

1 **Proposal of a new genome-based taxonomy for Cyanobacteria**

2

3 Juline M. Walter<sup>1,2</sup>, Felipe H. Coutinho<sup>1,2</sup>, Bas E. Dutilh<sup>2,4</sup>, Fabiano L. Thompson<sup>1,3\*</sup>,  
4 and Cristiane C. Thompson<sup>1\*</sup>

5 <sup>1</sup>*Laboratory of Microbiology, Institute of Biology, Federal University of Rio de*  
6 *Janeiro (UFRJ), Rio de Janeiro, Brazil.*

7 <sup>2</sup>*Radboud University Medical Centre, Radboud Institute for Molecular Life Sciences,*  
8 *Centre for Molecular and Biomolecular Informatics (CMBI), Nijmegen, The*  
9 *Netherlands.*

10 <sup>3</sup>*Center of Technology - CT2, SAGE-COPPE, Federal University of Rio de Janeiro*  
11 *(UFRJ), Rio de Janeiro, Brazil.*

12 <sup>4</sup>*Theoretical Biology and Bioinformatics, Utrecht University, Utrecht, The*  
13 *Netherlands*

14

15 \*Corresponding Authors: Av. Carlos Chagas Filho 373. Sala 102, Bloco A,  
16 CCS/IB/BIOMAR, Lab. de Microbiologia, Cidade Universitária, Rio de Janeiro, RJ,  
17 Brasil, CEP 21941-599, T/F: +552139386567, E-mails:  
18 [fabianothompson1@gmail.com](mailto:fabianothompson1@gmail.com) (FLT); [thompsoncristiane@gmail.com](mailto:thompsoncristiane@gmail.com) (CCT).

19

20 **Short/ Running Title:** Genomic Taxonomy of Cyanobacteria

21 **Keywords:** Microbial Ecology; Genomic Taxonomy; Genome-based Microbial  
22 Taxonomy; Cyanobacteria.

23

## 24     **ABSTRACT**

25     Cyanobacteria are major contributors to global biogeochemical cycles. The genetic  
26     diversity among Cyanobacteria enables them to thrive across many habitats. However,  
27     the taxonomy of Cyanobacteria remains unstable because of an inadequate taxonomic  
28     classification system. Taxa within Cyanobacteria have historically been classified  
29     using morphological traits, which is inadequate to categorize the diversity within this  
30     phylum. The aim of this study was to propose a new taxonomic framework for  
31     Cyanobacteria using whole-genome-based taxonomic analysis including *in silico*  
32     Genome-to-Genome hybridization (GGH), Average Amino acid Identity (AAI),  
33     Average Nucleic acid Identity (ANI), phylogenetic reconstruction using a set of  
34     conserved marker genes (MLST), and 16S rRNA gene sequences. Applying these  
35     genomic signatures to a set of 100 cyanobacterial genomes allowed 86 species and 43  
36     genera to be identified, among which 32 species and 19 genera were found to be  
37     novel. By exploring changes in the relative abundances of the analyzed genomes  
38     throughout diverse marine and freshwater ecosystems, we determined the ecological  
39     niches occupied by these taxa, adding another level to our proposed taxonomic  
40     scheme.

41

## 42     **INTRODUCTION**

43     Earth is home to nearly one trillion ( $10^{12}$ ) microbial species that have evolved  
44     over approximately 4 billion years (Locey & Lennon 2016). Cyanobacteria emerged  
45     approximately 3 billion years ago, ushering Earth's transition from anoxygenic to  
46     oxygenic conditions through photosynthesis (Shih et al. 2013). Throughout their

47 evolution, Cyanobacteria became diverse and widely distributed among bacterial phyla.  
48 They are highly heterogeneous organisms comprising unicellular and multicellular,  
49 photosynthetic and non-photosynthetic, free-living and symbiotic organisms, with  
50 genomes sizes ranging from 1 to 10 Mb (Shih et al. 2013). A novel non-photosynthetic  
51 Cyanobacteria, named Melainabacteria, was recently discovered (Di Rienzi et al. 2013;  
52 Soo et al. 2014).

53 Cyanobacteria members (also known as the Cyanophyceae, Cyanophyta,  
54 cyanoprokaryota, blue-green algae or blue-green bacteria) share similar ecological  
55 niches and metabolic features with eukaryotic algae and have been named according to  
56 the Botanical Code (Kauff & Büdel 2010). The inclusion of Cyanobacteria in taxonomic  
57 schemes of Bacteria was only proposed in 1978 by Stanier and coworkers (Stanier et al.  
58 1978). This proposal has conflicted with the botanical nomenclature (Oren 2004; Oren  
59 & Garrity 2014). More than two decades were necessary to include Cyanobacteria in a  
60 Note to General Consideration 5 (1999) under the rules of the International Committee  
61 on Systematic Bacteriology (ICSB)/International Committee on Systematic of  
62 Prokaryotes (ICSP) (Labeda 2000; Tindall 1999; De Vos & Trüper 2000).  
63 Nomenclature has also been discussed by several authors without reaching a consensus  
64 (Hoffmann et al. 2005; Oren et al. 2009; Oren & Tindall 2005). As a result, more than  
65 50 genera of Cyanobacteria have been described since 2000, and many of them remain  
66 unrecognized in the List of Prokaryotic Names with Standing in Nomenclature, LPSN,  
67 <<http://www.bacterio.net>> (Parte 2014) or in databases (e.g., NCBI).

68 The traditional morphologic classification of Cyanobacteria is subjective (Boone  
69 & Castenholz 2001; Rippka et al. 1979). Molecular phylogenetic studies indicate that all

70 subsections and genera based on morphology are polyphyletic (Gugger & Hoffmann  
71 2004; Hugenholtz et al. 2016; Schirrmeyer et al. 2011). Molecular studies revealed that  
72 morphological and phylogenetic diversity do not match (Gugger & Hoffmann 2004;  
73 Hugenholtz et al. 2016). Several studies have analyzed the molecular diversity of  
74 Cyanobacteria, but the taxonomic structure of *Prochlorococcus* and *Synechococcus* has  
75 only recently been elucidated using genomics, even though they are the most abundant  
76 Cyanobacteria in marine environments and major contributors to global primary  
77 productivity (Coutinho et al. 2015; Coutinho et al. 2016; Thompson et al. 2013). A  
78 recent study proposed that there should be 170 Cyanobacteria genera based on 16S  
79 rRNA sequences only (Kozlov et al. 2016). Farrant et al. (2016) delineated 121  
80 *Prochlorococcus* and 15 *Synechococcus* ecologically significant taxonomic units  
81 (ESTUs) in the global ocean using single-copy *petB* sequences (encoding cytochrome  
82 *b<sub>6</sub>*) and environmental cues. Out of 136 ESTUs, only 11 (three for *Prochlorococcus* and  
83 eight for *Synechococcus*) were found to be abundant in the *Tara* Ocean metagenomes  
84 (Farrant et al. 2016).

85         The advent of new high-throughput sequencing technologies has generated large  
86 amounts of sequence data, and major breakthroughs regarding the diversity, evolution,  
87 and taxonomy of Cyanobacteria have occurred that previously could not have been  
88 achieved by analysis of morphology or phylogenetic markers alone (Coenye et al. 2005;  
89 Farrant et al. 2016; Gevers et al. 2005; Shih et al. 2013). Strains from the same species  
90 share at least 98.8% 16S rRNA gene sequence similarity, 95% of AAI and ANI, and  
91 70% *in silico* DDH (Konstantinidis & Tiedje 2005; Thompson et al. 2013). Strains of  
92 the same genus form monophyletic branches (Qin et al. 2014; Yarza et al. 2008).

93 However, the species and genera definitions have been applied to only a few  
94 cyanobacterial genera in practice.

95 Researchers are no longer convinced of the scientific founding of the  
96 Cyanobacteria taxonomy as it is practiced today, as whole genome sequence (WGS) has  
97 become standard practice (Garrity 2016; Hugenholtz et al. 2016; Thompson et al. 2015).  
98 Moreover, Cyanobacteria taxonomy does not address the growing gap between  
99 taxonomy and microbial ecology. The challenge to keep up with progress in  
100 environmental and evolutionary microbiology is contrasted with the needs of clinical  
101 microbiologists and epidemiologists.

102 The aim of the present study was to examine the taxonomic structure of  
103 Cyanobacteria using a genomics approach to taxonomy. We analyzed 100  
104 Cyanobacteria genomes by means of phylogenetic reconstruction (i.e., 16S rRNA and  
105 MLSA) and genomic signatures (Average Amino acid Identity (AAI) and Average  
106 Nucleotide Identity (ANI), and *in silico* DNA-DNA hybridization (DDH)). In this  
107 study, we propose the delineation of 86 species and 43 genera within Cyanobacteria,  
108 which were further classified according to their ecological traits as inferred through  
109 metagenomics.

110

## 111 **METHODS**

### 112 **Genome selection**

113 A total of 100 complete genomes publicly available as of January 2016 were  
114 retrieved from RefSeq, GenBank and Geba databases for analysis (Table 1). Draft  
115 genomes were not considered. Completeness of all genomes selected was assessed with

116 CheckM (Parks et al. 2015), and the genomes that were at least 90% complete were  
117 used for further analysis. All genomes used in this paper are listed in Table 1.

118

### 119 **Annotation and genomic taxonomy**

120 All genomes were annotated using Prokka version 1.11 (Seemann 2014), with  
121 default settings. Genomic similarity between the 100 strains was determined by  
122 Average Amino Acid Identity (AAI) (Konstantinidis & Tiedje 2005), Average  
123 Nucleotide Identity (ANI), and *in silico* DNA-DNA Hybridization (DDH) or Genome-  
124 to-Genome Distance (GGD) (Auch et al. 2010; Auch et al. 2010; Meier-Kolthoff et al.  
125 2013). GGD were calculated using the Genome-to-Genome Distance Calculator tool,  
126 version 2.1 (Meier-Kolthoff et al. 2013) (<http://ggdc.dsmz.de/>). Genomes were  
127 considered to belong to the same species if they displayed AAI and ANI ranging from  
128 95 to 100% and 70 to 100% GGD. To verify the taxonomic similarity, the genomic  
129 content GGD values were plotted in heatmaps, which were generated in R (R  
130 Development Core Team 2011), using the heatmap.2 {gplots} package.

131

### 132 **Phylogenetic analysis**

133 To establish the phylogenetic structure of the phylum Cyanobacteria,  
134 phylogenetic trees were constructed using the 16S rRNA gene sequences and the  
135 concatenated alignments of a set of conserved genes, most of which encode ribosomal  
136 proteins.

137 **Ribosomal RNA sequences.** The small subunit ribosomal RNA (16S rRNA)  
138 sequences from all cyanobacterial strains for which whole genome sequence data are

139 publically available (exception see below, thus  $N = 97$ ), as well as 16S rRNA gene  
140 sequences from additional type-strains available ( $N = 14$ ) were all analyzed. The  
141 sequences were retrieved from the ARB SILVA database (Pruesse et al. 2007; Quast et  
142 al. 2013). Whenever sequences were not available, they were retrieved directly from the  
143 genomes using RNAMMER 1.2 Server (Lagesen et al. 2007). Sequences were aligned  
144 through MUSCLE v. 3.8 (Edgar 2004), with default settings, and Gblocks 0.91b  
145 (Castresana 2000; Talavera & Castresana 2007) was used for alignment curation. Using  
146 MEGA 6 (Tamura et al. 2013), best-fitting nucleic acid substitution models were  
147 calculated through the MLModelTest feature. Models were ranked based on their  
148 Bayesian Information Criterion (BIC) scores as described by Tamura et al. (2013). The  
149 model with the lowest BIC score was selected and used for further phylogenetic  
150 analysis. The phylogenetic inference was obtained using the Maximum Likelihood  
151 method based on the Kimura 2 parameter method with the Gamma distributed rate  
152 variation (K2+G) as the nucleotide substitution model, which was estimated from the  
153 data. The support branches of tree topology were checked by 1,000 bootstrap replicates.  
154 The 16S rRNA gene alignments were used to estimate the degree of genetic distance  
155 between strains through the Tajima-Nei method.

156 *Gloeobacter violaceus* PCC 7421 was set as the outgroup in both trees. Trees  
157 were visualized with FigTree, version 1.4.2 (Rambaut 2015). Due to incomplete or  
158 partial sequences, *Synechococcus* sp. CB0101 was omitted from these analyses.  
159 *Planktothrix mougeotii* NIVA-CYA 405 as well as *Planktothrix prolifica* NIVA-CYA  
160 540 were not included in the phylogenetic analyses because 16S rRNA sequences are  
161 not currently available for these strains (and not retrievable from their genomes).

162 The type-strains of each taxa were included in the 16S phylogenetic tree to  
163 confirm the phylogenetic relatedness of the cyanobacterial genomes. Designations of  
164 type strain and type species were not available for *Chaemaesiphon minutus* PCC6605,  
165 *Pleurocapsa* sp. PCC7319, *Rivularia* sp. PCC7116, *Synechocystis* sp. PCC7509,  
166 *Trichodesmium erythraeum* IMS01, *Xenococcus* sp. PCC7305, cyanobacterium ESFC-  
167 1, and cyanobacterium JSC-12. *Geitlerinema* sp. PCC7105 is the reference strain for  
168 marine species of *Geitlerinema*, and PCC73106 is the reference strain for *Gloeocapsa*  
169 (Sarma 2012).

170

171 **Conserved marker genes: multilocus sequence analysis (MLSA).** A  
172 phylogenetic tree was generated using 31 conserved gene sequences previously  
173 validated as phylogenetic markers for (cyano)bacteria (Shih et al. 2013; Wu & Eisen  
174 2008). The sequences of these proteins were mined using the *AutoMated Phylogenomic*  
175 *infeRence Application* - AMPHORA2 tool (Wu & Scott 2012), through default settings  
176 for the Bacteria option, and with a cut-off value of  $1.e^{-10}$ . Individual alignments were  
177 performed for each of the 31 gene sets through MUSCLE v. 3.8 with default settings  
178 (Edgar 2004). All alignments were then concatenated. Only OTUs with complete  
179 sequences for all 31 genes were used in the phylogenetic analysis. A Maximum  
180 Likelihood tree was constructed using RaxML v. 7 (Stamatakis 2006) and the  
181 Dayhoff+G likelihood model. One thousand bootstrap replications were calculated to  
182 evaluate the relative support of the branches. Trees were visualized with FigTree,  
183 version 1.4.2 (Rambaut 2015).

184

## 185 **Abundance of Cyanobacterial genomes across aquatic environments**

186 Marine and freshwater metagenomes were retrieved to determine the abundance of  
187 100 cyanobacterial genomes across the Earth. A set of 191 marine metagenomes from  
188 the *Tara* Ocean project were retrieved for analysis along with their associated metadata  
189 (Sunagawa et al. 2015). Eight freshwater metagenomes were retrieved for analysis from  
190 the Caatinga biome microbial community project along with their associated metadata  
191 (Lopes et al. 2016).

192 Metagenome reads were mapped to a database containing the 100 analyzed  
193 cyanobacterial genomes through Bowtie2 (Langmead & Salzberg 2012) using *-very-*  
194 *sensitive-local* and *-a* options. Abundance of genomes across samples was calculated  
195 based on the number of mapped reads as described (Iverson et al. 2012). Next,  
196 Spearman correlation coefficients (R, or Spearman's rho) were calculated for the  
197 abundance of each genome across samples and the levels of measured environmental  
198 parameters. Metagenomes were compared based on the relative abundances of the 100  
199 analyzed genomes within them using non-metric multidimensional scaling (NMDS).

200

## 201 **RESULTS**

### 202 **Phylogenetic reconstruction using 16S rRNA and conserved protein genes**

203 The 16S rRNA phylogenetic tree was reconstructed based on the ML method  
204 (Figure 1). The 16S rRNA topology showed incongruences on the taxonomy of  
205 Cyanobacteria. The MLSA tree (Figure S1) based on conserved genes sequences gave a  
206 higher resolution than the 16S rRNA phylogenetic analysis. In total, 43 branches (<

207 90% 16S rRNA, and < 90% MLSA sequences similarity) corresponding to genera were  
208 recognized.

209

### 210 **Genomic taxonomy: cyanobacterial species and genus delimitation**

211 A total of 43 genera (19 new genera) and 86 species (32 new species) were  
212 delineated based on the phylogenetic reconstructions (Figures 1, 2, and S1) and AAI,  
213 ANI and DDH analyses (Figure 3, Supplemental Information Tables S1 and S2). The  
214 genus and species cut-off delimitation were > 65% and > 95% AAI similarity,  
215 respectively. Forty-three genera were distinguished, among which 19 are new.

216 In total, 47 genera and 40 species were taxonomically reclassified and/or re-named.  
217 From a total of the 100 genomes used in this study, 69 were previously classified to the  
218 species level, whereas the remaining 31 had incomplete taxonomic classification (i.e.,  
219 only sp. or unclassified). Thus, we found that 61% (N = 61) of all analyzed genomes  
220 required reassignment at one or more ranks to reconcile existing taxonomic  
221 classifications with our genomic taxonomy (Figures 2 and S1). Over the next section,  
222 we highlight four specific cases to exemplify taxonomic issues that could be resolved  
223 through our genome-driven approach. These cases illustrate how the use of genomic  
224 data is fundamental for the classification of Cyanobacteria.

225 **Case I.** *Oscillatoria* group. Analysis of the five genomes of *Oscillatoria*  
226 distinguished three genera, based on the genomic signatures (i.e., GGDH, AAI, 16S,  
227 and MLSA): i. *Oscillatoria acuminata* PCC 6304 type strain formed a separate group;  
228 ii. *Oscillatoria* PCC 10802 formed a separate divergent group, corresponding to a new  
229 genus named *Somacatellium* (*S. hydroxylic* PCC 10802); and iii. *Oscillatoria* strains

230 PCC 7112, PCC 6407, PCC 6506 and *Microcoleus vaginatus* FGP-2 formed a new  
231 genus named *Toxinema* (*T. nigroviridis* PCC 7112, *T. oscillati* PCC 6407, *T. oscillati*  
232 PCC 6506, and *T. vaginatus* FGP-2).

233 **Case II.** *Leptolyngbya* group. Examination of the genomic signatures (i.e.,  
234 GGDH, AAI, 16S, and MLSA) of the five *Leptolyngbya* strains showed that i. the  
235 strains were polyphyletic, forming different phylogenetic branches. *Leptolyngbya*  
236 *boryana* PCC 6306 type strain forms a separate group with cyanobacterium JSC-12,  
237 while the rest of the *Leptolyngbya* strains cluster apart; ii. strain PCC 7376 forms a  
238 group along with *Synechococcus euryhalinus* PCC 7002, representing a new genus  
239 *Symphothece* (*S. fragile* PCC 7376 and *S. euryhalinus* PCC 7002); iii. strain PCC 7375  
240 forms a new genus named *Acrophormium* (*A. splendidus* PCC 7375); iv. strain PCC  
241 7104 forms a new genus named *Allonema* (*A. longislandicus* PCC 7104); and v. strains  
242 PCC 6406 and BDU (*Lyngbya confervoides*) form a new genus named *Euryforis* (*E.*  
243 *eilemai* PCC 6406 and *E. confervoides* BDU).

244 **Case III.** *Arthrospira* group. Examination of the genomic signatures (i.e.,  
245 GGDH, AAI, 16S, and MLSA) of the five *Arthrospira* strains indicated that i. *A.*  
246 *platensis* C1 should be considered a new species, named *A. sesilensis*; ii. strain 8005  
247 belongs to a new species, named *A. nitrilium*; and iii. the type strain of *Arthrospira*  
248 *platensis* (PCC 7345) formed a tight cluster with NIES-39 and Paraca.

249 **Case IV.** *Synechococcus* group. Examination of the genomic signatures (i.e.,  
250 GGDH, AAI, 16S, and MLSA) of the nine *Synechococcus* strains indicated that i. *S.*  
251 *elongatus* PCC 6301 type strain forms a separate group with *S. elongatus* PCC 7942; ii.  
252 strain PCC 6312 forms a new genus named *Stenotopis* (*S. californii* PCC 6312); iii.

253 strain PCC 7335 belongs to a new genus named *Coccusdissimilis* (*C. mexicanus* PCC  
254 7335); iv. strains JA23Ba213 and JA33Ab formed a new genus named *Leptococcus* (*L.*  
255 *springii* JA23Ba213 and *L. yellostonii* JA33Ab); v. strain PCC 7336 formed a new  
256 genus named *Eurycoccus* (*E. berkleyi* PCC 7336); vi. strain PCC 7502 belonged to a  
257 new genus named *Leptovivax* (*L. bogii* PCC 7502); and vii. strain PCC 7002 formed a  
258 tight group with strain PCC 7376 (*Leptolyngbya* sp.), representing a new genus named  
259 *Symphothece* (*S. euryhalinus* PCC 7002 and *S. fragile* PCC 7376).

260

### 261 **Delineating ecological niches of Cyanobacteria**

262 Our genomic analysis was complemented by an ecological characterization of the  
263 analyzed strains. Correlating the relative abundances of the analyzed genomes with  
264 environmental parameters measured at *Tara* Oceans samples (Sunagawa et al, 2015)  
265 revealed associations between Cyanobacteria and physical, chemical and biological  
266 variables of their habitat (Figure 4). This allowed us to classify these organisms into  
267 three major ecological groups (Eco\_A, Eco\_B and Eco\_C) (Figure 5) according to their  
268 response to changing environmental conditions, which relates to their preferred  
269 ecological niches. Interestingly, the phylogeny of these Cyanobacteria does not seem to  
270 be directly associated with their niche. Often closely related organisms (i.e., of the same  
271 genus) show drastically distinct associations with environmental parameters, leading  
272 them to be assigned to different ecological clusters.

273 Members of Eco\_A were characterized by positive correlations with the  
274 concentration of nitrogen and phosphorus; weak positive correlations with minimum  
275 generation time, silicate and depth; and by negative correlations with temperature,

276 microbial cell abundance, oxygen availability, and salinity. Meanwhile, members of  
277 Eco\_B were characterized by strong positive correlations with the concentration of  
278 nitrogen and phosphorus; positive correlations (stronger than those presented by  
279 Eco\_A) with minimum generation time, silicate and depth; and by negative correlations  
280 (also stronger than those presented by Eco\_A) with temperature, microbial cell  
281 abundance (in particular with autotroph cell density), oxygen availability, and salinity.  
282 As suggested by correlation analyzes (Figure 4), NMDS revealed Eco\_B to be more  
283 abundant in cold and eutrophic waters, while Eco\_C exhibited the opposite pattern and  
284 was more abundant in warm and oligotrophic environments (Figures 6 and 7). In turn,  
285 Eco\_A was more abundant at intermediate conditions between these polar opposites and  
286 was shown to be more abundant in samples with higher cell densities and NO<sub>2</sub>  
287 concentrations (Figures 6 and 7). Finally, members of Eco\_C were characterized by  
288 negative correlations with the concentration of nitrogen and phosphorus and positive  
289 correlations with temperature and autotroph cell abundance. NMDS of freshwater  
290 samples suggested that Eco\_C has a preference for habitats with higher concentrations  
291 of POC, phosphorus, ammonia and nitrate, while Eco\_A displayed a preference for  
292 higher pH and DOC, nitrite and total nitrogen concentrations (Figures S3 and S4). We  
293 also investigated the abundance of the ecological clusters in freshwater environments.  
294 Unfortunately, there is no currently available large-scale dataset of freshwater  
295 metagenomes with associated metadata (comparable to the *Tara* Oceans dataset).  
296 Therefore, we could not define freshwater ecological clusters and chose to extrapolate  
297 the classification obtained from the analyses of the marine dataset. In freshwater

298 metagenomes, Eco\_B was the dominant group in all of the analyzed samples (Figure  
299 S2).

300

## 301 **DISCUSSION**

302 Here, we produce an ecological map that spans members of the phylum  
303 Cyanobacteria, coupled with a novel taxonomic classification. This approach allows us  
304 to place the novel classification scheme into a meaningful ecological framework,  
305 linking taxonomy with environmental function. The biogeographical distribution  
306 profiles of *Prochlorococcus*, *Parasynechococcus*, *Prolificoccus* and *Expansiococcus*  
307 strains throughout aquatic ecosystems has been extensively researched, revealing the  
308 diversity of ecological niches among these closely related lineages (Farrant et al. 2016;  
309 Johnson et al. 2006; Kashtan et al. 2014; Pittera et al. 2014; Rocap et al. 2003; Scanlan  
310 2003; Scanlan & West 2002).

311

## 312 **WGS as the basic unit for Cyanobacteria genomic taxonomy (CGT)**

313 New Cyanobacteria taxonomy comes with the acknowledgement that WGS are  
314 the basic unit for CGT. This scenario configures the Renaissance of the Cyanobacteria  
315 taxonomy. Comparative genomics allows for identification of sequence clusters with  
316 high genotypic resolution based on variation in protein coding genes distributed across  
317 the genomes. Analyses of environmental metagenomes and microbiomes have shown  
318 that microbial communities consist of genotypic clusters of closely related organisms  
319 (Farrant et al. 2016). These clusters display cohesive environmental associations and  
320 dynamics that differentiate them from other clusters co-existing in the same samples. In

321 light of new concepts, restlessness is mounting with the inability to define the microbial  
322 species itself. Evolution studies on closely related bacteria show rapid and highly  
323 variable gene fluxes in evolving microbial genomes, suggesting that extensive gene loss  
324 and horizontal gene transfer leading to innovation are the dominant evolutionary  
325 processes (Batut et al. 2014; Puigbò et al. 2014). CGT will solve the problem of the  
326 frequent observation that even closely related genomes can have high gene content  
327 variation that gives phenotypic variation. CGT is completely adjusting to the genomics  
328 era, addressing the needs of its users in microbial ecology and clinical microbiology, in  
329 a new paradigm of open access (Beiko 2015). CGT will provide a predictive operational  
330 framework for reliable automated and openly available identification and classification  
331 (Thompson et al. 2015).

332

### 333 **Proposals for cyanobacterial taxonomy**

334 A main gap exists and is growing each day between the formal taxonomy of  
335 cyanobacteria and the forest of acronyms and numbers in the different databases.  
336 Indeed, the nameless OTUs, strains, isolates and WGS sequences (Beiko 2015; Kozlov  
337 et al. 2016) form the great majority of data in private and public databases. There is a  
338 need to re-examine the Cyanobacteria prokaryote species, taking into account all  
339 recently developed concepts, e.g., the gene flow unit, OTU, ETUs and CTU in the  
340 context of a pragmatic genome-based taxonomic scheme. The type can be a culture,  
341 DNA or a WGS. The CGT system should maintain all of the existing information,  
342 integrating it with new data on DNA, genomes, isolates/strains, cultured and uncultured,  
343 “Candidatus” cases and reconstructed genomes from metagenomes (Brown et al. 2015;

344 Hugerth et al. 2015). The international initiatives of Geba is currently working on  
345 determining the WGS of all type strains of known microbial species to shorten this gap  
346 (more than eleven thousand genomes).

347 High-throughput genome sequence technologies (e.g., Illumina) are causing a  
348 revolution in microbial diversity studies. Through the process of binning of contigs or  
349 scaffolds derived from the same strain, complete genomes can be reconstructed. Recent  
350 studies have obtained dozens of new metagenome-assembled genomes from complex  
351 environmental samples (Almstrand et al. 2016; Brown et al. 2015; Haroon et al. 2016;  
352 Hugerth et al. 2015; Pinto et al. 2016). The abundance of these genomes across different  
353 environments can now be inferred from metagenomics, including their metabolic and  
354 ecological potential. It is clear that a new system is required to allow for precise  
355 taxonomic identification of these new genomes.

356

## 357 **CONCLUSIONS**

358 The present study introduces a main and a new informative, reproducible, and  
359 digital taxonomic framework for Cyanobacteria to facilitate ecological studies. New  
360 isolates can be identified and classified using this framework. Future studies should aim  
361 to determine the evolutionary and physiological basis for niche occupancy of known  
362 and newly discovered Cyanobacteria, as well as linking their niche to other important  
363 ecological variables not analyzed here such as phage susceptibility, light utilization  
364 strategies and inter-species interactions.

365

## 366 **Formal description of new genera and species**

367 **Description of *Pseudogeitlerinema* gen. nov.**

368 *Pseudogeitlerinema* (Pseudo.gei.tle.ri.ne.ma. Gr. adj. *pseud*, false; *geitlerinema*, other  
369 cyanobacteria genus [cf. genus *Geitlerinema*]; M. L. fem. n. *Pseudogeitlerinema*).

370

371 Description of *Pseudogeitlerinema shalloid* (formerly known as *Geitlerinema* sp.)  
372 (shal.loid. Gr. *-oid*, resembling, having the appearance of; *shalloid* referring to the  
373 appearance of very shallow adjacent cells): This species is characterized by  $\beta$ -  
374 carboxysome. Type strain is PCC 7407. The genome of this strain contains 4,68 Mbp  
375 (GC = 58,46%) harboring 3,727 coding DNA sequences.

376

377 **Description of *Somacatellium* gen. nov.**

378 *Somacatellium* (Som.a.ca.te.llium. Gr. n. *soma* (*somatiko*), body; L. fem. n. *catella*, a  
379 small chain; *Somacatellium* referring to a small chain filamentous body).

380

381 *Somacatellium hydroxylic* (formerly known as *Oscillatoria* sp.) (hy.drox.y.lic. Gr. *-ic*  
382 (*ikos*), relating to or having some characteristic of; *hydroxylic* referring to the hydroxyl-  
383 proline and hydroxyl aspartic acid compounds the new cinnamycin variant found in this  
384 strain): This species is characterized by  $\beta$ -carboxysome. Type strain is PCC 10802. The  
385 genome of this strain contains 8,59 Mbp (GC = 54,1%) harboring 7,012 coding DNA  
386 sequences.

387

388 **Description of *Tapinonema* gen. nov.**

389 *Tapinonema* (*ta.pi.no.nema*. Gr. *tapino* small, modest, weak; Gr. n. *nema*, thread;  
390 *Tapinonema*, small filament).

391

392 Description of *Tapinonema colecalium* (formerly known as cyanobacterium)  
393 (*col.e.ca.lium*. L. *cole*, inhabit; Gr. *calo*, beautiful; *-ium*, quality or relationship;  
394 *colecalium*, beautiful inhabitant): This species is characterized by  $\beta$ -carboxysome. Type  
395 strain is JSC-12, isolated from freshwater habitat. The genome of this strain contains  
396 5,52 Mbp (GC = 47,49%) harboring 5,024 coding DNA sequences.

397

398 **Description of *Toxinema* gen. nov.**

399 *Toxinema* (*tox.i.nema*. N.L. n. *toxi* toxins related to the strains; Gr. neut. n. *nema* thread,  
400 filament; N.L. neut. n. *Toxinema* a filament with release toxins).

401

402 Description of *Toxinema nigroviridis* comb. nov. (formerly known as *Oscillatoria*  
403 *nigroviridis*) (*nig.ro.vi.ri.dis*. Gr. adj. *nigro*, distinctus; Gr. adj. *viridis*, campus, a green  
404 field; *Oscillatoria* is other genus of cyanobacteria [cf. genus *Oscillatoria*): This species  
405 is characterized by  $\beta$ -carboxysome. Type strain is PCC7112, isolated from a soil habitat  
406 in USA. The genome of this strain contains 7,47 Mbp (GC = 45,87%) harboring 6,925  
407 coding DNA sequences.

408

409 Description of *Toxinema vaginatus* comb. nov. (formerly known as *Microcoleus*  
410 *vaginatus*) [cf. genus species *Microcoleus vaginatus*): This species is characterized by  
411  $\beta$ -carboxysome. Type strain is FGP-2, isolated from soil habitat in Canyonlands

412 National Park, UT, USA. The genome of this strain contains 6,69 Mbp (GC = 46,04%)  
413 harboring 5,519 coding DNA sequences.

414

415 Description of *Toxinema oscillati* (formerly known as *Oscillatoria* sp.) (os.ci.lla.ti. Gr.  
416 *oscillare*, to swing, *Oscillatoria* is other genus of cyanobacteria [cf. genus  
417 *Oscillatoria*): This species is characterized by  $\beta$ -carboxysome. Type strain is PCC  
418 6407, isolated from freshwater habitat. The genome of this strain contains 6,89 Mbp  
419 (GC = 43,43%) harboring 5,693 coding DNA sequences.

420

421 **Description of *Stenotopis* gen. nov.**

422 *Stenotopis* (Ste.no.to.pis. Gr. adj. *stenos*, narrow; Gr. n. *topos*, place; *Stenotopis*  
423 referring a living within a narrow range of places).

424

425 Description of *Stenotopis californii* comb. nov. (formerly known as *Synechococcus*  
426 *californii*) (Ste.no.to.pis. Gr. adj. *stenos*, narrow; Gr. n. *topos*, place; *Stenotopis*  
427 referring a living within a narrow range of places): This species is characterized by  $\beta$ -  
428 carboxysome. Type strain is PCC 6312, isolated from freshwater habitat in California,  
429 USA. The genome of this strain contains 3.72 Mbp (GC = 48.49%) harboring 3,795  
430 coding DNA sequences.

431

432 **Description of *Euryforis* gen. nov.**

433 *Euryforis* (Eur.y.fo.rys. Gr. masc. n. *eury*, broad; N.L. masc. n. *foris*, outside; N.L.  
434 masc. n. *Euryforis* a *habitat* widespread genus).

435

436 Description of *Euryforis confervoides* comb. nov. (formerly known as *Lyngbya*  
437 *confervoides*) [cf. species *L. confervoides*]: This species is characterized by  $\beta$ -  
438 carboxysome. Type strain is BDU, isolated from a NA habitat in NA. The genome of  
439 this strain contains 8,79 Mbp (GC = 55,63%) harboring 8,370 coding DNA sequences.

440

441 Description of *Euryforis eilemai* (formerly known as *Leptolyngbya* sp.) (ei.le.mai. N.L.  
442 gen. n. *eilemai*, referring to the presence of sheath): This species is characterized by  $\beta$ -  
443 carboxysome. Type strain is PCC 6406, isolated from freshwater habitat in California,  
444 USA. The genome of this strain contains 5,77 Mbp (GC = 55,18%) harboring 5,156  
445 coding DNA sequences.

446

447 **Description of *Allonema* gen. nov.**

448 *Allonema* (All.o.nema. Gr. adj. *allos*, other; Gr. n. *nema*, thread; M. L. fem. n.  
449 *Allonema*, other filament).

450

451 *Allonema longislandicus* (formerly known as *Leptolyngbya* sp.) (long.is.lan.di.cus.  
452 referring to the habitat collected, Long Island, NY, USA): This species is characterized  
453 by  $\beta$ -carboxysome. Type strain is PCC 7104, isolated from marine habitat rockshore,  
454 Montauk Point, Long Island, NY, USA. The genome of this strain contains 6,89 Mbp  
455 (GC = 57,69%) harboring 6,414 coding DNA sequences.

456

457 **Description of *Coccusdissimilis* gen. nov.**

458 *Coccusdissimilis* (Co.ccus.di.ssi.mi.lis. Gr. n. *kokkos*, grain or kernel; L. adj. *dissimilis*,  
459 different; *Coccusdissimilis*, a different coccus).

460

461 Description of *Coccusdissimilis mexicanus* comb. nov. (formerly known as  
462 *Synechococcus mexicanus*) (*coccusdissimilis*. Gr. n. *kokkos*, grain or kernel; L. adj.  
463 *dissimilis*, different; *Coccusdissimilis*, a different coccus): This species is characterized  
464 by  $\beta$ -carboxysome. Type strain is PCC 7335, isolated from marine habitat in Snail shell,  
465 intertidal zone, Puerto Penasco, Mexico. The genome of this strain contains 5.97 Mbp  
466 (GC = 48.20%) harboring 5,702 coding DNA sequences.

467

468 **Description of *Cyclospexia* gen. nov.**

469 *Cyclospexia* (Cy.clos.pex.ia. Gr. n. *kyklos*, cycle; Gr. *pexia* (adj. *-pectic*), fixation; n.  
470 *Cyclospexia* involved in N<sub>2</sub> fixation cycle).

471

472 *Cyclospexia valenium* (formerly known as cyanobacterium) (*val.e.nium*. L. *valens*,  
473 strong; L. suff. *-icus -a -um*, suffix used with the sense of pertaining to; *valenium*  
474 referring to an organism restricted to few types of environmental conditions, such as  
475 extremophylic habitat): This species is characterized by  $\beta$ -carboxysome. Type strain is  
476 ESFC-1, isolated from marine habitat in extremophylic mat communities, Elkhorn  
477 Slough estuary, CA, USA. The genome of this strain contains 5.62 Mbp (GC = 46.51%)  
478 harboring 4,857 coding DNA sequences.

479

480 **Description of *Paraspirulina* gen. nov.**

481 *Paraspirulina* (Par.as.pi.ru.li.na. Gr. adj. *para*, to bring forth, to bear, alongside;  
482 *Spirulina*, other genus of cyanobacteria [cf. genus *Spirulina*]; M.L. fem. n.  
483 *Paraspirulina*).

484

485 Description of *Paraspirulina subsalsa* comb. nov. (formerly known as *Spirulina*  
486 *subsalsa*) [c.f. *S. subsalsa*]: This species is characterized by  $\beta$ -carboxysome. Type strain  
487 is PCC 9445. The genome of this strain contains 5,32 Mbp (GC = 47,39%) harboring  
488 4,580 coding DNA sequences.

489

490 **Description of *Symphothece* gen. nov.**

491 *Symphothece* (Sym.pho.the.ce. Gr. *sympho*, to grow together; Gr. fem. n. *theke*, case,  
492 envelope; *Symphothece*, cells-box that growing together).

493

494 Description of *Symphothece euryhalinus* comb. nov. (formerly known as *Synechococcus*  
495 *euryhalinus*) [c.f. *S. euryhalinus*]: This species is characterized by  $\beta$ -carboxysome. Type  
496 strain is PCC 7002. The genome of this strain contains 3.41 Mbp (GC = 49.16%)  
497 harboring 3,121 coding DNA sequences.

498

499 Description of *Symphothece fragile* (formerly known as *Leptolyngbya* sp.) (fra.gi.le.  
500 *fragile*, sensitive, delicate): This species is characterized by  $\beta$ -carboxysome. Type strain  
501 is PCC 7376, isolated from marine habitat in Limestone, Crystal Cave, Bermuda. The  
502 genome of this strain contains 5.12 Mbp (GC = 43.87%) harboring 4,601 coding DNA  
503 sequences.

504

505 **Description of *Leptococcus* gen. nov.**

506 *Leptococcus* (Lept.o.co.ccus. Gr. adj. *leptos*, delicate or thin; N.L. masc. n. *coccus* [from  
507 Gr. Masc. n. *kokkos*, grain, seed, kernel]; N.L. masc. n. *Leptococcus*, a delicate coccus).

508

509 Description of *Leptococcus springii* com. nov. (formerly known as *Synechococcus*  
510 *springii*) [c.f. *S. springii*]: This species is characterized by  $\beta$ -carboxysome. Type strain  
511 is JA23Ba213, isolated from freshwater habitat in Octopus Spring, Yellowstone Park,  
512 USA. The genome of this strain contains 3.05 Mbp (GC = 58.50%) harboring 3,064  
513 coding DNA sequences.

514

515 Description of *Leptococcus yellowstonii* comb. nov. (formerly known as *Synechococcus*  
516 *yellowstonii*) [c.f. *S. yellowstonii*]: This species is characterized by  $\beta$ -carboxysome.  
517 Type strain is JA33Ab, isolated from freshwater habitat in Octopus Spring, Yellowstone  
518 Park, USA. The genome of this strain contains 2.93 Mbp (GC = 60.20%) harboring  
519 3,036 coding DNA sequences.

520

521 **Description of *Eurycoccus* gen. nov.**

522 *Eurycoccus* (Eur.y.co.ccus. Gr. adj. *eury*, wide, broad; N.L. masc. n. *coccus* [from Gr.  
523 Masc. n. *kokkos*, grain, seed, kernel]; N.L. masc. n. *Eurycoccus*, a wide coccus).

524

525 Description of *Eurycoccus berkleyi* comb. nov. (formerly known as *Synechococcus*  
526 *berkleyi*) [c.f. *S. berkleyi*]: This species is characterized by  $\beta$ -carboxysome. Type strain

527 is PCC 7336, isolated from marine habitat in Sea Water Tank, Berkeley University, CA,  
528 USA. The genome of this strain contains 5.07 Mbp (GC = 53.70%) harboring 5,093  
529 coding DNA sequences.

530

531 **Description of *Paraleptovivax* gen. nov.**

532 *Paraleptovivax* (Pa.ra.lep.to.vi.vax. *para*, to bring forth, to bear, alongside; Gr. adj.  
533 *leptos*, delicate or thin; L. adj. *vivax*, long-lived, tenacious of life; *Leptovivax*, delicate  
534 long-lived).

535

536 Description of *Paraleptovivax allomegium* (formerly known as *Pseudoanabaena* sp.)  
537 (all.o.me.gium. Gr. adj. *allos*, other; Gr. nom. neut. adj. *mega*, big; Gr. *-ium*, quality or  
538 relationship): This species is characterized by  $\beta$ -carboxysome. Type strain is PCC 6802,  
539 isolated from freshwater habitat in California, USA. The genome of this strain contains  
540 5.62 Mbp (GC = 47.83%) harboring 5,363 coding DNA sequences.

541

542 **Description of *Leptovivax* gen. nov.**

543 *Leptovivax* (Lept.o.vi.vax. Gr. adj. *leptos*, delicate or thin; L. adj. *vivax*, long-lived,  
544 tenacious of life; *Leptovivax*, delicate long-lived).

545

546 Description of *Leptovivax bogii* (formerly known as *Synechococcus bogii*) [c.f. *S.*  
547 *bogii*]: This species is characterized by  $\beta$ -carboxysome. Type strain is PCC 7502,  
548 isolated from a sphagnum bog habitat. The genome of this strain contains 3.58 Mbp  
549 (GC = 40.60%) harboring 3,703 coding DNA sequences.

550

551 **Description of *Calotaxis* gen. nov.**

552 *Calotaxis* (Ca.lo.ta.xis. Gr. *calo*, beautiful; Gr. *taxis*, arrangement; *Calotaxis*, beautiful  
553 arrangement of the filament).

554

555 Description of *Calotaxis gracile* (formerly known as *Pseudoanabaena* sp.) (gra.ci.le, L.  
556 neut. adj. *gracile*, slender): This species is characterized by  $\beta$ -carboxysome. Type strain  
557 is PCC 7367, isolated from marine habitat in intertidal zone, Mexico. The genome of  
558 this strain contains 4.55 Mbp (GC = 46.31%) harboring 3,960 coding DNA sequences.

559

560 **Description of *Acrophormium* gen. nov.**

561 *Acrophormium* (A.cro.phor.mium. Gr. *acro*, high, top; Gr. *phorm*, basket, mat; *-um*,  
562 suffix used with the sense of pertaining to; *Acrophormium*, living on the top of the mat).

563

564 Description of *Acrophormium splendidus* (formerly known as *Leptolyngbya* sp.)  
565 (splen.di.dus. L. adj. *splendidus*, brilliant): This species is characterized by  $\beta$ -  
566 carboxysome. Type strain is PCC 7375, isolated from marine habitat in Woods Hole,  
567 Massachusetts, USA. The genome of this strain contains 9.42 Mbp (GC = 47.62%)  
568 harboring 8,366 coding DNA sequences.

569

570 **Description of *Eurycolium* gen. nov.**

571 *Eurycolium* (Eur.y.co.lium. Gr. adj. *eury*, wide, broad; L. *cole*, inhabit; Gr. *-ium*,  
572 quality or relationship, *Eurycolium* referring to the spread inhabiting trait in marine  
573 habitats).

574

575 Description of *Eurycolium pastoris* (formerly known as *Prochlorococcus pastoris*) [c.f.  
576 *P. pastoris*]: This species is characterized by  $\alpha$ -carboxysome. Type strain is CCMP  
577 1986 (or MED4), isolated from marine habitat. The genome of this strain contains 1,65  
578 Mbp (GC = 30,8%) harboring 1,777 coding DNA sequences.

579

580 Description of *Eurycolium tetisii* (formerly known as *Prochlorococcus tetisii*) [c.f. *P.*  
581 *tetisii*]: This species is characterized by  $\alpha$ -carboxysome. Type strain is MIT 9515,  
582 isolated from marine habitat. The genome of this strain contains 1,7 Mbp (GC =  
583 30,79%) harboring 1,784 coding DNA sequences.

584

585 Description of *Eurycolium neptunius* (formerly known as *Prochlorococcus neptunius*)  
586 [c.f. *P. neptunius*]: This species is characterized by  $\alpha$ -carboxysome. Type strain is MIT  
587 9312, isolated from marine habitat. The genome of this strain contains 1,7 Mbp (GC =  
588 31,21%) harboring 1,815 coding DNA sequences.

589

590 Description of *Eurycolium ponticus* (formerly known as *Prochlorococcus ponticus*) [c.f.  
591 *P. ponticus*]: This species is characterized by  $\alpha$ -carboxysome. Type strain is MIT 9301,  
592 isolated from marine habitat. The genome of this strain contains 1,64 Mbp (GC =  
593 31,34%) harboring 1,774 coding DNA sequences.

594

595 Description of *Eurycolium nereus* (formerly known as *Prochlorococcus nereus*) [c.f. *P.*  
596 *nereus*]: This species is characterized by  $\alpha$ -carboxysome. Type strain is MIT 9202,  
597 isolated from marine habitat. The genome of this strain contains 1,69 Mbp (GC =  
598 31,1%) harboring 1,795 coding DNA sequences.

599

600 Description of *Eurycolium chisholmii* (formerly known as *Prochlorococcus chisholmii*)  
601 [c.f. *P. chisholmii*]: This species is characterized by  $\alpha$ -carboxysome. Type strain is  
602 AS9601, isolated from marine habitat. The genome of this strain contains 1,66 Mbp  
603 (GC = 31,32%) harboring 1,769 coding DNA sequences.

604

605 **Description of *Prolificoccus* gen. nov.**

606 *Prolificoccus* (Pro.li.fi.co.ccus. L. *prolificus*, productive, abundant, numerous;  
607 *Prolificoccus*, referring to an abundant coccus).

608

609 Description of *Prolificoccus proteus* comb. nov. (formerly known as *Prochlorococcus*  
610 *proteus*) [c.f. *P. proteus*]: This species is characterized by  $\alpha$ -carboxysome. Type strain  
611 is NATL2A, isolated from a marine habitat in Northern Atlantic (10 m depth). The  
612 genome of this strain contains 1,84 Mbp (GC = 35,12%) harboring 1,930 coding DNA  
613 sequences.

614

615 **Description of *Expansiococcus* gen. nov.**

616 *Expansiococcus* (Ex.pan.sio.co.ccus. L. *expansio*, increase, expansion; *Expansiococcus*,  
617 referring to an organism that occupy large extensions).

618

619 Description of *Expansiococcus swingsii* comb. nov. (formerly known as  
620 *Prochlorococcus swingsii*) [c.f. *P. swingsii*]: This species is characterized by  $\alpha$ -  
621 carboxysome. Type strain is MIT 9313, isolated from a marine habitat in Gulf Stream  
622 (135 m depth). The genome of this strain contains 2,41 Mbp (GC = 50,74%) harboring  
623 2,339 coding DNA sequences.

624

#### 625 **Formal description of new species (within known genera)**

##### 626 *Arthrospira* genus

627 Description of *Arthrospira nitrilium* (formerly known as *Arthrospira* sp.) (ni.tri.lium.  
628 N.L. gen. n. *nitrilium*, referring to the capacity to utilize nitriles (R-C $\equiv$ N) as the sole  
629 source of nitrogen): this species is characterized by  $\beta$ -carboxysome and the  
630 phycobilisome pigmentation of this strain has not been characterized. Type strain is  
631 PCC 8005, isolated from a freshwater habitat in alkaline salt lakes. The genome of this  
632 strain contains 6,27 Mbp (GC = 44,7%) harboring 5,171 coding DNA sequences.

633

634 Description of *Arthrospira sesilensis* (formerly known as *Arthrospira platensis*)  
635 (ses.i.len.sis. N.L. gen. n. *sesilensis*, referring to the lack of gliding): this species is  
636 characterized by  $\beta$ -carboxysome and the phycobilisome pigmentation of this strain has  
637 not been characterized. Type strain is PCC 9438 (or C1), isolated from a freshwater

638 habitat in alkaline salt lakes. The genome of this strain contains 6 Mbp (GC = 44,69%)  
639 harboring 4,852 coding DNA sequences.

640

641 ***Geitlerinema* genus**

642 Description of *Geitlerinema catellasis* (formerly known as *Geitlerinema* sp.)  
643 (cat.e.lla.sis. L. fem. n. *catella*, referring to the filamentous small chain): This species is  
644 characterized by  $\alpha$ -carboxysome. Type strain is PCC 7105, isolated in USA. The  
645 genome of this strain contains 6,15 Mbp (GC = 51,59%) harboring 4,735 coding DNA  
646 sequences.

647

648 ***Coleofasciculus* genus**

649 Description of *Coleofasciculus microcolis* (formerly known as *Microcoleus* sp.)  
650 (mi.cro.co' lis. Gr. adj. *mikros* small (or thin); Gr. n. *koleos* sheath; M.L. masc. n.  
651 *microcolis* referring to a small or thin sheath): This species is characterized by  $\beta$ -  
652 carboxysome. Type strain is PCC 7113, isolated from a soil habitat in San Francisco,  
653 California, USA. The genome of this strain contains 7,47 Mbp (GC = 46,21%)  
654 harboring 6,734 coding DNA sequences.

655

656 ***Chroococidiopsis* genus**

657 Description of *Chroococidiopsis thermolimnetic* (formerly known as *Gloeocapsa* sp.)  
658 (ther.mo.lim.ne.tic. Gr. adj. *thermos*, hot ; N.L. fem. adj. *limnetica* (from Gr. n. *limnê*),  
659 lake; N.L. gen. n. *thermolimnetic*, referring to the habitat, freshwater thermal springs):  
660 This species is characterized by  $\beta$ -carboxysome. Type strain is PCC 7428, isolated from

661 freshwater habitat in moderate hot spring. The genome of this strain contains 5,43 Mbp  
662 (GC = 43,27%) harboring 5,254 coding DNA sequences.

663

664 Description of *Chroococciopsis synechocystis* (formerly known as *Synechocystis* sp.)  
665 (syn.e.cho.cys.tis, *synechocystis* referring to *Synechocystis* genus): This species is  
666 characterized by  $\beta$ -carboxysome. Type strain is PCC 7509, isolated from soil habitat in  
667 Switzerland. The genome of this strain contains 4,9 Mbp (GC = 41,67%) harboring  
668 4,859 coding DNA sequences.

669

#### 670 *Lyngbya* genus

671 Description of *Lyngbya hydrogenis* (formerly known as *Lyngbya aestuarii*)  
672 (hydr.o.ge.nis. *hydrogenis*, referring to the powerful hydrogen production of this strain):  
673 This species is characterized by  $\beta$ -carboxysome. Type strain is BL-J. The genome of  
674 this strain contains 6,87 Mbp (GC = 41,16%) harboring 5,597 coding DNA sequences.

675

#### 676 *Anabaena* genus

677 Description of *Anabaena mossii* (formerly known as *Anabaena* sp.) (mos.si. *mossi*,  
678 referring to the isolation area, intertidal zone, Moss Beach, CA, USA): This species is  
679 characterized by  $\beta$ -carboxysome. Type strain is PCC 7108, isolated from marine habitat  
680 in intertidal zone, Moss Beach, CA, USA. The genome of this strain contains 5,9 Mbp  
681 (GC = 38,78%) harboring 5,169 coding DNA sequences.

682

683 Description of *Anabaena stagnale* (formerly known as *Cylindrospermum stagnale*) [c.f.  
684 *C. stagnale*]: This species is characterized by  $\beta$ -carboxysome. Type strain is PCC 7417,  
685 isolated from soil habitat in Stockholm, Sweden. The genome of this strain contains  
686 7,61 Mbp (GC = 42,2%) harboring 6,127 coding DNA sequences.

687

688 Description of *Anabaena nostocii* (formerly known as *Nostoc* sp.) [c.f. genus *Nostoc*]:  
689 This species is characterized by  $\beta$ -carboxysome. Type strain is PCC 7524, isolated from  
690 freshwater habitat in hot spring, Amparai District, Maha Oya, Sri Lanka. The genome of  
691 this strain contains 7,61 Mbp (GC = 42,2%) harboring 5,326 coding DNA sequences.

692

693 Description of *Anabaena reyesii* (formerly known as *Nostoc* sp.) (*rey.es.si. reyesi*,  
694 referring to the isolation area, Point Reyes Peninsula, CA, USA): This species is  
695 characterized by  $\beta$ -carboxysome. Type strain is PCC 7107, isolated from freshwater  
696 habitat in Point Reyes Peninsula, CA, USA. The genome of this strain contains 6,32  
697 Mbp (GC = 40,36%) harboring 5,200 coding DNA sequences.

698

699 Description of *Anabaena welwii* (formerly known as *Fischerella* sp.) (*wel.wii. welwii*,  
700 referring to the presence of welwitindolinone gene cluster, a hapalindole-type family of  
701 natural products): This species is characterized by  $\beta$ -carboxysome. Type strain is PCC  
702 9431. The genome of this strain contains 7,16 Mbp (GC = 40,19%) harboring 6,104  
703 coding DNA sequences.

704

705 Description of *Anabaena hapalii* (formerly known as *Fischerella* sp.) (hap.a.lii. *hapalii*,  
706 referring to the presence of hapalindole gene cluster, a hapalindole-type family of  
707 natural products): This species is characterized by  $\beta$ -carboxysome. Type strain is PCC  
708 9339. The genome of this strain contains 8 Mbp (GC = 40,16%) harboring 6,720 coding  
709 DNA sequences.

710

711 Description of *Anabaena sesquitii* (formerly known as *Fischerella* sp.) (ses.qui.tii.  
712 *sesquitii*, referring to the presence of sesquiterpene gene cluster, encoding three  
713 proteins): This species is characterized by  $\beta$ -carboxysome. Type strain is JSC-11. The  
714 genome of this strain contains 5,38 Mbp (GC = 41,05%) harboring 4,627 coding DNA  
715 sequences.

716

717 Description of *Anabaena peptidasii* (formerly known as *Fischerella* sp.) (pep.ti.da.sii.  
718 *peptidasii*, referring to the presence of genes encoding peptidases, such as M16,  
719 characteristic of the group II bacteriocin gene clusters): This species is characterized by  
720  $\beta$ -carboxysome. Type strain is PCC 9605, isolated from soil (limestone) habitat in  
721 Jerucham, Har Rahama, Israel. The genome of this strain contains 8,08 Mbp (GC =  
722 42,61%) harboring 7,060 coding DNA sequences.

723

#### 724 ***Xenococcus* genus**

725 Description of *Xenococcus lajollai* (formerly known as *Xenococcus* sp.) (la.jo.llai.  
726 *lajollai*, referring to the isolation area, La Jolla Aquarium, CA, USA): This species is  
727 characterized by  $\beta$ -carboxysome. Type strain is PCC 7305, isolated from marine habitat

728 in Aquarium, La Jolla, CA, USA. The genome of this strain contains 5,92 Mbp (GC =  
729 39,68%) harboring 4,992 coding DNA sequences.

730

731 ***Pleurocapsa* genus**

732 Description of *Pleurocapsa penascus* (formerly known as *Pleurocapsa* sp.) (pe.nas.cus.  
733 *penascus*, referring to the isolation area, Puerto Penasco): This species is characterized  
734 by  $\beta$ -carboxysome. Type strain is PCC 7319, isolated from marine habitat in Arizona  
735 Station, Gulf of California, Puerto Penasco, Mexico. The genome of this strain contains  
736 7,38 Mbp (GC = 38,74%) harboring 4,516 coding DNA sequences.

737

738 ***Microcoleus* genus**

739 Description of *Microcoleus franciscanus* (formerly known as *Microcoleus* sp.)  
740 (fran.cis.ca.nus. *franciscanus*, referring to the isolation area, San Francisco): This  
741 species is characterized by  $\beta$ -carboxysome. Type strain is PCC 7113, isolated from soil  
742 habitat in San Francisco, California, USA. The genome of this strain contains 7,47 Mbp  
743 (GC = 46,21%) harboring 6,734 coding DNA sequences.

744

745 ***Dactylococcopsis* genus**

746 Description of *Dactylococcopsis halotolerans* (formerly known as *Halotheca* sp.)  
747 (ha.lo.to.le.rans. *Gr. n. hals, halos*, salt, sea; L. pres. part. *tolerans*, tolerating; M. L.  
748 part. adj. *Halotolerans* salt-tolerating): This species is characterized by  $\beta$ -carboxysome.  
749 Type strain is PCC 7418, isolated from freshwater habitat in Solar Lake, Israel. The

750 genome of this strain contains 4,18 Mbp (GC = 42,92%) harboring 3,663 coding DNA  
751 sequences.

752

### 753 *Geminocystis* genus

754 Description of *Geminocystis stanieri* comb. nov. (formerly known as *Cyanobacterium*  
755 *stanieri*) [c.f. *C. stanieri*]: This species is characterized by  $\beta$ -carboxysome. Type strain  
756 is PCC 7202, isolated from freshwater habitat in thermal spring, alkaline pod. The  
757 genome of this strain contains 3,16 Mbp (GC = 38,66%) harboring 2,886 coding DNA  
758 sequences.

759

### 760 *Planktothrix* genus

761 Description of *Planktothrix stereotis* comb. nov. (formerly known as *Planktothrix*  
762 *agardhii*) (ster.eo.tis. Gr. *stereos*, solid; *stereotis*, referring to a solid and straight  
763 filament): This species is characterized by  $\beta$ -carboxysome. Type strain is NIVA-CYA  
764 126, isolated from freshwater habitat. The genome of this strain contains 5,04 Mbp (GC  
765 = 39,57%) harboring 4,188 coding DNA sequences.

766

### 767 ABBREVIATIONS

768 adj. adjective; *Gr.* Greek; *L.* Latin; *n.* noun; neut. neutro; c.f. confer, conferre; v. verb;  
769 CDS protein-coding sequence; NA "Not Available" or "Not Applicable" or "Not  
770 Announced".

771

### 772 ACKNOWLEDGMENTS

773 The authors thankfully acknowledge National Counsel of Technological and  
774 Scientific Development (CNPq), Coordination for the Improvement of Higher  
775 Education Personnel (CAPES), and Rio de Janeiro Research Foundation (FAPERJ) for  
776 their financial support for this study. This paper is part of the D. Sc. requirements of  
777 JMW at the Biodiversity and Evolutionary Biology Graduate Program of the Federal  
778 University of Rio de Janeiro (UFRJ), and was developed within the Science Without  
779 Borders Program (CNPq, 207751/2014-5).

780

#### 781 **SEQUENCE DATA**

782 All publicly available sequence data used in this paper was retrieved from the  
783 RefSeq (<https://www.ncbi.nlm.nih.gov/refseq/>) and GenBank, as part of the  
784 International Nucleotide Sequence Database Collaboration, and also from the Geba  
785 database, produced by the US Department of Energy Joint Genome Institute  
786 (<http://www.jgi.doe.gov/>) in collaboration with the user community.

787

#### 788 **REFERENCES**

- 789 Almstrand, R. et al., 2016. Draft Genome Sequence of a Novel Desulfobacteraceae Member  
790 from a Sulfate-Reducing Bioreactor Metagenome. *Genome announcements*, 4(1),  
791 pp.e01540-15. Available at: <http://genomea.asm.org/content/4/1/e01540-15.full>.
- 792 Auch, A.F., von Jan, M., et al., 2010. Digital DNA-DNA hybridization for microbial species  
793 delineation by means of genome-to-genome sequence comparison. *Standards in genomic  
794 sciences*, 2(1), pp.117–34. Available at: <http://standardsingenomics.org/content/2/1/117/>.
- 795 Auch, A.F., Klenk, H.-P. & Göker, M., 2010. Standard operating procedure for calculating  
796 genome-to-genome distances based on high-scoring segment pairs. *Standards in genomic  
797 sciences*, 2(1), pp.142–8. Available at: <http://standardsingenomics.org/content/2/1/142/>.
- 798 Batut, B. et al., 2014. Reductive genome evolution at both ends of the bacterial population size  
799 spectrum. *Nature Reviews Microbiology*, 12, pp.841–850.
- 800 Beiko, R.G., 2015. Microbial malaise: How can we classify the microbiome? *Trends in  
801 Microbiology*, 23(11), pp.671–679. Available at:  
802 <http://dx.doi.org/10.1016/j.tim.2015.08.009>.
- 803 Boone, D.R. & Castenholz, R.W., 2001. *Bergey's Manual of Systematic Bacteriology Volume*

- 804 *One : The Archaea and the Deeply Branching and Phototrophic Bacteria* 2nd ed. G.  
805 Garrity, D. R. Boone, & R. W. Castenholz, eds., New York, NY: Springer-Verlag New  
806 York.
- 807 Brown, C.T. et al., 2015. Unusual biology across a group comprising more than 15% of domain  
808 Bacteria. *Nature*, 523(7559), pp.208–211. Available at:  
809 <http://www.nature.com/doi/10.1038/nature14486>.
- 810 Castresana, J., 2000. Selection of Conserved Blocks from Multiple Alignments for Their Use in  
811 Phylogenetic Analysis. *Molecular Biology and Evolution*, 17(4), pp.540–552. Available at:  
812 <http://www.ncbi.nlm.nih.gov/pubmed/10742046> \n<http://mbe.oxfordjournals.org/cgi/content/abstract/17/4/540>.
- 814 Coenye, T. et al., 2005. Towards a prokaryotic genomic taxonomy. *FEMS Microbiology*  
815 *Reviews*, 29(2), pp.147–167.
- 816 Coutinho, F. et al., 2016. Comparative genomics of Synechococcus and proposal of the new  
817 genus Parasynechococcus. *PeerJ*, 4(May), p.e1522. Available at:  
818 <https://peerj.com/articles/1522>.
- 819 Coutinho, F.H. et al., 2015. Niche distribution and influence of environmental parameters in  
820 marine microbial communities: a systematic review. *PeerJ*, 3, p.e1008. Available at:  
821 <https://peerj.com/articles/1008>.
- 822 Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high  
823 throughput. *Nucleic Acids Research*, 32(5), pp.1792–1797.
- 824 Farrant, G.K. et al., 2016. Delineating ecologically significant taxonomic units from global  
825 patterns of marine picocyanobacteria.
- 826 Garrity, G.M., 2016. The Future of Whole Genome Sequencing for public health and the clinic.  
827 *Journal of Clinical Microbiology*, 54(8), p.JCM.01082-16. Available at:  
828 <http://jcm.asm.org/lookup/doi/10.1128/JCM.01082-16>.
- 829 Gevers, D. et al., 2005. Re-evaluating prokaryotic species. *Nat.Rev.Microbiol.*, 3(1740–1526  
830 (Print)), pp.733–739. Available at: C:\Karsten\PDFs\Grundlagen-PDFs\Grund-  
831 2005\Gevers et al.-Re-evaluating prokaryotic species.pdf.
- 832 Gugger, M.F. & Hoffmann, L., 2004. Polyphyly of true branching cyanobacteria  
833 (Stigonematales). *International Journal of Systematic and Evolutionary Microbiology*,  
834 54(2), pp.349–357.
- 835 Haroon, M.F. et al., 2016. Data Descriptor : A catalogue of 136 microbial draft genomes from  
836 Red Sea metagenomes. *Scientific Data*, pp.1–6.
- 837 Hoffmann, L., Komárek, J. & Kaštovský, J., 2005. System of cyanoprokaryotes (cyanobacteria)  
838 – state in 2004. *Algological Studies*, 117(1), pp.95–115.
- 839 Hugenholtz, P., Skarshewski, A. & Parks, D.H., 2016. Genome-Based Microbial Taxonomy  
840 Coming of Age. *Cold Spring Harbor perspectives in biology*, p.cshperspect.a018085-  
841 Available at:  
842 <http://cshperspectives.cshlp.org/content/early/2016/03/17/cshperspect.a018085.abstract>.
- 843 Hugerth, L.W. et al., 2015. Metagenome-assembled genomes uncover a global brackish  
844 microbiome. *Genome biology*, 16(279). Available at:  
845 <http://biorxiv.org/content/early/2015/04/23/018465.abstract>.
- 846 Iverson, V. et al., 2012. Untangling Genomes from Metagenomes: Revealing an Uncultured  
847 Class of Marine Euryarchaeota. *Science*, 335(6068), pp.587–590.
- 848 Johnson, Z.I. et al., 2006. Partitioning Among Prochlorococcus Ecotypes Along Environmental  
849 Gradients. *Science*, 311(March), pp.1737–1740.
- 850 Kashtan, N. et al., 2014. Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations  
851 in Wild Prochlorococcus. *Science (New York, N.Y.)*, 344(April), pp.416–420.
- 852 Kauff, F. & Büdel, B., 2010. Phylogeny of cyanobacteria: an overview. *Progress in Botany*, 72,  
853 pp.209–224.

- 854 Konstantinidis, K.T. & Tiedje, J.M., 2005. Genomic insights that advance the species definition  
855 for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of*  
856 *America*, 102(7), pp.2567–72. Available at:  
857 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=549018&tool=pmcentrez&ren](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=549018&tool=pmcentrez&rendertype=abstract)  
858 [dertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=549018&tool=pmcentrez&rendertype=abstract).
- 859 Konstantinidis, K.T. & Tiedje, J.M., 2005. Towards a genome-based taxonomy for prokaryotes.  
860 *Journal of Bacteriology*, 187(18), pp.6258–6264.
- 861 Kozlov, A.M. et al., 2016. Phylogeny-aware Identification and Correction of Taxonomically  
862 Mislabeled Sequences. *bioRxiv*, 2(13), pp.1–12. Available at:  
863 <http://biorxiv.org/lookup/doi/10.1101/042200>.
- 864 Labeda, D.P., 2000. International Committee on Systematic Bacteriology IXth International  
865 (IUMS) Congress of Bacteriology and Applied. *International Journal of Systematic and*  
866 *Evolutionary Microbiology*, Minutes of(50), pp.2245–2247.
- 867 Lagesen, K. et al., 2007. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes.  
868 *Nucleic Acids Research*, 35(9), pp.3100–3108.
- 869 Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature*  
870 *Methods*, 9(4), pp.357–359.
- 871 Locey, K.J. & Lennon, J.T., 2016. Scaling laws predict global microbial diversity. *Proceedings*  
872 *of the National Academy of Sciences*, Early Edit.
- 873 Lopes, F.A.C. et al., 2016. Microbial Community Profile and Water Quality in a Protected Area  
874 of the Caatinga Biome. *Plos One*, 11(2), p.e0148296. Available at:  
875 <http://dx.plos.org/10.1371/journal.pone.0148296>.
- 876 Meier-Kolthoff, J.P. et al., 2013. Genome sequence-based species delimitation with confidence  
877 intervals and improved distance functions. *BMC bioinformatics*, 14, p.60. Available at:  
878 <http://www.ncbi.nlm.nih.gov/pubmed/23432962>.
- 879 Oren, Aharon; Tindall, B.J., 2005. Nomenclature of the  
880 cyanophyta/cyanobacteria/cyanoprokaryotes under the International Code of  
881 Nomenclature of Prokaryotes. *Algological Studies*, 117(14), pp.39–52.
- 882 Oren, A., 2004. A proposal for further integration of the cyanobacteria under the Bacteriological  
883 Code. *International Journal of Systematic and Evolutionary Microbiology*, 54(5),  
884 pp.1895–1902.
- 885 Oren, A. & Garrity, G.M., 2014. Proposal to change General Consideration 5 and Principle 2 of  
886 the International Code of Nomenclature of Prokaryotes. *International Journal of*  
887 *Systematic and Evolutionary Microbiology*, 64(PART 1), pp.309–310.
- 888 Oren, A., Komárek, J. & Hoffmann, L., 2009. Nomenclature of the  
889 Cyanophyta/Cyanobacteria/Cyanoprokaryotes – What has happened since IAC  
890 Luxembourg? *Algological Studies*, 130(1), pp.17–26.
- 891 Parks, D.H. et al., 2015. CheckM: assessing the quality of microbial genomes recovered from  
892 isolates, single cells, and metagenomes. *Genome research*, 25(7), pp.1043–55. Available  
893 at:  
894 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4484387&tool=pmcentrez&re](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4484387&tool=pmcentrez&rendertype=abstract)  
895 [ndertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4484387&tool=pmcentrez&rendertype=abstract).
- 896 Parte, A.C., 2014. LPSN - List of prokaryotic names with standing in nomenclature. *Nucleic*  
897 *Acids Research*, 42(D1), pp.613–616.
- 898 Pinto, A.J. et al., 2016. Draft Genome Sequences of Two Novel. *Genome announcements*, 4(1),  
899 pp.4–5.
- 900 Pittera, J. et al., 2014. Connecting thermal physiology and latitudinal niche partitioning in  
901 marine *Synechococcus*. *The ISME journal*, pp.1–16. Available at:  
902 <http://www.ncbi.nlm.nih.gov/pubmed/24401861>.
- 903 Pruesse, E. et al., 2007. SILVA: A comprehensive online resource for quality checked and

- 904 aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*,  
905 35(21), pp.7188–7196.
- 906 Puigbò, P. et al., 2014. Genomes in turmoil: quantification of genome dynamics in prokaryote  
907 supergenomes. *BMC Biology*, 12(1), p.66. Available at:  
908 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4166000&tool=pmcentrez&re](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4166000&tool=pmcentrez&rendertype=abstract)  
909 [ndertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4166000&tool=pmcentrez&rendertype=abstract).
- 910 Qin, Q.L. et al., 2014. A proposed genus boundary for the prokaryotes based on genomic  
911 insights. *Journal of Bacteriology*, 196(12).
- 912 Quast, C. et al., 2013. The SILVA ribosomal RNA gene database project: Improved data  
913 processing and web-based tools. *Nucleic Acids Research*, 41(D1), pp.590–596.
- 914 Rambaut, A., 2015. FigTree v1.4.2: Tree figure drawing tool. Available at:  
915 <http://tree.bio.ed.ac.uk/software/figtree/>.
- 916 Di Rienzi, S.C. et al., 2013. The human gut and groundwater harbor non-photosynthetic bacteria  
917 belonging to a new candidate phylum sibling to Cyanobacteria. *eLife*, 2013(2), pp.1–25.
- 918 Rippka, R. et al., 1979. Generic Assignments, Strain Histories and Properties of Pure Cultures  
919 of Cyanobacteria. *Journal of General Microbiology*, 111(1), pp.1–61.
- 920 Rocap, G. et al., 2003. Genome divergence in two Prochlorococcus ecotypes reflects oceanic  
921 niche differentiation. *Nature*, 424(6952), pp.1042–1047. Available at:  
922 <http://www.nature.com/doi/10.1038/nature01947>.
- 923 Sarma, T.A., 2012. *Handbook of Cyanobacteria*, Boca Raton, FL: CRC Press.
- 924 Scanlan, D.J., 2003. Physiological diversity and niche adaptation in marine Synechococcus.  
925 *Advances in Microbial Physiology*, 47, pp.1–64.
- 926 Scanlan, D.J. & West, N.J., 2002. Molecular ecology of the marine cyanobacteria genera  
927 Prochlorococcus and Synechococcus. *FEMS Microbiology Ecology*, 40, pp.1–12.
- 928 Schirmer, B.E. et al., 2011. Evolution of cyanobacterial morphotypes: Taxa required for  
929 improved phylogenomic approaches. *Communicative & integrative biology*, 4(4), pp.424–  
930 7. Available at:  
931 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3181511&tool=pmcentrez&re](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3181511&tool=pmcentrez&rendertype=abstract)  
932 [ndertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3181511&tool=pmcentrez&rendertype=abstract).
- 933 Shih, P.M. et al., 2013. Improving the coverage of the cyanobacterial phylum using diversity-  
934 driven genome sequencing. *Proceedings of the National Academy of Sciences of the*  
935 *United States of America*, 110(3), pp.1053–8. Available at:  
936 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3549136&tool=pmcentrez&re](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3549136&tool=pmcentrez&rendertype=abstract)  
937 [ndertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3549136&tool=pmcentrez&rendertype=abstract).
- 938 Soo, R.M. et al., 2014. An expanded genomic representation of the phylum cyanobacteria.  
939 *Genome Biology and Evolution*, 6(5), pp.1031–1045.
- 940 Stamatakis, A., 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with  
941 thousands of taxa and mixed models. *Bioinformatics*, 22(21), pp.2688–2690.
- 942 Stanier, R.Y. et al., 1978. Proposal to Place the Nomenclature of the Cyanobacteria (Blue-Green  
943 Algae) Under the Rules of the International Code of Nomenclature of Bacteria.  
944 *International Journal of Systematic Bacteriology*, 28(2), pp.335–336.
- 945 Sunagawa, S. et al., 2015. Structure and function of the global ocean microbiome. *Science*,  
946 348(6237), p.1261359. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/25999513>.
- 947 Talavera, G. & Castresana, J., 2007. Improvement of phylogenies after removing divergent and  
948 ambiguously aligned blocks from protein sequence alignments. *Systematic biology*, 56(4),  
949 pp.564–77. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17654362>.
- 950 Tamura, K. et al., 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0.  
951 *Molecular Biology and Evolution*, 30(12), pp.2725–2729.
- 952 Thompson, C.C. et al., 2013. Genomic Taxonomy of the Genus Prochlorococcus. *Microbial*  
953 *Ecology*, 66(4).

- 954 Thompson, C.C. et al., 2013. Microbial genomic taxonomy. *BMC genomics*, 14, p.913.  
955 Available at:  
956 [http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3879651&tool=pmcentrez&re](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3879651&tool=pmcentrez&rendertype=abstract)  
957 [ndertype=abstract](http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3879651&tool=pmcentrez&rendertype=abstract).  
958 Thompson, C.C. et al., 2015. Microbial taxonomy in the post-genomic era: Rebuilding from  
959 scratch? *Archives of Microbiology*, 197(3), pp.359–370.  
960 Tindall, B.J., 1999. Proposals to update and make changes to the Bacteriological Code.  
961 *International Journal of Systematic and Evolutionary Microbiology*, (49), pp.1309–1312.  
962 De Vos, P. & Trüper, H.G., 2000. Judicial Commission of the International Committee on  
963 Systematic Bacteriology. *International Journal of Systematic and Evolutionary*  
964 *Microbiology*, Minutes of(50), pp.2239–2244.  
965 Wu, M. & Eisen, J. a, 2008. A simple, fast, and accurate method of phylogenomic inference.  
966 *Genome biology*, 9(10), p.R151.  
967 Wu, M. & Scott, A.J., 2012. Phylogenomic analysis of bacterial and archaeal sequences with  
968 AMPHORA2. *Bioinformatics*, 28(7), pp.1033–1034.  
969 Yarza, P. et al., 2008. The All-Species Living Tree project: A 16S rRNA-based phylogenetic  
970 tree of all sequenced type strains. *Systematic and Applied Microbiology*, 31(4), pp.241–  
971 250.

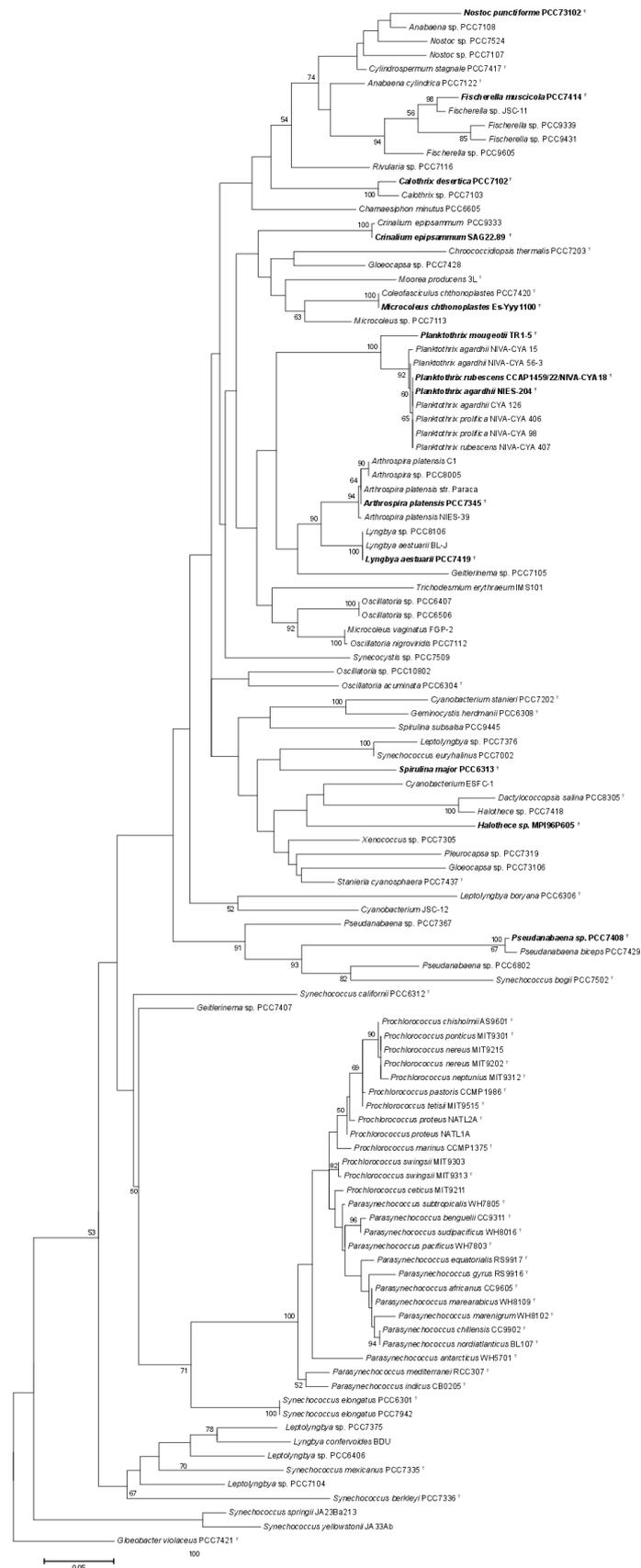
972  
973

#### 974 **COMPETING INTERESTS**

975 Fabiano Thompson is an Academic Editor for PeerJ.

976  
977  
978  
979

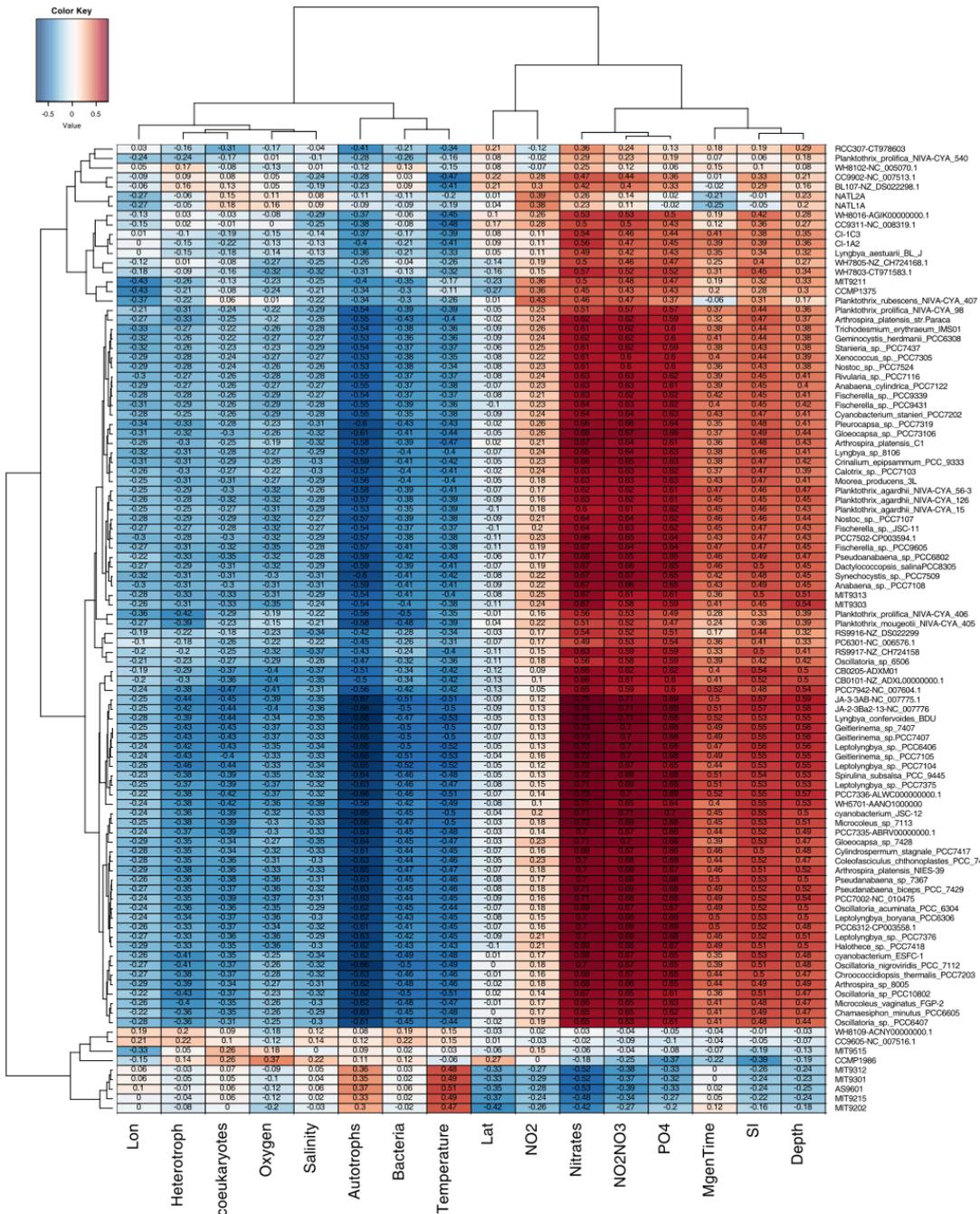
#### **FIGURES AND FIGURE LEGENDS**



981 **Figure 1. Ribosomal phylogenetic reconstruction of the Cyanobacteria phylum.**  
982 Tree was constructed through ML using the Kimura 2-parameter method, and GTR+G  
983 substitution model. Tree was inferred from 110 16S rRNA gene sequences (~ 1,400 bp).  
984 The species cutoff was 98.8 % similarity (Thompson et al., 2014). The percentages of  
985 replicate trees in which the associated taxa clustered together in the bootstrap test (1,000  
986 replicates) are shown next to the branches. Nodes supported with a bootstrap of  $\geq 50$  %  
987 are indicated. Bold names indicate the included type-strains. The unit of measure for the  
988 scale bars is the number of nucleotide substitutions per site. *Gloeobacter violaceus* PCC  
989 7421 sequence was designated as outgroup.

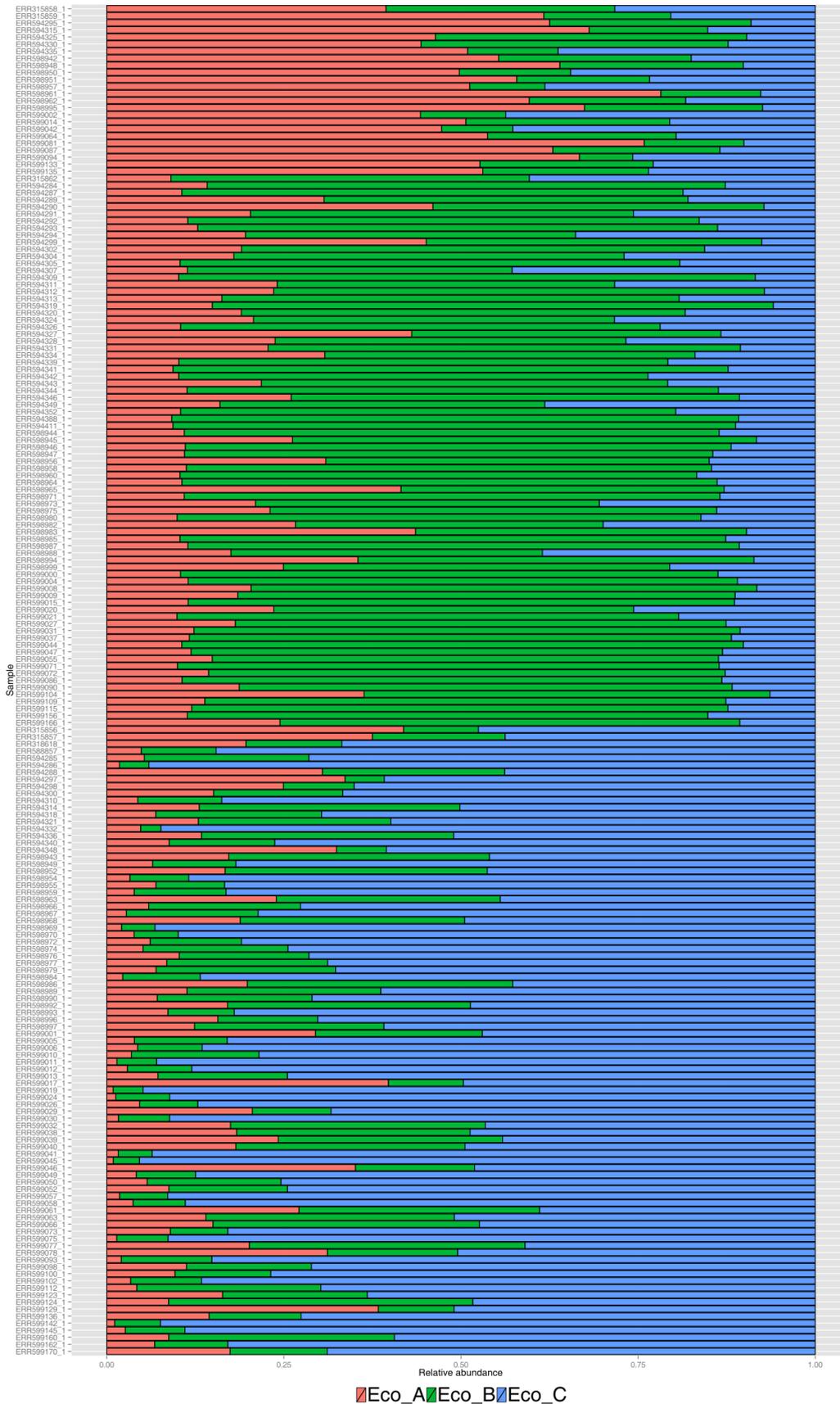




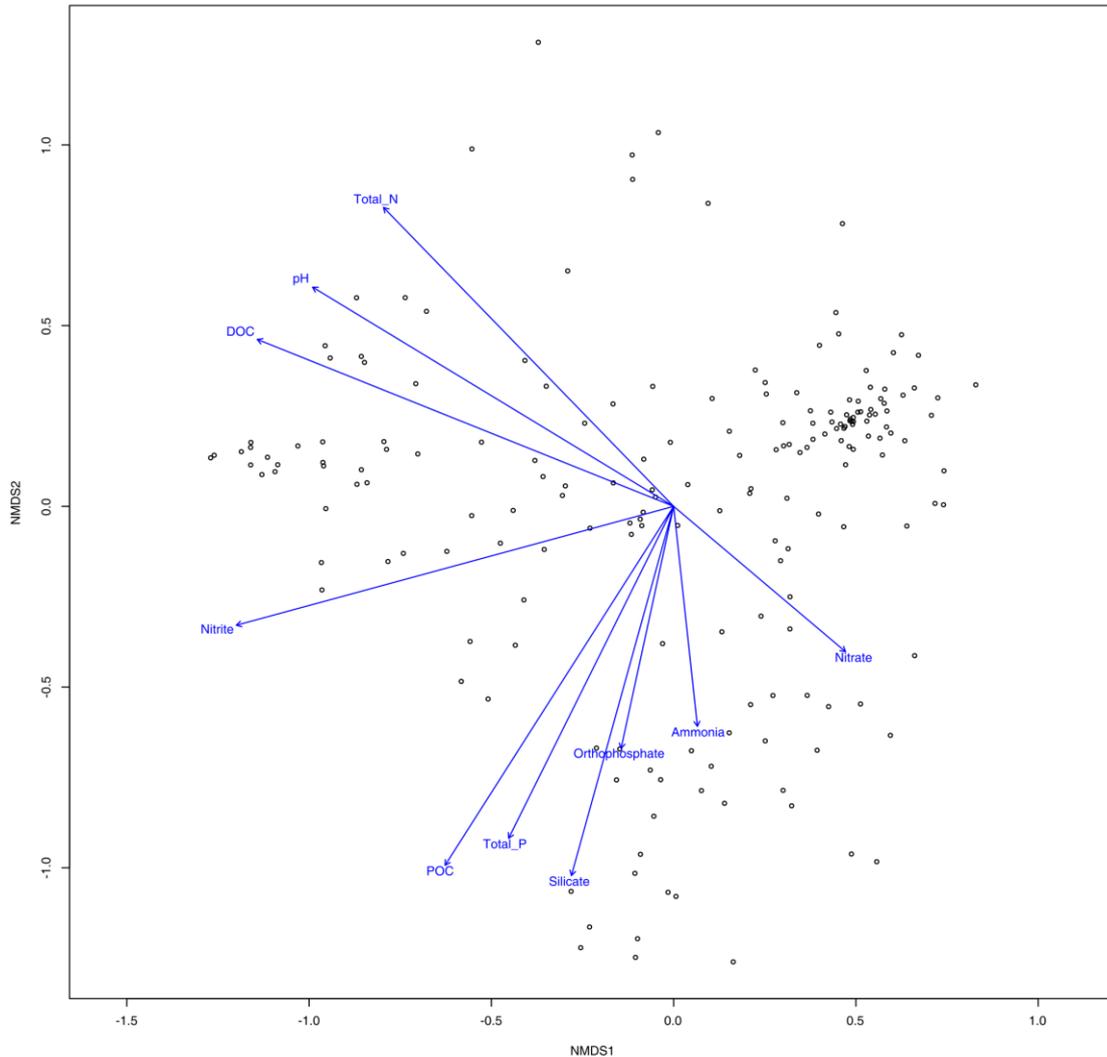


1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014

**Figure 4. Abundance and distribution profiles of cyanobacterial strains across global metagenomes.** The distributions are displayed as a hierarchically-clustered heatmap. Ambiguous reads were corrected as describing in Method section.

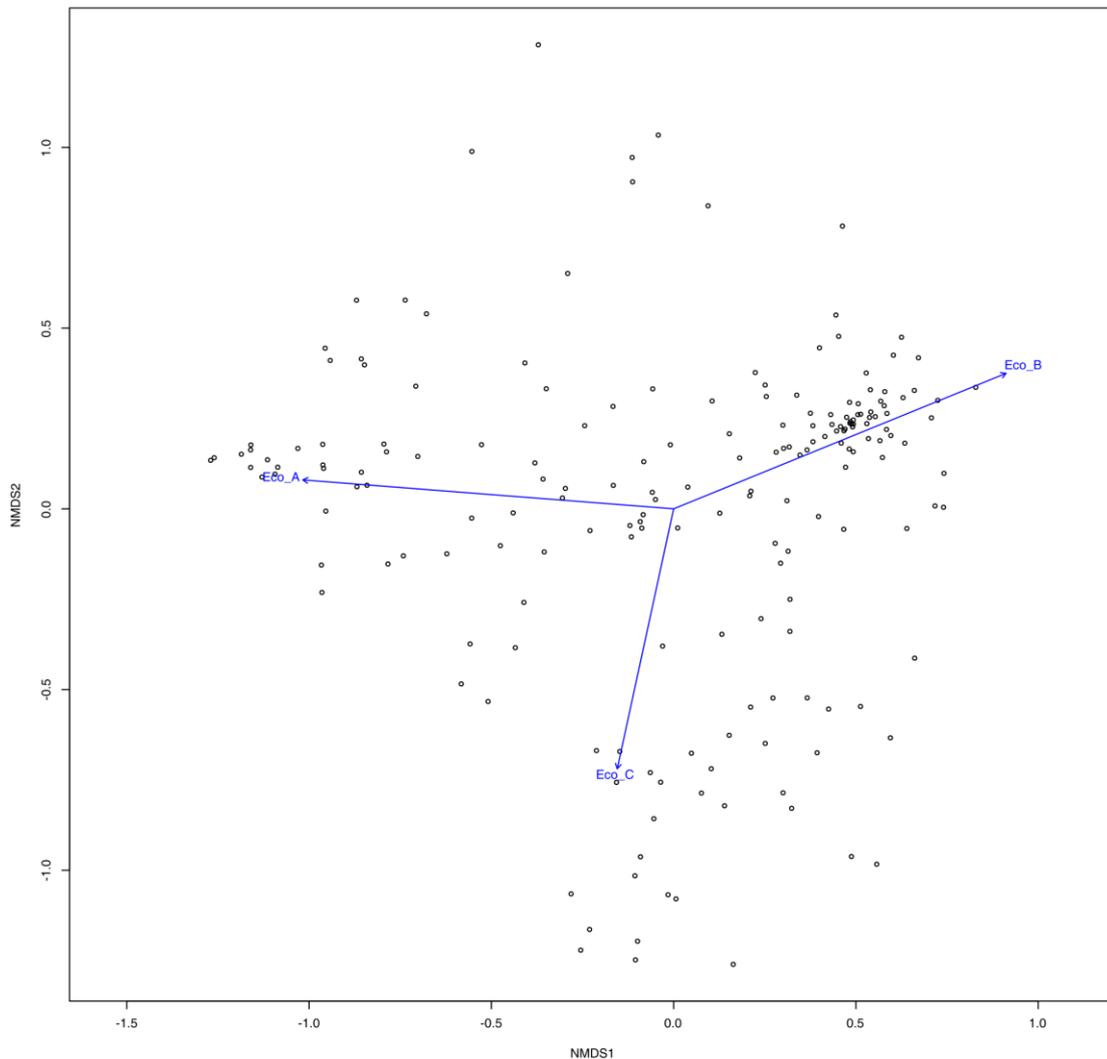


1016 **Figure 5. Abundance and distribution of Eco clusters across global marine**  
1017 **metagenomes.** Relative abundance of Eco-A, Eco-B and Eco-C in a global scale.  
1018 Metagenomes (N = 191) from marine stations presents the global diversity.  
1019  
1020



1021  
1022  
1023  
1024  
1025  
1026  
1027  
1028

**Figure 6. Non-metric multidimensional scaling (NMDS) analysis of the marine metagenomes and environmental parameters.** Ordination plot of physicochemical parameters and the community structure of metagenomes from all stations. Distances of samples were used in generating the NMDS. The lengths of the lines represent the strength of the correlation. Dots indicate the metagenomes samples.



1029  
1030

1031 **Figure 7. Non-metric multidimensional scaling (NMDS) analysis of the marine**  
1032 **metagenomes and Eco clusters.** Ordination plot of Eco clusters and the community  
1033 structure of metagenomes from all stations. Distances of samples were used in  
1034 generating the NMDS. The lengths of the lines represent the strength of the correlation.  
1035 Dots indicate the metagenomes samples.

1036  
1037

1038  
1039

## TABLES AND TABLE LEGENDS

Bacterial Strain <sup>1</sup>	Strain <sup>1</sup>	NCBI or JGI Reference Sequence	New Genus Proposal	New Species Proposal	Habitat	Type Source / Place	# Contigs <sup>Δ</sup>	Lenght (Mbp) <sup>Δ</sup>	% mol GC <sup>Δ</sup>	# CDS	Completeness *	Carboxysome
<i>Anabaena cylindrica</i>	PCC 7122 <sup>T</sup>	<a href="#">NC_019771.1</a>			Freshwater	Cambridge, UK Intertidal zone,	7	7.06	38.79	6182	99.44	β
<i>Anabaena sp.</i>	PCC 7108	<a href="#">NZ_AJWF00000000.1</a>		<i>A. mossi</i>	Marine (coastal)	Moss Beach, CA, USA	3	5.9	38.78	5169	99.63	β
<i>Arthrospira platensis</i>	C1 <sup>2</sup>	<a href="#">NZ_CM001632.1</a>		<i>A. sesilensis</i>	Freshwater	Alkaline salt lakes	63	6	44.69	4852	99.71	β
<i>Arthrospira platensis</i>	NIES-39	<a href="#">NC_016640.1</a>			Freshwater	Alkaline salt lakes	1	6.78	44.27	6676	99.13	β
<i>Arthrospira platensis</i>	Paraca	<a href="#">NZ_ACSK00000000.3</a>			Freshwater	Alkaline salt lakes	239	6.49	44.31	5436	99.34	β
<i>Arthrospira sp.</i>	PCC 8005	<a href="#">NZ_FO818640.1</a>		<i>A. nitrilium</i>	Unknown	Unknown	119	6.27	44.7	5171	99.93	β
<i>Calotrix sp.</i>	PCC 7103	<a href="#">NZ_ALVJ00000000.1</a>	<i>Anabaena</i>	<i>A. calotrixus</i>	Unknown	Unknown Berkeley, CA,	12	11.58	38.55	9371	99.39	β
<i>Chamaesiphon minutus</i>	PCC 6605	<a href="#">NC_019697.1</a>			Freshwater	USA	1	6.28	45.73	5956	99.48	β
<i>Chroococciopsis thermalis</i>	PCC 7203 <sup>T</sup>	<a href="#">NC_019695.1</a>			Soil	Greifswald, Germany Salt marsh in Woods Hole, Massachusetts,	3	6.68	44.47	5618	99.63	β
<i>Coleofasciculus chthonoplastes</i>	PCC 7420 <sup>3</sup> <sup>T</sup>	<a href="#">NZ_ABR00000000.1</a>			Marine (coastal)	USA	142	8.65	45.43	7100	99.37	β
<i>Crinalium epipsammum</i>	PCC 9333	<a href="#">NC_019753.1</a>			NA	NA Extremophylic mat communities, Elkhorn Slough estuary, CA, USA	1	5.31	40.16	5002	99.48	β
<i>Cyanobacterium</i>	ESFC-1	<a href="#">NZ_ARCP00000000.1</a>	<i>Cyclospexia</i>	<i>C. valenium</i> <i>T.</i>	Marine (coastal)	USA	52	5.62	46.51	4857	99.59	β
<i>Cyanobacterium</i>	JSC-12	<a href="#">NZ_CM001633.1</a>	<i>Tapinonema</i>	<i>coecalium</i>	Freshwater	NA Thermal springs,	20	5.52	47.49	5024	99.29	β
<i>Cyanobacterium stanieri</i>	PCC 7202 <sup>T</sup>	<a href="#">CP003940.1</a>	<i>Geminocystis</i>	<i>G. stanieri</i>	Freshwater	alkaline pod Stockholm,	1	3.16	38.66	2886	99.52	β
<i>Cylindrospermum stagnale</i>	PCC 7417 <sup>T</sup>	<a href="#">NC_019757.1</a>	<i>Anabaena</i>	<i>A. stagnale</i>	Soil	Sweden	4	7.61	42.2	6127	99.78	β
<i>Dactylococcopsis salina</i>	PCC 8305 <sup>T</sup>	<a href="#">NC_019780.1</a>			Freshwater	Solar Lake, Israel	1	3.78	42.44	3412	99.55	β
<i>Fischerella sp.</i>	JSC-11	<a href="#">NZ_AGIZ00000000.1</a>	<i>Anabaena</i>	<i>A. sesquiti</i>	NA	NA	34	5.38	41.05	4627	99.76	β

<i>Fischerella</i> sp.	PCC 9339	<a href="#">NZ_ALVS00000000.1</a>	<i>Anabaena</i>	<i>A. hapalii</i>	NA	NA	13	8	40.16	6720	99.76	β
<i>Fischerella</i> sp.	PCC 9431	<a href="#">ALVX00000000.1</a>	<i>Anabaena</i>	<i>A. welwii</i>	NA	NA	8	7.16	40.19	6104	99.76	β
<i>Fischerella</i> sp.	PCC 9605	<a href="#">NZ_ALVT00000000.1</a>	<i>Anabaena</i>	<i>A. peptidasii</i>	Soil	Limestone, Jerucham, Har Rahama, Israel	12	8.08	42.61	7060	100	β
<i>Geitlerinema</i> sp.	PCC 7105	<a href="#">NZ_ANFQ00000000.1</a>		<i>G. catellasis</i>	NA	USA	8	6.15	51.59	4735	93.75	β
<i>Geitlerinema</i> sp.	PCC 7407	<a href="#">NC_019703.1</a>	<i>Pseudogeitlerinema</i>	<i>P. shalloid</i>	Unknow	Unknow Lake near Madison, Wisconsin, USA	1	4.68	58.46	3727	99.87	β
<i>Geminocystis herdmanii</i>	PCC 6308 <sup>T</sup>	<a href="#">NZ_ALVO00000000.1</a>		<i>C. thermolimn</i>	Freshwater	Moderate hot spring Sphagnum bog, Switzerland	1	4.26	34.28	3887	99.78	β
<i>Gloeocapsa</i> sp.	PCC 7428	<a href="#">NC_019745.1</a>	<i>Chroococciopsis</i>	<i>etic</i>	Freshwater	Solar Lake, Israel	1	5.43	43.27	5254	99.78	β
<i>Gloeocapsa</i> sp.	PCC 73106	<a href="#">NZ_ALVY00000000.1</a>		<i>G. sphagnus</i>	Freshwater	Lake near Madison, Wisconsin, USA	228	4.025	41.11	3704	98.84	β
<i>Halothece</i> sp.	PCC 7418	<a href="#">NC_019779.1</a>	<i>Dactylococcopsis</i>	<i>halotolerans</i>	Freshwater	Rock at shoreline, Montauk Point, Long Island, NY, USA	1	4.18	42.92	3663	99.48	β
<i>Leptolyngbya boryana</i>	PCC 6306 <sup>T</sup>	<a href="#">NZ_ALVM00000000.1</a>		<i>A. longislandicus</i>	Freshwater	Plankton, Woods Hole, Massachusetts, USA	5	7.26	47.02	6827	99.41	β
<i>Leptolyngbya</i> sp.	PCC 7104 <sup>4</sup>	<a href="#">NZ_ALVP00000000.1</a>	<i>Allonema</i>	<i>us</i>	Marine (coastal)	Massachusetts, USA	2	6.89	57.69	6414	99.18	β
<i>Leptolyngbya</i> sp.	PCC 7375	<a href="#">NZ_ALVN00000000.1</a>	<i>Acrophormium</i>	<i>splendidus</i>	Marine (coastal)	Limestone, Crystal Cave, Bermuda	5	9.42	47.62	8366	99.73	β
<i>Leptolyngbya</i> sp.	PCC 7376	<a href="#">NC_019683.1</a>	<i>Symphothece</i>	<i>S. fragile</i>	Marine (coastal)	California, USA	1	5.12	43.87	4601	99.42	β
<i>Leptolyngbya</i> sp.	PCC 6406	<a href="#">NZ_ALVV00000000.2</a>	<i>Euryforis</i>	<i>E. eilemai</i>	Freshwater		3	5.77	55.18	5156	98.64	β
<i>Lyngbya aestuarii</i>	BL-J	<a href="#">NZ_AUZH00000000.1</a>		<i>L. hydrogenis</i>	NA		432	6.87	41.16	5597	99.74	β
<i>Lyngbya confervoides</i>	BDU	<a href="#">NZ_JTHE00000000.1</a>	<i>Euryforis</i>	<i>E.</i>	NA		298	8.79	55.63	8370	99.34	β

				<i>confervoides</i>								
<i>Lyngbya sp.</i>	PCC 8106	<a href="#">NZ_AAVU00000000.1</a>		<i>L. aestuarii</i>	Marine (coastal)	NA	110	7.03	41.11	5854	99.3	β
<i>Microcoleus sp.</i>	PCC 7113	<a href="#">NC_019738.1</a>		<i>M. franciscanus</i>	Soil	San Francisco, California, USA	1	7.47	46.21	6734	99.56	β
<i>Microcoleus vaginatus</i>	FGP-2	<a href="#">NZ_AFJC00000000.1</a>	<i>Toxinema</i>	<i>T. vaginatus</i>	Soil	Canyonlands National Park, UT, USA	40	6.69	46.04	5519	99.67	β
<i>Moorea producens</i>	3L <sup>5T</sup>	<a href="#">NZ_AEPQ00000000.1</a>			NA	NA	287	8.38	43.68	6979	98.56	β
<i>Nostoc sp.</i>	PCC 7107	<a href="#">NC_019676.1</a>	<i>Anabaena</i>	<i>A. reyesii</i>	Freshwater	Point Reyes Peninsula, California, USA	1	6.32	40.36	5200	99.26	β
<i>Nostoc sp.</i>	PCC 7524	<a href="#">NC_019684.1</a>	<i>Anabaena</i>	<i>A. nostocii</i>	Freshwater	Hot spring, Amparai District, Maha Oya, Sri Lanka	3	6.71	41.53	5326	99.33	β
<i>Oscillatoria acuminata</i>	PCC 6304 <sup>T</sup>	<a href="#">NC_019693.1</a>		<i>T.</i>	Soil	NA	1	7.68	47.6	6004	99.71	β
<i>Oscillatoria nigroviridis</i>	PCC 7112	<a href="#">NC_019729.1</a>	<i>Toxinema</i>	<i>nigroviridis</i>	Soil	USA	1	7.47	45.87	6925	99.78	β
<i>Oscillatoria sp.</i>	PCC 10802	<a href="#">NZ_ANKO00000000.1</a>	<i>Somacatellium</i>	<i>S. hydroxylic</i>	NA	NA	9	8.59	54.1	7012	100	β
<i>Oscillatoria sp.</i>	PCC 6506	<a href="#">NZ_CACA00000000.1</a>	<i>Toxinema</i>	<i>T. oscillati</i>	NA	NA	377	6.67	43.4	6007	99.12	β
<i>Oscillatoria sp.</i>	PCC 6407 <sup>6</sup>	<a href="#">NZ_ALVI00000000.1</a>	<i>Toxinema</i>	<i>T. oscillati</i>	Freshwater	NA	12	6.89	43.43	5693	99.56	β
<i>Parasynechococcus nordiatlanticus</i>	BL107 <sup>T</sup> •	<a href="#">NZ_DS022298.1</a>			Marine	Blanes Bay, Mediterranean Sea, 1,800 m	1	2.29	54.2	2322	99.46	α
<i>Parasynechococcus sudiatlanticus</i>	CB0101 <sup>T</sup> •	<a href="#">NZ_ADXL00000000.1</a>			Marine	Chesapeake Bay, Baltimore, Maryland, USA	94	2.69	64.2	2757	99.73	α
<i>Parasynechococcus indicus</i>	CB0205 <sup>T</sup> •	<a href="#">NZ_ADXM00000000.1</a>			Marine	Chesapeake Bay, Baltimore, Maryland, USA	78	2.43	63	2473	99.18	α
<i>Parasynechococcus benguelii</i>	CC9311 <sup>T</sup> •	<a href="#">NC_008319.1</a>			Marine	California current, Pacific, coastal, 95 m	1	2.61	52.4	2627	99.73	α
<i>Parasynechococcus</i>	CC9605 <sup>T</sup> •	<a href="#">NC_007516</a>			Marine	California	1	2.51	59.2	2583	99.73	α

## africanus

					current, Pacific, oligotrophic, 51 m									
					California									
<i>Parasynechococcus chillensis</i>	CC9902 <sup>T</sup> •	<a href="#">NC_007513</a>		Marine	current, Pacific, oligotrophic, 5 m	1	2.23	54.2	2289		99.46			α
<i>Parasynechococcus mediterranei</i>	RCC307 <sup>T</sup> •	<a href="#">NC_009482.1</a>		Marine	Mediterranean Sea, 15 m	1	2.22	60.8	2348		99.64			α
<i>Parasynechococcus gyrus</i>	RS9916 <sup>T</sup> •	<a href="#">NZ_DS022299.1</a>		Marine	Gulf of Aqaba, Red Sea, 10 m	1	2.66	59.8	2603		99.73			α
<i>Parasynechococcus equatorialis</i>	RS9917 <sup>T</sup> •	<a href="#">NZ_CH724158.1</a>		Marine	Gulf of Aqaba, Red Sea, 10 m	1	2.58	64.4	2575		99.46			α
<i>Parasynechococcus antarcticus</i>	WH5701 <sup>T</sup> •	<a href="#">NZ_CH724159– NZ_CH724167</a>		Marine	Long Island Sound, Connecticut, USA	116	3.28	65.4	2917		99.46			α
<i>Parasynechococcus pacificus</i>	WH7803 <sup>T</sup> •	<a href="#">NC_009481</a>		Marine	Sargasso Sea, 25 m	1	2.37	60.2	2439		99.18			α
<i>Parasynechococcus subtropicalis</i>	WH7805 <sup>T</sup> •	<a href="#">NZ_CH724168.1</a>		Marine	Sargasso Sea	3	2.63	57.6	2595		99.73			α
<i>Parasynechococcus sudipacificus</i>	WH8016 <sup>T</sup> •	<a href="#">AGIK00000000.1</a>		Marine	Woods Hole, MA, USA	16	2.69	54.1	2990		99.18			α
<i>Parasynechococcus marenigrum</i>	WH8102 <sup>T</sup> •	<a href="#">NC_005070.1</a>		Marine	Sargasso Sea	1	2.43	59.4	2461		99.46			α
<i>Parasynechococcus marearabicus</i>	WH8109 <sup>T</sup> •	<a href="#">ACNY00000000.1</a>		Marine	Sargasso Sea	1	2.12	60.1	2661		99.32			α
<i>Planktothrix agardhii</i>	NIVA-CYA 126	<a href="#">NZ_CM002803.1</a>	<i>P. stereotis</i>	Freshwater	NA	13	5.04	39.57	4188		100			β
<i>Planktothrix agardhii</i>	NIVA-CYA 15	<a href="#">NZ_AVFS00000000.1</a>		Freshwater	NA	238	5.38	39.48	4606		100			β
<i>Planktothrix agardhii</i>	NIVA-CYA 56-3	<a href="#">NZ_AVFY00000000.1</a>		Freshwater	NA	185	5.48	39.48	4674		99.78			β
<i>Planktothrix mougeotii</i>	NIVA-CYA 405	<a href="#">NZ_AVFU00000000.1</a>	<i>P. agardhii</i>	Freshwater	NA	240	5.46	39.47	4697		99.56			β
<i>Planktothrix prolifica</i>	NIVA-CYA 406	<a href="#">NZ_AVFV00000000.1</a>	<i>P. agardhii</i>	Freshwater	NA	375	5.62	39.51	4873		100			β
<i>Planktothrix prolifica</i>	NIVA-CYA 540	<a href="#">NZ_AVFX00000000.1</a>	<i>P. agardhii</i>	Freshwater	NA	157	5.5	39.48	4710		99.78			β
<i>Planktothrix prolifica</i>	NIVA-CYA	<a href="#">NZ_AVFZ00000000.1</a>	<i>P. agardhii</i>	Freshwater	NA	346	5.61	39.52	4862		99.78			β

	98											
	NIVA-CYA											
<i>Planktothrix rubescens</i>	407	<a href="#">NZ_AVFW00000000.1</a>		<i>P. agardhii</i>	Freshwater	NA	219	5.39	39.46	4658	100	β
						Arizona Station, Gulf of California,						
<i>Pleurocapsa sp.</i>	PCC 7319	<a href="#">NC_019689.1</a>		<i>P. penascus</i>	Marine (coastal)	Puerto Penasco, Mexico	10	7.38	38.74	4516	99.56	β
				<i>E.</i>		Arabian Sea, 50						
<i>Prochlorococcus chisholmii</i>	AS9601 T °	<a href="#">NC_008816.1</a>	<i>Eurycolium</i>	<i>chisholmii</i>	Marine	m	1	1.66	31.32	1769	99.64	α
	MIT9202 T °					South Pacific, 79						
<i>Prochlorococcus nereus</i>		<a href="#">NZ_ACDW00000000.1</a>	<i>Eurycolium</i>	<i>E. nereus</i>	Marine	m	1	1.69	31.1	1795	98.78	α
						Equatorial						
<i>Prochlorococcus marinus</i>	MIT9211	<a href="#">NC_009976.1</a>			Marine	Pacific, 83 m	1	1.68	38.01	1748	99.73	α
						Equatorial						
<i>Prochlorococcus nereus</i>	MIT9215 °	<a href="#">NC_009840.1</a>	<i>Eurycolium</i>	<i>E. nereus</i>	Marine	Pacific, surface	1	1.73	31.15	1840	99.73	α
	MIT9301 T °					Sargasso Sea, 90						
<i>Prochlorococcus ponticus</i>		<a href="#">NC_009091.1</a>	<i>Eurycolium</i>	<i>E. ponticus</i>	Marine	m	1	1.64	31.34	1774	99.46	α
						Sargasso Sea, 100						
<i>Prochlorococcus swingsii</i>	MIT9303 °	<a href="#">NC_008820.1</a>	<i>Expansiococcus</i>	<i>E. swingsii</i>	Marine	m	1	2.68	50.01	2504	100	α
	MIT9312 T °			<i>E.</i>		Gulf Stream, 135						
<i>Prochlorococcus neptunius</i>		<a href="#">NC_007577.1</a>	<i>Eurycolium</i>	<i>neptunius</i>	Marine	m	1	1.7	31.21	1815	99.73	α
	MIT9313 T °					Gulf Stream, 135						
<i>Prochlorococcus swingsii</i>		<a href="#">NC_005071.1</a>	<i>Expansiococcus</i>	<i>E. swingsii</i>	Marine	m	1	2.41	50.74	2339	99.46	α
	MIT9515 T °					Equatorial						
<i>Prochlorococcus tetisii</i>		<a href="#">NC_008817.1</a>	<i>Eurycolium</i>	<i>E. tetisii</i>	Marine	Pacific, 15 m	1	1.7	30.79	1784	100	α
						Northern						
<i>Prochlorococcus proteus</i>	NATL1A °	<a href="#">NC_008819.1</a>	<i>Prolificoccus</i>	<i>P. proteus</i>	Marine	Atlantic, 30 m	1	1.86	34.98	2204	99.73	α
	NATL2A T °					Northern						
<i>Prochlorococcus proteus</i>		<a href="#">NC_007335.2</a>	<i>Prolificoccus</i>	<i>P. proteus</i>	Marine	Atlantic, 10 m	1	1.84	35.12	1930	99.45	α
	CCMP1375					Sargasso Sea, 120						
<i>Prochlorococcus marinus</i>	T	<a href="#">NC_005042.1</a>			Marine	m	1	1.75	36.44	1883	100	α
						Mediterranean						
<i>Prochlorococcus marinus</i>	CCMP1986	<a href="#">NC_005072.1</a>	<i>Eurycolium</i>	<i>E. marinus</i>	Marine	Sea, 5 m	1	1.65	30.8	1777	99.46	α
<i>Pseudanabaena biceps</i>	PCC 7429	<a href="#">NZ_ALWB00000000.1</a>			Freshwater	NA	464	5.47	43.18	4774	99.29	β
					Marine	Intertidal zone,						
<i>Pseudanabaena sp.</i>	PCC 7367	<a href="#">NC_019701.1</a>	<i>Leptocystis</i>	<i>L. gracile</i>	(coastal)	Mexico	1	4.55	46.31	3960	98.23	β
						California, USA						
<i>Pseudoanabaena sp.</i>	PCC 6802	<a href="#">ALVK00000000.1</a>	<i>Paraleptovivax</i>	<i>P.</i>	Freshwater		6	5.62	47.83	5363	99.76	β

<i>allomegium</i>												
<i>Rivularia sp.</i>	PCC 7116	<a href="#">NC_019678.1</a>	<i>Anabaena</i>	<i>A. rivularialis</i>	Marine (coastal)	Baja California, USA	3	8.72	37.53	6612	99.78	β
<i>Spirulina subsalsa</i>	PCC 9445	<a href="#">NZ_ALVR00000000.1</a>	<i>Paraspirulina</i>	<i>P. subsalsa</i>	NA	NA	10	5.32	47.39	4580	99.56	β
<i>Stanieria cyanosphaera</i>	PCC 7437 <sup>T</sup>	<a href="#">NC_019748.1</a>			Freshwater	Havana, Cuba	6	5.54	36.22	4895	99.56	β
<i>Synechococcus elongatus</i>	PCC 6301 <sup>T</sup>	<a href="#">NC_006576.1</a>			Freshwater	NA	1	2.7	55.5	2576	99.73	β
<i>Synechococcus elongatus</i>	PCC 7942	<a href="#">NC_007604.1</a>			Freshwater	NA	2	2.74	55.46	2655	100	β
<i>Synechococcus springii</i>	JA23Ba213 <sup>T</sup> •	<a href="#">NC_007776</a>	<i>Leptococcus</i>	<i>L. springii</i>	Freshwater	Octopus Spring, Yellowstone Park, USA	1	3.05	58.5	3064	100	β
<i>Synechococcus yellowstonii</i>	JA33Ab <sup>T</sup> • PCC 6312 <sup>T</sup>	<a href="#">NC_007775.1</a>	<i>Leptococcus</i>	<i>L. yellowstonii</i>	Freshwater	Octopus Spring, Yellowstone Park, USA	1	2.93	60.2	3036	100	β
<i>Synechococcus californii</i>	• PCC 7002 <sup>T</sup>	<a href="#">NC_019680.1</a>	<i>Stenotopis</i>	<i>S. californii</i>	Freshwater	California, USA	2	3.72	48.49	3795	99.29	β
<i>Synechococcus euryhalinus</i>	•	<a href="#">NC_010475.1</a>	<i>Symphothece</i>	<i>S. euryhalinus</i>	Unknown	Unknown Snail shell, intertidal zone, Puerto Penasco, Mexico	7	3.41	49.16	3121	100	β
<i>Synechococcus mexicanus</i>	PCC 7335 <sup>T</sup> •	<a href="#">ABRV00000000.1</a>	<i>Coccosdissimilis</i>	<i>C. mexicanus</i>	Marine (coastal)	Sea Water Tank, Berkeley University, CA, USA	11	5.97	48.2	5702	98.91	β
<i>Synechococcus berkleyi</i>	PCC 7336 <sup>T</sup> •	<a href="#">ALWC00000000.1</a>	<i>Eurycoccus</i>	<i>E. berkleyi</i>	Marine (coastal)	USA	1	5.07	53.7	5093	100	β
<i>Synechococcus bogii</i>	PCC 7502 <sup>T</sup> •	<a href="#">CP003594.1</a>	<i>Leptovivax</i>	<i>L. bogii</i>	Sphagnum bog	NA	3	3.58	40.6	3703	99.76	β
<i>Synechocystis sp.</i>	PCC 7509	<a href="#">ALVU00000000.2</a>	<i>Chroococciopsis</i>	<i>C. synechocystis</i>	Soil	Rock scraping, Switzerland	4	4.9	41.67	4859	99.67	β
<i>Trichodesmium erythraeum</i>	IMS01 <sup>T</sup>	<a href="#">NC_008312.1</a>			Marine (coastal)	NA	1	7.75	34.14	4358	99.71	β
<i>Xenococcus sp.</i>	PCC 7305	<a href="#">NZ_ALVZ00000000.1</a>		<i>X. lajollai</i>	Marine (coastal)	Aquarium, La Jolla, CA, USA	234	5.92	39.68	4992	99.78	β
<i>Gloeobacter violaceum</i>	PCC 7421	<a href="#">NC_005125.1</a>			Soil	Calcareous (chalky) rock,	1	4.66	62	4511	99.15	β

Switzerland

1040

1041

**Add informations:**

<sup>1</sup> Cyanobacterial genomes used in Komarek et al. (2014) paper and available at public database in January 2016 were retrieved for this study.

Type strains or type species are indicated with overwritten T at the end of the name.

<sup>2</sup> *Arthrospira platensis* is also called *Spirulina platensis*.

<sup>3</sup> *Coleofasciculus chthonoplastes* PCC 7420 is also called *Microcoleus chthonoplastes* PCC 7420.

<sup>4</sup> *Leptolyngbya* sp. PCC 7104 is also called *Nodosilinea nodulosa* PCC 7104.

<sup>5</sup> *Moorea producens* 3L is also called *Moorea producta* 3L.

<sup>6</sup> *Oscillatoria* sp. PCC 6407 is also called *Kamptonema formosum* PCC 6407, and even *O. formosa* PCC 6407.

<sup>7</sup> Outgroup used in the phylogenetic analysis.

• New taxonomic identification proposed by Coutinho et al. 2016.

° New taxonomic classification proposed by Thompson et al. 2013.

Δ Number of contigs, total length and GC content values were obtained using QUASt tool.

\* Values using CheckM tool.

1042 **Table 1. Details of all cyanobacterial genomes included in this study.** Ecological and  
1043 molecular features were indicated, such as environment sampling, as well as number of  
1044 contigs, genome size, GC % content, completeness score, and carboxysome type.  
1045 Overwritten T indicates Type strain or Type species.