Insights on the taxonomy and ecogenomics of the *Synechococcus* collective

Vinícius W. Salazar[1,2], *Cristiane C. Thompson*[3], *Diogo A. Tschoeke*[4], *Jean Swings*[5], *Marta Mattoso*[2], *Fabiano L. Thompson*[1,3]*

[1]Center of Technology-CT2, SAGE-COPPE, [2]Department of Systems and Computer Engineering, COPPE, [3]Institute of Biology, [4]Department of Biomedical Engineering, COPPE, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil. [5]Laboratory of Microbiology, Ghent University, Ghent, Belgium. Corresponding author: *fabianothompson1@gmail.com

ABSTRACT

The genus *Synechococcus* (also named *Synechococcus* collective, SC) is a major contributor to global primary productivity. It is found in a wide range of aquatic ecosystems. *Synechococcus* is metabolically diverse, with some lineages thriving in polar and nutrient-rich locations, and other in tropical riverine waters. Although many studies have discussed the ecology and evolution of *Synechococcus*, there is a paucity of knowledge on the taxonomic structure of SC. Only a few studies have addressed the taxonomy of SC, and this issue still remains largely ignored. Our aim was to establish a new classification system for SC. Our analyses included comparing GC% content, genome size, pairwise Average Amino acid Identity (AAI) values, phylogenomics and gene cluster profiles of 170 publicly available SC genomes. All analyses were consistent with the discrimination of 11 genera, from which 2 are newly proposed (*Lacustricoccus* and *Synechospongium*). The new classification is also consistent with the habitat distribution (seawater, freshwater and thermal environments) and reflects the ecological and evolutionary relationships of SC. We provide a practical and consistent classification scheme for the entire *Synechococcus* collective.

INTRODUCTION

Synechococcus was first described by Carl Nägeli in the mid-19th century (Nägeli 1849) and ever since *S. elongatus* has been considered its type species (holotype). *Synechococcus* were regarded mostly as freshwater bacteria related to the *Anacystis* genus (Ihlenfeldt & Gibson, 1975), which is considered a heterotypic synonym for the *Synechococcus* genus. Species later described as *Synechococcus* were also found in thermal springs and microbial mats (Copeland, 1936, Inman, 1940). With the subsequent discovery of marine *Synechococcus* (Waterbury et al. 1979), which were classified as such based on the defining characters of cyanobacteria, described by Stanier (1971), the genus aggregated organisms with distinct ecological and physiological characteristics. The first analysis of the complete genome of a marine *Synechococcus* (Palenik et al. 2003) already displayed several differences to their freshwater counterparts, such as nickel- and cobalt- (as opposed to iron) based enzymes, reduced regulatory mechanisms and motility mechanisms.

Cyanobacteria of the genus *Synechococcus* are of vital importance, contributing to aquatic ecosystems at a planetary scale (Zwirglmaier et al. 2008, Huang et al. 2012). Along with the closely related *Prochlorococcus*, it is estimated that these organisms are responsible for at least one quarter of global primary productivity (Flombaum et al. 2013), therefore being crucial to the regulation of all of Earth's ecosystems (Bertilsson et al. 2003). Both of these taxa are globally abundant, but while *Prochlorococcus* is found in a more restricted latitudinal range, *Synechococcus* is more widely distributed, being found in freshwater ecosystems, hot spring microbial mats, polar regions, and nutrient-rich waters (Farrant et al. 2016, Sohm et al. 2016, Lee et al. 2019). This demonstrates the metabolic diversity of *Synechococcus*, which has served as a model organism for biotechnological applications (Hendry et al. 2016). Genomic studies deepened our understanding of the unique adaptions of different lineages in the group, regarding their light utilization (Six et al. 2007), nutrient and metal uptake (Palenik et al. 2006) and motility strategies (Dufresne et al. 2008). By analysing the composition of *Synechococcus* genomes, Dufresne and colleagues (2008) identified two distinct lifestyles in marine *Synechococcus* lineages, corresponding to coastal or open ocean habitats, and although there might be an overlap in geographical distribution, niche partitioning is affected by the presence and absence of genes. These insights were mostly restricted to marine *Synechococcus* genomes, and by then, freshwater strains still had their taxonomy status relatively poorly characterized. With these early genomic studies, clear separations started to show between the freshwater type species *Synechococcus elongatus* PCC 6301 and marine lineages such as WH8102 and WH8109. Gene sequences identified as Synechospongium appear in numerous ecological studies as a major component of different sponge species (Erwin & Tacker, 2008). However, this genus has not been formally described, having an uncertain taxonomic position. Despite remarkable ecological and physiological differences within the *Synechococcus* and the successful identification of distinct genomic clades (Ahlgren & Rocap 2012, Mazard et al. 2012, Farrant et al 2016, Sohm et al 2016), the taxonomy of the Synechococcus collective (SC) remained largely unresolved.

A first attempt to unlock the taxonomy of SC was performed by Coutinho et al (2016ab). They compared 24 *Synechococcus* genomes and i. proposed the creation of the new genus *Parasynechococcus* to encompass the marine lineages and ii. described 15 new species (Coutinho et al. 2016b). The description of these new species was attributed to the genetic diversity within these genomes, approaching the problem of classifying all of them under the same name (an issue previously raised by Shih et al. 2013). The new nomenclature also highlighted the genetic difference between marine *Parasynechococcus* and freshwater *Synechococcus*. Walter et al (2017) further elucidates this difference and propose 12 genera for the SC. However, the limited number genomes examined in this previous study hampered a more fine-grained taxonomic analysis of the *Synechococcus* collective.

The present work performs a comprehensive genomic taxonomy analyses using 170 presently available genomes. By combining several genome-level analysis (GC% content, genome size, AAI, phylogenetic reconstruction, gene cluster profiling), we propose splitting the *Synechococcus* collective into 11 clearly separated genera, including two new genera (*Lacustricoccus* and *Synechospongium*). Genus level definition of prokaryotic organisms has been based on the use of AAI (Konstantinidis & Tiedje 2005, Thompson et al. 2013). Modified versions of AAI have also been employed in defining genus level boundaries (Qin et al. 2014) and evolutionary rates across taxonomic ranks (Hugenholtz et al 2016, Parks et al 2018). Therefore, genera were broadly defined based on an AAI cutoff and supported by further genomic analysis, such as the phylogenomic trees, required to confirm genus level definitions (Chun et al. 2018). Based on the presently available data of *Synechococcus* genomes, we propose a new genome-based taxonomy for the group, splitting the *Synechococcus* collective into 10 clearly separated genera, and the creation of two new genera.

METHODS

Data acquisition and processing

All *Synechococcus* genomes (n=229) were downloaded from NCBI Assembly database (Kitts et al. 2015) in February 2020 using the Python package “NCBI Genome Download” (https://github.com/kblin/ncbi-genome-download) and querying for the genus “*Synechococcus*”. The metadata table with NCBI Entrez data generated by the package was used as a template for the metadata master table (Table S1). To ensure a standardized treatment of each genome data, instead of using the preexisting files from the assembly directories available at NCBI, only assembly files (containing complete chromosomes, scaffolds, or contigs) were used for analysis.

Quality assurance

To infer the completeness of each genome, we used CheckM v1.0.12 (Parks et al. 2015) with the "taxonomy_wf" workflow and default settings. The workflow is composed of three steps: i) "taxon_set", where a taxonomic-specific marker gene set is generated from reference genomes of the selected taxon (in this case, the genus *Synechococcus*), ii) "analyse", where the marker genes are identified in the genomes, and iii) "qa", where genomes are assessed for contamination and completeness based on the presence/absence of the marker genes. CheckM results were then parsed with the Pandas v0.25.1 package (McKinney 2011) in a Jupyter Notebook (Ragan-Kelley et al. 2014). Results for completeness and contamination were then added to the master metadata table (Table S1). For all further analyses, we only used genomes with at least 50% completeness and less than 10% contamination as inferred by CheckM. We also removed 9 genomes that did not bin with any other genomes at a 70% AAI cutoff. Thus, 50 "low quality" and 9 "singleton" genomes were discarded, leaving 170 genomes for downstream analyses.

GC content and genome size

GC content and genome size statistics were calculated from contigs files downloaded from NCBI using Python functions and are displayed in the metadata table (Table S1). The data was aggregated with Pandas to produce the values in Figure 1 and Table 1. For plotting, the libraries Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2018) were used.

AAI analysis

Comparative Average Amino acid Identity (AAI) analysis was carried out with the CompareM package (https://github.com/dparks1134/CompareM) v0.0.23. To do so, we ran CompareM's "aai_wf", which utilizes protein coding sequences (CDS) predicted with Prodigal (Hyatt et al. 2007), performs all-vs-all reciprocal sequence similarity search with Diamond (Buchfink et al. 2014) and computes pairwise AAI values based on the orthologous fraction shared between genes of the two genomes. The command was run on default settings, with parameters for defining homology being >30% sequence similarity and >70% alignment length. The output table from the AAI analysis was then imported into a Jupyter Notebook a symmetrical distance table was constructed using Pandas v0.25.1. This table is the transformed into a one-dimensional condensed distance matrix using the "squareform()" function from the SciPy library (Jones et al. 2001), "spatial" package. This resulting matrix is subjected to clustering with the "linkage()" function (SciPy library, "cluster" package) with the "method='complete'", "metric='cityblock'" and "optimal_ordering=True" parameters. A more in-depth explanation of these parameters can be found in the SciPy documents page (https://docs.scipy.org/doc/scipy/reference/index.html). The resulting array is used as input into a customized function based on SciPy's "dendrogram()" function.

For our analysis, we performed a hierarchical clustering of pairwise AAI values between all 139 genomes, defining a >70% cutoff for genera (Figure 2). This cutoff is empirically defined by previous studies

(Thompson et al. 2013, Rodriguez & Konstantinidis 2014, Qin et al. 2014). Genomes which didn't cluster with any other genomes based on this criterium were removed from downstream analyses.

Names for each genera were maintained the same as in Walter et al (2017). An exception to that are the newly-named *Synechospongium* gen. nov. and *Lacustricoccus* gen. nov. Species were defined at a >5% AAI cutoff (based on Thompson et al. 2013). New species were left unnamed. To define a type genome for each species, we used the following criteria, in order of priority: Whether the genome had already been used as a type genome; Genome completeness; Genome release date; Genome source (with a preference for single-cell, then isolate, then metagenome-augmented genomes).

Phylogenetic trees

To build the phylogenetic trees, we used the GToTree package (Lee, 2019) with default parameters. Two trees were generated, the first (Figure 3, panel A) using 251 Cyanobacteria marker genes and the second (Figure 3, panel B) using 74 Bacteria marker genes. The input dataset consisted of the 170 quality-filtered *Synechococcus* genomes with the addition of a *Prochlorococcus marinus* genome (strain CCMP1375, Genbank accession GCA_000007925.1) to serve as the root for each tree. The genomes were searched against a Hidden Markov Model of the marker genes using HMMER3 (Eddy, 2011). From the 171 genomes, 162 and 160 genomes were respectively retained in the first and second tree after GToTree's default settings quality control. A concatenated protein alignment from the marker genes was constructed using Muscle (Edgar, 2004) and subsequently trimmed using TrimAl (Capella-Gutiérrez et al. 2009). The alignment was then used to construct a tree using Fast Tree 2 (Price et al. 2010) with default parameters and the pairwise distance matrix using MEGA 6.0 (Tamura, 2013). All processing was done with GNU Parallel (Tange 2018). Trees were rendered using ETE 3 (Huerta-Cepas et al. 2016).

CyCOG profiles and *k*-means analysis.

Cyanobacterial Clusters of Orthologous Groups profiles were determined by aligning the proteome profiles predicted with Prodigal (see the "AAI analysis" section above) against the NCBI COG database (Galperin et al. 2014) using Diamond in using the parameters 'evalue=10e-6' and 'max_target_alignments=1'. The resulting hits table was filtered against the CyCOG database (Berube et al. 2018), preserving only COGs from cyanobacterial-related genomes. To minimize false negatives gene occurrences, stricter constraints on genome quality were used, and only genomes with at least 95% completeness (as estimated by CheckM) were kept in the CyCOG table. The resulting table (Table S2) was converted to binary form (1 if a CyCOG product was present in a genome and 0 if it was not) and used to plot Figure 4 (CyCOG profiles).

K-means analyses were conducted with the implementation available in the SciPy cluster package using the resulting CyCOG table. Values used for *k* were 2, 3, and 4 and the resulting clusters are displayed in Table 2.

Data and code availability

Whole genome data can be downloaded directly from NCBI Assembly database using the accession codes available in Table S1, in the “assembly_accession” column. We recommend using the above cited “NCBI Genome Download” package to facilitate this. Data generated from CompareM and GToTree and code used for the analysis (in the format of Jupyter notebooks) are available in the following GitHub repository: https://github.com/vinisalazar/SynechococcusGT. Users are encouraged to recreate and examine the figures using Jupyter and the available data. The repository’s “Issues” tab may be used for any further data and/or code requests.

RESULTS & DISCUSSION

Synechococcus collective GC% content and genome size

Genomic diversity within the *Synechococcus* collective (SC) was observed at several scales, including GC% content and genome size (bp). The sheer span of these two features between genera of the SC indicates marked differences between them. The genome size varies from 0.99 to 3.47 megabase pairs (Mbps), and GC content varies from 49.12% to 69.2% (Figure 1a). However, when the SC is split into several genera, these GC content and genome size values become more consistent (Figure 1bc; Table 1) and closer to proposed ranges for taxonomic grouping (Meier-Kolthoff et al. 2014). Genetically homogeneous genera, such as *Enugrolinea*, *Synechococcus* and *Leptococcus* form clusters of very low variability in GC content and genome size (Figure 1a). Interestingly, the variability is not so low in the new genera *Synechospongium* (57.89% to 63.05% GC content and 1.31 to 2.27 Mbp) and *Lacustricoccus* (51.9% to 52.6% GC content and 1.47 to 2.67 Mbp).

Delimitation SC genera by Average Amino acid Identity (AAI)

The AAI analyses discriminated 11 genera (Figure 2). Genomes sharing >70% AAI were grouped into genera. Certain genera (e.g. *Lacustricoccus* and *Synechococcus*) are homogeneous, having at maximum 9.9% AAI difference. Meanwhile other genera (e.g. *Pseudosynechococcus* and *Parasynechococcus*) are very heterogeneous, having up to 29.1% AAI variation. Heterogeneous genera are mostly marine lineages, and display the highest number of genomes (47 and 41, respectively) (Table 1). They are considered oceanic generalists, living in both low and high temperature environments (Walter et al. 2017). In contrast, the freshwater *Lacustricoccus* (previously *Synechococcus lacustris*; Cabello-Yevez et al. 2017, 2018), the thermophilic *Leptococcus*, isolated from Yellowstone hot springs (Becraft et al. 2011), and the *Synechospongium* gen nov. (previously Candidatus *Synechococcus spongiarum*), a symbiont to marine sponges (Usher et al. 2004, Erwin & Thacker 2008, Slaby & Hentschel 2017), appear all to have a more cohesive genome structure at the genus level. The genome previously classified as *Synechococcus lividus* PCC 6715, considered a thermophilic *Synechococcus*, was reclassified as the previously described genus *Thermosynechococcus* (Nakamura et al. 2002), thus enforcing the need to classify novel or earlier *Synechococcus* genomes into a new taxonomic framework. The AAI dendrogram also illustrates the difference between the major ecogenomic groups, which include: Marine/oceanic (*Parasynechococcus* and *Pseudosynechococcus*), Marine/coastal (*Magnicoccus*, *Regnicoccus*, *Lacustricoccus* and *Inmanicoccus*), Symbiont (*Synechospongium*), and freshwater/thermal (*Synechococcus* and *Enugrolinea* as freshwater representatives and *Thermosynechococcus* and *Leptococcus* as thermal representatives). The terms "Marine/oceanic" and "Marine/coastal" can also respectively be exchanged "high temperature/low nutrient" and "low temperature/high nutrient" environments.

Phylogenomic structure of the SC

Genera delimited by AAI analyses were also found by phylogenetic analyses (Figure 3). Both the 251 cyanobacterial marker gene tree and the 74 bacterial marker genes tree depict the eleven genera observed in the AAI dendrogram. The trees support the same groups discriminated in the AAI figure. However, the AAI was superior to discriminate the closely related genera *Magnicoccus* and *Regnicoccus*. These genera group together in both phylogenetic trees, but group separately in the AAI dendrogram (Figure 2). Despite sharing similar ecological characteristics, being sourced from coastal, estuarine-influenced waters, *Magnicoccus* and *Regnicoccus* have distinct GC% and genome size, reinforcing their status as separated genera. The two newly proposed genera (*Lacustricoccus* and *Synechospongium*) form monophyletic branches in both phylogenetic reconstructions, giving strong support for our proposal to formally create these new genera.

CyCOG profiles and *k*-means analyses.

Distinct profiles of Cyanobacterial Clusters of Orthologous Groups (CyCOGs) could be observed for each genus (Figure 4). It is possible to observe similar patterns of presence/absence of CyCOG products within each genus (Figure 4), and when subjected to *k*-means analysis, these patterns represent the same major groups identified in the AAI (Figure 2) and phylogenomic (Figure 3) analyses. Grouping into *k*-means is show in Table 2. When k = 2, the division is broad, between the Marine groups (including the Symbiont *Synechospongium*) and Freshwater/thermal. When k is raised to 3, the division is between Marine, Symbiont and Freshwater/thermal. When k = 4, the division is between Marine, Symbiont, Freshwater and Thermal genera. For each respective k value, the data shows that: i) The broadest ecogenomic divide is between genomes of marine and freshwater/thermal environments; ii) the Symbiont group is then separated, suggesting that its symbiotic lifestyle has led to a different pattern of CyCOG presence/absence within the Marine group (Slaby & Hentschel, 2017) and iii) Within the Freshwater/thermal group, the Freshwater and Thermal group display distinct patterns. There was little difference within genera of the Marine/oceanic and Marine/coastal groups. This was perhaps surprisingly, as some genomes from these groups come from very different environments, such as the *Regnicoccus* genome which are sourced from both temperate estuarine waters (the type species WH 5701 was isolated from the Long Island Sound, USA) (Fuller et al. 2003) and extreme environments such as the Ace Lake, in the Vestfold Hills of Antarctica (strain SynAce01) (Powell et al. 2005). The new genus *Lacustricoccus* is also surprisingly grouped within the Marine/coastal group, as genomes from this genus were sourced from brackish water reservoirs (Cabello-Yevez et al. 2017, 2018).

CONCLUSION

It is timely to establish a genome-based taxonomy for SC (Gevers et al. 2005, Stackebrandt 2006). With the advent of next generation sequencing and increasingly available sequence data, there has been a transition from the former paradigm of a 'polyphasic' taxonomy towards a genomic taxonomy (Thompson et al. 2015). Examining prokaryotic taxonomy using the organisms' whole genome would be able to capture meaningful relationships and define monophyletic groups, capturing their rate of evolution across taxonomic ranks (Hugenholtz et al. 2016, Parks et al. 2018). In their large-scale analysis, Parks and colleagues (2018) examined over 18000 genomes and divide the *Synechococcus* in at least 5 genera, but, these authors do not delve further into the detailed taxonomic analyses of the taxon. To the best of our knowledge, there is not a consensus on whether the *Synechococcus* form a monophyletic clade. This may be the case for specific marine or freshwater lineages, but when examined in the context of the *Cyanobacteria* phylum, the genus as presently classified is paraphyletic or polyphyletic as demonstrated here (Walter et al. 2017). Our advanced genomic taxonomy analyses demonstrate the heterogeneous nature of the SC collective. This study brings new insights into the taxonomic structure of SC collective with the evident distinction of 11 genera. We anticipate that this newly proposed taxonomic structure will be useful for further environmental surveys and ecological studies (Arevalo et al. 2019), including those targeting the identification of populations, ecotypes and species.

ACKNOWLEDGEMENTS

The authors thank CAPES and CNPq for funding.

REFERENCES

Ahlgren, N.A. & Rocap, G. 2012. Diversity and distribution of marine Synechococcus: multiple gene phylogenies for consensus classification and development of qPCR assays for sensitive measurement of clades in the ocean. *Front. Microbiol.* 3:213.

Becraft, E.D., Cohan, F.M., Kühl, M., Jensen, S.I. & Ward, D.M. 2011. Fine-scale distribution patterns of Synechococcus ecological diversity in microbial mats of Mushroom Spring, Yellowstone National Park. *Appl. Environ. Microbiol.* 77:7689–97.

Bertilsson, S., Berglund, O., Karl, D.M. & Chisholm, S.W. 2003. Elemental composition of marine Prochlorococcus and Synechococcus: Implications for the ecological stoichiometry of the sea.

Berube, P.M., Biller, S.J., Hackl, T., Hogle, S.L., Satinsky, B.M., Becker, J.W., Braakman, R. et al. 2018. Data descriptor: Single cell genomes of Prochlorococcus, Synechococcus, and sympatric microbes from diverse marine environments. *Sci. Data*. 5:1–11.

Buchfink, B., Xie, C. & Huson, D.H. 2014. Fast and sensitive protein alignment using DIAMOND.

Cabello-Yeves, P.J., Haro-Moreno, J.M., Martin-Cuadrado, A.B., Ghai, R., Picazo, A., Camacho, A. & Rodriguez-Valera, F. 2017. Novel Synechococcus genomes reconstructed from freshwater reservoirs. *Front. Microbiol.*

Cabello-Yeves, P.J., Picazo, A., Camacho, A., Callieri, C., Rosselli, R., Roda-Garcia, J.J., Coutinho, F.H. et al. 2018. Ecological and genomic features of two widespread freshwater picocyanobacteria. *Environ. Microbiol.*

Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*.

Copeland, J.J. 1936. YELLOWSTONE THERMAL MYXOPHYCEAE. *Ann. N. Y. Acad. Sci.*

Coutinho, F.H., Dutilh, B.E., Thompson, C.C. & Thompson, F.L. 2016. Proposal of fifteen new species of Parasynechococcus based on genomic, physiological and ecological features. *Arch. Microbiol.* 198:973–86.

Coutinho, F., Tschoeke, D.A., Thompson, F. & Thompson, C. 2016. Comparative genomics of Synechococcus and proposal of the new genus Parasynechococcus. *PeerJ*. 4:e1522.

Chun, J., Oren, A., Ventosa, A., Christensen, H., Arahal, D.R., da Costa, M.S., Rooney, A.P. et al. 2018. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int. J. Syst. Evol. Microbiol.*

Dufresne, A., Ostrowski, M., Scanlan, D.J., Garczarek, L., Mazard, S., Palenik, B.P., Paulsen, I.T. et al. 2008. Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol.* 9:R90.

Eddy, S.R. 2011. Accelerated profile HMM searches. *PLoS Comput. Biol.*

Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–7.

Erwin, P.M. & Thacker, R.W. 2008. Cryptic diversity of the symbiotic cyanobacterium Synechococcus spongiarum among sponge hosts. *Mol. Ecol.* 17:2937–47.

Farrant, G.K., Doré, H., Cornejo-Castillo, F.M., Partensky, F., Ratin, M., Ostrowski, M., Pitt, F.D. et al. 2016. Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria. *Proc. Natl. Acad. Sci.* 113:E3365–74.

Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincon, J., Zabala, L.L., Jiao, N., Karl, D.M. et al. 2013. Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Proc. Natl. Acad. Sci.* 110:9824–9.

Fuller, N.J., Marie, D., Partensky, F., Vaulot, D., Post, A.F. & Scanlan, D.J. 2003. Clade-specific 16S ribosomal DNA oligonucleotides reveal the predominance of a single marine Synechococcus clade throughout a stratified water column in the Red Sea. *Appl. Environ. Microbiol.* 69:2430–43.

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., & Koonin, E. V. 2014. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Research*, *43*(D1), D261--D269.

Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E. et al.

2005. Reevaluating prokaryotic species. *Nat. Rev. Microbiol.* 3:733–9.

Hendry, J.I., Prasannan, C.B., Joshi, A., Dasgupta, S. & Wangikar, P.P. 2016. Metabolic model of Synechococcus sp. PCC 7002: Prediction of flux distribution and network modification for enhanced biofuel production. *Bioresour. Technol.* 213:190–7.

Huang, S., Wilhelm, S.W., Harvey, H.R., Taylor, K., Jiao, N. & Chen, F. 2012. Novel lineages of Prochlorococcus and Synechococcus in the global oceans. *ISME J.* 6:285–97.

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msw046

Hugenholtz, P., Skarshewski, A. & Parks, D.H. 2016. Genome-based microbial taxonomy coming of age. *Cold Spring Harb. Perspect. Biol.* 8:a018085.

Hunter, J.D. 2007. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*

Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W. & Hauser, L.J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 11:119.

Ihlenfeldt, M.J.A. & Gibson, J. 1975. Phosphate utilization and alkaline phosphatase activity in Anacystis nidulans (Synechococcus). *Arch. Microbiol.*

Inman, O.L. 1940. STUDIES ON THE CHLOROPHYLLS AND PHOTOSYNTHESIS OF THERMAL ALGAE FROM YELLOWSTONE NATIONAL PARK, CALIFORNIA, AND NEVADA. *J. Gen. Physiol.*

Jones, E., Oliphant, T., Peterson, P. & others 2001. SciPy: Open source scientific tools for Python.

Kent, A.G., Baer, S.E., Mouginot, C., Huang, J.S., Larkin, A.A., Lomas, M.W. & Martiny, A.C. 2019. Parallel phylogeography of Prochlorococcus and Synechococcus. *ISME J.*

Kitts, P.A., Church, D.M., Thibaud-Nissen, F., Choi, J., Hem, V., Sapojnikov, V., Smith, R.G. et al. 2015. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* 44:D73–80.

Konstantinidis, K.T. & Tiedje, J.M. 2005. Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol.* 187:6258–64.

Lee, M.D. 2019. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics*.

Lee, M.D., Ahlgren, N.A., Kling, J.D., Walworth, N.G., Rocap, G., Saito, M.A., Hutchins, D.A. et al. 2019. Marine Synechococcus isolates representing globally abundant genomic lineages demonstrate a unique evolutionary path of genome reduction without a decrease in GC content. *Environ. Microbiol.* 21:1677–86.

Mazard, S., Ostrowski, M., Partensky, F. & Scanlan, D.J. 2012. Multi-locus sequence analysis, taxonomic resolution and biogeography of marine Synechococcus. *Environ. Microbiol.*

McKinney, W. 2011. pandas: a foundational Python library for data analysis and statistics. *Python High Perform. Sci. Comput.* 14.

Meier-Kolthoff, J.P., Klenk, H.P. & Göker, M. 2014. Taxonomic use of DNA G+C content and DNA-DNA hybridization in the genomic age. *Int. J. Syst. Evol. Microbiol.* 64:352–6.

Nägeli, C. 1849. Gattungen einzelliger Algen, physiologisch und systematisch bearbeitet. *Neue Denkschriften der Allg. Schweizerischen Gesellschaft für die Gesammten Naturwissenschaften*. 10:1–139.

Nakamura, Y., Kaneko, T., Sato, S., Ikeuchi, M., Katoh, H., Sasamoto, S., Watanabe, A., Iriguchi, M., Kawashima, K., Kimura, T., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Nakazaki, N., Shimpo, S., Sugimoto, M., Takeuchi, C., … Tabata, S. 2002. Complete genome structure of the thermophilic cyanobacterium Thermosynechococcus elongatus BP-1. *DNA Research*. https://doi.org/10.1093/dnares/9.4.123

Palenik, B., Brahamsha, B., Larimer, F.W., Land, M., Hauser, L., Chain, P., Lamerdin, J. et al. 2003.

The genome of a motile marine Synechococcus. *Nature*.

Palenik, B., Ren, Q., Dupont, C.L., Myers, G.S., Heidelberg, J.F., Badger, J.H., Madupu, R. et al. 2006. Genome sequence of Synechococcus CC9311: Insights into adaptation to a coastal environment. *Proc. Natl. Acad. Sci. U. S. A.*

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P. & Tyson, G.W. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 25:1043–55.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.-A. & Hugenholtz, P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36:996–1004.

Powell, L.M., Bowman, J.P., Skerratt, J.H., Franzmann, P.D. & Burton, H.R. 2005. Ecology of a novel Synechococcus clade occurring in dense populations in saline Antarctic lakes. *Mar. Ecol. Prog. Ser.* 291:65–80.

Price, M.N., Dehal, P.S. & Arkin, A.P. 2010. FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS One*. 5.

Qin, Q.L., Xie, B. Bin, Zhang, X.Y., Chen, X.L., Zhou, B.C., Zhou, J., Oren, A. et al. 2014. A proposed genus boundary for the prokaryotes based on genomic insights. *J. Bacteriol.* 196:2210–5.

Ragan-Kelley, M., Perez, F., Granger, B., Kluyver, T., Ivanov, P., Frederic, J. & Bussonnier, M. 2014. The Jupyter/IPython architecture: a unified view of computational research, from interactive exploration to communication and publication. *In AGU Fall Meeting Abstracts*.

Rodriguez-R, L.M. & Konstantinidis, K.T. 2014. Bypassing Cultivation To Identify Bacterial Species Culture-independent genomic approaches identify credibly distinct clusters, avoid cultivation bias, and provide true insights into microbial species. *Microbe*. 9:111–7.

Shih, P.M., Wu, D., Latifi, A., Axen, S.D., Fewer, D.P., Talla, E., Calteau, A. et al. 2013. Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc. Natl. Acad. Sci. U. S. A.*

Six, C., Thomas, J.C., Garczarek, L., Ostrowski, M., Dufresne, A., Blot, N., Scanlan, D.J. et al. 2007. Diversity and evolution of phycobilisomes in marine Synechococcus spp.: A comparative genomics study. *Genome Biol.*

Slaby, B.M. & Hentschel, U. 2017. Draft Genome Sequences of "Candidatus Synechococcus spongiarum," cyanobacterial symbionts of the mediterranean sponge Aplysina aerophoba. *Genome Announc.* 5:e00268--17.

Sohm, J.A., Ahlgren, N.A., Thomson, Z.J., Williams, C., Moffett, J.W., Saito, M.A., Webb, E.A. et al. 2016. Co-occurring Synechococcus ecotypes occupy four major oceanic regimes defined by temperature, macronutrients and iron. *ISME J.* 10:333–45.

Stackebrandt, E. 2006. Defining taxonomic ranks. *Prokaryotes Vol. 1 Symbiotic Assoc. Biotechnol. Appl. Microbiol.* 29–57.

Stanier, R.Y., Kunisawa, R., Mandel, M. & Cohen-Bazire, G. 1971. Purification and properties of unicellular blue-green algae (order Chroococcales). *Bacteriol. Rev.*

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. 2013. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.*

Tange, O. 2011. GNU Parallel: the command-line power tool. *USENIX Mag.*

Thompson, C.C., Chimetto, L., Edwards, R.A., Swings, J., Stackebrandt, E. & Thompson, F.L. 2013. Microbial genomic taxonomy. *BMC Genomics*. 14.

Thompson, C.C., Amaral, G.R., Campeão, M., Edwards, R.A., Polz, M.F., Dutilh, B.E., Ussery, D.W. et al. 2015. Microbial taxonomy in the post-genomic era: Rebuilding from scratch? *Arch. Microbiol.*

Usher, K.M., Toze, S., Fromont, J., Kuo, J. & Sutton, D.C. 2004. A new species of cyanobacterial symbiont from the marine sponge Chondrilla nucula. *Symbiosis*. 36:183–92.

Walter, J.M., Coutinho, F.H., Dutilh, B.E., Swings, J., Thompson, F.L. & Thompson, C.C. 2017. Ecogenomics and taxonomy of Cyanobacteria phylum. *Front. Microbiol.* 8.

Waskom, M. 2018. Seaborn: statistical data visualization.

Waterbury, J.B., Watson, S.W., Guillard, R.R.L. & Brand, L.E. 1979. Widespread occurrence of a unicellular, marine, planktonic, cyanobacterium.

Zwirglmaier, K., Jardillier, L., Ostrowski, M., Mazard, S., Garczarek, L., Vaulot, D., Not, F. et al. 2008. Global phylogeography of marine Synechococcus and Prochlorococcus reveals a distinct partitioning of lineages among oceanic biomes. *Environ. Microbiol.* 10:147–61.

FIGURES AND TABLES

Table 1: Genera of the *Synechococcus* collective. In total eleven genera, from which two are proposed in the present study (*Lacustricoccus* and *Synechospongium*). Type genomes were chosen based on specific criteria (see Methods section - Description criteria). Additional information for all genomes can be found in Table S1. GC% and genome size (Mbp) values are shown for means ± standard deviation.

Genus	# genomes	# species*	Type Genome	NCBI name	Lifestyle	GC content (%)	Size (Mbps)
Parasynechococcus	47	22	*Parasynechococcus africanus* CC9605	*Synechococcus* sp.	Marine (oceanic)	58.14 ± 3.02	1.96 ± 0.46
Pseudosynechococcus	41	21	*Pseudosynechococcus subtropicalis* WH 7805	*Synechococcus* sp.	Marine (oceanic)	56.43 ± 3.19	2.22 ± 0.48
Synechospongium gen. nov.	28	7	*Synechospongium spongiarum* 15L	Candidatus *Synechococcus spongiarum*	Symbiont	61.56 ± 1.14	1.86 ± 0.28
Enugrolinea	12	3	*Enugrolinea euryhalinus* PCC 7002	*Synechococcus* sp.	Freshwater	49.26 ± 0.1	3.33 ± 0.11
Regnicoccus	9	7	*Regnicoccus antarcticus* WH 5701	*Synechococcus* sp.	Marine (coastal)	65.36 ± 2.46	2.79 ± 0.51
Inmanicoccus	8	5	*Inmanicoccus mediterranei* RCC307	*Synechococcus* sp.	Marine (coastal)	61.04 ± 1.55	1.78 ± 0.27
Leptococcus	8	2	*Leptococcus yellowstonii* JA-3-3Ab	*Synechococcus* sp.	Thermophilic	56.34 ± 2.74	3.06 ± 0.1
Thermosynechococcus	6	5	*Thermosynechococcus elongatus* BP-1	*Thermosynechococcus elongatus*	Thermophilic	53.65 ± 0.27	2.61 ± 0.06
Synechococcus	5	2	*Synechococcus elongatus* PCC 6301	*Synechococcus elongatus*	Freshwater	55.27 ± 0.25	2.75 ± 0.08
Lacustricoccus gen. nov.	3	2	*Lacustricoccus lacustris* TousA	*Synechococcus lacustris*	Brackish	51.81 ± 0.72	1.98 ± 0.62
Magnicoccus	3	2	*Magnicoccus sudiatlanticus* CB0101	*Synechococcus* sp.	Marine (coastal)	63.43 ± 0.56	2.53 ± 0.23

* Several genomes were added to species that were previously defined (in Walter et al 2017) by a single genome. These include, but are not limited to: *Pseudosynechococcus sudipacificus*, *Parasynechococcus marenigrum*, *Inmanicoccus mediterranei*, and, most notably, *Enugrolinea euryhalinus* and *Leptococcus yellowstonii*, respectively with 8 and 7 genomes. In addition to the support of previous species groups, our analysis also expands upon existing genera by proposing new, robust species groups inside of them, specially in *Parasynechoccocus*, with 3 new species (with type genomes N32, CC9616, and KORDI-49), containing a total of 16 genomes, and *Pseudosynechococcus*, with 5 new species (with type genomes MITS9504, MITS9508, AG-673-F03, BS55D, and UW105), and a total of 20 genomes. Type species for each species group are noted by a "T" character besides their name (Figure 2). The discovery of these new species can be attributed to a surge of newly available *Synechococcus* high quality whole genome data, obtained mainly from single-cell sequencing (Berube et al. 2018, Kent et al. 2019).

Table 2: *k*-means groups of CyCOG products. Using the CyCOG presence/absence table, genomes for each genus were clustered using the *k*-means algorithm with *k* values of 2, 3 and 4. All genomes within a genus fell into the same group, therefore it was possible to depict rows as genera instead of individual genomes. As the *k* values increases, it is possible to identify divides within the genera that correspond to ecogenomic groups.

Genus	**2-means**	**3-means**	**4-means**
Leptococcus	Freshwater/Thermal	Freshwater/Thermal	Thermal
Thermosynechococcus	Freshwater/Thermal	Freshwater/Thermal	Thermal
Synechococcus	Freshwater/Thermal	Freshwater/Thermal	Freshwater
Enugrolinea	Freshwater/Thermal	Freshwater/Thermal	Freshwater
Synechospongium	Seawater	Symbiont	Symbiont
Regnicoccus	Seawater	Seawater	Seawater
Pseudosynechococcus	Seawater	Seawater	Seawater
Parasynechococcus	Seawater	Seawater	Seawater
Magnicoccus	Seawater	Seawater	Seawater
Lacustricoccus	Seawater	Seawater	Seawater
Inmanicoccus	Seawater	Seawater	Seawater

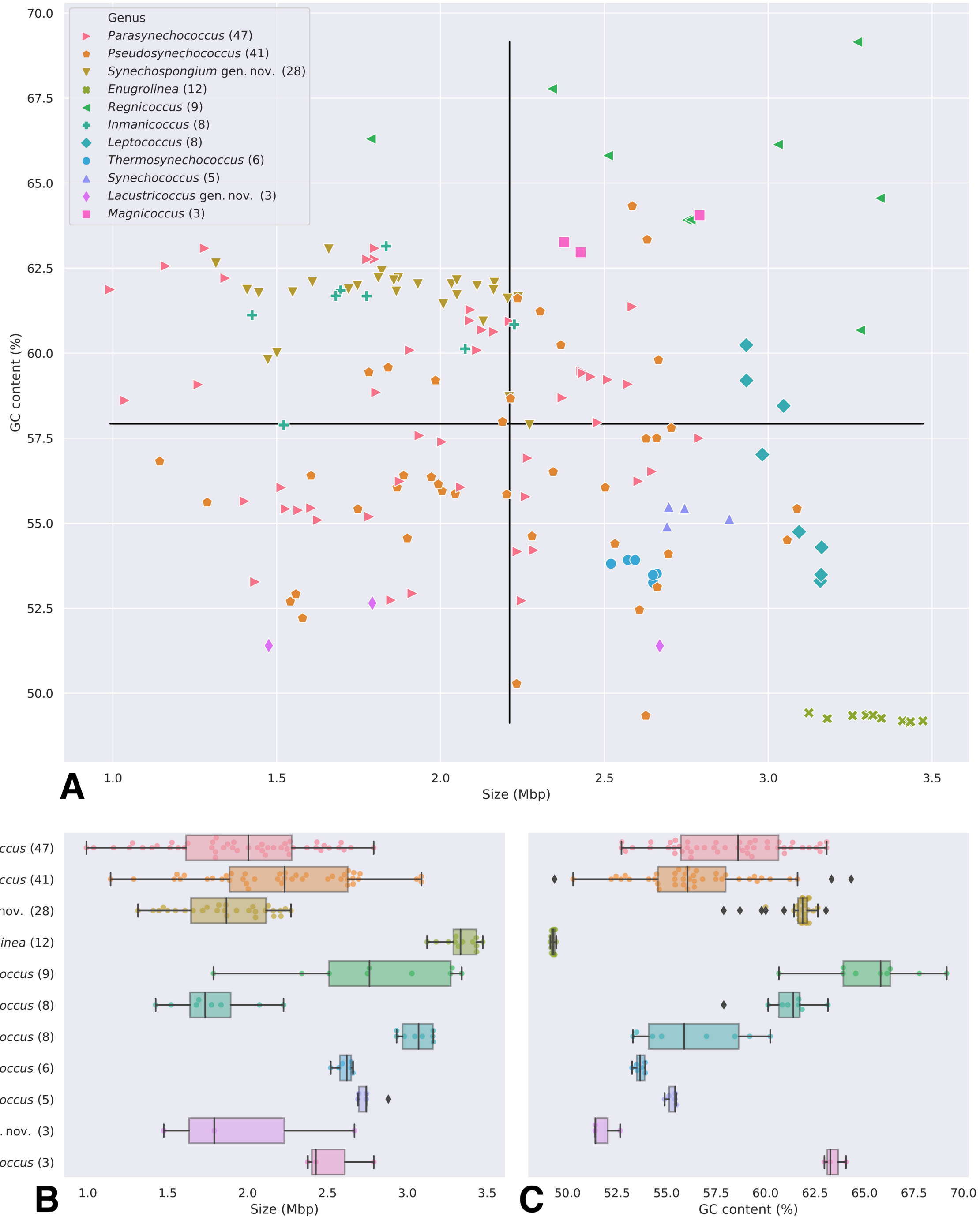
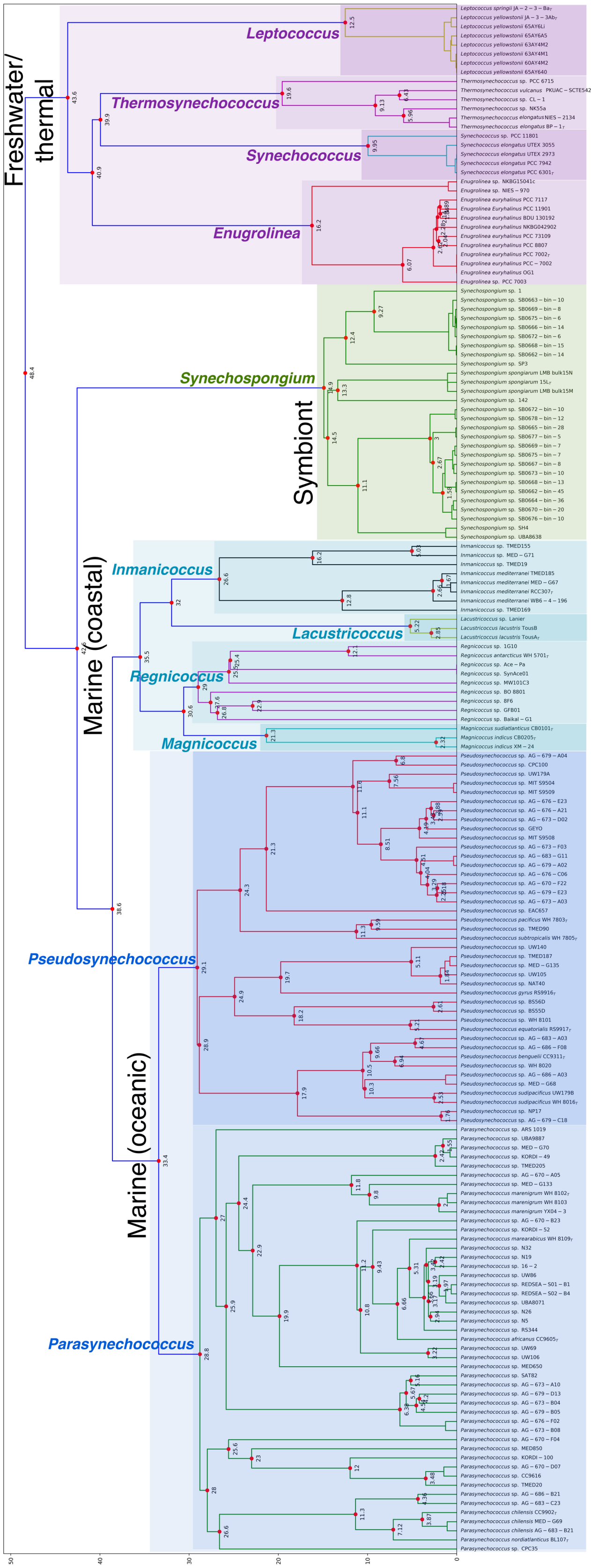


Figure 1: GC content and genome size charts. A. Scatter plot of GC content and genome size (in megabases). Black lines indicate the median for all genomes. Genera with lower genetic variability (as shown in the AAI dendrogram) cluster together in small GC/size ranges (with the exception of *Synechospongium* gen. nov.). The genera with most genomes (*Parasynechococcus* and *Pseudosynechococcus*) display a variable GC/size range but still there are no outliers. **B** and **C.** Box plots of genome size (**B**) and GC content (**C**) for each genus. Outliers are shown in diamond shapes. Error bars represent the 1st and 4th quartiles, boxes represent 2nd and 3rd quartiles and the median.

Figure 2: Hierarchical clustering of pairwise AAI values between all *Synechococcus* genomes. New proposed genera are shown within a >70% AAI cutoff. Dotted values show AAI 'dissimilarity' values (e.g. 100 minus the AAI value for the pairwise comparison). Dotted values < 1.5 were omitted. Species were defined at a >5% AAI cutoff (Thompson et al. 2013). Type genomes for each SLB are signaled with a "T" character next to the strain name, based on defined criteria (see Methods section). New species were left named as "sp.". Economic groups are labeled and highlighted in either blue, cyan, green, or purple.


Freshwater/ thermal
Leptococcus
Thermosynechococcus
Synechococcus
Enugrolinea
Synechospongium
Symbiont
Inmanicoccus
Lacustricoccus
Regnicoccus
Magnicoccus
Marine (coastal)
Pseudosynechococcus
Parasynechococcus
Marine (oceanic)
Leptococcus springii JA − 2 − 3 − Ba$_T$
Leptococcus yellowstonii JA − 3 − 3Ab$_T$
Leptococcus yellowstonii 65AY6Li
Leptococcus yellowstonii 65AY6A5
Leptococcus yellowstonii 63AY4M2
Leptococcus yellowstonii 63AY4M1
Leptococcus yellowstonii 60AY4M2
Leptococcus yellowstonii 65AY640
Thermosynechococcus sp. PCC 6715
Thermosynechococcus vulcanus PKUAC − SCTE542$_T$
Thermosynechococcus sp. CL − 1
Thermosynechococcus sp. NK55a
Thermosynechococcus elongatus NIES − 2134
Thermosynechococcus elongatus BP − 1$_T$
Synechococcus sp. PCC 11801
Synechococcus elongatus UTEX 3055
Synechococcus elongatus UTEX 2973
Synechococcus elongatus PCC 7942
Synechococcus elongatus PCC 6301$_T$
Enugrolinea sp. NKBG15041c
Enugrolinea sp. NIES − 970
Enugrolinea euryhalinus PCC 7117
Enugrolinea Euryhalinus PCC 11901
Enugrolinea euryhalinus BDU 130192
Enugrolinea euryhalinus NKBG042902
Enugrolinea euryhalinus PCC 73109
Enugrolinea euryhalinus PCC 8807
Enugrolinea euryhalinus PCC 7002$_T$
Enugrolinea euryhalinus PCC − 7002
Enugrolinea euryhalinus OG1
Enugrolinea sp. PCC 7003
Synechospongium sp. 1
Synechospongium sp. SB0663 − bin − 10
Synechospongium sp. SB0669 − bin − 8
Synechospongium sp. SB0675 − bin − 6
Synechospongium sp. SB0666 − bin − 14
Synechospongium sp. SB0672 − bin − 6
Synechospongium sp. SB0668 − bin − 15
Synechospongium sp. SB0662 − bin − 14
Synechospongium sp. SP3
Synechospongium spongiarum LMB bulk15N
Synechospongium spongiarum 15L$_T$
Synechospongium spongiarum LMB bulk15M
Synechospongium sp. 142
Synechospongium sp. SB0672 − bin − 10
Synechospongium sp. SB0678 − bin − 12
Synechospongium sp. SB0665 − bin − 28
Synechospongium sp. SB0677 − bin − 5
Synechospongium sp. SB0669 − bin − 7
Synechospongium sp. SB0675 − bin − 7
Synechospongium sp. SB0667 − bin − 8
Synechospongium sp. SB0673 − bin − 10
Synechospongium sp. SB0668 − bin − 13
Synechospongium sp. SB0662 − bin − 45
Synechospongium sp. SB0664 − bin − 36
Synechospongium sp. SB0670 − bin − 20
Synechospongium sp. SB0676 − bin − 10
Synechospongium sp. SH4
Synechospongium sp. UBA8638
Inmanicoccus sp. TMED155
Inmanicoccus sp. MED − G71
Inmanicoccus sp. TMED19
Inmanicoccus mediterranei TMED185
Inmanicoccus mediterranei MED − G67
Inmanicoccus mediterranei RCC307$_T$
Inmanicoccus mediterranei WB6 − 4 − 196
Inmanicoccus sp. TMED169
Lacustricoccus sp. Lanier
Lacustricoccus lacustris TousB
Lacustricoccus lacustris TousA$_T$
Regnicoccus sp. 1G10
Regnicoccus antarcticus WH 5701$_T$
Regnicoccus sp. Ace − Pa
Regnicoccus sp. SynAce01
Regnicoccus sp. MW101C3
Regnicoccus sp. BO 8801
Regnicoccus sp. 8F6
Regnicoccus sp. GFB01
Regnicoccus sp. Baikal − G1
Magnicoccus sudiatlanticus CB0101$_T$
Magnicoccus indicus CB0205$_T$
Magnicoccus indicus XM − 24
Pseudosynechococcus sp. AG − 679 − A04
Pseudosynechococcus sp. CPC100
Pseudosynechococcus sp. UW179A
Pseudosynechococcus sp. MIT S9504
Pseudosynechococcus sp. MIT S9509
Pseudosynechococcus sp. AG − 676 − E23
Pseudosynechococcus sp. AG − 676 − A21
Pseudosynechococcus sp. AG − 673 − D02
Pseudosynechococcus sp. GEYO
Pseudosynechococcus sp. MIT S9508
Pseudosynechococcus sp. AG − 673 − F03
Pseudosynechococcus sp. AG − 683 − G11
Pseudosynechococcus sp. AG − 679 − A02
Pseudosynechococcus sp. AG − 676 − C06
Pseudosynechococcus sp. AG − 670 − F22
Pseudosynechococcus sp. AG − 679 − E23
Pseudosynechococcus sp. AG − 673 − A03
Pseudosynechococcus sp. EAC657
Pseudosynechococcus pacificus WH 7803$_T$
Pseudosynechococcus sp. TMED90
Pseudosynechococcus subtropicalis WH 7805$_T$
Pseudosynechococcus sp. UW140
Pseudosynechococcus sp. TMED187
Pseudosynechococcus sp. MED − G135
Pseudosynechococcus sp. UW105
Pseudosynechococcus sp. NAT40
Pseudosynechococcus gyrus RS9916$_T$
Pseudosynechococcus sp. BS56D
Pseudosynechococcus sp. BS55D
Pseudosynechococcus sp. WH 8101
Pseudosynechococcus equatorialis RS9917$_T$
Pseudosynechococcus sp. AG − 683 − A03
Pseudosynechococcus sp. AG − 686 − F08
Pseudosynechococcus benguelii CC9311$_T$
Pseudosynechococcus sp. WH 8020
Pseudosynechococcus sp. AG − 686 − A03
Pseudosynechococcus sp. MED − G68
Pseudosynechococcus sudipacificus UW179B
Pseudosynechococcus sudipacificus WH 8016$_T$
Pseudosynechococcus sp. NP17
Pseudosynechococcus sp. AG − 679 − C18
Parasynechococcus sp. ARS 1019
Parasynechococcus sp. UBA9887
Parasynechococcus sp. MED − G70
Parasynechococcus sp. KORDI − 49
Parasynechococcus sp. TMED205
Parasynechococcus sp. AG − 670 − A05
Parasynechococcus sp. MED − G133
Parasynechococcus marenigrum WH 8102$_T$
Parasynechococcus marenigrum WH 8103
Parasynechococcus marenigrum YX04 − 3
Parasynechococcus sp. AG − 670 − B23
Parasynechococcus sp. KORDI − 52
Parasynechococcus marearabicus WH 8109$_T$
Parasynechococcus sp. N32
Parasynechococcus sp. N19
Parasynechococcus sp. 16 − 2
Parasynechococcus sp. UW86
Parasynechococcus sp. REDSEA − S01 − B1
Parasynechococcus sp. REDSEA − S02 − B4
Parasynechococcus sp. UBA8071
Parasynechococcus sp. N26
Parasynechococcus sp. N5
Parasynechococcus sp. RS344
Parasynechococcus africanus CC9605$_T$
Parasynechococcus sp. UW69
Parasynechococcus sp. UW106
Parasynechococcus sp. MED650
Parasynechococcus sp. SAT82
Parasynechococcus sp. AG − 673 − A10
Parasynechococcus sp. AG − 679 − D13
Parasynechococcus sp. AG − 673 − B04
Parasynechococcus sp. AG − 679 − B05
Parasynechococcus sp. AG − 676 − F02
Parasynechococcus sp. AG − 673 − B08
Parasynechococcus sp. AG − 670 − F04
Parasynechococcus sp. MED850
Parasynechococcus sp. KORDI − 100
Parasynechococcus sp. AG − 670 − D07
Parasynechococcus sp. CC9616
Parasynechococcus sp. TMED20
Parasynechococcus sp. AG − 686 − B21
Parasynechococcus sp. AG − 683 − C23
Parasynechococcus chilensis CC9902$_T$
Parasynechococcus chilensis MED − G69
Parasynechococcus chilensis AG − 683 − B21
Parasynechococcus nordiatlanticus BL107$_T$
Parasynechococcus sp. CPC35
50
40
30
20
10
0

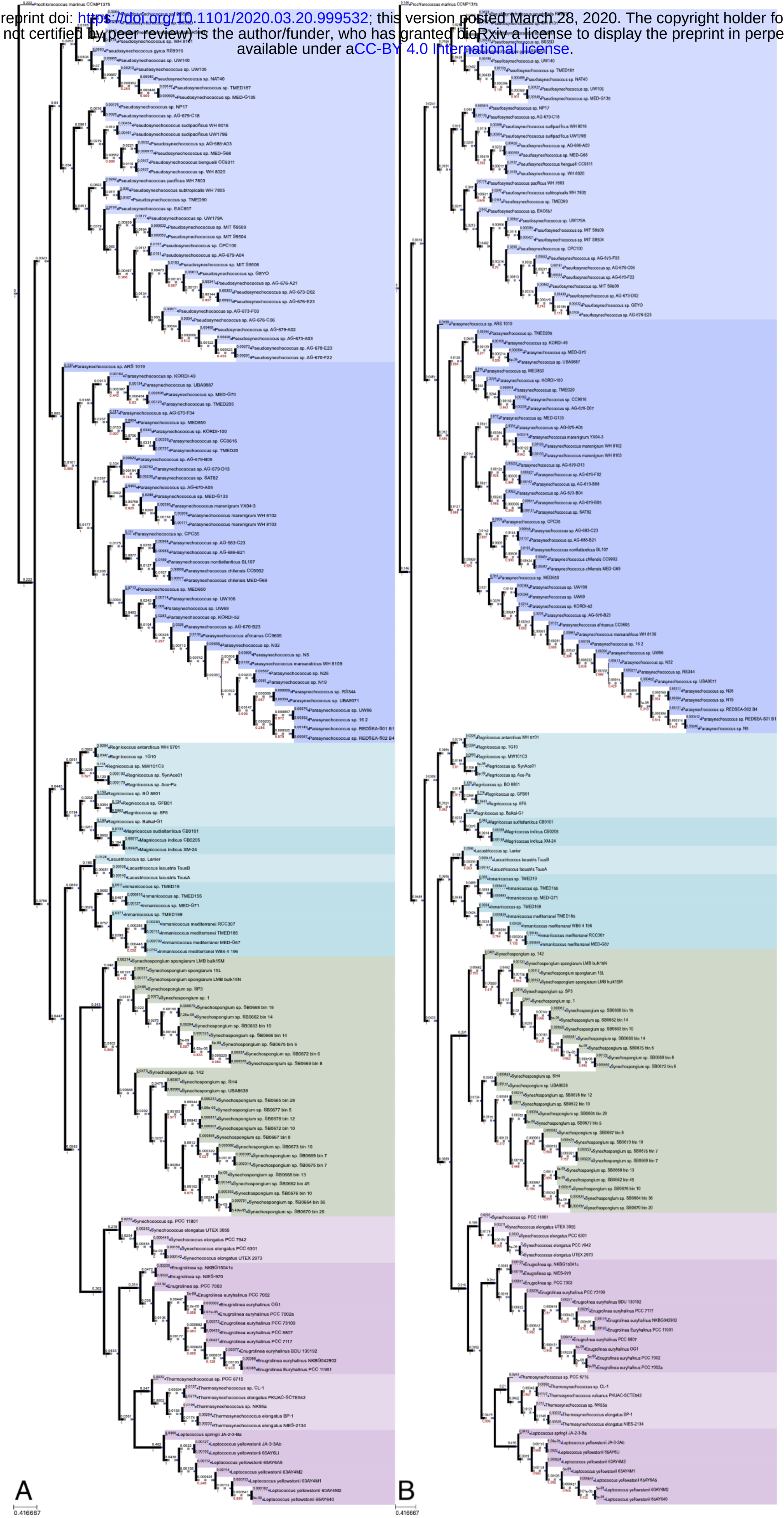

A
0.416667
B
0.416667

Figure 3: Phylogenetic trees of *Synechococcus*-related genera. Built from the concatenated protein alignment of A) 251 cyanobacterial marker genes and B) 74 bacterial marker genes. *Prochlorococcus marinus* CCMP 1375 is rooted as the outgroup. Red values show branch support and black values show substitutions per site. Ecogenomic groups are highlighted in either blue (Marine/oceanic), cyan (Marine/coastal), green (Symbiont), or purple (Freshwater/thermal).

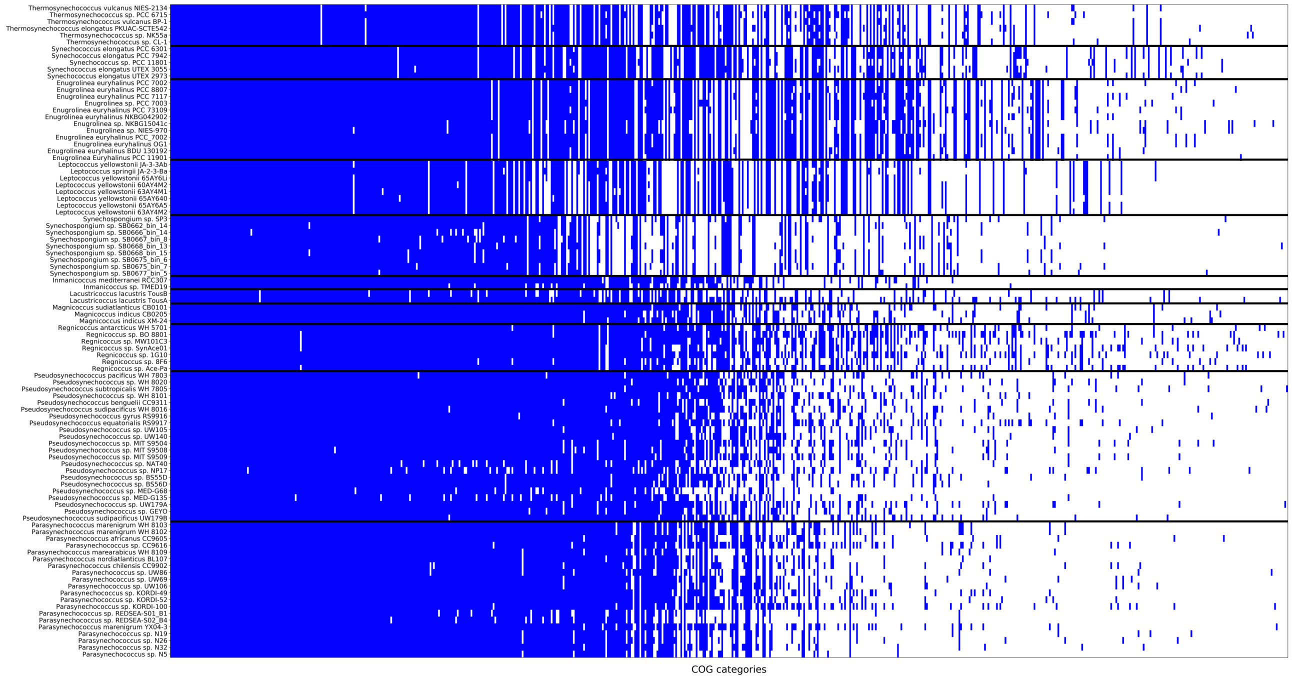


Figure 4: Presence/absence of CyCOG products. Blue bars represent presence of a CyCOG product and white bars its absence for each genome. Different genera are separated by black bars. The data used to generate this figure is in Table S2.