGLFHIA-----
PCC 7120_CBPV          ---IIGIHLL-LLGAGAWLLVAKALFWG------VVGGHIWVGILCIGGGFWHIL-----
PCC 7203_CBPV          ---ILGIHLM-LLGLGALLLVAKAMFWG------IVGGHIWVGGILIGGGIFHIL-----
PCC 7375_CBPV          ---IIGIHLV-LLGLGAWLLVAKAMFWG------IVGGHLWVGLMCVLGGIWHIA-----
PCC 6406_CBPV          ---ILGIHLM-LLGIGALLLVLKGAYFG------LVGGHFWVALLCLGGGFFHIM-----
PCC 6406_CPBPV         ---ILGIHLV-LLGVGALLLVVKATTFG------VVGGHLWIGAIAILGGIWHIR-----
PCC 6406_CBPVI         ---ILGHHLA-ILGLGALLLVVKATAFG------LVGGHIYVAVLLLLGGAWHIL-----
PCC 7203_CBPVI         ---ILGNHLI-FLGIGALLLVAKAMFFG------VVGGHIYVGALLIVAGIWHMI-----

                              Helix 5                           Helix 6
BP-1_CP43              -LSYSLGA-----LSMMGFIATCFVWFN------WLATSHFVLAFF-FLVGHLWHAG
BP-1_CP47              -LSSSIAA-----VFFAAFVVAGTMWYG------WFTFAHAVFALL-FFFGHIWHGA
BP-1_PsaA              -LTTSWHAQLAINLAMMGSLSIIVAQHM------SLFTHHMWIGGF-LVVGGAAHGA
BP-1_PsaB              -YNNSLHFQLGWHLACLGVITSLVAQHM------ALYTHHQYIAGF-LMVGAFAHGA
BP-1_IsiA_CBPIII       -LSYSLGG-----IALAGFVAAYFCAVN------WLANAHFFLAFF-FLQGHLWHAL
CCMP1375_PcbC_dCBP     -LSYSLAG-----VAYCAFVAAFWCATN------WLSNVHFYLGFF-FLQGHLWHAL
MIT9313_PcbA_dCBP      -LSTSLAG-----AAFIAFVAAFWASMN------WLSNFHFYLGFF-YLQGHFWHGL
CCMP1986_PcbA_dCBP     -LSWSLAG-----IGWMAIIAAFWSASN------WLANVHYYFGFF-FIQGHLWHAL
PCC 9006_PcbC_CBPI     -LSYSLGA-----VGWMGLLSGFFVRYC------GAAAVQYILGVL-LLVGHVWHAT
MBIC11017_PcbC_CBPI    -LSWSVAS-----VGFMGISSSLFIRYC------GAATLQLILGLVWMLGGGLWHGL
P1_PcbA_CBPII          -LAYALGG-----LAIMGFTAAVYCAFN------WLCNVHFFLAFF-VLQGHLWHAL
PCC 7120_CBPIV         -LSYSLGA-----LSLMSLIAAYFVSIN------WLANAHFWLGFF-FLQGHLWHAL
PCC 6406_CBPIV         -LSYSIGA-----VSLMAFVATLFVSVN------WLANAHFWLGFF-FLQGHLFHAL
PCC 7375_CBPIV         -LAYSLGA-----LSLMTFVATLFVSVN------WLANAHFWLGFF-FLQGHLWHTL
PCC 7203_CBPIV         -LSYSLGA-----LALMGFIATLFVSVN------WLANTHFWLAFF-FLQGHIWHAL
PCC 7120_CBPV          -LSYSLAA-----LAYMGLLAAYFVTVN------WLATSHFALAIV-FLSGHIWHAL
PCC 7203_CBPV          -LAYSIGA-----VAYMGFFAAYFASVN------WLVSFHFVLAVI-FLLGHIWHAL
PCC 7375_CBPV          -LSYSLGA-----LAYMGIFAGYFVTVN------WLAAFHFAFGGL-LLAGHLWHAI
PCC 6406_CBPV          -LSYSLGA-----LAIAGLSVAVFVSVN------ALASVHAGLGFL-ALLGHLWHAC
PCC 6406_CPBPV         -LAYSQAA-----LAYMGFFAAYFVWVN------WLMLFHVVFASL-LLAGHFWHGL
PCC 6406_CBPVI         -LSYSLFG-----IALAGFAASYYCGFN------WLANAHFYLAFF-FLQGGLWHFQ
PCC 7203_CBPVI         -LSYSLFS-----LALTGFAGSYFCGFN------WLANTYFYLSFF-TLQGSLWHFG
```

Figure S6. Newly characterized CBP clades have conserved residues for potentially binding chlorophyll. Alignment of the transmembrane helices of CBP proteins and similar light-harvesting proteins. Amino acids highlighted in green (histidine) and yellow (glutamine) correspond to putative chlorophyll-binding residues. Organisms are named accordingly to the Strain ID of Table S1.

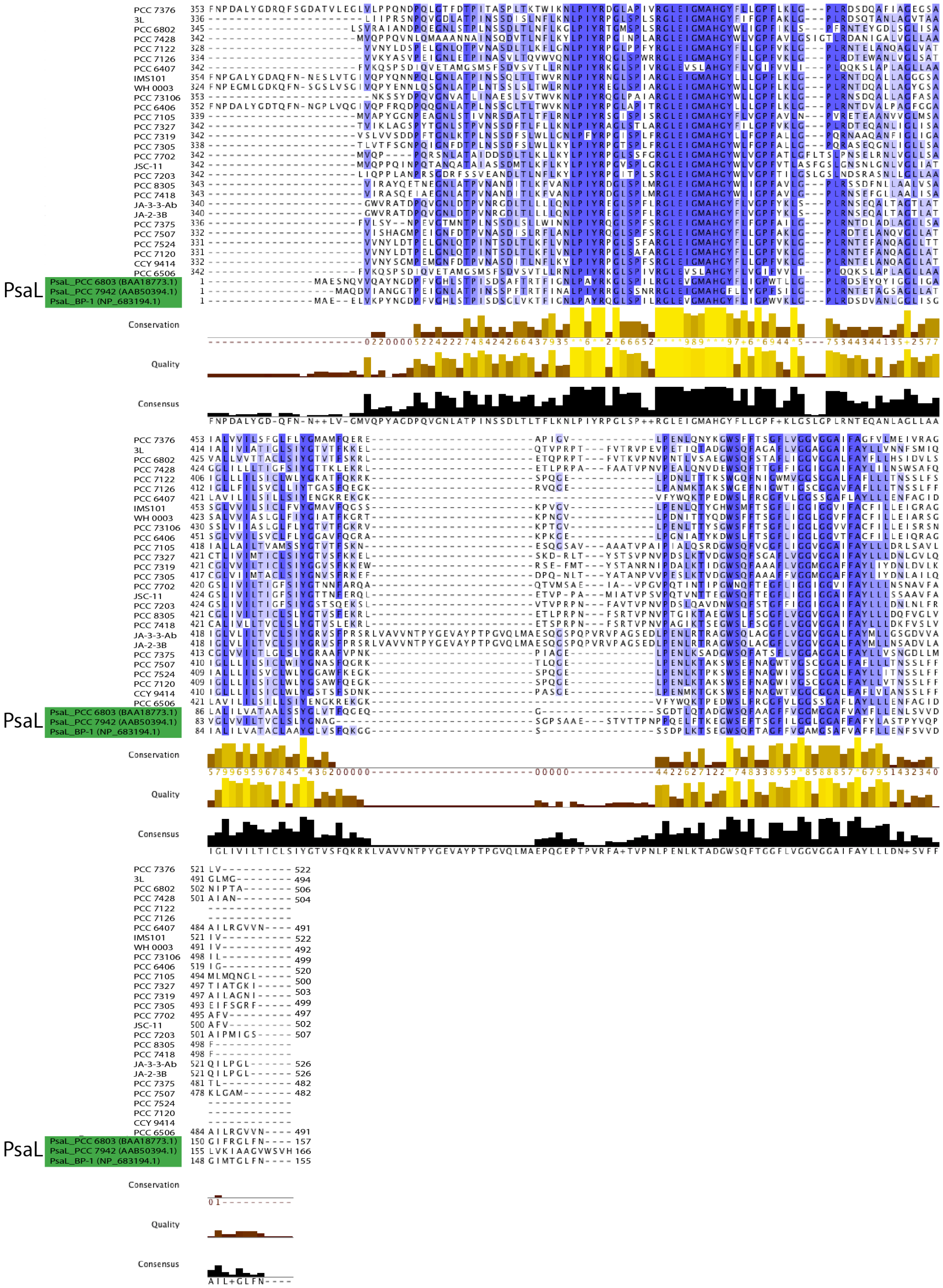


Figure S7. The C-terminal PsaL-like domain of CBPV proteins is homologous to PsaL. Alignment of the C-terminal PsaL-like domain of CBPV proteins containing full-length PsaL domains to the canonical PsaL of PSI (highlighted green accessions mark the amino acid sequences of the PsaL subunits of *Synechoccoccus elongatus* PCC 7942, *Synechocystis* sp. PCC 6803, and *Thermosynechococcus elongatus* BP-1. Organisms are named accordingly to the Strain ID of Table S1.

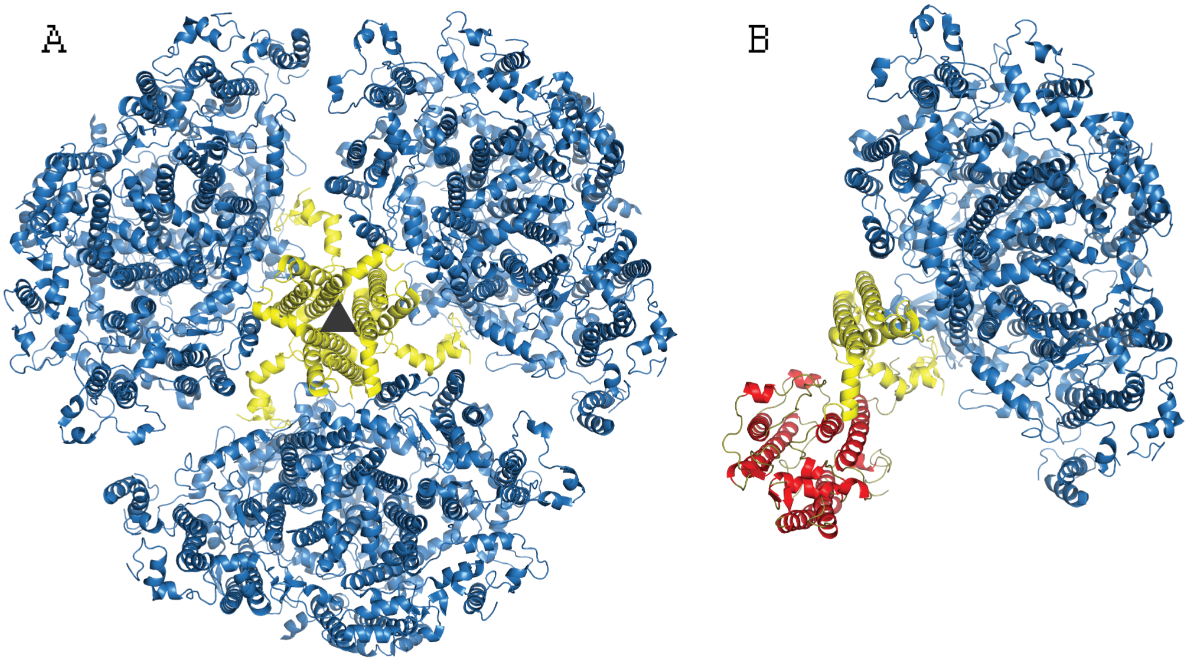


Figure S8. Comparison of trimeric Photosystem I to proposed CBPV-Photosystem I complex model. A, Top view of trimeric Photosystem I structure of *Thermosynechococcus elongatus* from the Protein Data Bank structure, 1JB0 (PDB ID). The threefold symmetry axis is denoted by the black triangle in the center. PsaL subunits are highlighted in yellow. **B,** Top view of proposed model of CBPV from *Chroococcidiopsis thermalis* PCC 7203 (Chro_2988) interacting with the Photosystem I monomer from the upper right of the trimer. Replacing the PsaL subunit (yellow) of a monomeric PSI with the PsaL-like domain of CBPV would preclude trimer formation, potentially resulting in monomerization of Photosystem I. The CBP domain (first six helices) is highlighted in red, whereas the monomeric Photosystem I, excluding the PsaL subunit, is highlighted in yellow.

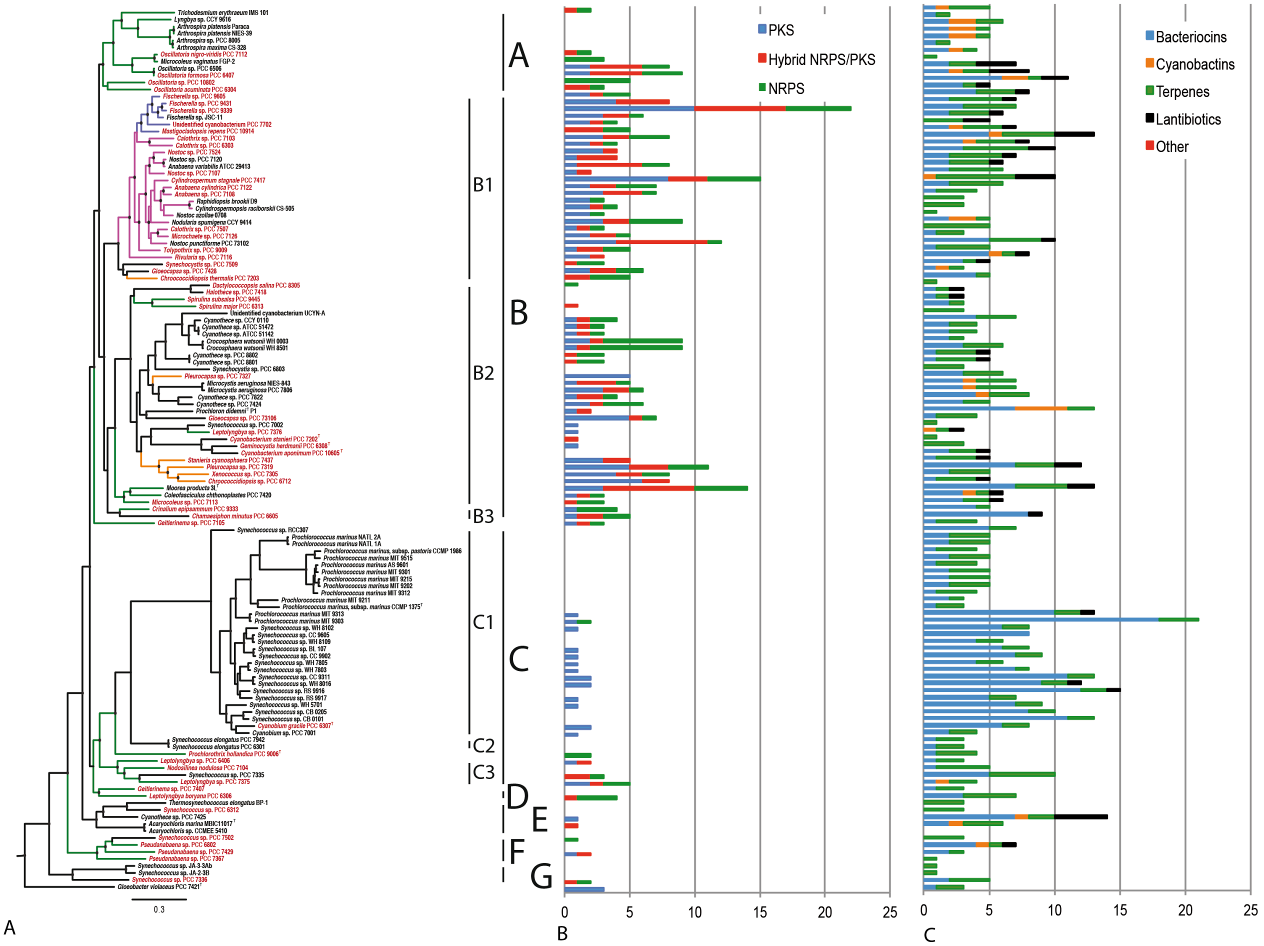


Figure S9. | Distribution of the ribosome dependent and nonribosomal encoded peptide and polyketide biosynthetic pathways in Cyanobacteria. A, Cyanobacterial Tree as in Fig. 1, **B**, Distribution of the nonribosomal peptide and polyketide gene clusters (number and occurrence within each genome), **C**, Distribution of the gene clusters involved in ribosome-dependent synthesis of diverse peptides (number and occurrence within each genome).

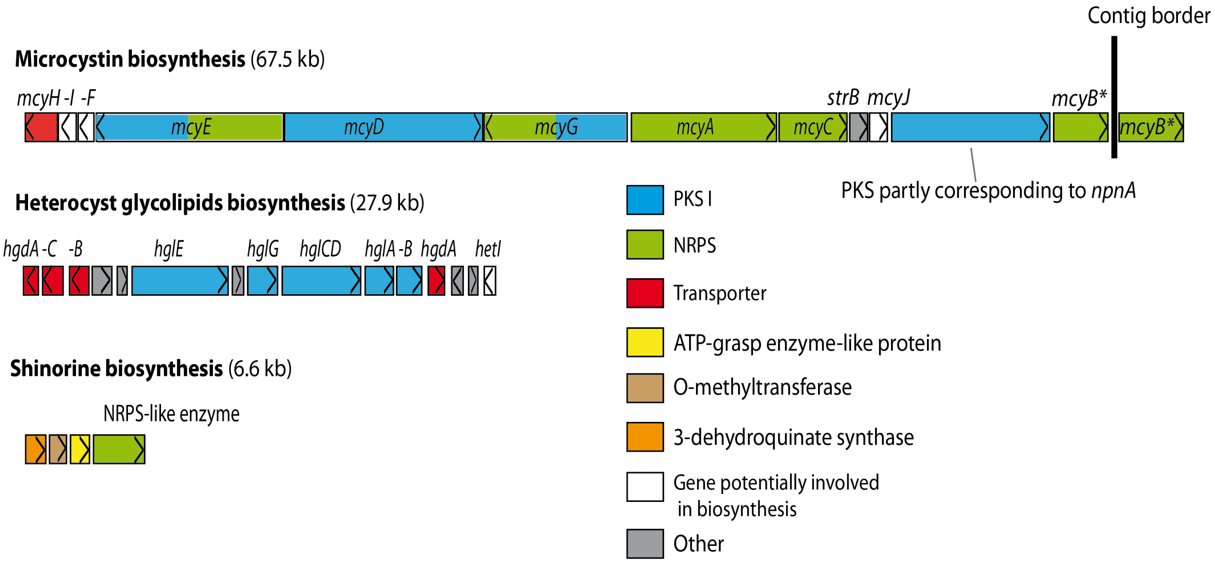

Contig border
Microcystin biosynthesis (67.5 kb)
mcyH -I -F
mcyE
mcyD
mcyG
mcyA
mcyC
strB mcyJ
mcyB*
mcyB*
PKS partly corresponding to npnA
Heterocyst glycolipids biosynthesis (27.9 kb)
hgdA -C -B
hglE
hglG
hglCD
hglA -B
hgdA
hetI
Shinorine biosynthesis (6.6 kb)
NRPS-like enzyme
PKS I
NRPS
Transporter
ATP-grasp enzyme-like protein
O-methyltransferase
3-dehydroquinate synthase
Gene potentially involved in biosynthesis
Other

Fig. S10. Predicted genetic potential for production of already kwon secondary metabolites found in the genome of *Fischerella* sp. PCC 9339. The identities of the sequence are estimated at the amino-acid level (% AASI). The putative microcystin gene cluster has 79.8% AASI to the one of *Anabaena* sp. 90 (25) and 88.5% AASI to the partial one retrieved from *Hapalosiphon hibernicus* BZ-3-1(26). Note the additional PKS gene, which on 2/3 of its length with 77.5% AASI corresponds to *NpnA* gene of the nostophycin gene cluster in *Nostoc* sp. 152 (27). The putative heterocyst glycolipids gene cluster has 67% AASI to the gene cluster required for synthesis and deposition of envelope glycolipids in *Nostoc* sp. PCC 7120 (28). Note the presence of two *hgdA* and the combination of *hglC* and *hglD* into a single gene in the heterocyst producing *Fischerella* sp. PCC 9339. The putative shinorine gene cluster is 70% AASI to the one identified in *Anabaena variabilis* ATCC 29413 (29).

Tables S1-S9

Table S1. 126 Cyanobacteria included in this study

Details on the strains are available in the Dataset S1.

^{T} indicates Type strain or Type species, for genome status: F, finished, D, draft, P, permanent draft.

Strain	Strain ID	Genome size (Mb)	% mol GC	No of scaffolds (chromosome / plasmid) - status	NCBI Project ID	References
Subsection I						
Acaryochloris sp.	CCMEE 5410	7.88	47.1	511 - D	16707	(30)
Acaryochloris marina	MBIC11017^{T}	8.36	47	10 (1/9) - F	12997	(31)
Chamaesiphon minutus	PCC 6605	6.76	45.7	3 - P	158825	This study
Crocosphaera watsonii	WH 0003	5.89	37.7	1126 - D	61839	(32)
Crocosphaera watsonii	WH 8501	6.24	37.1	323 - D	10651	(33)
Cyanobacterium aponinum	PCC 10605^{T}	4.18	34.9	2 - F	158691	This study
Cyanobacterium stanieri	PCC 7202^{T}	3.16	38.7	1 - F	39697	This study
Cyanobium gracile	PCC 6307^{T}	3.34	68.7	1 - F	158695	This study
Cyanobium sp.	PCC 7001	2.83	68.7	2 - D	19301	
Cyanothece sp.	ATCC 51142	5.46	37.9	6 (2/4) - F	20319	(34)
Cyanothece sp.	ATCC 51472	5.46	37.9	7 - F	59973	(35)
Cyanothece sp.	CCY 0110	5.88	36.7	163 - D	18951	
Cyanothece sp.	PCC 7424	6.55	38.5	7 (1/6) - F	20479	(35)
Cyanothece sp.	PCC 7425	5.79	50.7	4 (1/3) - F	28337	(35)
Cyanothece sp.	PCC 7822	7.84	39.9	7 (1/6) - F	28535	(35)
Cyanothece sp.	PCC 8801	4.79	39.8	4 (1/3) - F	20503	(35)
Cyanothece sp.	PCC 8802	4.8	39.8	5 (1/4) - F	28339	(35)
Dactylococcopsis salina	PCC 8305	3.78	42.4	1 - F	158703	This study
Geminocystis herdmanii	PCC 6308^{T}	4.26	34.3	1 - P	62511	This study
Gloeobacter violaceus	PCC 7421^{T}	4.66	62	1 - F	9606	(36)
Gloeocapsa sp.	PCC 73106	4.03	41.1	228 - D	159497	This study

Gloeocapsa sp.	PCC 7428	5.88	43.4	5 - F	158831	This study
Halothece sp.	PCC 7418	4.18	42.9	1 - F	40817	This study
Microcystis aeruginosa	NIES-843	5.84	42.3	1 - F	27835	(37)
Microcystis aeruginosa	PCC 7806	5.2	42	118 - D	15702	(38)
Prochlorococcus marinus	AS9601	1.67	31.3	1 - F	13548	(39)
Prochlorococcus marinus	MIT9202	1.69	31.1	1 - D	19343	
Prochlorococcus marinus	MIT9211	1.69	38	1 - F	13551	(39)
Prochlorococcus marinus	MIT9215	1.74	31.2	1 - F	18633	(39)
Prochlorococcus marinus	MIT9301	1.64	31.3	1 - F	15746	(39)
Prochlorococcus marinus	MIT9303	2.68	50	1 - F	13496	(39)
Prochlorococcus marinus	MIT9312	1.71	31.2	1 - F	13910	(40)
Prochlorococcus marinus	MIT9313	2.41	50.7	1 - F	220	(41)
Prochlorococcus marinus	MIT9515	1.7	31	1 - F	13617	(39)
Prochlorococcus marinus	NATL1A	1.86	35	1 - F	15660	(39)
Prochlorococcus marinus	NATL2A	1.84	35.1	1 - F	13911	(39)
Prochlorococcus marinus, subsp. *marinus*	CCMP1375^{T}	1.75	36.4	1 - F	419	(42)
Prochlorococcus marinus, subsp. *pastoris*	CCMP1986	1.66	30.8	1 - F	213	(41)
Prochloron didemni (metagenome)	P1	6.2	42	100 - D	13452	(43)
*Synechococcus elongatu*s	PCC 6301	2.7	55.5	1 - F	13282	(44)
Synechococcus elongatus	PCC 7942	2.74	55.4	2 (1/1) - F	10645	
Synechococcus sp.	BL107	2.28	54.2	6 - D	13559	(45)
Synechococcus sp.	CB0101	2.69	64.2	94 - D	46501	
Synechococcus sp.	CB0205	2.43	63	78 - D	46503	
Synechococcus sp.	CC9311	2.61	52.5	1 - F	12530	(46)
Synechococcus sp.	CC9605	2.51	59.2	1 - F	13643	(45)
Synechococcus sp.	CC9902	2.23	54.2	1 - F	13655	(45)
Synechococcus sp.	JA-2-3B	3.05	58.5	1 - F	16252	(47)
Synechococcus sp.	JA-3-3Ab	2.93	60.2	1 - F	16251	(47)
Synechococcus sp.	PCC 6312	3.72	48.5	2 - F	158717	This study

Synechococcus sp.	PCC 7002	3.41	49.2	7 (1/6) - F	28247	
Synechococcus sp.	PCC 7335	5.97	48.2	11 - F	19377	
Synechococcus sp.	PCC 7336	5.14	53.7	2 - F	158719	This study
Synechococcus sp.	PCC 7502	3.58	40.6	3 - F	159509	This study
Synechococcus sp.	RCC307	2.22	60.8	1 - F	13654	(45)
Synechococcus sp.	RS9916	2.66	59.8	4 - D	13557	(45)
Synechococcus sp.	RS9917	2.58	64.5	9 - D	13555	(45)
Synechococcus sp.	WH 5701	3.04	65.4	135 - D	13554	(45)
Synechococcus sp.	WH 7803	2.37	60.2	1 - F	13642	(45)
Synechococcus sp.	WH 7805	2.62	57.6	13 - F	13553	(45)
Synechococcus sp.	WH 8016	2.71	54.1	1 - F	61805	
Synechococcus sp.	WH 8102	2.43	59.4	1 - F	230	(48)
Synechococcus sp.	WH 8109	2.12	60.1	1 - F	37911	
Synechocystis sp.	PCC 6803	3.95	47.4	5 (1/4) - F	60	(49)
Synechocystis sp.	PCC 7509	4.77	41.6	174 - D	159501	This study
Thermosynechococcus elongatus	BP-1	2.59	53.9	1 - F	308	(50)
Unidentified cyanobacterium (symbiont)	UCYN-A	1.44	31.1	1 - F	30917	(51)
Subsection II						
Chroococcidiopsis sp.	PCC 6712	5.7	35.3	3 - F	158687	This study
Chroococcidiopsis thermalis	PCC 7203	6.69	44.5	3 - F	38119	This study
Pleurocapsa sp.	PCC 7319	7.39	38.7	10 - P	158813	This study
Pleurocapsa sp.	PCC 7327	4.99	45.2	1 - F	158829	This study
Stanieria cyanosphaera	PCC 7437	5.55	36.2	6 - F	158877	This study
Xenococcus sp.	PCC 7305	5.93	39.7	234 - D	159499	This study
Subsection III						
Arthrospira maxima	CS-328	6	44.8	129 - D	29085	
Arthrospira platensis	NIES-39	6.79	44.3	1 - F	42161	(52)
Arthrospira platensis	Paraca	5,00	44.3	1820 - D	34793	
Arthrospira sp.	PCC 8005	6.15	44.7	119 - D	40633	(53)

Coleofasciculus chthonoplastes	PCC 7420	8.68	45.4	57 - D	19325	
Crinalium epipsammum	PCC 9333	5.62	40.2	9 - F	158835	This study
Geitlerinema sp.	PCC 7105	6.15	51.6	8 - P	158727	This study
Geitlerinema sp.	PCC 7407	4.68	58.5	1 - F	158833	This study
Leptolyngbya boryana	PCC 6306	7.26	47	5 - P	158729	This study
Leptolyngbya sp.	PCC 6406	5.61	55.2	377 - P	159511	This study
Leptolyngbya sp.	PCC 7375	9.42	47.6	5 - P	43137	This study
Leptolyngbya sp.	PCC 7376	5.13	43.9	1 - F	43487	This study
Lyngbya sp.	CCY 9616	7.04	41.1	110 - D	13409	
Microcoleus sp.	PCC 7113	7.97	46.2	9 - F	158839	This study
Microcoleus vaginatus	FGP-2	6.7	46	40 - P	47601	(54)
Moorea producta	3L^{T}	8.48	43.7	161 - D	60895	(55)
Nodosilinea nodulosa	PCC 7104	6.89	57.7	2 - P	62311	This study
Oscillatoria acuminata	PCC 6304	7.8	47.6	3 - F	158709	This study
Oscillatoria formosa	PCC 6407	6.89	43.4	12 - P	158733	This study
Oscillatoria nigro-viridis	PCC 7112	8.27	45.8	6 - F	158711	This study
Oscillatoria sp.	PCC 10802	8.59	54.1	9 - P	158815	This study
Oscillatoria sp.	PCC 6506	6.68	43.4	377 - D	49445	(56)
Prochlorothrix hollandica	PCC 9006^{T}	5.65	54.4	13 - P	158811	This study
Pseudanabaena sp.	PCC 6802	5.62	47.8	6 - P	158731	This study
Pseudanabaena sp.	PCC 7367	4.89	46.2	2 - F	158713	This study
Pseudanabaena sp.	PCC 7429	5.48	43.2	464 - D	158837	This study
Spirulina major	PCC 6313	5.05	53.5	2 - F	158715	This study
Spirulina subsalsa	PCC 9445	5.32	47.4	2 - F	158827	This study
Trichodesmium erythraeum	IMS101	7.75	34.1	1 - F	318	
Subsection IV						
Anabaena cylindrica	PCC 7122	7.06	38.8	7 - F	43355	This study
Anabaena sp.	PCC 7108	5.89	38.8	3 - F	158737	This study
Anabaena variabilis	ATCC 29413	7.11	41.4	5 (2/3) - F	10642	
Calothrix sp.	PCC 6303	6.96	39.8	4 - F	158041	This study

Calothrix sp.	PCC 7103	11.58	38.6	12 - P	159495	This study
Calothrix sp.	PCC 7507	7.02	42.3	1 - F	158683	This study
Cylindrospermopsis raciborskii	CS-505	3.88	40.2	93 - D	40109	(57)
Cylindrospermum stagnale	PCC 7417	7.61	42.2	4 - P	158809	This study
Microchaete sp.	PCC 7126	5.74	42.2	3 - P	158817	This study
Nodularia spumigena	CCY 9414	5.32	41.3	204 - D	13447	
Nostoc azollae (endosymbiont)	708	5.49	38.4	3 (1/2) - F	30807	(58)
Nostoc punctiforme	PCC 73102	9.06	41.4	6 (1/5) - F	216	
Nostoc sp.	PCC 7107	6.33	40.4	1 - F	158705	This study
Nostoc sp.	PCC 7120	7.21	41.3	7 (1/6) - F	244	(59)
Nostoc sp.	PCC 7524	6.72	41.5	3 - F	158707	This study
Raphidiopsis brookii	D9	3.19	40.1	47 - D	40111	(57)
Rivularia sp.	PCC 7116	8.73	37.5	3 - F	63147	This study
Tolypothrix sp.	PCC 9009	8.18	41.2	204 - D	63425	This study
Subsection V						
Fischerella sp.	JSC-11	5.38	41.1	34 - D	61093	
Fischerella sp.	PCC 9339	8.4	40.1	95 - P	159505	This study
Fischerella sp.	PCC 9431	7.14	40.2	36 - P	158821	This study
Fischerella sp.	PCC 9605	8.2	42.6	36 - P	158819	This study
Mastigocladopsis repens	PCC 10914	6.31	43.5	23 - P	158735	This study
Unidentified cyanobacterium*	PCC 7702	4.9	42.4	4 - P	158823	This study

*PCC 7702 corresponds to the high temperature forms (HTF) of cyanobacteria found in hot springs, at temperatures higher than 50 °C (up to 62°C), and originally thought to be related to "*Mastigocladus laminosus*". The morphology of this HTF strain is variable from unicellular to very short filaments, and consequently, impossible to identify at the genus level. Furthermore, PCC 7702 strain is unable to fix nitrogen under aerobic conditions but contains *nif* genes.

Table S2. Improvement of phylogenetic diversity with the addition of the CyanoGEBA dataset measured by Tree Imbalance

Phylogenetic Diversity Metric

CyanoGEBA set	Random set	Fold Improvement
10.82	5.28±0.37	1.92-2.20

Tree Imbalance

Average Colless's Imbalance (n=1000)	Genomes prior to this study	All Genomes, including CyanoGEBA
Uniform Speciation	0.093	0.059
Equiprobable Speciation	0.30	0.24

Table S3. Novel* proteins in CyanoGEBA genomes

*lacking similarity to any protein in GenBank

CyanoGEBA genome	Number of novel proteins coding genes	% of novel protein coding gene
Anabaena cylindrica PCC 7122	338	5.40
Anabaena sp. PCC 7108	291	5.57
Calothrix sp. PCC 6303	370	6.33
Calothrix sp. PCC 7103	1153	11.16
Calothrix sp. PCC 7507	375	6.00
Chamaesiphon minutus PCC 6605	704	10.94
Chroococcidiopsis sp. PCC 6712	334	6.45
Chroococcidiopsis thermalis PCC 7203	339	5.62
Crinalium epipsammum PCC 9333	372	7.35
Cyanobacterium aponinum PCC 10605	138	3.82
Cyanobacterium stanieri PCC 7202	97	3.30
Cyanobium gracile PCC 6307	212	6.16
Cylindrospermum stagnale PCC 7417	486	7.21
Dactylococcopsis salina PCC 8305	199	5.40
Fischerella sp. PCC 9339	505	7.40
Fischerella sp. PCC 9431	360	5.90
Fischerella sp. PCC 9605	626	8.78
Geitlerinema sp. PCC 7105	412	7.63
Geitlerinema sp. PCC 7407	162	4.14
Geminocystis herdmanii PCC 6308	168	4.00
Gloeocapsa sp. PCC 73106	171	4.12
Gloeocapsa sp. PCC 7428	251	4.73
Halothece sp. PCC 7418	133	3.39
Leptolyngbya boryana PCC 6306	736	10.65
Leptolyngbya sp. PCC 6406	468	8.92
Leptolyngbya sp. PCC 7375	1137	13.46
Leptolyngbya sp. PCC 7376	342	7.35
Mastigocladopsis repens PCC 10914	409	7.17
Microchaete sp. PCC 7126	336	6.37
Microcoleus sp. PCC 7113	458	6.71
Nodosilinea nodulosa PCC 7104	480	7.42
Nostoc sp. PCC 7107	220	3.97
Nostoc sp. PCC 7524	253	4.45
Oscillatoria acuminata PCC 6304	419	6.87
Oscillatoria formosa PCC 6407	110	8.76
Oscillatoria nigro-viridis PCC 7112	508	13.37
Oscillatoria sp. PCC 10802	937	1.55

Pleurocapsa sp. PCC 7319	452	6.70
Pleurocapsa sp. PCC 7327	221	4.73
Prochlorothrix hollandica PCC 9006	492	10.20
Pseudanabaena sp. PCC 6802	525	9.64
Pseudanabaena sp. PCC 7367	357	8.89
Pseudanabaena sp. PCC 7429	406	8.42
Rivularia sp. PCC 7116	437	6.29
Spirulina major PCC 6313	247	5.54
Spirulina subsalsa PCC 9445	216	4.67
Stanieria cyanosphaera PCC 7437	255	5.06
Synechococcus sp. PCC 6312	313	8.25
Synechococcus sp. PCC 7336	472	9.90
Synechococcus sp. PCC 7502	256	6.98
Synechocystis sp. PCC 7509	247	5.19
Tolypothrix sp. PCC 9009	636	8.53
Xenococcus sp. PCC 7305	396	7.30
Unidentified cyanobacterium PCC 7702	170	3.89

Table S4. Prediction of CRISPR loci in CyanoGEBA genomes

CyanoGEBA genome	**Number of spacer-direct repeat units**	**Number of CRISPR loci**
Anabaena cylindrica PCC 7122	367	13
Anabaena sp. PCC 7108	95	7
Calothrix sp. PCC 6303	72	6
Calothrix sp. PCC 7103	178	13
Calothrix sp. PCC 7507	336	10
Chamaesiphon minutus PCC 6605	59	3
Chroococcidiopsis sp. PCC 6712	47	5
Chroococcidiopsis thermalis PCC 7203	64	2
Crinalium epipsammum PCC 9333	113	6
Cyanobacterium aponinum PCC 10605	166	10
Cyanobacterium stanieri PCC 7202	15	2
Cyanobium gracile PCC 6307	0	0
Cylindrospermum stagnale PCC 7417	191	10
Dactylococcopsis salina PCC 8305	0	0
Fischerella sp. PCC 9339	26	7
Fischerella sp. PCC 9431	18	4
Fischerella sp. PCC 9605	11	2
Geitlerinema sp. PCC 7105	650	15
Geitlerinema sp. PCC 7407	23	1

Geminocystis herdmanii PCC 6308	33	2
Gloeocapsa sp. PCC 73106 *	50	4
Gloeocapsa sp. PCC 7428	98	3
Halothece sp. PCC 7418	443	4
Leptolyngbya boryana PCC 6306	80	5
Leptolyngbya sp. PCC 6406 *	168	9
Leptolyngbya sp. PCC 7375	188	12
Leptolyngbya sp. PCC 7376	6	1
Mastigocladopsis repens PCC 10914	0	0
Microchaete sp. PCC 7126	88	4
Microcoleus sp. PCC 7113	72	10
Nodosilinea nodulosa PCC 7104	75	4
Nostoc sp. PCC 7107	252	14
Nostoc sp. PCC 7524	278	6
Oscillatoria acuminata PCC 6304	279	10
Oscillatoria formosa PCC 6407	95	10
Oscillatoria nigro-viridis PCC 7112	304	9
Oscillatoria sp. PCC 10802	531	18
Pleurocapsa sp. PCC 7319	68	1
Pleurocapsa sp. PCC 7327	100	4
Prochlorothrix hollandica PCC 9006	237	8
Pseudanabaena sp. PCC 6802	77	2
Pseudanabaena sp. PCC 7367	160	7
Pseudanabaena sp. PCC 7429 *	610	14
Rivularia sp. PCC 7116	256	15
Spirulina major PCC 6313	102	7
Spirulina subsalsa PCC 9445	625	17
Stanieria cyanosphaera PCC 7437	74	4
Synechococcus sp. PCC 6312	154	4
Synechococcus sp. PCC 7336	285	8
Synechococcus sp. PCC 7502	62	2
Synechocystis sp. PCC 7509 *	6	1
Tolypothrix sp. PCC 9009 *	201	15
Xenococcus sp. PCC 7305 *	37	5
Unidentified cyanobacterium PCC 7702	8	2

* These genomes are not finished and currently contain more than 100 scaffolds. The number of spacer-direct repeat units and CRISPR loci therefore may be underestimated.

Table S5. Comparative genomics of morphological transitions
Events of morphological transition are shown in Fig. 1. For each event, the set of genes involved in one genome or in genomes belonging to one subsection (genome in) were compared those of genomes of another subsection (genome out). Genomes are annotated by the Strain ID as in Table S1.

Morphological transition (Genomes in *vs* out)	**Evolutionary transition (Subsection to Subsection)**	**Number of genes**
Event 1 (PCC 7367, PCC 7429, PCC 6802 *vs* PCC 7502)	III to I	88
Event 2 (PCC 6406, PCC 7104, PCC 7375 *vs* PCC 7335)	III to I	674
Event 3 (PCC 9006, PCC 6406, PCC 7104, PCC 7375, PCC 6306, PCC 7407 *vs* subclade C1 and C2)	III to I	32
Event 4 (PCC 7002, PCC 7202, PCC 6308, and PCC 10605 *vs* PCC 7376)	I to III	3172
Event 5 (NIES-843, PCC 7806, PCC 7822, and PCC 7424 *vs* PCC 7327)	I to II	2531
Event 6 (PCC 7428, PCC 7509 *vs* PCC 7203)	I to II	3783
Event 7 (PCC 7203, PCC 7428, PCC 7509 *vs* Subsection IV and V)	I to IV and V	9
Event 8 (Subsection V *vs* Subsection IV)	IV to V	0

Table S6. Homologous proteins lost during the reversion of filamentous to unicellular morphology in both Event 2 and Event 3.

Query locus tag in Event 2	Top hit locus tag in Event 3	Query annotation
Pro9006DRAFT_1077	Lepto6406DRAFT_00007290	Arsenite-activated ATPase ArsA
Pro9006DRAFT_3818	Lepto6406DRAFT_00024510	HAS barrel domain.
Pro9006DRAFT_3344	Lepto6406DRAFT_00009900	Hypothetical protein
Pro9006DRAFT_0305	Lepto6406DRAFT_00035530	Hypothetical protein
Pro9006DRAFT_0620	Lepto6406DRAFT_00010140	Hypothetical protein
Pro9006DRAFT_4432	Lepto6406DRAFT_00049190	Highly conserved protein containing a thioredoxin domain
Pro9006DRAFT_1144	Lepto6406DRAFT_00002010	Asparaginase
Pro9006DRAFT_2144	Lepto6406DRAFT_00041660	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain
Pro9006DRAFT_3622	Lepto6406DRAFT_00019670	Iron-sulfur cluster binding protein, putative
Pro9006DRAFT_0863	Lepto6406DRAFT_00033060	Hypothetical protein
Pro9006DRAFT_2707	Lepto6406DRAFT_00005310	Hypothetical protein
Pro9006DRAFT_1113	Lepto6406DRAFT_00025920	Hypothetical protein
Pro9006DRAFT_0892	Lepto6406DRAFT_00019960	Hypothetical protein
Pro9006DRAFT_0326	Lepto6406DRAFT_00025500	Hypothetical protein
Pro9006DRAFT_4045	Lepto6406DRAFT_00040530	Alpha-amylase/alpha-mannosidase
Pro9006DRAFT_1882	Lepto6406DRAFT_00043930	Hypothetical protein
Pro9006DRAFT_4594	Lepto6406DRAFT_00035370	Hypothetical protein
Pro9006DRAFT_1996	Lepto6406DRAFT_00016810	Hypothetical protein
Pro9006DRAFT_1710	Lepto6406DRAFT_00003520	Hypothetical protein
Pro9006DRAFT_2550	Lepto6406DRAFT_00031350	Polyketide cyclase / dehydrase and lipid transport.
Pro9006DRAFT_2845	Lepto6406DRAFT_00014640	Hypothetical protein
Pro9006DRAFT_0040	Lepto6406DRAFT_00005410	Hypothetical protein
Pro9006DRAFT_1334	Lepto6406DRAFT_00025630	Hypothetical protein
Pro9006DRAFT_1711	Lepto6406DRAFT_00003510	Hypothetical protein
Pro9006DRAFT_1895	Lepto6406DRAFT_00032390	Hypothetical protein
Pro9006DRAFT_2407	Lepto6406DRAFT_00028690	FOG: GAF domain
Pro9006DRAFT_4751	Lepto6406DRAFT_00041610	Hypothetical protein
Pro9006DRAFT_1359	Lepto6406DRAFT_00013970	Uncharacterized conserved protein
Pro9006DRAFT_1554	Lepto6406DRAFT_00014960	Uncharacterized protein conserved in bacteria

Table S7. Increase in number of cyanobacterial proteins improves prediction of eukaryotic nuclear genes that resulted from Endosymbiotic Gene Transfer.

Eukaryote	**Number of genes predicted without CyanoGEBA genomes**	**Number of genes predicted including CyanoGEBA genomes**	**% increase with CyanoGEBA**
Arabidopsis (plant)	3811	4339	14%
Physcomitrella (plant)	2941	3300	12%
Micromonas (green algae)	1472	1643	12%
Cyanidioschyzon (red algae)	711	777	9%
Ectocarpus (brown algae)	1891	2156	14%
Emiliana (haptophyte)	4397	5151	17%
Phaeodactylum (diatom)	1425	1610	13%
Thalassiosira (diatom)	1436	1637	14%
Cyanophora (glaucophyte)	2417	2739	13%
Average			**13%**

Table S8. COG functional category distribution of nuclear genes that are of cyanobacterial descent

Functional category of Cluster of Orthologous Group (COG) from cyanobacterial genomes retrieved in the nuclear genomes of diverse photosynthetic eukaryotes. The latter are indicated as followed: 1, *Arabidopsis*; 2, *Physcomitrella;* 3, *Micromonas;* 4, *Cyanidioschyzon;* 5, *Ectocarpus;* 6, *Emiliana;* 7, *Thalassiosira;* 8, *Phaeodactylum;* 9, *Cyanophora*

COG	1	2	3	4	5	6	7	8	9
RNA processing and modification	0%	0%	0%	0%	0%	0%	0%	0%	0%
Chromatin structure and dynamics	0%	0%	0%	0%	0%	0%	0%	0%	0%
Energy production and conversion	4%	7%	5%	5%	4%	4%	5%	5%	4%
Cell cycle control, cell division, chromosome partitioning	0%	1%	1%	1%	1%	1%	0%	1%	1%
Amino acid transport and metabolism	5%	6%	6%	9%	4%	4%	6%	7%	5%
Nucleotide transport and metabolism	1%	1%	1%	2%	1%	2%	1%	1%	1%
Carbohydrate transport and metabolism	7%	9%	7%	7%	5%	6%	6%	6%	5%
Coenzyme transport and metabolism	3%	4%	6%	7%	4%	4%	5%	5%	4%
Lipid transport and metabolism	8%	5%	3%	4%	3%	4%	4%	4%	2%
Translation, ribosomal structure and biogenesis	4%	5%	6%	7%	3%	4%	5%	5%	3%
Transcription	2%	3%	2%	2%	3%	2%	2%	2%	3%
Replication, recombination and repair	2%	2%	3%	5%	3%	4%	3%	3%	5%
Cell wall/membrane/ envelope biogenesis	6%	6%	5%	6%	4%	4%	4%	5%	3%
Cell motility	1%	0%	1%	0%	3%	1%	1%	1%	1%
Posttranslational modification, protein turnover, chaperones	7%	7%	9%	8%	8%	7%	9%	8%	6%
Inorganic ion transport and metabolism	4%	4%	4%	5%	4%	6%	4%	5%	4%
Secondary metabolites biosynthesis, transport and catabolism	6%	4%	5%	3%	4%	7%	4%	5%	3%
General function prediction only	21%	18%	20%	16%	25%	19%	23%	19%	23%
Function unknown	12%	11%	12%	9%	13%	13%	11%	11%	10%
Signal transduction mechanisms	4%	4%	3%	2%	5%	3%	4%	4%	14%
Intracellular trafficking, secretion, and vesicular transport	2%	2%	2%	2%	4%	2%	2%	2%	2%
Defense mechanisms	0%	0%	0%	0%	0%	1%	0%	0%	1%
Extracellular structures	0%	0%	0%	0%	0%	0%	0%	0%	0%
Cytoskeleton	0%	0%	0%	0%	0%	0%	0%	0%	0%

Table S9. Sequencing information of CyanoGEBA organisms
The finishing efforts are indicated as followed: MF, manual finishing; AF, autofinishing. Submit indicates that the genome sequence has been submitted to NCBI to obtain the BioProject number.

CyanoGEBA Organism	454 Libraries	454 Total Reads	454 Total Mb	Illumina Libraries	Illumina Total Reads	Illumina Total bp	Finishing efforts	Nb of contigs / scaffolds	IMG Taxon ID
Anabaena cylindrica PCC 7122	(1) 454 STD TIT, (3) 454 PE (9138 kb, 3178 kb, NA)	1,079,579	361.9	(1) ILL STD	180,472,451	6,497,008,236	MF	7 / 7	2503982047
Anabaena sp. PCC 7108	(1) 454 STD TIT, (2) 454 PE (11344kb, 4036 kb)	727,027	181.2	(1) ILL STD	60,554,068	4,602,109,168	AF	13 / 3	2506485002
Calothrix sp. PCC 6303	(1) 454 STD TIT, (2) 454 PE (9829 kb, 4087.8 kb)	1,303,031	461.6	(1) ILL STD	115,161,558	8,752,278,408	MF	4 / 4	2503982036
Calothrix sp. PCC 7103	(0) 454 STD TIT, (2) 454 PE (5331 kb, 6844 kb)	640,339	216.1	(1) ILL STD	37,899,348	2,880,350,448	AF	67 / 12	2507262048
Calothrix sp. PCC 7507	(1) 454 STD TIT, (2) 454 PE (5438 kb, 2730 kb)	672,159	258.3	(1) ILL STD	42,042,292	3,195,214,192	MF	1 / 1	2505679032
Chamaesiphon minutus PCC 6605	(1) 454 STD TIT, (1) 454 PE (6916 kb)	976,084	247.3	(1) ILL STD	60,314,630	4,583,911,880	MF	3 / 3	2510436000
Chroococcidiopsis sp. PCC 6712	(1) 454 STD TIT, (3) 454 PE (2604 kb, 12,305 kb, 2694 kb)	1,269,117	353.5	(1) ILL STD	36,438,868	1,311,799,248	AF	18 / 3	2505679029
Chroococcidiopsis thermalis PCC 7203	(1) 454 STD TIT, (1) 454 PE (8583 kb)	788,934	272.3	(3) ILL STD	32,800,000	1,180,704,000	MF	3 / 3	2503538021
Crinalium epipsammum PCC 9333	(1) 454 STD TIT, (1) 454 PE (8063)	230,731	128.8	(1) ILL STD	30,965,529	1,114,759,044	MF	9 / 9	2504643013
Cyanobacterium aponinum PCC 10605	(2) 454 STD TIT, (2) 454 PE (NA, NA)	519,034	145	(1) ILL STD	43,225,758	3,285,157,608	MF	2 / 2	2503707009
Cyanobacterium stanieri PCC 7202	(1) 454 STD TIT, (1) 454 PE (8540 kb)	754,375	252.4	(1) ILL STD	2,050,270	366,482,655	MF	1 / 1	2503283023
Cyanobium gracile PCC 6307	(1) 454 STD TIT, (1) 454 PE (7784 kb)	356,894	159	(1) ILL STD	66,080,366	5,022,107,816	MF	1 / 1	2508501011

Cylindrospermum stagnale PCC 7417	(1) 454 STD TIT, (2) 454 PE (6956 kb, 4374 kb)	1,662,064	379.2	(1) ILL STD	74,952,294	5,696,374,344	AF	10 / 4	2509601025
Dactylococcopsis salina PCC 8305	(1) 454 STD TIT, (1) 454 PE (7217 kb)	976,293	246.7	(1) ILL STD	29,937,544	1,077,751,584	MF	1 / 1	2509276056
Fischerella sp. PCC 9339	-	-	-	(1) ILL STD, (1) ILL PE	31,117,314	4,667,600,000	none	171 / 95	2516653082
Fischerella sp. PCC 9431	-	-	-	(1) ILL STD, (1) ILL PE (6617 kb)	560,072,428	81,357,230	none	201 / 36	2512875027
Fischerella sp. PCC 9605	-	-	-	(1) ILL STD, (1) ILL PE (2209 kb)	45,267,538	6,790,130,000	none	49 / 36	2516143000
Geitlerinema sp. PCC 7105	(1) 454 STD TIT, (2) 454 PE (10539 kb, 4458 kb)	1,285,347	304.2	(1) ILL STD	116,062,307	7,311,925,341	AF	288 / 8	2510065011
Geitlerinema sp. PCC 7407	(1) 454 STD TIT, (1) 454 PE (4018 kb)	292,666	167.4	(1) ILL STD	37,618,333	2,858,993,308	MF	1 / 1	2503538020
Geminocystis herdmanii PCC 6308	-	-	-	(1) ILL STD	64,203,930	4,882,710,000	AF	11 /1	2509601046
Gloeocapsa sp. PCC 73106	(1) 454 STD TIT, (2) 454 PE / (8550 kb and 7666 kb)	481,442	297.2	(1) ILL STD	62,560,585	4,754,604,460	none	228/ 228	2508501033
Gloeocapsa sp. PCC 7428	(1) 454 STD TIT, (1) 454 PE/ (9786 kb)	129,654	226.5	(1) ILL STD	31,204,529	576,136,120	MF	5 / 5	2503754017
Halothece sp. PCC 7418	(0) 454 STD TIT, (2) 454 PE (2627 kb, 9799 kb)	902,827	216.1	(1) ILL STD	257,227,056	19,549,256,256	MF	1 / 1	2503538028
Leptolyngbya boryana PCC 6306	-	-	-	(1) ILL STD	9,298,704	6,649,250,000	AF	11 / 5	2509601031
Leptolyngbya sp. PCC 6406	(1) 454 STD TIT, (2) 454 PE (8212 kb)	1,049,271	273.4	(1) ILL STD	86,532,372	6,576,460,272	none	377 / 377	2517572073
Leptolyngbya sp. PCC 7375	(1) 454 STD TIT, (1) 454 PE/ (12811 kb)	228,442	170	(1) ILL STD	22,675,741	816,326,676	AF	40 / 5	2509601039
Leptolyngbya sp. PCC 7376	-	-	-	(1) ILL STD, (1) ILL PE (2481 kb)	529,092,128	79,363,820,000	MF	1 / 1	2503754048
Mastigocladopsis repens PCC 10914	(1) 454 STD TIT, (2) 454 PE (9610 kb, 3964 kb)	1,444,337	316.7	(1) ILL STD	25,286,224	910,304,064	none	325 / 23	2517093042

Microchaete sp. PCC 7126	(1) 454 PE/ (117346 kb)	735,764	109.6	(1) ILL STD	69,022,092	5,245,678,992	AF	5 / 3	2509601027
Microcoleus sp. PCC 7113	(2) 454 STD TIT, (3) 454 PE (4283 kb, 7800 kb, NA)	626,176	201.3	(1) ILL STD	57,251,139	4,351,086,564	MF	9 / 9	2509276031
Nodosilinea nodulosa. PCC 7104	(1) 454 STD TIT, (4) 454 PE (2798 kb, 24356kb, 22893 kb, 11125 kb)	1,921,672	486.1	(1) ILL STD	25,897,163	932,297,868	AF	62 / 2	2509601026
Nostoc sp. PCC 7107	(1) 454 STD TIT, (2) 454 PE (1695 kb, 4068 kb)	2,132,299	546.3	(1) ILL STD	62,447,094	4,745,979,144	MF	1 / 1	2503707008
Nostoc sp. PCC 7524	(1) 454 STD TIT, (2) 454 PE (11786 kb, 11762 kb)	681,222	256.3	(1) ILL STD	17,798,114	640,732,104	MF	3 / 3	2509601032
Oscillatoria acuminata PCC 6304	(0) 454 STD TIT, (1) 454 PE (8203 kb)	652,065	129.4	(1) ILL STD	67,180,232	5,105,697,632	MF	3 / 3	2509276028
Oscillatoria formosa PCC 6407	(1) 454 STD TIT, (2) 454 PE	1,050,403	253.9	(1) ILL STD	25,052,472	901,888,992	AF	259 / 12	2508501075
Oscillatoria nigro-viridis PCC 7112	(1) 454 STD TIT, (2) 454 PE (8172 kb , 6631 kb)	1,446,977	433.8	(1) ILL STD	46,329,519	3,521,043,444	AF	108 / 6	2503982035
Oscillatoria sp. PCC 10802	(1) 454 STD TIT, (2) 454 PE	499,658	244.6	(1) ILL STD	70,039,722	5,323,018,872	MF	6 / 9	2509276047
Pleurocapsa sp. PCC 7319	(1) 454 STD TIT, (1) 454 PE (12243 kb)	1,020,605	299.4	(1) ILL STD	31,122,538	2,365,312,888	AF	30 / 10	2509601013
Pleurocapsa sp. PCC 7327	(1) 454 STD TIT, (2) 454 PE (1525 kb, 7351 kb)	1,361,678	352.7	(1) ILL STD	145,035,126	11,022,669,576	MF	1 / 1	2509276061
Prochlorothrix hollandica PCC 9006	(1) 454 STD TIT, (2) 454 PE (5749 kb, 8122 kb)	830,913	198.7	(1) ILL STD	112,562,730	8,554,767,480	AF	233 / 13	2509276045
Pseudanabaena sp. PCC 6802	(1) 454 STD TIT, (2) 454 PE (4119 kb, 12050 kb)	1,300,658	271.9	(1) ILL STD	31,942,889	1,149,944,004	AF	28 / 6	2506783054
Pseudanabaena sp. PCC 7367	(0) 454 STD TIT, (1) 454 PE (10442 kb)	396,482	75.9	(1) ILL STD	82,635,242	6,280,278,392	MF	2 / 2	2504643012
Pseudanabaena sp. PCC 7429	(1) 454 STD TIT, (2) 454 PE (9299 kb, 3092 kb)	613,351	198.8	(1) ILL STD	83,683,990	6,359,983,240	AF	517 / 464	2504557005

Rivularia sp. PCC 7116	(1) 454 STD TIT, (3) 454 PE (3104 kb, 22486 kb, 22469 kb)	1,240,665	224	(1) ILL STD	47,209,745	1,699,550,820	MF	3 / 3	2510065008
Spirulina major PCC 6313	(1) 454 STD TIT, (1) 454 PE (5378 kb)	487,235	257.8	(1) ILL STD	87,627,634	6,659,700,184	AF	10 / 2	2506520014
Spirulina subsalsa PCC 9445	(1) 454 STD TIT, (1) 454 PE (13930 kb)	404,680	198	(1) ILL STD	61,669,554	4,686,886,104	AF	10 / 2	2506520011
Stanieria cyanosphaera PCC 7437	(0) 454 STD TIT, (1) 454 PE (7497 kb)	378,359	74.8	(1) ILL STD	86,083,820	6,542,370,320	MF	6 / 6	2503754019
Synechococcus sp. PCC 6312	(1) 454 STD TIT, (1) 454 PE (4402 kb)	823,816	251.4	(1) ILL STD	72,440,844	5,505,504,144	MF	2 / 2	2509276030
Synechococcus sp. PCC 7336	(1) 454 STD TIT, (2) 454 PE (4179 kb, 22856 kb)	949,313	199.2	(1) ILL STD	44,507,806	3,382,593,256	AF	9 / 2	2506520048
Synechococcus sp. PCC 7502	(1) 454 STD TIT, (3) 454 PE (1102 kb, 9022 kb, 9794 kb)	573,805	166.2	(1) ILL STD	86,633,080	6,150,948,680	MF	8 / 3	2508501041
Synechocystis sp. PCC 7509	-	-	-	(1) ILL STD	3,753,429	5,832,004,000	none	174 / 174	2517572074
Tolypothrix sp. PCC 9009	(0) 454 STD TIT, (1) 454 PE (7854 kb)	920,752	178.4	(1) ILL STD	72,204,518	5,487,543,368	AF	167 / 204	2504756053
Xenococcus sp. PCC 7305		-	-	(1) ILL STD	9,298,704	7,029,000,000	none	234 / 225	2508501034
Unidentified cyanobacterium PCC 7702	-	-	-	(1) ILL STD, (1) ILL PE	45,267,538	6,790,130,000	none	49 / 4	2512564012

Dataset S1. Metadata information for all cyanobacteria used in this study
This table is available separately as SI_DatasetS1xls.

Dataset S2. Distribution of protein orthologs involved in cell division and cell differentiation
A. Putative protein orthologs involved in cell division and morphogenesis B. Putative protein orthologs involved in cell differentiation. The 126 genomes were taken into account for the establishment of the core genes *(C)* for the cell division process, whereas only the organisms belonging to Subsections IV and V were considered to define the core genes for heterocyst differentiation. Seed proteins were downloaded from the cyanobase (http://genome.kazusa.or.jp/cyanobase) or used according to Campbell et al, 2003 (60), Lehner et al, 2011 (61), and Zhou et al, 2002 (62).

This table is available separately as SI_DatasetS2.xls.

Dataset S3. Predicted eukaryotic genes of cyanobacterial descent
This dataset is available separately as SI_DatasetS3.xls.

References:

1. Bennett S (2004) Solexa Ltd. *Pharmacogenomics* 5:433-438.
2. Margulies M*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.
3. Zerbino DR & Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821-829.
4. Ewing B & Green P (1998) Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res* 8:186-194.
5. Ewing B, Hillier L, Wendl MC, & Green P (1998) Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res* 8:175-185.
6. Gordon D, Abajian C, & Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Res* 8:195-202.
7. Han C & Chain P (2006) Finishing repeat regions automatically with Dupfinisher. *Proceeding of the 2006 international conference on bioinformatics & computational biology.*, eds Arabnia HR & Valafar H (CSREA Press), pp 141-146.
8. Gnerre S*, et al.* (2010) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* 108:1513-1518.
9. Hyatt D*, et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 11:119.
10. Pati A*, et al.* (2010) GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nat Methods* 7:455-457.
11. Lowe TM & Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25:955–964.
12. Pruesse E*, et al.* (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35:7188-7196.
13. Markowitz VM*, et al.* (2009) IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics* 25:2271-2278.
14. Wu M & Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* 9:R151.
15. Katoh K, Kuma K-i, Toh H, & Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33:511-518.
16. Eddy SR (1998) Profile hidden markov models. *Bioinformatics* 14:755-763.
17. Sonnhammer E & Hollich V (2005) Scoredist: A simple and robust protein sequence distance estimator. *BMC Bioinformatics* 6:108.
18. Guindon S*, et al.* (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Sys Biol* 59:307-321.
19. Abascal F, Zardoya R, & Posada D (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21:2104-2105.
20. Colless DH (1982) Phylogenetics: The theory and practice of phylogenetic systematics. *Syst Zool* 31:100-104.
21. Maddison WP & Maddison DR (2011) Mesquite: A modular system for evolutionary analysis. *Version 2.75* http://mesquiteproject.org.

22. Lima T*, et al.* (2009) HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res* 37:D471-D478.

23. Grissa I, Vergnaud G, & Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* 35:W52-W57.

24. Bland C*, et al.* (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 8:1-8.

25. Rouhiainen L*, et al.* (2004) Genes Coding for hepatotoxic heptapeptides (Microcystins) in the cyanobacterium *Anabaena* strain 90. *Appl Environ Microbiol* 70:686-692.

26. Fewer D*, et al.* (2007) Recurrent adenylation domain replacement in the microcystin synthetase gene cluster. *BMC Evol Biol* 7:183.

27. Fewer DP*, et al.* (2011) Nostophycin biosynthesis is directed by a hybrid polyketide synthase-nonribosomal peptide synthetase in the toxic cyanobacterium *Nostoc* sp. strain 152. *Appl Environ Microbiol* 77:8034-8040.

28. Fan Q*, et al.* (2005) Clustered genes required for synthesis and deposition of envelope glycolipids in Anabaena sp. strain PCC 7120. *Mol Microbiol* 58:227-243.

29. Balskus EP & Walsh CT (2010) The genetic and molecular basis for sunscreen biosynthesis in cyanobacteria. *Science* 329:1653-1656.

30. Miller S, Wood A, Blankenship RE, Kim M, & Ferriera S (2011) Dynamics of gene duplication in the genomes of chlorophyll *d*-producing cyanobacteria: Implications for the ecological niche. *Genome Biol Evol* 3:601-613.

31. Swingley W*, et al.* (2008) Niche adaptation and genome expansion in the chlorophyll d-producing cyanobacterium *Acaryochloris marina*. *Proc Natl Acad Sci USA* 12:2005-2010.

32. Bench S, Ilikchyan I, Tripp H, & Zehr J (2011) Two strains of *Crocosphaera watsonii* with highly conserved genomes are distinguished by strain-specific features. *Front Microbiol* 2:261.

33. Shi T, Ilikchyan I, Rabouille S, & Zehr J (2010) Genome-wide analysis of diel gene expression in the unicellular N_2-fixing cyanobacterium *Crocosphaera watsonii* WH 8501. *ISME J* 4:621-632.

34. Welsh E*, et al.* (2008) The genome of *Cyanothece* 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle. *Proc Natl Acad Sci USA* 105:15094-15099.

35. Bandyopadhyay A*, et al.* (2011) Novel metabolic attributes of the genus *Cyanothece*, comprising a group of unicellular nitrogen-fixing cyanobacteria. *mBio* 2:e00214-00211. doi:00210.01128/mBio.00214-00211.

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

37. Kaneko T*, et al.* (2007) Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. *DNA Res* 14:247-256.

38. Frangeul L*, et al.* (2008) Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics* 9:274.

39. Kettler G, *et al.* (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* 3:e231. doi:210.1371/journal.pgen.0030231.
40. Coleman M, *et al.* (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311:1768-1770.
41. Rocap G, *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424:1042-1047.
42. Dufresne A, *et al.* (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci USA* 100:10020-10025.
43. Donia M, *et al.* (2011) Complex microbiome underlying secondary and primary metabolism in the tunicate-*Prochloron* symbiosis. *Proc Natl Acad Sci USA* 108:E1423-1432.
44. Sugita C, *et al.* (2007) Complete nucleotide sequence of the freshwater unicellular cyanobacterium *Synechococcus elongatus* PCC 6301 chromosome: gene content and organization. *Photosynth Res* 93:55-67.
45. Dufresne A, *et al.* (2008) Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol* 9:R90. doi:10.1186/gb-2008-1189-1185-r1190.
46. Palenik B, *et al.* (2006) Genome sequence of *Synechococcus* CC9311: Insights into adaptation to a coastal environment. *Proc Natl Acad Sci USA* 103:13555-13559.
47. Bhaya D, *et al.* (2007) Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *ISME J* 1:703-713.
48. Palenik B, *et al.* (2003) The genome of a motile marine *Synechococcus*. *Nature* 424:1037-1042.
49. Kaneko T, *et al.* (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC 6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* 3:109-136.
50. Nakamura Y, *et al.* (2002) Complete genome structure of the thermophilic cyanobacterium *Thermosynechococcus elongatus* BP-1. *DNA Res* 9:123-130.
51. Tripp H, *et al.* (2010) Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* 464:90-94.
52. Fujisawa T, *et al.* (2010) Genomic structure of an economically important cyanobacterium, *Arthrospira* (Spirulina) *platensis* NIES-39. *DNA Res* 17:85-103.
53. Janssen P, *et al.* (2010) Genome sequence of the edible cyanobacterium *Arthrospira* sp. PCC 8005. *J Bacteriol* 192:2465-2466.
54. Starkenburg S, *et al.* (2011) Genome of the cyanobacterium *Microcoleus vaginatus* FGP-2, a photosynthetic ecosystem engineer of arid land soil biocrusts worldwide. *J Bacteriol* 193:4569-4570.
55. Jones A, *et al.* (2011) Genomic insights into the physiology and ecology of the marine filamentous cyanobacterium *Lyngbya majuscula*. *Proc Natl Acad Sci USA* 108:8815-8820.

56. Méjean A*, et al.* (2010) The genome sequence of the cyanobacterium *Oscillatoria* sp. PCC 6506 reveals several gene clusters responsible for the biosynthesis of toxins and secondary metabolites. *J Bacteriol* 192:5264-5265.

57. Stucken K*, et al.* (2010) The smallest known genomes of multicellular and toxic cyanobacteria: comparison, minimal gene sets for linked traits and the evolutionary implications. *PLoS ONE* 5:e9235. doi:9210.1371/journal.pone.0009235.

58. Ran L*, et al.* (2010) Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. *PLoS ONE* 5:e11486. doi:11410.11371/journal.pgen.0030231.

59. Kaneko T*, et al.* (2001) Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res* 8:227-253.

60. Campbell EL, Wong FCY, & Meeks JC (2003) DNA binding properties of the HrmR protein of *Nostoc punctiforme* responsible for transcriptional regulation of genes involved in the differentiation of hormogonia. *Mol Microbiology* 47(2):573-582.

61. Lehner J*, et al.* (2011) The morphogene AmiC2 is pivotal for multicellular development in the cyanobacterium *Nostoc punctiforme*. *Mol Microbiol* 79:1655-1669.

62. Zhou R & Wolk C (2002) Identification of an akinete marker gene in *Anabaena variabilis*. *J Bacteriol* 184:2529-2532.