A new genomic taxonomy system for the *Synechococcus* collective

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

1 Center of Technology-CT2, SAGE-COPPE, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil. 2 Department of Systems and Computer Engineering, COPPE, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil. 3 Department of Biomedical Engineering, COPPE, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil. 4 Laboratory of Microbiology, Ghent University, Ghent, Belgium. 5 Institute of Biology, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil. Corresponding author: *Email: fabianothompson1@gmail.com, thompsoncristiane@gmail.com*, Address: Edifício do Centro de Ciências da Saúde, UFRJ - Av. Carlos Chagas Filho, 373 - Sala A1-050 - Cidade Universitária da Universidade Federal do Rio de Janeiro, Rio de Janeiro - RJ, 21941-902, Phone number: +55 21 39386567

**Summary**

Cyanobacteria of the genus *Synechococcus* are major contributors to global primary productivity and are found in a wide range of aquatic ecosystems. This *Synechococcus* collective (SC) is metabolically diverse, with some lineages thriving in polar and nutrient-rich locations and others in tropical or riverine waters. Although many studies have discussed the ecology and evolution of the SC, there is a paucity of knowledge on its taxonomic structure. Thus, we present a new taxonomic classification framework for the SC based on recent advances in microbial genomic taxonomy. Phylogenomic analyses of 1085 cyanobacterial genomes demonstrate that organisms classified as *Synechococcus* are polyphyletic at the order rank. The SC is classified into 15 genera, which are placed into five distinct orders within the phylum Cyanobacteria: i. *Synechococcales* (*Cyanobium*, *Inmanicoccus*, *Lacustricoccus* gen. nov., *Parasynechococcus*, *Pseudosynechococcus*, *Regnicoccus*, *Synechospongium* gen. nov., *Synechococcus*, and *Vulcanococcus*); ii. *Cyanobacteriales* (*Limnothrix*); iii. *Leptococcales* (*Brevicoccus* and *Leptococcus*); iv. *Thermosynechococcales* (*Stenotopis* and *Thermosynechococcus*), and v. *Neosynechococcales* (*Neosynechococcus*). The newly proposed classification is consistent with habitat distribution patterns (seawater, freshwater, brackish, and thermal environments) and reflects the ecological and evolutionary relationships of the SC.

**Introduction**

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1462-2920.15173

This article is protected by copyright. All rights reserved.
Cyanobacteria of the genus *Synechococcus* are of vital importance and contribute 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), which makes them crucial for 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. It is found in tropical oceans waters, freshwater ecosystems, hot spring microbial mats, and polar and nutrient-rich waters (Farrant *et al.*, 2016, Sohm *et al.*, 2016, Lee *et al.*, 2019).

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

Genomic studies have deepened our understanding of the unique adaptions of different lineages in the *Synechococcus* collective (SC) in regard to 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 SC genomes, Dufresne *et al.* (2008) identified two distinct lifestyles in marine *Synechococcus* lineages corresponding to coastal or open ocean habitats. Gupta and Mathews (2010) demonstrate freshwater and
seawater clades for distinct phylogenomic groups (e.g. the freshwater type species *Synechococcus elongatus* PCC 6301 and marine lineages WH8102 and WH8109).

It was also possible to identify distinct clades of Cyanobacteria that are directly related to the different groups of *Synechococcus*. Clade A is a deep-branching clade grouping the diazotrophic *Synechococcus* JA-3-3Ab and JA-2-3-Ba as well as *Gleobacter violaceous* PCC 7421. Clade B contains the thermophilic *Thermosynechococcus*, the marine *Synechococcus* PCC 7002, and genera such as *Microcystis*, *Nostoc*, and *Synechocystis*. Clade C contains marine *Synechococcus* lineages that are grouped monophyletically with *Prochlorococcus*. This phylogenetic structure is consistent with other studies (Shih et al., 2013; Dvořák et al., 2014a; Walter et al., 2017). The results from these studies lead to two main conclusions: i) organisms classified as *Synechococcus* are polyphyletic; and ii) to understand the phylogenetic history of groups within the SC, they must be analysed in the context of the phylum Cyanobacteria.

A robust classification system for the SC is still lacking. Coutinho et al. (2016ab) made a first attempt to unlock the taxonomy of SC. They compared 24 *Synechococcus* genomes by using an alignment of 607 orthologous proteins. They proposed the creation of the new genus *Parasynechococcus* to encompass the marine lineages and described 15 new species (Coutinho et al., 2016b). The description of these new species was attributed to the genetic diversity within these genomes. Walter et al. (2017) presented a phylogenetic analysis of 100 cyanobacterial genomes, elucidated the different SC lineages, and proposed 12 genera.

These studies demonstrate that SC is polyphyletic. However, the limited number of genomes examined at the time limited the depth of the analysis. They also do not include more recently sequenced strains of the sponge symbiont lineages Candidatus *Synechococcus spongiarum* (Usher et al., 2004a; Usher, 2008; Slaby and Hentschel, 2017) or the freshwater lineages of *Synechococcus lacustris* (Cabello-Yeves et al., 2017). The genomes pertaining to these lineages are particularly relevant to the taxonomic review of the SC because they represent understudied ecogenomic groups (the marine sponge symbiont and freshwater
ecotypes). Thus, it is important to establish a genome-based taxonomy for these organisms (Gevers et al., 2005; Stackebrandt, 2006; Arevalo et al., 2019).

With increasingly available sequence data, there has been a transition from the paradigm of polyphasic taxonomy towards genomic taxonomy (Thompson et al., 2015). Currently, the systematics of the SC according to the list of prokaryotic names with standing in nomenclature comprise four genera (Neosynechococcus, Parasynechococcus, Synechococcus, and Thermosynechococcus) and 11 species (Candidatus Synechococcus aeruginosus, Ca. S. calcipolaris, S. cedrorum, S. elongatus, Ca. S. feldmannii, Ca. S. lacustris, S. lividus, Ca. S. spongiarum, S. vulcanus, Neosynechococcus sphagnicola, and Thermosynechococcus vulcanus) (Parte, 2018). Examining prokaryotic taxonomy using the organisms’ whole genome could reveal meaningful relationships and define monophyletic groups, which would reveal their rate of evolution across taxonomic ranks (Hugenholtz et al., 2016; Parks et al., 2018). In a large-scale analysis, Parks et al. (2018) examined over 94,000 genomes and divided Synechococcus into at least 5 genera. However, they did not delve further into detailed taxonomic analyses of the taxon.

The aim of the present study is to propose a more robust taxonomic classification system for the SC to account for their diversity within the Cyanobacteria phylum. We combined several genome-level analyses (GC% content, genome size, average amino acid identity clustering, phylogenetic reconstruction, and gene functional profiling), from which we propose splitting the SC into 15 clearly separated genera, including two new genera: Lacustricoccus gen. nov. and Synechospongium gen. nov. Genus-level definition of prokaryotic organisms has been based on the average amino acid identity (AAI) (Konstantinidis and 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 cut-off and supported by further genomic analysis, such as phylogenomic trees, which are required to confirm genus-level definitions (Chun et al., 2018).
Results and discussion

We identified the genomes of the SC as polyphyletic at the order level (Figure 1). SC genomes were identified in five different orders and 15 different genera (Table 1). Phylogenomic reconstruction of 251 cyanobacterial marker genes (Figure 1) and genomic taxonomy analyses with GTDB-tk show higher rank relationships between SC genomes, while AAI hierarchical clustering delineates genera at the >70% threshold (Figure 2). AAI values support the phylogenetic relationships provided by concatenated protein alignment with GTDB-tk and GToTree but are able to resolve genus-level differences more finely (Qin et al., 2014) and reproduce previously reported genera distinctions in the SC (Walter et al., 2017).

FIGURE 1

Order Synechococcales

The most abundant order in the SC is Synechococcales, with 167 out of a total of 652 SC genomes (Figure 1 and Figure 2). The order contains the type species for the genus, *S. elongatus* PCC 6301 (Sugita et al., 2007). However, *S. elongatus* and related genomes are in the family Synechococcaceae, which is composed of genomes sourced exclusively from freshwater. A vast majority of SC genomes are in another family in this order, the family *Cyanobiaceae*, which is composed of eight SC genera, and also the monophyletic branch of *Prochlorococcus*, which accounts for most of the genomes in the order (483 out of 652).

Many SC genomes are grouped in the same family as *Prochlorococcus*. The genera *Parasynechococcus* and *Pseudosynechococcus* are more related to it than to other SC genomes, including the type species *S. elongatus* PCC 6301. The phylogenetic placement of the *Prochlorococcus* collective makes it a large sub-branch of the SC, and even if it represents a monophyletic group, the genetic divergence within the *Prochlorococcus* clade suggests that it should be split into multiple genera (Tschoeke et al., 2020).
genus *Cyanobium* is also a part of the SC and lends its name to the family *Cyanobioiaceae*. It is closely related to the genera *Regnicoccus* and *Vulcanococcus*.

Therefore, the order Synechococcales could be considered the “genuine” SC clade: it possesses most genomes and includes the families *Cyanobioiaceae* (with eight SC genera *Prochlorococcus*) and *Synechococciaceae* (with the genus *Synechococcus*). While the latter is typical of freshwater settings, the genera of *Cyanobioiaceae* have a more variable habitat distribution. *Prochlorococcus*, *Parasynechococcus*, and *Pseudosynechococcus* are mostly restricted to oligotrophic ocean waters. *Regnicoccus* is halotolerant and found in estuaries, such as Long Island Sound, USA, in the case of the type species *Regnicoccus antarcticus* WH 5701 (Fuller *et al.*, 2003), as well as extreme environments such as the saline Ace Lake in the Vestfold Hills of Antarctica (strain SynAce01) (Powell *et al.*, 2005). *Cyanobium* is primarily sourced from freshwater environments but is also halotolerant (Budinoff and Hollibaugh, 2007). The recently described *Vulcanococcus* was isolated from freshwater (the pink pigmented type-species, *Vulcanococcus limneticus* LL, was isolated from Lake Albano, Italy). It has been reported as phylogenetically close to marine/brackish lineages (Di Cesare *et al.*, 2018). In fact, genomes sourced from coastal marine waters, such as the strain CB0101, also cluster with *Vulcanococcus*, also providing evidence for its halotolerance.

The three genera *Regnicoccus*, *Cyanobium*, and *Vulcanococcus* are distinguished in the present genomic taxonomy analyses (with GTDB-tk and GTDB r89). They form a monophyletic group that also shares a phenotypic feature of tolerating wide salinity ranges, so it is possible that they might constitute a separate family from *Cyanobioiaceae*. The two genera *Inmanicoccus* and *Lacustricoccus* gen. nov. share a similar characteristic: they also form a monophyletic grouping, but *Lacustricoccus* was described as a new species from freshwater sources (Ca. *Synechococcus lacustris*, Cabello-Yepez *et al.*, 2017; 2018), and *Inmanicoccus* was isolated from coastal waters (the type species *Inmanicoccus mediterranei* RCC 307 was isolated from a depth of 15 m in the Mediterranean Sea).
<table>
<thead>
<tr>
<th>Order</th>
<th>#</th>
<th>Family</th>
<th>#</th>
<th>Genus</th>
<th>Type species</th>
<th>Accession</th>
<th>Ecotype</th>
<th>Source of type species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synechococcales</td>
<td>167 (652)</td>
<td>Cyanobacteria</td>
<td>161 (644)</td>
<td>Synechococcus</td>
<td>Paraxechnococcus africanus CC9605</td>
<td>GCA_000012625.1</td>
<td>Marine</td>
<td>California current – 51m</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pseudosynechococcus</td>
<td>Pseudosynechococcus subtricalis WH 7805</td>
<td>GCA_000153285.1</td>
<td>Marine</td>
<td>Sargasso sea – 25m</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Synechococcus</td>
<td>Synechococcus spongiam 15L</td>
<td>GCA_001007635.1</td>
<td>Marine</td>
<td>Aplysia aerophoba sponge, Slovenia</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inmanicoccus</td>
<td>Inmanicoccus mediterranei RCC307</td>
<td>GCA_000063525.1</td>
<td>Marine</td>
<td>Mediterranean Sea – 15m</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cyanobium</td>
<td>Cyanobium gracile PCC 6307</td>
<td>GCA_000316515.1</td>
<td>Halotolerant</td>
<td>Lake in Wisconsin, USA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Regnicoccus</td>
<td>Regnicoccus antarcticus WH 5701</td>
<td>GCA_000153045.1</td>
<td>Halotolerant</td>
<td>Long Island Sound, USA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Volcanococcus</td>
<td>Volcanococcus limneticus LL</td>
<td>GCA_002252705.1</td>
<td>Halotolerant</td>
<td>Lake Albano, Italy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lacustricococcus</td>
<td>Lacustricoccus lacustris Tous</td>
<td>GCA_003011125.1</td>
<td>Freshwater</td>
<td>Lake in Tous, Spain – 12m</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Synechococcaceae</td>
<td>6</td>
<td>Synechococcus</td>
<td>Synechococcus elongatus PCC 6301</td>
<td>GCA_000010065.1</td>
<td>Freshwater</td>
<td>Pasteur Culture Collection</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>14 (358)</td>
<td>Limnothrichaceae</td>
<td>14</td>
<td>Limnothrix</td>
<td>Limnothrix euryhalinus PCC 7002</td>
<td>GCA_000019485.1</td>
<td>Halotolerant</td>
<td>Pasteur Culture Collection</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Leptococcales</td>
<td>10</td>
<td>Leptococcus</td>
<td>Leptococcus yellowstoni JA-3-3-Ab</td>
<td>GCA_000013205.1</td>
<td>Hot spring</td>
<td>Octopus Spring, Yellowstone National Park, USA – 51-61 ºC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Brevicoccus</td>
<td>Brevicoccus berkleyi PCC 7336</td>
<td>GCA_000332275.1</td>
<td>Marine</td>
<td>Sea water, Berkeley, California, USA</td>
</tr>
<tr>
<td>Thermosynechococcales</td>
<td>8 (10)</td>
<td>Thermosynechococcaceae</td>
<td>8 (10)</td>
<td>Thermosynechococcus</td>
<td>Thermosynechococcus elongatus BP-1</td>
<td>GCA_000011345.1</td>
<td>Hot spring</td>
<td>Hot spring in Beppu, Jappan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stenotopis</td>
<td>1 (10)</td>
<td>Stenotopis californii</td>
<td>Stenotopis californii PCC 6312</td>
<td>GCA_000316685.1</td>
<td>Freshwater</td>
<td>Fresh water, California, USA</td>
</tr>
<tr>
<td>Neosynechococcales</td>
<td>1 (2)</td>
<td>Neosynechococcaceae</td>
<td>1 (2)</td>
<td>Neosynechococcus</td>
<td>Neosynechococcus sphagnicola CAUP A 1101</td>
<td>GCA_000775285.1</td>
<td>Peat bog</td>
<td>Peat moss in Slovakia</td>
</tr>
<tr>
<td>Classified as other cyanobacteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pseudanabaenales</td>
<td>1 (14)</td>
<td>Pseudanabaenaaceae</td>
<td>1 (14)</td>
<td>Leptovivax</td>
<td>Leptovivax bogii PCC 7502</td>
<td>GCA_000317085.1</td>
<td>Peat bog</td>
<td>Peat bog in Switzerland</td>
</tr>
<tr>
<td>Phormidesmiiales</td>
<td>1 (18)</td>
<td>Phormidesmiaceae</td>
<td>1 (18)</td>
<td>Phormidesmis</td>
<td>Phormidesmis mexicanus PCC 7335</td>
<td>GCA_00155595.1</td>
<td>Marine</td>
<td>Snail Shell, Intertidal Zone, Puerto Peñasco,</td>
</tr>
</tbody>
</table>

Table 1

This article is protected by copyright. All rights reserved.
The last genus in the Cyanobiaceae is *Synechospongium* gen. nov., which was previously described as Ca. *Synechococcus spongiarum* (Usher et al., 2004; Slaby and Hentschel, 2017), a symbiont of marine sponges. This genus constitutes the only one in a branch of Cyanobiaceae. *Synechospongium* genomes are reportedly genetically distinct from other *Synechococcus* and are found in marine sponges worldwide (for a detailed discussion, see Erwin and *Tha*cock, 2008).

**FIGURE 2 (VERTICAL HALF OF PAGE)**

*Order Cyanobacteriales*

A result that stands out is the phylogenetic distance between Synechococcales and Cyanobacteriales, which is the second most abundant order containing organisms of the SC (Figure 1). This order comprises many well-studied organisms, such as the genera *Oscillatoria*, *Spirulina*, *Microcystis*, and *Nostoc*. Twelve genomes presently classified as *Synechococcus* sp. were determined to belong to the genus *Limnothrix* in the family *Limnotrichaceae*, as they cluster with two described *Limnothrix* genomes. The correct taxonomic classification of these organisms is particularly important as they may be valuable for biotechnological applications (e.g., *L. euryhalinus* PCC 11901, presented by Włodarczyk et al. (2020) as *Synechococcus* sp.) or have unique metabolic profiles, such as sulphide pathways (Liu et al., 2020) and positive responses to high salinity reported for *L. euryhalinus* PCC 7002 (Ludwig and Bryant et al., 2012).

*Order Leptococcales and Thermosynechococcales.*

In contrast to the orders Synechococcales and Cyanobacteriales, which are restricted to marine and freshwater organisms, the orders Thermosynechococcales and Leptococcales include thermophilic representatives. Leptococcales can be divided into two separate families: *Leptococcaceae* contains 10
genomes of the thermophilic *Leptococcus* isolated from hot springs in Yellowstone National Park (Becraft et al., 2011), and *Brevicoccaceae* has a single genome (*Brevicoccus berkleyi* PCC 7336) isolated from sea water. Although ecological differences have recently been documented for the *Leptococcus* strains (Becraft et al., 2020), they do not include *B. berkleyi* PCC 7336. It is surprising to see *B. berkleyi* grouped with *Leptococcus*, even at the order level. *B. berkleyi* was isolated in 1971 from a marine water tank from (presumably) the Pacific Ocean (see NCBI Biosample entry SAMN02261338), whereas the latter has a highly specific distribution in the hot springs of Yellowstone.

The order Thermosynechococcales also comprises hot spring organisms, but it is phylogenetically separate from the Leptococcales. It is formed by two SC genera: *Thermosynechococcus*, with two previously described species (*T. elongatus* and *T. vulcanus*) (Nakamura et al., 2002), and *Stenotopis* (*S. californii* PCC 6312) (Walter et al., 2017). A genome previously classified as *Synechococcus lividus* PCC 6715 and considered a thermophilic *Synechococcus*, was reclassified as *Thermosynechococcus*, thus enforcing the need for a new taxonomic framework.

Order Neosynechococcales.

The single genome *Neosynechococcus sphagnicola* CAUP A 1101 has an exclusive position in the order Neosynechococcales. GTDB-tk analysis shows that it clusters at the family level with another cyanobacterial genome with an uncertain taxonomic position that is presently classified as *Leptolynbya* sp. Hensonii (Table S1). *N. sphagnicola* was sourced from a peat bog site, and in the description, it had was already been reported as a divergent lineage of the SC (Dvořák et al., 2014b). In this context, this study further analyses the SC into in more detail to gain a better understanding of its ecogenomic features.

*Genome-level divergences of genera in the SC*
Genomic diversity was observed within the SC on several scales, including GC% content and genome size (in base pairs) (Figure 3). The sheer span of these two features between genera of the SC indicates marked differences between them. The genome size among genera varies from 0.99 to 5.97 megabase pairs (Mbps), and the GC content varies from 43.87% to 69.14% (Figure 3a). However, when the SC is split into several genera, the GC content and genome size become more consistent (Figure 3bc) and closer to proposed ranges for taxonomic grouping (Meier-Kolthoff et al., 2014). Genetically homogeneous genera (with higher intragenus AAI values), such as *Limnothrix* and *Synechococcus*, form clusters of very low variability in GC content and genome size (Figure 3a).

**FIGURE 3**

*Functional profiles and in silico diagnostic phenotypes.*

Annotation of Clusters of Orthologous Group (COG) proteins showed differences in functional profiles between orders, genera, and ecotypes (Figure 4). These *in silico* phenotypic differences may be for distinguishing the niches of different taxa. Four clusters were found among the genomes examined. Cluster 1 is a freshwater *Synechococcus* cluster with the thermophilic genera *Leptococcus* and *Thermosynechococcus*, which suggests a marked phenotypic difference between marine genera of the family *Cyanobiaceae* and *Synechococcus*. Cluster 2 comprises the halotolerant or freshwater genera *Lacustricoccus*, *Regnicoccus*, *Cyanobium*, and *Vulcanococcus* (family *Cyanobiaceae*), which suggests functional similarity between them and a phenotypic difference between these genera and marine genera in the same family. Cluster 3 comprises the four marine genera of the order Synechococcales: *Parasynechococcus*, *Pseudosynechococcus*, *Inmanicoccus*, and *Synechospongium*, which are late branching close to *Prochlorococcus*.

Cluster 4 is a heterogeneous group comprising *Neosynechococcus*, *Brevicoccus*, *Limnothrix*, and *Stenotopsis* (Figure 4). These genera belong to different orders which have contrasting phylogenetic and
ecological characteristics. *Brevicoccus* is classified in the early branching order Leptococcales (Figure 1), albeit in its own family, but it does not share the highly adapted thermophilic lifestyle of the *Leptococcus* genus in the same order. Similarly, *Stenotopis* is in the early branching Thermosynechococcales (Figure 1), but it is sourced from freshwater. *Neosynechococcus* displays the highest number of genes in several COG categories, which may be related to its larger genome size. This genus is a divergent lineage of the SC, with unique ecological and phenotypical characteristics, as reported in its original description (Dvořák et al., 2014b). The in silico phenotype differences between *Limnothrix* and *Synechococcus* reinforces the need to rename organisms in the *Limnothrix* presently classified as “*Synechococcus* sp.”. Thus, the partitioning of functional profiles of COG categories was consistent with both the taxonomy and ecology, displaying useful diagnostic phenotypical differences across both taxonomic ranks and ecotypes.

Cluster 3 contains genera that have a smaller number of proteins for the transport and metabolism of inorganic ions and for signal transduction mechanisms than other genera (Figure 4). The lack of these genes may reflect the smaller genome size of these four genera than others (Figure 3). The sponge symbiont *Synechospongium* also has a lower number of genes for cell wall/membrane/envelope biogenesis, carbohydrate transport and metabolism, and energy production and conversion than any other genera. The lack of these genes may be the result of a symbiotic lifestyle in sponge holobionts (Burgsdorf et al., 2015).

Iron transport mediated by siderophores, eukaryotic-like proteins, and defence mechanisms, like CRISPR-Cas (clustered regularly interspaced short palindromic repeats-associated proteins) are present in *Synechospongium* and absent or rare in other genomes of the SC (Burgsdorf et al., 2015). *Synechospongium* lacks O-antigen coding genes. Thus, CRISPR-Cas may be an alternative to fight against phage infections in the sponge holobiont. Genomes of *Lacustricoccus* have been reported as having a high content of transporters and membrane proteins (“Cell wall/membrane/envelope biogenesis” category) compared to marine SC lineages, especially *Parasynechococcus* and *Pseudoxynechococcus* (Cabello-Yeves et al., 2017).
This trend is shared between the other three in the cluster and is consistent with the freshwater niche preference.

**Conclusion**

The present study analysed 1085 cyanobacterial genomes and provided a new taxonomic system for the SC, representing five orders of the phylum Cyanobacteria. Although there is no doubt that the organisms classified as *Synechococcus* are polyphyletic, there is not yet a consensus on how to classify them. The results of this study and those of modern genomic taxonomy systems, such as the GTDB, show that most of the SC genomes are restricted to the order Synechococcales and that this order is composed of the genus *Synechococcus* and eight other genera of the SC, along with *Prochlorococcus*. Therefore, it makes sense to reframe the SC as a single order of the phylum *Cyanobacteria* comprising the freshwater genus *Synechococcus*, eight genera found in variable environments, and the marine *Prochlorococcus*. Phylogenetic relationships within this order (especially at the family level) pose an important research challenge as they may reveal important insights about the evolution of these organisms, which are crucial to the regulation of Earth’s ecosystems.

Reframing the order Synechococcales as the SC means that genera present in the four other orders (Cyanobacteriales, Leptococcales, Thermosynechococcales, Neosynechococcales) should no longer be regarded as “*Synechococcus*-like”. This is especially true for the genus *Limnothrix* as it shares habitats and halotolerance capacity with genera from the order Synechococcales, making it easy to mistake one for the other. The other three genera are easily distinguished by their unique distribution patterns based on whether they are either thermophilic, present in hot springs, or found
in peat bog sites (for Neosynechococcus). However, Thermosynechococcus and Neosynechococcus have similar nomenclatures to Synechococcus.

Another research challenge is the taxonomic description and ecogenomic understanding of the new SC species. Considering the current species-level threshold (>95% AAI), the results of this study show the presence of a hundred SC species that require further investigation. Our genomic analyses demonstrate the heterogeneous nature of the SC regarding genome organization, functional profiles, and ecotypes. Furthermore, they bring new insights into the taxonomic structure of SC collective with the evident distinction of 15 genera and how they are positioned in the phylum Cyanobacteria. This newly proposed taxonomic structure will be useful for further environmental surveys and ecological studies of population-level genetic diversity, including those targeting the identification of ecotypes and species.

Experimental procedures

Data acquisition and processing

A total of 1085 genomes classified as the phylum Cyanobacteria were analysed in the present study. All Synechococcus genomes (n=240, hereafter referred to as “SC genomes”) were downloaded from the NCBI Assembly database (Kitts et al., 2015) in May 2020 using the Python package “NCBI Genome Download” (https://github.com/kblin/ncbi-genome-download) by querying for the genus “Synechococcus” with the “--fuzzy-genus” parameter. The metadata table with NCBI Entrez data generated by the package was used as a template for the new metadata table (Table S1). To ensure a standardized treatment of each genome data, only assembly files were used (FASTA format containing complete chromosomes, scaffolds, or contigs) instead of pre-existing files from the assembly directories available from the NCBI.
Quality assurance

To infer the completeness of each SC 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 package Pandas v0.25.1 (McKinney 2011) in Jupyter Notebook (Ragan-Kelley et al., 2014). The results for completeness and contamination were then added to the metadata table (Table S1). For all further analyses, we only used genomes with at least 50% completeness and less than 10% contamination according to CheckM.

Genomic taxonomy analyses and cyanobacterial reference genomes

SC genomes were processed by GTDB-Tk v1.0.2 (Chaumeil et al., 2020) in order to determine their taxonomic status (Table S1). All genomes classified as the class “Cyanobacteria” by GTDB (Parks et al., 2020) (due to being the common clade of all SC genomes) were downloaded in May 2020 from NCBI Assembly by querying their accession numbers with the aforementioned NCBI Genome Download package. The metadata from NCBI Entrez, GTDB, and CheckM from both of these genomes and SC genomes were concatenated, as shown in Table S1. The same quality filter criteria of at least 50% completeness and less than 10% contamination were applied to these genomes.

Phylogenetic reconstruction
To build the phylogenetic tree, we used the GToTree package (Lee, 2019) with default parameters and the “Cyanobacterial” profile. The input dataset consisted of the 1085 quality-filtered assemblies including the SC and other cyanobacterial genomes. Genomes that were not classified by GTDB as the class “Cyanobacteriia” were removed. Assemblies were searched against a Hidden Markov Model of 251 cyanobacterial marker genes using HMMER3 (Eddy, 2011). From the 1085 genomes, 35 were retained in the tree after applying GToTree’s default settings for 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 (File S1). All processing was done with GNU Parallel (Tange 2018). Trees were rendered with FigTree v1.4.4 (https://github.com/rambaut/figtree/).

AAI analysis

Comparative AAI analysis was carried out with the package CompareM v0.0.25 (https://github.com/dparks1134/CompareM). To do so, we ran the “aai_wf” command in CompareM, which utilizes protein coding sequences (CDS) predicted with Prodigal (Hyatt et al., 2007), performs an 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 with the default settings and parameters for defining homology of >30% sequence similarity and >70% alignment length.

The output table from the AAI analysis (Table S2) was then imported into Jupyter Notebook, and a symmetrical distance table was constructed using Pandas v0.25.1. This table was transformed into a one-dimensional condensed distance matrix using the “squareform()” function from the “spatial” package of the Scipy library (Jones et al., 2001). The resulting matrix was subjected to clustering with the “linkage()” function (SciPy library, “cluster” package) with the parameters “method=’complete’”, “metric=’cityblock’”
and “optimal_ordering=True”. A more in-depth explanation of these parameters can be found on the SciPy documents page (https://docs.scipy.org/doc/scipy/reference/index.html). The resulting array was used as input for a customized function based on SciPy’s “dendrogram()” function.

We performed hierarchical clustering of pairwise AAI values between all genomes in the SC with a >70% cut-off defined for genera within the same family. This cut-off was empirically defined by previous studies (Thompson et al., 2013; Rodriguez and Konstantinidis, 2014; Qin et al., 2014). Four genomes did not cluster with any other SC genomes based on this criterion: *Stenotopis californii* PCC 6312, *Neosynechococcus sphagnicola* CAUP A 1101, *Leptovivax bogii* PCC 7502, and *Phormidesmis mexicanus* PCC 7335. These genomes were classified as “singletons”. *S. californii* PCC 6312 was classified as a separate genus in the family Thermosynechococcaceae.

The second genome is positioned exclusively in the order Neosynechococcales and family Neosynechococcaceae, which comprises it and a separate genus with uncertain taxonomic position that was previously classified as *Leptolyngbya* sp. Hensonii (see Results and discussion section). The latter two were classified in other cyanobacterial orders. The first was *Leptovivax bogii* PCC 7335 and was classified in its own genus in the Pseudanabaenaceae family (order Pseudanabaenales), while the second was *Phormidesmis mexicanus* PCC 7335, which was previously described as *Coccusdissimilis mexicanus* PCC 7335 (Walter et al., 2017). Because these two genomes are positioned in existing cyanobacterial orders, they were not considered as SC genomes, and we recommend that they be treated as misclassified bacteria of these orders. The threshold range for setting a cyanobacterial order is 50 to 70% according to the Relative Evolutionary Divergence (RED) value, as established by the GTDB-tk methodology (Parks et al., 2018; Chaumeil et al., 2019).

**Nomenclature**
The same names for each genus were maintained as those used by Walter et al. (2017) except for *Enugrolinea* (which was reclassified as *Limnothrix*), *Magnicoccus* (reclassified as *Vulcanococcus*), and the newly-named *Synechospongium* gen. nov. and *Lacustricoccus* gen. nov. *Limnothrix* genomes were previously classified as *Enugrolinea*, but seeing as the *Limnothrix* nomenclature is older (Meffert, 1987), the latter name was kept. The same happened for *Vulcanococcus*, which was previously classified as *Magnicoccus*. Genera and species were defined at AAI cut-offs of >70% and >95%, respectively (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.

**GC content and genome size**

GC content and genome size statistics were calculated from assembly files downloaded from NCBI using Python functions and are displayed in the metadata table (Table S1). The libraries Matplotlib v3.2.1 (Hunter, 2007) and Seaborn v0.10.1 (Waskom, 2018) were used for plotting. For the GC and genome size plot (Fig. 3), we fitted genome sizes based on completeness using simple extrapolation (e.g. if the genome is 50% complete, it is assumed that its length is 50% of the true value). Figure S1 provides the unadjusted values.

**COG functional profiling**

To find useful diagnostic *in silico* phenotypes that can be applied to differentiate different genera, protein-coding sequences of quality-filtered SC genomes were annotated with the NCBI COG database (Galperin et al., 2015), which was downloaded in May 2020 using Diamond. For these analyses, a higher-quality cut-off was used to prevent false negatives while keeping only genomes with >95% completeness according to CheckM. Sequences that did not align with the reference database were ignored.
Counts of annotated proteins (hits) classified according to COG code were then transformed into a table. Hits with COG codes in the same broader COG category were summed, generating a table of functional categories per genome (Table S3). The mean of each category was then calculated by grouping by genus. The resulting matrix of grouped data was used to generate a distance matrix with the Bray-Curtis Index, which was used as input for a complete linkage cluster analysis, similarly to the AAI analysis section. The hierarchical clustering and matrix of grouped data were then plotted together. Processing was done with Pandas, Scipy, Matplotlib, and Seaborn with the aforementioned versions.

Data and code availability

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

Acknowledgements

The authors thank CAPES, FAPERJ, and CNPq for funding.

References


This article is protected by copyright. All rights reserved.


This article is protected by copyright. All rights reserved.


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


This article is protected by copyright. All rights reserved.


FIGURE AND TABLE CAPTIONS

Figure 1: Phylogenomic tree of 1050 cyanobacterial genomes. Values in brackets show number of genomes in branch. The order Leptococcales contains the families Leptococcaceae (with the genus Leptococcus (ten genomes)) and Brevicoccaceae (with the genus Brevicoccus (one genome)). Family Thermosynechococcaceae contains the genera Thermosynechococcus (seven genomes) and Stenotopis (one genome). Number of genomes in each branch may differ from Table 1 as a result of GToTree quality control step. Branches may contain additional genomes that are cyanobacterial genomes reclassified as groups in the SC, e.g. the family Thermosynechococcaceae contains two cyanobacterial genomes reclassified in that group.

Figure 2: AAI hierarchical clustering of SC genomes. Dotted values show difference in AAI between genomes. Genera are delimited at a >70% AAI cut-off.

Figure 3: GC% content and genome size ranges in the SC. A. Scatter plot of GC% content and genome size (in megabases). B, C. Box plots for genome size and GC% content, respectively. Genera are coloured by ecotypes with corresponding colours of Figure 1: purple for thermophilic, green for freshwater, olive for halotolerant, blue for marine, dark grey for peat bog. Light grey dots in panel A represent other cyanobacterial genomes (which are not contained in the SC). Genome sizes were fitted by estimated genome completeness (see Experimental Procedures - “GC content and genome size” section).
Figure 4: Heatmap and hierarchical clustering of COG functional profiling. Colours illustrate average number of genes per genus in each functional category. Auxiliary columns display ecotype and order. Clusters as mentioned in the text are labeled numerically in the dendrogram.

Table 1: Taxa of the Synechococcus collective. "#" symbol denotes number of genomes in that group. Accession numbers refer to NCBI Genbank database. Values in brackets show total of genomes in group (including genomes which are not classified as Synechococcus).
Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Salazar, VW; Tschoeke, DA; Swings, J; Cosenza, CA; Mattoso, M; Thompson, CC; Thompson, FL

Title:
A new genomic taxonomy system for the Synechococcus collective

Date:
2020-11

Citation:

Persistent Link:
http://hdl.handle.net/11343/298010