Ecogenomics and Taxonomy of Cyanobacteria Phylum

Juline M. Walter, Felipe H. Coutinho, Bas E. Dutilh, Jean Swings, Fabiano L. Thompson and Cristiane C. Thompson

Cyanobacteria are major contributors to global biogeochemical cycles. The genetic diversity among Cyanobacteria enables them to thrive across many habitats, although only a few studies have analyzed the association of phylogenomic clades to specific environmental niches. In this study, we adopted an ecogenomics strategy with the aim to delineate ecological niche preferences of Cyanobacteria and integrate them to the genomic taxonomy of these bacteria. First, an appropriate phylogenomic framework was established using a set of genomic taxonomy signatures (including a tree based on conserved gene sequences, genome-to-genome distance, and average amino acid identity) to analyse ninety-nine publicly available cyanobacterial genomes. Next, the relative abundances of these genomes were determined throughout diverse global marine and freshwater ecosystems, using metagenomic data sets. The whole-genome-based taxonomy of the ninety-nine genomes allowed us to identify 57 (of which 28 are new genera) and 87 (of which 32 are new species) different cyanobacterial genera and species, respectively. The ecogenomic analysis allowed the distinction of three major ecological groups of Cyanobacteria (named as i. Low Temperature; ii. Low Temperature Copiotroph; and iii. High Temperature Oligotroph) that were coherently linked to the genomic taxonomy. This work establishes a new taxonomic framework for Cyanobacteria in the light of genomic taxonomy and ecogenomic approaches.

Keywords: microbial ecology, ecological niches, charting biodiversity, genome-based microbial taxonomy, metagenome, high-throughput sequencing technology

INTRODUCTION

Earth is home to nearly one trillion ($10^{12}$) microbial species that have evolved over ~4 billion years (Locey and Lennon, 2016). Cyanobacteria emerged ~3 billion years ago, ushering Earth's transition from anoxicogenic to oxygenic conditions through photosynthesis (Schirrmeister et al., 2011a). Throughout their evolution, Cyanobacteria became one of the most diverse and widely distributed Prokaryotes, occupying many niches within terrestrial, planktonic, and benthic habitats. Their long history evolved in a broad heterogeneity comprising unicellular and multicellular, photosynthetic and non-photosynthetic (i.e., Melainabacteria) (Schirrmeister et al., 2011a; Di Rienzi et al., 2013; Soo et al., 2014), free-living, symbiotic, toxic and predatory organisms (Soo et al., 2015), with genomes sizes ranging from 1 to 10 Mb (Shih et al., 2013). Here we consider Cyanobacteria phylum as consisting only of oxygenic phototrophs.
Cyanobacteria (also known as the Cyanophyceae, Cyanophyta, cyanoprokaryota, blue-green algae or blue-green bacteria) share similar metabolic features with eukaryotic algae and have been named according to the Botanical Code (Kaufmann and Büdel, 2010). The inclusion of Cyanobacteria in taxonomic schemes of Bacteria was only proposed in 1978 by Stanier et al. (1978), and through time the bacterial taxonomic names have come into conflict with the botanical nomenclature (Oren, 2004; Oren and Garrity, 2014). More than two decades passed before a Note to General Consideration 5 (1999) was published for Cyanobacteria to be included under the rules of the International Committee on Systematic Bacteriology (ICSB)/International Committee on Systematic of Prokaryotes (ICSP) (Tindall, 1999; De Vos and Trüper, 2000; Labeda, 2000). Taxa nomenclature within this group has long been a topic of discussion, but currently there is no consensus (Hoffmann et al., 2005; Oren and Tindall, 2005; Oren et al., 2009; Oren and Ventura, 2017). As a result, more than 50 genera of Cyanobacteria have been described since 2000, and many of them remain unrecognized in the List of Prokaryotic Names with Standing in Nomenclature, LPSN, http://www.bacterio.net (Parte, 2014) or in databases (e.g., NCBI).

The Cyanobacteria form a challenging group for the microbiologists. Their traditional taxonomy based on morphologic traits does not reflect the results of phylogenetic analyses (Rippka et al., 1979; Boone and Castenholz, 2001; Gugger and Hoffmann, 2004; Schirrmieister et al., 2011b; Hugenholtz et al., 2016). The predominance of morphology assembled unrelated Cyanobacteria into polyphyletic species and genera and higher taxonomic categories which require revisions in the future (Komárek et al., 2014). The polyphylly is an indicative of the taxonomic mislabeling of many taxa. The 16S rRNA gene sequences were useful in charting and characterizing microbial communities (Kozlov et al., 2016) but this molecule lack sensitivity for evolutionary changes that occur in ecological dynamics, where microbial diversity is organized by physicochemical parameters (Choudoir et al., 2012; Becraft et al., 2015). Hence, the processes that shape cyanobacterial communities over space and time are less known. A recent study proposed that there should be 170 genera of Cyanobacteria based on 16S rRNA sequences only (Kozlov et al., 2016). Farrar et al. (2016) delineated 121 Prochlorococcus and 15 Synechococcus ecologically significant taxonomic units (ESTUs) in the global ocean using single-copy petB sequences (encoding cytochrome b6) and environmental cues.

High Throughput Sequencing (HTS) have revolutionized the practice of microbial systematics, providing an informative, reproducible, and portable tool to delineate species, reconstruct their evolutionary history, and infer ecogenomic features (Gevers et al., 2005; Konstantinidis and Tiedje, 2005a,b; Garrity and Oren, 2012; Gribaldo and Brochier-Armanet, 2012; Shih et al., 2013; Sutcliffe et al., 2013; Hugenholtz et al., 2016). This approach allows both cultured (Al-saari et al., 2015; Appolinario et al., 2016) and uncultured microorganisms (Iverson et al., 2012; Brown et al., 2015; Hugler et al., 2015) to be studied. The latter is especially important because the cyanobacterial cultivation in laboratory is another hurdle in the study of this group of bacteria.

Recommendations that nomenclature should agree with and reflect genomic information were stated during the pre-genomic era (Wayne et al., 1987), due nothing describes an organism better than its genome. Sequence-based methods to delimit prokaryotic species have emerged to define and to improve cut-offs criteria during the genomic era (Gevers et al., 2005; Konstantinidis and Tiedje, 2005a,b; Konstantinidis et al., 2006; Goris et al., 2007; Richter and Rossello-Mora, 2009; Auch et al., 2010a; Thompson et al., 2013a,b; Varghese et al., 2015), demonstrating a greater discriminatory power. Inexorable advances in methodologies will incorporate genomics into the taxonomy and systematics of the prokaryotes, boosting the credibility of taxonomy in the current post-genomic era (Coenye et al., 2005; Chun and Rainey, 2014). Up-to-date, while several groups have been analyzed through a genomic-wide view (Gupta et al., 2015; Adeolu et al., 2016; Hahnke et al., 2016; Ahn et al., 2017; Amin et al., 2017; Waite et al., 2017), many others have faced hurdles, such as Cyanobacteria. However, a genomic taxonomy approach has successfully been applied to elucidate the taxonomic structure of the two cyanobacterial genera, Prochlorococcus and Synechococcus (Thompson et al., 2013a; Coutinho et al., 2016a,b). As genomic taxonomy postulates numeric, non-subjective, cut-offs for taxa delimitation, strains were considered to belong to the same species when share at least 98.8% 16S rRNA gene sequence similarity, 95% of AAI, and 70% GGD (Konstantinidis and Tiedje, 2005a; Thompson et al., 2013a,b), while species from the same genus form monophyletic branches (Yarza et al., 2008; Qin et al., 2014). It is in agreement with the concept of species as a discrete, monophyletic and genomically homogeneous population of organisms that can be discriminated from other related populations by means of diagnostic properties (Rossello-Mora and Amann, 2001; Stackebrandt et al., 2002). The availability of whole-genomes opened the doors for an in-depth knowledge in microbial diversity and ecology, where the entire genomic pool may be applied to understanding the forces that govern community structure. The use of ecogenomic analysis postulates a reliable and scalable approach to delineate species and genera in order to reconstruct their evolution and to draw a global picture of possible ecological determinants (Di Rienzi et al., 2013; Soo et al., 2014; Spang et al., 2015; Thompson et al., 2015; Anantharaman et al., 2016; Garrity, 2016; Hug et al., 2016; Hugenholtz et al., 2016). Our hypothesis is that a phylogenomic framework will reflect ecologic groups found in nature.

To test this hypothesis, we first established a phylogenomic framework, using genomic signatures (i.e., a tree based on conserved gene sequences, average amino acid identity, and genome-to-genome distance), with the circumscription of species and genera. We then classified the genomes in three major groups according to their ecological traits as inferred through metagenomics and environmental metadata. Finally, we correlated the three disclosed ecogenomic groups (i. Low Temperature; ii. Low Temperature Copiotroph; and iii. High Temperature Oligotroph) with the circumscribed species and genera. We observed that the taxonomic delineation of species and genera is coherent with the ecogenomic groups.
MATERIALS AND METHODS

Genome Election
Cyanobacterial genomes publicly available in January 2016 were retrieved from RefSeq (NCBI Reference Sequence Database), GenBank and GEBa (Genomic Encyclopedia of Bacteria and Archaea) databases. Genome completeness was assessed with CheckM (Parks et al., 2015), and the genomes that were at least 90% complete and assembled in <500 contigs were used for further analyses. Ninety-nine genomes were selected based on that criterion, and they are listed in Table 1 (additional information on Table S2).

Annotation and Genomic Taxonomy
All genomes were annotated using Prokka version 1.11 (Seemann, 2014), with default settings, in order to avoid any possible bias. Genomic taxonomy of the ninety-nine cyanobacterial genomes was performed according to Thompson et al. (2013a) and Coutinho et al. (2016a) and are briefly described here. Average Amino acid Identity (AAI) and Genome-to-Genome Distance (GGD) were calculated as described previously (Konstantinidis and Tiedje, 2005a; Auch et al., 2010a,b; Meier-Kolthoff et al., 2013). GGD were calculated using the Genome-to-Genome Distance Calculator tool, version 2.1 under recommended settings (Meier-Kolthoff et al., 2013; http://ggdc.dsmz.de/), whereas AAI values were carried out through GenFaxo as previously described (Coutinho et al., 2016a). The species cut-offs delimitation were ≥95% AAI and ≥70% GGD, and ≥70% AAI for genus delimitation.

The Manhattan distances were calculated based on the percentage AAI values of every genome (genome-genome matrix) and was used as the input for making the hierarchical clustering using the hclust() function in R (R Development Core Team, 2011). This distance is able to indicate how far/close the genomes are located from each other. The heatmap was produced by heatmap2 [ggplot2] package in R, with background color of each panel mapping to percentage AAI values.

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

Ribosomal RNA Sequences
The small subunit ribosomal RNA (16S rRNA) sequences from all cyanobacterial strains for which whole genome sequence data are publicly available (exception see below, thus N = 97), as well as 16S rRNA gene sequences from additional type-strains available (N = 14) were all analyzed. The sequences were retrieved from the ARB SILVA database (Pruesse et al., 2007; Quast et al., 2013). Whenever sequences were not available, they were retrieved directly from the genomes using RNAmmer 1.2 Server (Lagesen et al., 2007). Sequences were aligned through MUSCLE v. 3.8 (Edgar, 2004), with default settings, and Gblocks 0.91b (Castresana, 2000; Talavera and Castresana, 2007) was used for alignment curation. Using MEGA 6 (Tamura et al., 2013), best-fitting nucleic acid substitution models were calculated through the MLModelTest feature. Models were ranked based on their Bayesian Information Criterion (BIC) scores as described by Tamura et al. (2013). The model with the lowest BIC score was selected and used for further phylogenetic analysis.

The phylogenetic inference was obtained using the Maximum Likelihood method based on the Kimura 2 parameter method with the Gamma distributed rate variation (K2+G) as the nucleotide substitution model, which was estimated from the data. The support branches of tree topology were checked by 1,000 bootstrap replicates. The 16S rRNA gene alignments were used to estimate the degree of genetic distance between strains through the Tajima-Nei method (Tajima and Nei, 1984).

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

The type-strains or the type-species of each taxa were included in the 16S phylogenetic tree to confirm the phylogenetic relatedness of the cyanobacterial genomes. Designations of type strain or type species were not available for Chaetoceros minusutus PCC6605, Pleurocapsa sp. PPC7319, Rivularia sp. PPC7116, Synechocystis sp. PCC7509, Trichodesmium erythraeum IMS01, Xenococcus sp. PCC7305, cyanobacterium ESFC-1, and cyanobacterium JSC-12. Geitlerinema sp. PCC7105 is the reference strain for marine species of Geitlerinema, and PCC73106 is the reference strain for Gloecocapsa (Sarma, 2012).

Conserved Marker Genes
A tree was generated using 31 conserved gene sequences previously validated as phylogenetic markers for (cyano) bacteria (Wu and Eisen, 2008, and recently used by Shih et al., 2013 and Komárek et al., 2014). The sequences of these proteins were mined using the AutoMated Phylogenomic inference Application—AMPHORA2 tool (Wu and Scott, 2012), through default settings for the Bacteria option, and with a cut-off value of 1.0–10. Individual alignments were performed for each of the 31 gene sets through MUSCLE v. 3.8 with default settings (Edgar, 2004). All alignments were then concatenated. Only genomes which present all the set of conserved genes were used in the phylogenetic analysis. A Maximum Likelihood tree was constructed using RaxML v. 7 (Stamatakis, 2006) and the Dayhoff+G likelihood model. One thousand bootstrap replications were calculated to evaluate the relative support of the branches. Trees were visualized with FigTree, version 1.4.2 (Rambaut, 2015).

Abundance of Cyanobacterial Genomes Across Aquatic Environments and Ecological Correlations
Marine and freshwater metagenomes were retrieved to determine the abundance of ninety-nine cyanobacterial genomes across the Earth. A set of 191 marine metagenomes from the Tara Ocean project were retrieved for analysis along with their
<table>
<thead>
<tr>
<th>Bacterial Strain⁶</th>
<th>Strain⁷</th>
<th>NCBI or JGI Reference Sequence</th>
<th>New genus proposal</th>
<th>New species proposal</th>
<th>Habitat</th>
<th>Type source/Place</th>
<th># Contigs</th>
<th>Lenght (Mbp)</th>
<th>% mol GC</th>
<th># CDS</th>
<th>Completeness¹</th>
<th>Carboxysome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arthrospira platensis</td>
<td>PCC 7108</td>
<td>NZ_AJW00000000.1</td>
<td>A. mosi</td>
<td>A. mosi</td>
<td>Marine (coastal)</td>
<td>Intertidal zone, Moss Beach, CA, USA</td>
<td>3</td>
<td>5.9</td>
<td>38.78</td>
<td>7,169</td>
<td>99.63</td>
<td>p</td>
</tr>
<tr>
<td>Anabaena sp.</td>
<td>PCC 7103</td>
<td>NZ_ALU00000000.1</td>
<td>C. wisconsinii</td>
<td>A. wisconsinii</td>
<td>Freshwater</td>
<td>Crawford Co., Wisconsin, USA</td>
<td>12</td>
<td>11.58</td>
<td>38.55</td>
<td>9,371</td>
<td>99.39</td>
<td>p</td>
</tr>
<tr>
<td>Anabaena sp.</td>
<td>PCC 7134</td>
<td>NZ_ACP00000000.1</td>
<td>C. palmae</td>
<td>A. palmae</td>
<td>Marine (coastal)</td>
<td>Salt marsh in Woods Hole, Massachusetts, USA</td>
<td>143</td>
<td>6.85</td>
<td>45.43</td>
<td>7,100</td>
<td>99.37</td>
<td>p</td>
</tr>
<tr>
<td>Arthrospira platensis</td>
<td>PCC 7100</td>
<td>NZ_ACP00000000.1</td>
<td>A. limicola</td>
<td>A. limicola</td>
<td>Marine (coastal)</td>
<td>San Francisco, California, USA</td>
<td>1</td>
<td>5.82</td>
<td>47.39</td>
<td>6,401</td>
<td>99.76</td>
<td>p</td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Strain</th>
<th>NCBI or JGI Reference Sequence</th>
<th>New genus proposal</th>
<th>New species proposal</th>
<th>Habitat</th>
<th>Type source/Place</th>
<th># Contigs</th>
<th>Length (Mbp)</th>
<th>% mol GC</th>
<th># CDS</th>
<th>Completeness*</th>
<th>Carboxysome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moorea producens</td>
<td>NZ_APQ000000000.1</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>287</td>
<td>8.38</td>
<td>43.68</td>
<td>9.97</td>
<td>98.56</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Nostoc sp.</td>
<td>NC_019676.1</td>
<td>Nostoc</td>
<td>N. reyesi</td>
<td>Freshwater</td>
<td>Point Reyes Peninsula, California, USA</td>
<td>1</td>
<td>6.32</td>
<td>43.36</td>
<td>9.20</td>
<td>99.26</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Nostoc sp.</td>
<td>NC_019684.1</td>
<td>Nostoc</td>
<td>N. ampulliformis</td>
<td>Freshwater</td>
<td>Hot spring, Ampara District, Maha Oya, Sri Lanka</td>
<td>3</td>
<td>6.71</td>
<td>41.53</td>
<td>9.32</td>
<td>99.33</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Oscillatoria aurantiaca</td>
<td>PCC 6304</td>
<td>Oscillatoria</td>
<td>M. rugulosa</td>
<td>Soil</td>
<td>NA</td>
<td>1</td>
<td>7.68</td>
<td>47.6</td>
<td>6.04</td>
<td>99.71</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Oscillatoria nigrobioides</td>
<td>PCC 7112</td>
<td>Oscillatoria</td>
<td>M. nigrobioides</td>
<td>Soil</td>
<td>USA</td>
<td>1</td>
<td>7.47</td>
<td>45.87</td>
<td>6.925</td>
<td>99.78</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Oscillatoria nigrobioides</td>
<td>PCC 10802</td>
<td>Oscillatoria</td>
<td>S. hydrophilic</td>
<td>NA</td>
<td>NA</td>
<td>9</td>
<td>8.59</td>
<td>54.1</td>
<td>7.012</td>
<td>100</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Oscillatoria sp.</td>
<td>PCC 6506</td>
<td>Oscillatoria</td>
<td>T. oscillatii</td>
<td>NA</td>
<td>NA</td>
<td>377</td>
<td>6.67</td>
<td>43.4</td>
<td>6.097</td>
<td>99.12</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Oscillatoria sp.</td>
<td>PCC 6405</td>
<td>Oscillatoria</td>
<td>T. oscillatii</td>
<td>Freshwater</td>
<td>NA</td>
<td>12</td>
<td>6.89</td>
<td>43.43</td>
<td>5.693</td>
<td>99.56</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Parasynechococcus</td>
<td>CC9605</td>
<td>Parasynechococcus</td>
<td>P. pacificus</td>
<td>Marine</td>
<td>California current, Pacific, oligotrophic, 5 m</td>
<td>1</td>
<td>2.51</td>
<td>59.2</td>
<td>2.583</td>
<td>99.73</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Parasynechococcus</td>
<td>CC9002</td>
<td>Parasynechococcus</td>
<td>P. pacificus</td>
<td>Marine</td>
<td>California current, Pacific, oligotrophic, 5 m</td>
<td>1</td>
<td>2.23</td>
<td>54.2</td>
<td>2.289</td>
<td>99.46</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Parasynechococcus</td>
<td>WH8109</td>
<td>Parasynechococcus</td>
<td>P. pacificus</td>
<td>Marine</td>
<td>Sargasso Sea</td>
<td>1</td>
<td>2.12</td>
<td>60.1</td>
<td>2.661</td>
<td>99.32</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Parasynechococcus</td>
<td>WH8102</td>
<td>Parasynechococcus</td>
<td>P. pacificus</td>
<td>Marine</td>
<td>Sargasso Sea</td>
<td>1</td>
<td>2.43</td>
<td>59.4</td>
<td>2.461</td>
<td>99.46</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Parasynechococcus</td>
<td>BL107</td>
<td>Parasynechococcus</td>
<td>P. pacificus</td>
<td>Marine</td>
<td>Blanes Bay, Mediterranean Sea, 1 800 m</td>
<td>1</td>
<td>2.29</td>
<td>54.2</td>
<td>2.322</td>
<td>99.46</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Parasynechococcus</td>
<td>CC9311</td>
<td>Parasynechococcus</td>
<td>P. pacificus</td>
<td>Marine</td>
<td>California current, Pacific, coastal, 95 m</td>
<td>1</td>
<td>2.61</td>
<td>52.4</td>
<td>2.627</td>
<td>99.73</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Parasynechococcus</td>
<td>RS9917</td>
<td>Parasynechococcus</td>
<td>P. pacificus</td>
<td>Marine</td>
<td>Gulf of Araha, Red Sea, 10 m</td>
<td>1</td>
<td>2.58</td>
<td>64.4</td>
<td>2.575</td>
<td>99.46</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Parasynechococcus</td>
<td>RS9916</td>
<td>Parasynechococcus</td>
<td>P. pacificus</td>
<td>Marine</td>
<td>Gulf of Araha, Red Sea, 10 m</td>
<td>1</td>
<td>2.66</td>
<td>59.8</td>
<td>2.603</td>
<td>99.73</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Parasynechococcus</td>
<td>WH7603</td>
<td>Parasynechococcus</td>
<td>P. pacificus</td>
<td>Marine</td>
<td>Sargasso Sea, 25 m</td>
<td>1</td>
<td>2.37</td>
<td>60.2</td>
<td>2.439</td>
<td>99.18</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Parasynechococcus</td>
<td>WH7605</td>
<td>Parasynechococcus</td>
<td>P. pacificus</td>
<td>Marine</td>
<td>Sargasso Sea</td>
<td>3</td>
<td>2.63</td>
<td>57.6</td>
<td>2.590</td>
<td>99.73</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Parasynechococcus</td>
<td>WH8016</td>
<td>Parasynechococcus</td>
<td>P. pacificus</td>
<td>Marine</td>
<td>Woods Hole, MA, USA</td>
<td>16</td>
<td>2.69</td>
<td>54.1</td>
<td>2.990</td>
<td>99.18</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Parasynechococcus</td>
<td>WH8701</td>
<td>Parasynechococcus</td>
<td>P. pacificus</td>
<td>Marine</td>
<td>Long Island Sound, Connecticut, USA</td>
<td>116</td>
<td>3.28</td>
<td>65.4</td>
<td>2.917</td>
<td>99.46</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Parasynechococcus</td>
<td>CB0026</td>
<td>Parasynechococcus</td>
<td>P. pacificus</td>
<td>Marine</td>
<td>Chesapeake Bay, Baltimore, Maryland, USA</td>
<td>78</td>
<td>2.43</td>
<td>63.0</td>
<td>2.473</td>
<td>99.18</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Parasynechococcus</td>
<td>CB0101</td>
<td>Parasynechococcus</td>
<td>P. pacificus</td>
<td>Marine</td>
<td>Chesapeake Bay, Baltimore, Maryland, USA</td>
<td>94</td>
<td>2.69</td>
<td>64.2</td>
<td>2.757</td>
<td>99.73</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Parasynechococcus</td>
<td>RCC307</td>
<td>Parasynechococcus</td>
<td>P. pacificus</td>
<td>Marine</td>
<td>Mediterranean Sea, 15 m</td>
<td>1</td>
<td>2.22</td>
<td>60.8</td>
<td>2.348</td>
<td>99.64</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>Planktothrix agardhii</td>
<td>NIVA-CYA 126/8</td>
<td>Planktothrix</td>
<td>P. agardhii</td>
<td>Freshwater</td>
<td>NA</td>
<td>13</td>
<td>5.04</td>
<td>39.57</td>
<td>4.188</td>
<td>100</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Planktothrix agardhii</td>
<td>NIVA-CYA 15</td>
<td>Planktothrix</td>
<td>P. agardhii</td>
<td>Freshwater</td>
<td>NA</td>
<td>238</td>
<td>5.38</td>
<td>39.48</td>
<td>4.606</td>
<td>100</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Planktothrix agardhii</td>
<td>NIVA-CYA 56/3</td>
<td>Planktothrix</td>
<td>P. agardhii</td>
<td>Freshwater</td>
<td>NA</td>
<td>185</td>
<td>5.48</td>
<td>39.48</td>
<td>4.674</td>
<td>99.78</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Planktothrix agardhii</td>
<td>NIVA-CYA 405</td>
<td>Planktothrix</td>
<td>P. agardhii</td>
<td>Freshwater</td>
<td>NA</td>
<td>240</td>
<td>5.46</td>
<td>39.47</td>
<td>4.697</td>
<td>99.56</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Planktothrix agardhii</td>
<td>NIVA-CYA 406</td>
<td>Planktothrix</td>
<td>P. agardhii</td>
<td>Freshwater</td>
<td>NA</td>
<td>375</td>
<td>5.62</td>
<td>39.51</td>
<td>4.873</td>
<td>100</td>
<td>$\beta$</td>
</tr>
</tbody>
</table>

(Continued)
TABLE 1 | Continued

<table>
<thead>
<tr>
<th>Bacterial Strain*</th>
<th>Strain*</th>
<th>NCBI or JGI Reference Sequence</th>
<th>New genus proposal</th>
<th>New species proposal</th>
<th>Habitat</th>
<th>Type source/Place</th>
<th># Contigs</th>
<th>Length (Mbp)</th>
<th>% mol GC</th>
<th>CDS</th>
<th>Completeness*</th>
<th>Carboxysome</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prochlorococcus</strong></td>
<td></td>
<td></td>
<td>E. chëtrëní</td>
<td>Marine</td>
<td>Arabian Sea, 50 m</td>
<td>1</td>
<td>1.66</td>
<td>31.32</td>
<td>1,769</td>
<td>99.64</td>
<td>α</td>
<td></td>
</tr>
<tr>
<td><strong>Prochlorococcus</strong></td>
<td></td>
<td></td>
<td>E. marinus</td>
<td>Marine</td>
<td>Mediterranean Sea, 5 m</td>
<td>1</td>
<td>1.65</td>
<td>30.8</td>
<td>1,777</td>
<td>99.46</td>
<td>α</td>
<td></td>
</tr>
<tr>
<td><strong>Prochlorococcus</strong></td>
<td></td>
<td></td>
<td>E. neptüinis</td>
<td>Marine</td>
<td>Gulf Stream, 135 m</td>
<td>1</td>
<td>1.71</td>
<td>31.21</td>
<td>1,815</td>
<td>99.73</td>
<td>α</td>
<td></td>
</tr>
<tr>
<td><strong>Prochlorococcus</strong></td>
<td></td>
<td></td>
<td>E. nerus</td>
<td>Marine</td>
<td>South Pacific, 79 m</td>
<td>1</td>
<td>1.69</td>
<td>31.1</td>
<td>1,795</td>
<td>98.78</td>
<td>α</td>
<td></td>
</tr>
<tr>
<td><strong>Prochlorococcus</strong></td>
<td></td>
<td></td>
<td>E. nerus</td>
<td>Marine</td>
<td>Equatorial Pacific, surface</td>
<td>1</td>
<td>1.73</td>
<td>31.15</td>
<td>1,840</td>
<td>99.73</td>
<td>α</td>
<td></td>
</tr>
<tr>
<td><strong>Prochlorococcus</strong></td>
<td></td>
<td></td>
<td>E. pontiús</td>
<td>Marine</td>
<td>Sargasso Sea, 90 m</td>
<td>1</td>
<td>1.64</td>
<td>31.34</td>
<td>1,774</td>
<td>99.46</td>
<td>α</td>
<td></td>
</tr>
<tr>
<td><strong>Prochlorococcus</strong></td>
<td></td>
<td></td>
<td>E. pontiús</td>
<td>Marine</td>
<td>Equatorial Pacific, 15 m</td>
<td>1</td>
<td>1.71</td>
<td>30.79</td>
<td>1,784</td>
<td>100</td>
<td>α</td>
<td></td>
</tr>
<tr>
<td><strong>Prochlorococcus</strong></td>
<td></td>
<td></td>
<td>P. proteus</td>
<td>Marine</td>
<td>Northern Atlantic, 30 m</td>
<td>1</td>
<td>1.86</td>
<td>34.98</td>
<td>2,204</td>
<td>99.73</td>
<td>α</td>
<td></td>
</tr>
<tr>
<td><strong>Prochlorococcus</strong></td>
<td></td>
<td></td>
<td>P. proteus</td>
<td>Marine</td>
<td>Northern Atlantic, 10 m</td>
<td>1</td>
<td>1.84</td>
<td>35.12</td>
<td>1,930</td>
<td>99.45</td>
<td>α</td>
<td></td>
</tr>
<tr>
<td><strong>Prochlorococcus</strong></td>
<td></td>
<td></td>
<td>P. cætus</td>
<td>Marine</td>
<td>Sargasso Sea, 120 m</td>
<td>1</td>
<td>1.75</td>
<td>36.44</td>
<td>1,883</td>
<td>100</td>
<td>α</td>
<td></td>
</tr>
<tr>
<td><strong>Prochlorococcus</strong></td>
<td></td>
<td></td>
<td>T. swëngi</td>
<td>Marine</td>
<td>Equatorial Pacific, 83 m</td>
<td>1</td>
<td>1.68</td>
<td>36.01</td>
<td>1,748</td>
<td>99.73</td>
<td>α</td>
<td></td>
</tr>
<tr>
<td><strong>Prochlorococcus</strong></td>
<td></td>
<td></td>
<td>T. swëngi</td>
<td>Marine</td>
<td>Sargasso Sea, 100 m</td>
<td>1</td>
<td>2.68</td>
<td>50.01</td>
<td>2,504</td>
<td>100</td>
<td>α</td>
<td></td>
</tr>
<tr>
<td><strong>Pseudoanabaena</strong></td>
<td></td>
<td></td>
<td>L. gracile</td>
<td>Marine (coastal)</td>
<td>Intertidal zone, México</td>
<td>1</td>
<td>4.55</td>
<td>46.31</td>
<td>3,960</td>
<td>98.23</td>
<td>β</td>
<td></td>
</tr>
<tr>
<td><strong>Pseudoanabaena</strong></td>
<td></td>
<td></td>
<td>P. alæmænæx</td>
<td>Freshwater</td>
<td>California, USA</td>
<td>6</td>
<td>5.62</td>
<td>47.83</td>
<td>5,363</td>
<td>99.76</td>
<td>β</td>
<td></td>
</tr>
<tr>
<td><strong>Rivulæra</strong></td>
<td></td>
<td></td>
<td>R. alæmænæx</td>
<td>Marine (coastal)</td>
<td>La Paz, Baja California Sur, México</td>
<td>3</td>
<td>8.72</td>
<td>37.53</td>
<td>6,612</td>
<td>99.78</td>
<td>β</td>
<td></td>
</tr>
<tr>
<td><strong>Synechococcus</strong></td>
<td></td>
<td></td>
<td>P. bistriætæ</td>
<td>marine</td>
<td>NA</td>
<td>10</td>
<td>5.32</td>
<td>47.39</td>
<td>4,580</td>
<td>99.56</td>
<td>β</td>
<td></td>
</tr>
<tr>
<td><strong>Synechococcus</strong></td>
<td></td>
<td></td>
<td>P. bistriætæ</td>
<td>marine</td>
<td>Havana, Cuba</td>
<td>6</td>
<td>5.54</td>
<td>36.22</td>
<td>4,895</td>
<td>99.56</td>
<td>β</td>
<td></td>
</tr>
<tr>
<td><strong>Synechococcus</strong></td>
<td></td>
<td></td>
<td>P. alæmænæx</td>
<td>Freshwater</td>
<td>NA</td>
<td>1</td>
<td>2.7</td>
<td>55.5</td>
<td>2,576</td>
<td>99.73</td>
<td>α</td>
<td></td>
</tr>
<tr>
<td><strong>Synechococcus</strong></td>
<td></td>
<td></td>
<td>P. alæmænæx</td>
<td>Freshwater</td>
<td>Octopus Spring, Yellowstone Park, USA</td>
<td>1</td>
<td>3.05</td>
<td>58.5</td>
<td>3,064</td>
<td>100</td>
<td>β</td>
<td></td>
</tr>
<tr>
<td><strong>Synechococcus</strong></td>
<td></td>
<td></td>
<td>P. alæmænæx</td>
<td>Marine (coastal)</td>
<td>Sea Water Tank, Berkeley, California, USA</td>
<td>11</td>
<td>5.97</td>
<td>48.2</td>
<td>5,702</td>
<td>98.91</td>
<td>β</td>
<td></td>
</tr>
<tr>
<td><strong>Synechococcus</strong></td>
<td></td>
<td></td>
<td>P. alæmænæx</td>
<td>Marine (coastal)</td>
<td>Sea Water Tank, Berkeley, California, USA</td>
<td>1</td>
<td>5.07</td>
<td>53.7</td>
<td>5,093</td>
<td>100</td>
<td>β</td>
<td></td>
</tr>
</tbody>
</table>

(Continued)
associated metadata (Sunagawa et al., 2015). Sample-associated environmental data were inferred across multiple depths at global scale of Tara’s metagenomics sampling: (i) surface water layer (5 m, s.d. = 0); and (ii) subsurface layer, including deep chlorophyll maximum zone (71 m, s.d. = 41 m) and mesopelagic zone (600 m, s.d. = 220 m) (Sunagawa et al., 2015). Eight freshwater metagenomes were retrieved for analysis from the Caatinga biome microbial community project along with their associated metadata (Lopes et al., 2016).

Metagenome reads were mapped to a database containing the ninety-nine analyzed cyanobacterial genomes through Bowtie2 (Langmead and Salzberg, 2012) using -very-sensitive-local and -a options. Abundance of genomes across samples was calculated based on the number of mapped reads as described by Iverson et al. (2012). Metagenomes were compared based on the relative abundances of the ninety-nine analyzed genomes within them using non-metric multidimensional scaling (NMDS).

Spearman correlation coefficients (R, or Spearman’s rho) were calculated for the abundance of each genome and the levels of measured environmental parameters across samples. Next, a dissimilarity matrix of Manhattan distances was calculated based on the Spearman correlation values of every genome. All correlations were used by this analysis regardless of the corrected p-value, as non-significant correlations are still ecologically informative as they indicate weak associations between microbes and environmental parameters. Finally, this dissimilarity matrix was used as input for hierarchical clustering using the complete linkage method within the hclust() function in R. The resulting dendrogram was visually inspected to define groups (i.e., ecogenomic groups) of organisms with similar correlation patterns which were named based on the main correlated feature.

The classification reassessment was made integrating the results of genomic taxonomy, phylogenomic analysis and ecogenomic signals through an accurately comparison.

**RESULTS**

**Phylogenomic Framework Reconstruction**

The tree based on conserved marker genes (Figure 1) revealed the topology with the presence of well-defined nodes in general with bootstrap support values greater than 50% over 1,000 replicates. The phylogenomic tree (Figure 1) gave a higher resolution than the 16S rRNA phylogenetic analysis (Figure S1 and Table S1), in the means that strains were better discriminated in the conserved marker genes tree (e.g., *Parasynechococcus* group, Figure 1 and Figure S1). The species assignations were considered correct when organisms located on the same phylogenetic branch as the corresponding type strains or type species presented the 16S rRNA sequence similarity higher than 98.8%, such as *Crinalium epipsum* SAG22.89 (Figure S1) and *Crinalium epipsum* PCC9333 (Figure S1 and Figure 1).

**Genomic Diversity of Cyanobacteria**

In total, we found 57 branches corresponding to genera based on the AAI and GGD analyses (Figure 2). The genus and species cut-off delimitation were ≥70% and ≥95% AAI similarity.
FIGURE 1 | Phylogenomic tree of the Cyanobacteria phylum with the proposed new names. Tree construction was performed using 100 genomes (ninety-nine used in this study plus the outgroup), based on a set of conserved marker genes. The numbers at the nodes indicate bootstrap values as percentages greater than 50%. Bootstrap tests were conducted with 1,000 replicates. The unit of measure for the scale bars is the number of nucleotide substitutions per site. The Gloeobacter violaceus PCC 7421 sequence was designated as outgroup. Capital letters indicate environmental source: F, freshwater; M, marine; P, peat bog (sphagnum); S, soil; T, thermal; and §, other habitat. New names are highlighted in red. Overwritten T indicates type strain or type species. Ecogenomic groups are depicted in different colors as indicated in the legend: Low Temperature group; Low Temperature Copiotroph group; and High Temperature Oligotroph group. Cases depicted in the Results section are in bold.
respectively. Thirty-three new genera and 87 species (of which 28 are new species) were circumscribed. From a total of ninety-nine genomes used in this study, 69 were previously classified to the species level, whereas the remaining 30 had incomplete taxonomic classification (i.e., only sp. or unclassified). In total, 13 genera (from a total of 33) and 38 species (from a total of 69) were taxonomically reclassified and/or re-named. Thus, we found that 71 of all analyzed genomes required reassignment at one or more ranks to reconcile existing taxonomic classifications with our new genomic taxonomy (Figure 2 and Figure S1).
Over the next section, we highlight four specific cases to exemplify cyanobacterial taxonomic issues that were resolved through our genome-driven approach (see Figure S2). These cases illustrate how the use of genomic taxonomy in Cyanobacteria provides relevant information (Data Sheet 1, Formal description of new genera and species).

Case I. Oscillatoria group. Analysis of the five genomes of Oscillatoria distinguished four genera, based on the genomic signatures (i.e., GGD, AAI, 16S, and conserved marker genes tree): (i) Oscillatoria acuminata PCC 6304 type strain formed a separate group; (ii) Oscillatoria sp. PCC 10802 formed a separate divergent group, corresponding to a new genus named Somacatellium (S. hydroxylic PCC 10802T); (iii) Oscillatoria nigroviridis strain PCC 7112 (closest related with Micrococus vaginatus FGP-2 type strain) belongs to the genus Micrococcus (M. nigroviridis PCC 7112T); and (iv) Oscillatoria strains PCC 6407 and PCC 6506 formed a new genus named Toxinema (T. oscillati PCC 6407T and T. oscillati PCC 6506).

Case II. Leptolyngbya group. The five Leptolyngbya strains were polyphyletic, forming different phylegentic branches. Thus, (i) Leptolyngbya boryana PCC 6306T type strain forms a separate group with cyanobacterium JSC-12, while the rest of the Leptolyngbya strains cluster apart; (ii) strain PCC 7376 forms a new genus named Enugrolinea (E. bermudensis PCC 7376T); (iii) strain PCC 7375 forms a new genus named Adonisia (A. splendidus PCC 7375T); (iv) strain PCC 7104 forms a new genus named Alunema (A. longislandicus PCC 7104T); and (v) strain PCC 6406 forms a new genus named Euryforis (E. eilemai PCC 6406T).

Case III. Arthrospira group. Examination of the four Arthrospira strains indicated that (i) A. platensis C1 should be considered a new species, named A. sesilensis (A. sesilensis C1T); (ii) strain PCC 8005 belongs to a new species, named A. nitrilium (A. nitrilium PCC 8005T); and (iii) the type strain of Arthrospira platensis (PCC 7345) formed a tight cluster along with NIES-39 and Paraca.

Case IV. Synechococcus group. The nine Synechococcus strains split in (i) S. elongatus PCC 6301T type strain forms a separate group with S. elongatus PCC 7942; (ii) strain PCC 6312 forms a new genus named Stenotrophi (S. californii PCC 6312T); (iii) strain PCC 7335 belongs to a new genus named Cocsudissimilis (C. mexicanus PCC 7335T); (iv) strains JA23Ba213 and JA33Ab formed a new genus named Leptococcus (L. springii JA23Ba213T and L. yellowstonii JA33AbT); (v) strain PCC 7336 formed a new genus named Eurycoccus (E. bermeyi PCC 7336T); (vi) strain PCC 7502 belonged to a new genus named Leptovivax (L. bogi PCC 7502T); and (vii) Synechococcus euryhalinus PCC 7002 represents a new genus named Enugrolinea (E. euryhalinus PCC 7002T).

Charting Ecological Groups of Cyanobacteria

Our phylogenomic analysis was complemented by an ecological characterization of the analyzed strains, providing essential insights into relations between taxonomy, phylonym, and ecological role (Beiko, 2015). Correlating the relative genome abundances with environmental parameters measured at Tara Oceans samples (Sunagawa et al., 2015) revealed associations between Cyanobacteria and physical, chemical and biological variables of their habitats (Figure 3). The ecogenomic analysis clustered genomes based on their profiles of correlations to environmental parameters. Three major ecogenomic groups were found: (a) Low Temperature; (b) Low Temperature Copiotroph; and (c) High Temperature Oligotroph (Figure 4 and Figure S3). Closely related species of the same genus showed tight associations with environmental parameters, grouped to the same ecogenomic group, such as Arthospira sesilensis C1T and A. nitrilium PCC 8005T, Eurycoccus pastoris CCMP1986T and E. teissi MIT9515T, and Pseudosynechococcus subtropicalis WH7805T and P. pacificus WH7803T (Figure 3).

In a few cases, closely related species showed different ecogenomic groups (P. agardhii NIVA-CYA-407 and P. agardhii NIVA-CYA-540 compared to other Planktothrix strains, and between Lyngbya aestuarii BL-J and L. limosa PCC 8106T) (Figure 3).

Members of the Low Temperature group were characterized by positive correlations with the concentration of nitrogen and phosphorus sources; weak positive correlations with minimum generation time, silicate and depth; and by negative correlations with temperature, microbial cell abundance, oxygen availability, and salinity (Figures 3, 4). Meanwhile, members of the Low Temperature Copiotroph group were characterized by strong positive correlations with the concentration of nitrogen and phosphorus; positive correlations (stronger than those presented by Low Temperature group) with minimum generation time, silicate and depth; and by negative correlations (also stronger than those presented by Low Temperature group) with temperature, microbial cell abundance (in particular with autotroph cell density), oxygen availability, and salinity (Figures 3, 4). Finally, members of High Temperature Oligotroph group were characterized by negative correlations with the concentration of nitrogen and phosphorus and positive correlations with temperature and autotroph cell abundance (Figures 3, 4).

As suggested by correlation analyses (Figures 4C,D), NMDS revealed the Low Temperature Copiotroph group to be more abundant in cold and eutrophic waters, while the High Temperature Oligotroph group exhibited the opposite pattern and was more abundant in warm and oligotrophic environments (Figures 4A,B). In turn, Low Temperature was more abundant at intermediate conditions between these polar opposites and was shown to be more abundant in samples with higher cell densities and NO2 concentrations.

We also investigated the abundance of the ecogenomic groups in freshwater environments. Unfortunately, there is no currently available large-scale dataset of freshwater metagenomes with associated metadata comparable to the Tara Oceans dataset. To define freshwater ecogenomic groups we chose to extrapolate the classification obtained from the analyses of the marine dataset. In freshwater metagenomes, the Low Temperature Copiotroph was the dominant group in all the analyzed samples (Figure S4A). NMDS of freshwater samples suggested that Low Temperature group displayed a preference for higher pH and DOC, nitrite and total nitrogen concentrations whereas the High Temperature Oligotroph group has a preference for habitats with higher concentrations of POC, phosphorus, ammonia and nitrate (Figures S4B,C).
DISCUSSION

The use of HTS technologies and environmental surveys have allowed studies that link phylogenomics and ecogenomics of Cyanobacteria. High-throughput genome sequence technologies are causing a revolution in microbial diversity studies. Recent studies have obtained dozens of new metagenome-assembled genomes from complex environmental samples (Brown et al., 2015; Hugelth et al., 2015; Almstrand et al., 2016; Haroon et al., 2016; Pinto et al., 2016). The abundance of these genomes across different environments can now be inferred from metagenomics, including their metabolic and ecological potential. It is clear

FIGURE 3 | Correlations between Cyanobacteria and environmental variables. Heatmap displays Spearman correlation scores between the abundance of cyanobacterial genomes and measured environmental parameters at Tara Ocean sampling sites. Correlations that showed q corrected $p < 0.05$ are marked with stars. Variables were grouped through the complete linkage clustering method using Manhattan distances as input. The proposed new genera and species names were adopted in this figure.
that a new system is required to allow for precise taxonomic identification of these new genomes.

**WGS as the Basic Unit for Cyanobacteria Genomic Taxonomy (CGT)**

Comparative genomic studies allow for identification of sequence groups with high genotypic similarity based on variation in protein coding genes distributed across the genomes. Analyses of environmental metagenomes and microbiomes have shown that microbial communities consist of genotypic clusters of closely related organisms (Farrant et al., 2016). These groups display cohesive environmental associations and dynamics that differentiate them from other groups co-existing in the same environment. In light of new concepts, restlessness is mounting with the inability to define the microbial species itself. Evolution studies on closely related bacteria show rapid and highly variable
gene fluxes in evolving microbial genomes, suggesting that extensive gene loss and horizontal gene transfer leading to innovation are the dominant evolutionary processes (Batut et al., 2014; Puigbó et al., 2014). CGT will solve the often-observed issue that even closely related genomes contain high gene content variation, that gives phenotypic variation. CGT is completely adjusting to the genomics era, addressing the needs of its users in microbial ecology and clinical microbiology, in a new paradigm of open access (Beiko, 2015). CGT will provide a predictive operational framework for reliable automated and openly available identification and classification (Thompson et al., 2015).

**Proposals for Cyanobacterial Taxonomy**

A main gap exists and is growing each day between the formal taxonomy of Cyanobacteria and the forest of acronyms and numbers in the different databases. Indeed, the nameless operational taxonomic units (OTUs), strains, isolates and WGS sequences (Beiko, 2015; Kozlov et al., 2016) form the great majority of data in private and public databases. There is a need to re-examine the Cyanobacteria prokaryote species, taking into account all recently developed concepts, e.g., the gene flow unit, OTU, ESTU and Candidate taxonomic unit (CTU) in the context of a pragmatic genome-based taxonomic scheme. The type species or strain can be a culture, DNA or a WGS. The CGT system should maintain all of the existing information, integrating it with new data on DNA, genomes, isolates/strains, cultured and uncultured, “Candidatus” cases and reconstructed genomes from metagenomes (Brown et al., 2015; Hugerth et al., 2015). The international initiatives of GBEA are currently working on determining the WGS of all type strains of known microbial species to shorten this gap (more than ten thousand genomes).

We strongly recommended that the modern taxonomy should be based on WGS. The enormous amount of unique gene sequences (e.g., 16S rRNA gene) databases should be always compared to the available genome-based phylogeny. Studies focusing on one specific taxa/group cannot be disregarded the phylogenetic analysis for the whole major taxa. It will avoid the inclusion of the previously erroneous taxa on the analysis. Furthermore, the anxiety to give a new name should be reconsidered. Proposes of new taxa where the phylogenetic relationship was not firmly established are frequently found (e.g., Rajaniemi et al., 2005).

**Ecogenomics and the Delineation of the Ecological Niches of Cyanobacteria**

Correlation analysis allowed us to characterize how the abundance of the analyzed genomes is associated with environmental parameters at both marine and freshwater habitats. These associations shed light on ecological interactions taking place within aquatic habitats that are responsible for delineating the ecological niches of Cyanobacteria. Our results showed that taxonomic affiliation and niche occupancy are coherently linked, i.e., closely related species of the same genus often shared correlation patterns, and consequently were assigned to the same ecogenomic group.

The identification of specific features responsible for defining niche occupancy among these organisms depends on extensive experimental data focusing on both physiological and morphological features, which is outside of our scope. Nevertheless, we speculate that some features are likely playing a role in this process:

1. **Transcriptional patterns**: The way in which Cyanobacteria regulate gene expression in response to changing environmental conditions is likely to play a role in defining which habitats are better suitable for growth of different species.

2. **Nutrient uptake and utilization**: Throughout the aquatic environment a myriad of gradients of nutrient abundance are formed (Stocker and Seymour, 2012). The cyanobacterial capacity for uptake and utilization of limiting nutrients (e.g., P, N and Fe) is associated with their ecological niches occupancy (Thompson et al., 2013a; Coutinho et al., 2016b; Farrant et al., 2016). Considering that significant associations were detected between the abundance of the analyzed genomes and the nutrients sources (phosphorus and nitrogen), we assume that the diversity and efficiency of their nutrient transporters plays a major role in defining the cyanobacterial affiliation to the proposed ecogenomic groups.

3. **Photosynthetic machinery and efficiency**: Cyanobacteria are remarkably diverse when considering their photosynthetic physiology. Species differ with regard their preferred light intensities and wavelengths which affects their photosynthetic efficiency (Moore et al., 1998; Ting et al., 2002). They also can be differentiated regarding their carboxysomes, sub-cellular structures where carbon fixation takes place (Yeates et al., 2008). To our knowledge, no study has consistently compared the photosynthetic yields of all the strains analyzed here, therefore we cannot determine if the proposed ecogenomic groups differ regarding this parameter. Nevertheless, distinctions regarding their requirements for efficient photosynthesis are likely linked to their patterns of niche occupancy.

**Ecogenomics, Global Changes, and Cyanobacterial Communities**

Over the past two centuries, human development has affected aquatic ecosystems due to nutrient over-enrichment (eutrophication), hydrologic alterations, global warming and ocean acidification. Temperature is one of the most important factors determining the taxonomic composition of marine microbial communities (Sunagawa et al., 2015). Our data shows that temperature is central for regulating the composition and functioning of cyanobacterial communities. Global warming can affect growth rates and bloom potentials of many taxa within this phylum (Fu et al., 2007; Paerl and Huisman, 2008; Flombaum et al., 2013; Pittera et al., 2014). Niche based models predict an increase in the absolute levels of organisms formerly classified as Prochlorococcus and Synechococcus due to global warming (Flombaum et al., 2013). Consequently, the functioning of the biogeochemical cycles in which these organisms are involved...
Anabaena, Trichodesmium, and Anabaena (Zehr, 2011). Both beneficial and deleterious effects of the ocean warming and associated phenomena (e.g., acidification) on the rates of growth and N₂ fixation have been reported (Hutchins et al., 2007; Shi et al., 2012; Fu et al., 2014), and recent laboratory and field experiments (Hong et al., 2017) showed that the acidification inhibit growth and N₂ fixation in T. erythraeum IMS101ᵀ due a decrease in cytosolic pH resulting biochemical cost of proton pumping across membranes. Rising temperatures might shift cyanobacterial community composition toward a state were diazotrophs are relatively less abundant. Because nitrogen is often a limiting nutrient to marine primary productivity (Tyrrell, 1999; Moore et al., 2013), alterations in the oceanic levels of nitrogen fixation could affect not only non-diazotrophic Cyanobacteria but also heterotrophic microbes as well as the higher trophic levels that are sustained by microorganisms. Furthermore, our findings suggest that changes in temperature can affect the contributions of Cyanobacteria to the global carbon pump (Flombaum et al., 2013; Biller et al., 2015). For example, the five strongest positive correlations with temperature between the High Temperature Oligotroph group involve the high-light adapted members of the Eurycolium genus (i.e., strains MIT9312ᵀ, MIT9301ᵀ, MIT9215, MIT9202ᵀ, and AS9601ᵀ). These are high-light adapted strains that display lower photosynthetic efficiency than their low-light adapted counterparts (Moore et al., 1998; Moore and Chisholm, 1999). Our results suggest that the relative abundance of high-light adapted strains would increase induced by the rising temperatures. In turn, these changes could affect the efficiency of carbon fixation in the ocean, a change that could also be influenced by the alterations in nitrogen fixation mentioned above.

CONCLUSIONS

The present study proposes a first attempt toward integrating taxonomy and ecogenomics, offering a compelling new perspective for the development of Cyanobacteria studies. Our results show that closely related genomes often share a niche and can be assigned to the same ecogenomic group. End-users of Cyanobacteria taxonomy may benefit from a more reproducible and portable taxonomic scheme. Future studies are needed to expand the evolutionary and physiological basis for the cyanobacterial niche occupancy, integrating other important ecological variables such as phage susceptibility, light utilization strategies, horizontal gene transfer, and inter-species interactions.

AUTHOR CONTRIBUTIONS

All authors contributed to the writing of the manuscript. JW, FC, BD, JS, FT, and CT designed and planned the study. JW and FC performed the bioinformatics analyses, analyzed the results, and compiled the data. All authors approved the final version of the manuscript.

FUNDING

This work was supported by the National Counsel of Technological and Scientific Development (CNPq), Coordination for the Improvement of Higher Education Personnel (CAPES), and Rio de Janeiro Research Foundation (FAPERJ).

ACKNOWLEDGMENTS

This paper is part of the D. Sc. requirements of JW at the Biodiversity and Evolutionary Biology Graduate Program of the Federal University of Rio de Janeiro (UFRJ), and was developed within the Science Without Borders Program (Oceanography and Environmental Impacts Coordination Program/CNPq, process no. 207751/2014-5). Sequence Data: All publicly available sequence data used in this paper was retrieved from the RefSeq (https://www.ncbi.nlm.nih.gov/refseq/) and GenBank, as part of the International Nucleotide Sequence Database Collaboration, and also from the GEBA database, produced by the US Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/) in collaboration with the user community.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2017.02132/full#supplementary-material

Figure S1 | Ribosomal phylogenetic reconstruction of the Cyanobacteria phylum. Tree was constructed through Mr. using the Kimura 2-parameter method, and GTR+G substitution model. Tree was inferred from 110 16S rRNA gene sequences (~1,400 bp). The species cut-off was 98.8% similarity (Thompson et al., 2015). The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test (1,000 replicates) are shown next to the branches. Nodes supported with a bootstrap of ≥ 50% are indicated. Overwritten T indicates type strain or type species of validly published species to assess their correct phylogenetic assignations. Bold names indicate the additional type strains or type species (only for 16S tree). The unit of measure for the scale bars is the number of nucleotide substitutions per site. Coleofascicularis chthonoplastes PCC
7420 is also called Microcoleus chthonoplastes PCC 7420. Gloeobacter violaceus PCC 7421 sequence was designated as outgroup.

**Figure S2 |** Heatmaps based on GGD metrics of specific cases. (A) Heatmap of GGD values between Oscillatoria group (case I), where Microcoleus vaginatus FGP-2 type strain was included to show the closest relationship with the PCC 7112 strain; (B) Heatmap of GGD values between Leptolyngbya group (case II); (C) Heatmap of GGD values between Arthospira group (case III); and (D) Heatmap of GGD values between Synechococcus group (case IV). The intraspecies limit is assumed as ≥70% GGD. The GGD values are associated with the respective thermal color scale located at the bottom left corner of the figure. The proposed new names were adopted in this figure.

**Figure S3 |** Abundance and distribution of ecogenomic clusters across global marine metagenomes. Relative abundance of Low Temperature group; Low Temperature Copiotroph group; and High Temperature oligotroph group at the global scale.

**Figure S4 |** Abundance and distribution of ecogenomic clusters across freshwater metagenomes. (A) Relative abundance of ecogenomic clusters in Caatinga biome (metagenomes, N = 8). (B) Non-metric multidimensional scaling (NMDS) analysis of the freshwater metagenomes and environmental parameters.

Ordination plot of physicochemical parameters. Dots indicate the metagenomes samples. Distances were calculated based on the Bray-Curtis Method. NMDS stress value = 0.15. (C) Non-metric multidimensional scaling (NMDS) analysis of the freshwater metagenomes and environmental parameters. Ordination plot of ecogenomic clusters. Dots indicate the metagenomes samples. Distances were calculated based on the Bray-Curtis Method. NMDS stress value = 0.15.

**Table S1 |** Estimates of genome relatedness of cyanobacterium strains. Values at the matrix indicates the intergenomic distances (i.e., evolutionary divergence between sequences). The numbers of base substitutions per site between sequences are shown. Analyses were conducted accordingly Tamura et al. method. The analysis involved 110 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 759 positions in the final dataset. Evolutionary analyses were conducted in MEGA6.

**Table S2 |** Details of all cyanobacterial genomes included in this study. Information of other classifications for comparison: NCBI (order and family), numerical identification and genera according to Kozol et al. (2016), and identification based in curated database Cyanotype v.1 (Ramos et al., 2017). Overwritten T indicates type strain or type species.

## REFERENCES


Antonie van Leeuwenhoek. 110, 1257–1269. doi: 10.1007/s10482-017-0848-0


Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling Editor declared a shared affiliation, though no other collaboration, with several of the authors [JW, FC, FT, CT].

Copyright © 2017 Walter, Coutinho, Dutilh, Swings, Thompson and Thompson. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.