MICROBIOLOGY

Microcystis pangenome reveals cryptic diversity within and across morphospecies

Haiyuan Cai[1], Christopher J. McLimans[1,2], Jessica E. Beyer[1,2], Lee R. Krumholz[3], K. David Hambright[1,2,4]*

***Microcystis*, a common harmful algal bloom (HAB) taxon, threatens water supplies and human health, yet species delimitation is contentious in this taxon, leading to challenges in research and management of this threat. Historical and common morphology-based classifications recognize multiple morphospecies, most with variable and diverse ecologies, while DNA sequence–based classifications indicate a single species with multiple ecotypes. To better delimit *Microcystis* species, we conducted a pangenome analysis of 122 genomes. Core- and non–core gene phylogenetic analyses placed 113 genomes into 23 monophyletic clusters containing at least two genomes. Overall, genome-related indices revealed that *Microcystis* contains at least 16 putative genospecies. Fifteen genospecies included at least one *Microcystis aeruginosa* morphospecies, and 10 genospecies included two or more morphospecies. This classification system will enable consistent taxonomic identification of *Microcystis* and thereby aid in resolving some of the complexities and controversies that have long characterized eco-evolutionary research and management of this important HAB taxon.**

INTRODUCTION

Ecologically, the widespread, toxigenic, bloom-forming, ecosystem-disruptive cyanobacterium *Microcystis* represents an intriguing enigma, due to its complex and controversial taxonomy (*1*), combined with its characteristically broad niche and cosmopolitan distribution (*2*–*5*). Variations in nutrient affinities, uptake rates, cell quotas, nitrogen metabolism, and toxin production have been observed within species (*4*, *6*, *7*), leading to the question of whether current morphological taxonomy represents accurate and meaningful species delimitation that can reliably inform *Microcystis* evolutionary history and niche adaptation. This enigma is typified by the most common species, *Microcystis aeruginosa*, but is also apparent in others, including *Microcystis viridis*, *Microcystis wesenbergii*, *Microcystis ichthyoblabe*, *Microcystis natans*, *Microcystis botrys*, *Microcystis firma*, *Microcystis flos-aquae*, *Microcystis novacekii*, *Microcystis panniformis*, and *Microcystis smithii* [e.g., (*8*, *9*)]. Colony and cell morphology, primary characteristics used for species delineation, and toxin production, the primary public health concern for temperate and tropical waters globally, have been shown to vary within *Microcystis* taxa and across environmental and seasonal conditions (*4*, *5*, *10*–*15*).

Taxonomy and systematics provide a foundational understanding of the biology and evolutionary relationships among organisms, without which our understanding of an organism's biology and ecology is severely impeded (*16*, *17*). Historically, taxonomy has been predominantly morphology based, although morphology does not necessarily reflect species boundaries as defined genetically (*16*, *18*). Working with a flawed species delineation will inevitably produce erroneous answers to evolutionary and biogeographical questions. For example, the whole-genome–based analysis of *Prochlorococcus* by Thompson and colleagues (*19*) upended the previous assumption of niche specialization and range restriction. On the basis of whole genomes, their analysis revealed that some species are generalists with wide geographical ranges, whereas some are specialists with narrow ranges. Regardless of which species concept is used (*20*, *21*), limitations of morphologically based taxonomy can become exacerbated by the low morphological complexity and small size of microbes (*16*). For example, it has long been known that morphology is unreliable for cyanobacterial species identification and systematics (*9*, *22*, *23*). *Microcystis* is no different. Because of the morphological and ecological incongruencies across *Microcystis* species, researchers have sought and used genetic-based taxonomic markers for decades [e.g., (*24*, *25*)].

Standard genetic markers in bacterial taxonomy include hypervariable regions within the 16*S* small subunit ribosomal RNA [rRNA; sensu, (*26*)], but the utility of the 16*S* rRNA genes for taxonomic resolution at the species level within *Microcystis* has proven uninformative due to high sequence similarities (>>97%) (*4*, *5*, *27*). Intergenic regions with higher levels of sequence divergence than 16*S* hypervariable regions, including the internal transcribed spacer (ITS) of the rRNA operon (*28*–*30*) and the intergenic spacer genes for phycocyanin biosynthesis (cpcBA-IGS) (*28*, *31*–*33*), have also previously been used in an effort to analyze *Microcystis* diversity and composition. As with 16*S* rRNA genes, these markers have proven insufficient with respect to satisfactory taxonomic or phylogenetic resolution beyond a single *Microcystis* genus. Classification using multilocus sequence typing based on seven single-copy housekeeping genes has been useful in sorting species belonging to this genus (*34*); however, the selected genes might not reflect overall genome evolution of *Microcystis*.

Whole-genome sequencing provides a potential solution to the lack of taxonomic resolution of traditional genetic markers (*19*, *35*, *36*). In instances for which full-length 16*S* rRNA sequence similarity is ≥98.7%, Chun *et al.* (*37*) suggest that species boundaries might be identified using whole-genome sequencing and an overall

[1]Plankton Ecology and Limnology Laboratory, Department of Biology, University of Oklahoma, Norman, OK, USA. [2]Program in Ecology and Evolutionary Biology, Department of Biology, University of Oklahoma, Norman, OK, USA. [3]Department of Microbiology and Plant Biology and Institute for Energy and the Environment, University of Oklahoma, Norman, OK, USA. [4]Geographical Ecology, Department of Biology, University of Oklahoma, Norman, OK, USA.
*Corresponding author. Email: dhambright@ou.edu

genome-related index (OGRI) based on average nucleotide identity (ANI) of <95 to 96% and digital DNA-DNA hybridization (dDDH) values of <70%. However, studies have shown that multiple *Microcystis* morphospecies form a single-species complex using this approach (*4*, *5*, *9*). Moreover, such universal criteria may not be robust for taxa with high variability in genome size and gene content, such as *Microcystis*, in which only a small part of each genome contains appropriate orthologous regions used in ANI calculations. Such high genome variability is primarily due to high genome plasticity and horizontal gene transfer (HGT), which has apparently played a substantial role in the evolution of *Microcystis* (*12*–*14*, *38*, *39*). As a complement to OGRI-based analyses, the use of a small group of vertically inherited genes (core genes) has the potential to create a more complete genealogical history in difficult to classify families of bacteria (*40*). Moreover, use of single-copy orthologous genes could minimize confounding effects of HGT (*41*) and thereby generate robust and resolved phylogenies (*42*). Further, phylogenies based on the complete genome data (i.e., core and noncore genes) should reflect both an organism's evolution and adaptation to specific habitats, thus providing a preferable framework for demarcating species (*40*, *43*).

Here, we used 122 published *Microcystis* genomes to create a robust phylogeny based on both core- and pan-gene phylogenies and identified putative genospecies using modified OGRI thresholds. We classified 113 genomes into 23 monophyletic clusters and identified at least 16 putative genospecies. To aid in the application of this new classification scheme when whole-genome sequences are not available, we propose 11 marker genes that can reliably place new *Microcystis* strains within the classification scheme. Ultimately, this new and emerging *Microcystis* taxonomy will focus further research on better characterizing *Microcystis* ecotypes, niches, and evolutionary history.

RESULTS

The *Microcystis* pangenome

The 122 *Microcystis* genomes ranged in size between 3.88 and 5.89 million base pairs (Mbp) and were comprised on average of 354 contigs, 42.7% G + C content, and 4597 protein-coding sequences (table S1). Inclusion of potentially incomplete draft genomes in our analysis did not appear to affect these estimates, as the nine complete genomes ranged in size from 4.30 to 5.87 Mbp, with an average of 42.5% G + C content, and 4999 protein-coding sequences. The pangenome contained 21,880 nonredundant genes, of which the core, accessory, and unique genomes were composed of 1639, 14,562, and 5679 genes, respectively (Fig. 1A). The rarefaction curve showed that more genes would be identified as more genomes are added, but the rate of new gene discovery decelerates as more genomes are added (43 new genes are predicted for the 123rd genome) (Fig. 1B). The rate of reduction in core gene numbers also slows but is projected to be much less affected by the addition of new genomes (e.g., the 123rd genome would have ~two fewer core genes) (Fig. 1B).

Among the core genes, 1452 single-copy genes (table S2) were used to identify core genomic relationships and evolutionary history. These single-copy genes included B vitamin biosynthetic genes, such as *thiCEG* [thiamine (B_1)-phosphate synthase], *ribAB* [riboflavin (B_2) biosynthesis], *pdxJ* [pyridoxine (B_6) 5′-phosphate synthase], and *cobD* [cobalamin (B_{12}) biosynthesis]. A set of organic and inorganic compound transporter genes were also present, including *pstABC* (phosphate transport system), *amtB* (an ammonium transporter), *cmpBCD* (bicarbonate transport system), *cysTW* (sulfate transport system), *corA* (cobalt/magnesium transport protein), *znuABC* (high-affinity zinc uptake system), *dppB* (dipeptide transport system permease protein), *glnQ* (glutamine transport adenosine triphosphate–binding protein), and *livFH* (high-affinity branched-chain amino acid transport system).

Microcystis phylogeny

The phylogeny based on sequence similarity of the single-copy core genes suggested the existence of 23 conserved, monophyletic clusters containing 113 genomes (Fig. 2A). Nine genomes, including the outgroup, failed to group with any other genome. The presence of 23 single-copy core gene clusters was corroborated by the pan-gene presence-absence tree (Fig. 2B). There were some differences in topology between the single-copy core and pan-gene trees, but they had identical cluster membership, including the same nine outliers. All 23 clusters included at least two genomes, were monophyletic, genomically coherent across core- and pan-gene phylogenies, and were supported by high bootstrap values. Moreover, ANI and dDDH analyses (fig. S2) revealed that 16 of the 23 clusters could be defined by thresholds of 0.970 and 0.750, respectively, with maximum between cluster ANI and dDDH values of 0.969 and 0.746, respectively (table S3).

Twenty-two of 23 clusters and 15 of the 16 putative genospecies contained at least one genome previously described as *M. aeruginosa* [as identified in the National Center for Biotechnology Information (NCBI) database]. Twelve clusters and 10 genospecies included at least two morphospecies. Genospecies C was composed solely of *M. panniformis*. However, some *M. panniformis* genomes also aligned with genospecies D (containing *M. aeruginosa* and unidentified *Microcystis* sp.). Genospecies B, H, and I contained both *M. aeruginosa* and *M. wesenbergii*. Genospecies J and M, as well as cluster 20, contained both *M. aeruginosa* and *M. flos-aquae*. Strains identified as *M. viridis* appeared in genospecies G and cluster 21. The nine genomes that did not cluster in either tree included six strains identified as *M. aeruginosa* (the outgroup Ma_SC_T_19800800_S464, Ma_W13_11, Ma_MB_F_20061100_S20D, Ma_QC_C_20070703_M131, Ma_NIES4325, and Ma_KW), one *M. wesenbergii* strain (Mw_MB-S_200031200_S109D), the only high-quality *M. novacekii* strain (Mn_MB_F_20050700_S1D), and one unclassified *Microcystis* strain (Msp_0824).

On the basis of the reported geographic origins of each strain, five genospecies revealed some level of geographic fidelity (Fig. 3). All strains within genospecies A, B, and P were isolated from lakes in North America (Canada or United States). However, North American strains are also grouped into genospecies D, E, F, G, H, I, J, M, and O, with 12 genomes falling outside the 16 putative genospecies. Genomes within genospecies K and N were isolated from lakes in East Asia, but East Asian genomes also grouped with genospecies D, G, H, I, J, L, and O, with nine genomes falling outside the 16 putative genospecies. Strains from genospecies C were all isolated from lakes in Brazil, but three Brazilian genomes did not cluster with any of the 16 putative genospecies. Only genospecies F and G appear to be more cosmopolitan, with genospecies F associated with North America, Africa, and Europe and genospecies G associated with North America, East Asia, Europe, and Australia. While there was

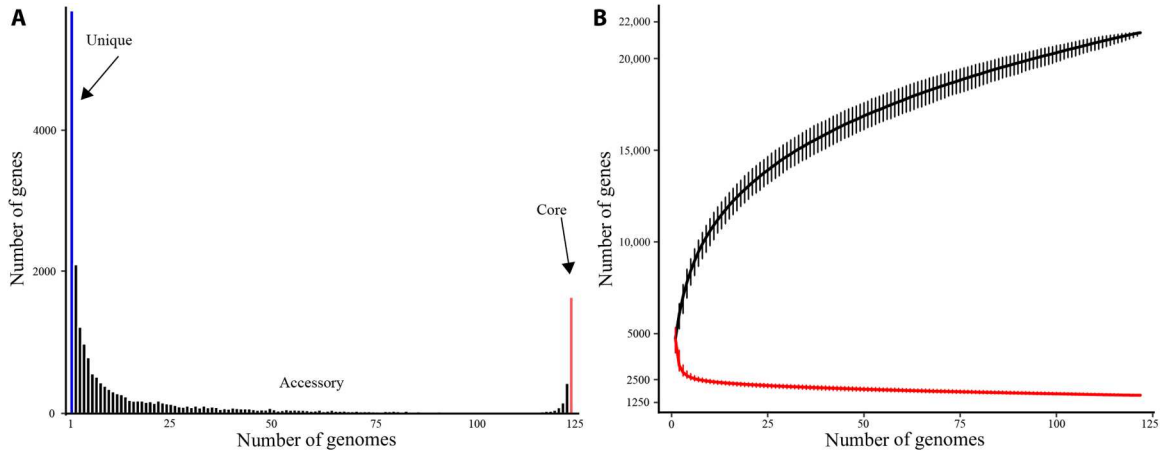

Fig. 1. Pangenome analysis of 122 *Microcystis*. (**A**) The frequencies of pan-genes (*y* axis) are plotted as a function of the number of genomes sharing those genes (*x* axis). Blue indicates unique genes, black indicates accessory genes, and orange indicates core genes. (**B**) Rarefaction curves created with PEPPAN for the accumulations of pan genes and core genes of 122 *Microcystis* genomes from 1000 random permutations. The gene accumulation curve of the increasing number of genomes fitted the power law ($n = \kappa N^{\alpha}$) with exponent $\alpha \pm 95\%$ confidence interval (CI) = 0.847 ± 0.003; core genes decreased as the number of genomes increases and followed a power law ($n = \kappa N^{\alpha}$) with exponent $\alpha \pm 95\%$ CI = −0.172 ± 0.008. Error bars indicate 95% CIs for 1000 random permutations.

some evidence for geographic relationships among some of the putative genospecies, ANI, dDDH, and core gene genetic distance revealed little relationship to geographic distance, with geographic distance explaining less than 7% of the variance in ANI, dDDH, or core gene genetic distance values (fig. S3).

Putative secondary metabolite biosynthetic gene clusters (smBGCs) constituted part of the *Microcystis* accessory genome and an average of 6.6% of the total genome across all 122 *Microcystis* strains. Seven general smBGC types were consistently identified by antiSMASH, although their distribution varied across genospecies (Fig. 3). For example, most genospecies and groups, except C, O, 8, and 19, had cytotoxic cyanobactin-type smBGCs (piricyclamide, microcyclamide, and viridisamide A). However, heterocyst glycolipid synthase-like PKS (hglE-KS)–type heterocyst glycolipid synthases were detected only in genospecies K, O, and P. Serine protease–inhibiting microviridins (microviridin B, K, and K139) were detected in all genospecies except B, I, and J. The cytotoxic nonribosomal peptide synthetase/polyketide synthase (NRPS/PKS)–type gene clusters (aeruginosin and microcystin) were found in all genospecies and groups except genospecies B, H, I, K, L, and M and group 7. The first and third most abundant smBGCs across all genomes were terpene-type ($N = 377$) and bacteriocin-type ($N = 208$) gene clusters that did not correspond to any known smBGCs in the antiSMASH database (fig. S4).

Marker genes

Analysis of genetic markers standard to bacterial taxonomy (16*S* rRNA and its various hypervariable regions, ITS of the rRNA operon, and the cpcBA-IGS) confirmed the low power of these traditional species markers in resolving *Microcystis* taxonomy. Linear correlations between the genome-based parameters (i.e., ANI, dDDH, and core gene similarity) and the 16*S* rRNA sequence similarities were weak (Table 1). Complete 16*S* rRNA genes (including hypervariable regions, e.g., V3 and V4) were detected in only 61 of 122 genomes with sequence similarities ranging from 99.3 to 100% (table S4). Notably, some genomes from different genospecies had identical 16*S* rRNA sequences (c.f., strain Mp_FACHB 1757 in genospecies D, Ma_PCC7806SL in genospecies F, and strain Ma_PCC9808 in genospecies G) (table S4). Like the 16*S* sequences, both the ITS ($N = 61$ complete sequences) and the cpcBA-IGS ($N = 122$ complete sequences) showed weak correlations with ANI, dDDH, and core gene similarity (ITS: $r^2 = 0.242$, 0.237, and 0.220; cpcBA-IGS: $r^2 = 0.429$, 0.449, and 0.411) (see fig. S5 for cpcBA-IGS phylogeny).

Among the 1452 single-copy core genes, 303 were longer than 1400 bp (i.e., comparable in size to the 16*S* rRNA gene). From these genes, we identified 11 genes that could resolve trees similar to the core gene tree and thus serve as taxonomically diagnostic marker genes (table S5). The ratio of variable sites to conserved sites (*44*) varied from 0.148 to 0.250, indicating that the gene sequences contain both highly conserved regions for primer design and hypervariable regions for phylogenetical analysis. Hypervariable regions were separated by highly conserved regions (fig. S6), allowing us to generate two sets of primers for each gene designed to amplify 1000- to 1500-bp fragments and 300- to 400-bp hypervariable fragments, respectively (table S5). These 11 genes included three transporter genes (*amtB*, *dppA*, and *yccS*), two exonuclease genes (*recJ* and *sbcC*), and two ligase genes (*glnA* and *murC*). All genes reflected overall genome evolution indicated by concordance with ANI, dDDH, and the core gene similarities (Table 1). Correlations between the 11 marker genes and the ANI, dDDH, and core gene similarity ranged from 59 to 83%, while correlations of the 16*S* rRNA gene to the same measures ranged from 9 to 11%. Moreover, the sequence similarity of marker genes ranged from 92.2 to 100%, compared with the similarity of 16*S* rRNA gene sequences that ranged from 99.3 to 100%, indicating that our 11 marker genes have diverged at a higher rate than the 16*S* rRNA gene, thus ultimately allowing more accurate identification of small-scale differences between *Microcystis* genospecies. Phylogenetic trees produced from these genes produced monophyletic groupings similar to those of the core gene– and the pangenome-based phylogenies (i.e., cluster assignment accuracy of >89%). The *sbcC* gene showed the highest concordance with ANI, dDDH, and core gene similarities, although none of the 11 marker genes individually

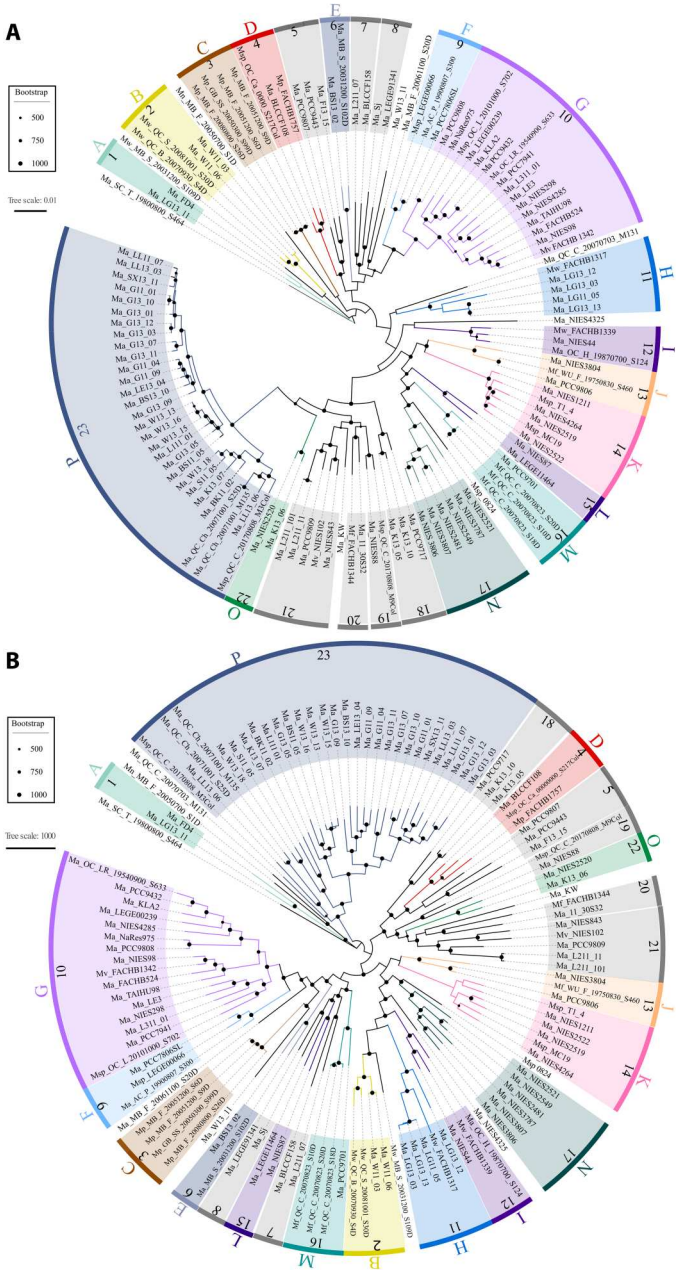

Fig. 2. *Microcystis* core- and pan-gene phylogenies. Twenty-three genomically coherent, monophyletic clusters identified by comparing core phylogeny and pan-genome-based dDDH/ANI clustering. (**A**) Rooted maximum likelihood phylogenetic tree inferred for 122 *Microcystis* genomes using the concatenated alignment of 1452 single-copy core genes. A total of 1,330,000 nucleotide positions was used, and *M. aeruginosa* Ma_SC_T_19800800_S464 was used as an outgroup, as identified through a previous comparison of all *Microcystis* genomes to other cyanobacteria (see fig. S1). (**B**) Phylogeny based on the binary gene presence and absence matrix of all pan-genes. Bootstrap support values were calculated from 1000 replicates; values above 500 are indicated. Lettered color segments indicate genospecies supported by ANI and dDDH clustering; numbered gray segments indicate clusters that were not supported by ANI and dDDH.

generate 100% cluster fidelity compared with the core gene tree. However, bootstrapped trees inferred from concatenated sequences [see (*45*)] of pairs of the 11 genes (table S6) revealed that the concatenation of four pairs of these marker genes (*acrB* + *glnA*, *amtB* + *sbcC*, *glnA* + *helY*, and *helY* + *sbcC*) could each generate phylogenetic trees with 100% cluster fidelity compared with the core gene phylogeny. The utility of the four gene pairs was confirmed by classifying 106 *Microcystis* MAGs from Lake Champlain and Pampulha Reservoir (fig. S7) (*14*). Even for cases in which none of these gene pairs were available, application of single marker genes to *Microcystis* MAGs used in in the studies of Pérez-Carrascal et al. (*14*) and Cook et al. (*3*) were able to provide first approximations of *Microcystis* identity and diversity (table S7).

DISCUSSION

Accurate classification of *Microcystis* species is essential for biological inquiry and will likely affect future management of *Microcystis*-based harmful algal blooms (HABs). Unfortunately, traditional morphospecies and DNA sequence approaches have provided contradictory and insufficient detail for advancing the study and understanding of *Microcystis*-based HABs (*9*, *46*). As in past evaluations [e.g., (*4*, *5*, *28*)], our work found that standard bacterial taxonomic markers (complete and hypervariable regions of the 16*S* rRNA gene, ITS, and cpcBA-IGS sequences) were not able to consistently resolve the 23 monophyletic, genomically coherent clusters that were identified through our whole-genome approach. Moreover, quantitative polymerase chain reaction (PCR) approaches for quantification of toxic genotypes using abundances of microcystin (*mcy*) genes fail to inform management of public health risk as *mcy* gene numbers do not correlate with environmental microcystin concentrations (*47*–*49*). Whole-genome–based taxonomy has shown good potential for distinguishing toxic and nontoxic *Microcystis* strains, except for those in genospecies G and J and group 8, in which both ecotypes are included. Although whole-genome based taxonomy is also of little direct value to public health risk management, our classification scheme offers a strong foundation for robust characterization of species-specific ecologies, which can then inform management.

Our analysis of sequence similarities found that at least 16 of the 23 clusters met putative genospecies criteria, with as many as 30 or more total possible genospecies to be delimited as more genomes become available. We posit that ANI and dDDH can be used in combination to classify *Microcystis* genospecies using the following rule: Any *Microcystis* strain exhibiting ANI and dDDH values of ≥0.970 and 0.750, respectively, to a strain from a valid genospecies, belongs to that genospecies, as long as the strain also exhibits ANI and dDDH values of <0.970 and 0.750 compared with other genospecies. For example, when a new *Microcystis* genome M054S2 (*14*) was placed into the core gene tree it joined existing genospecies G, which was supported with these criteria with the lowest within group ANI and dDDH of 0.985 and 0.867, respectively, and the maximum between group ANI and dDDH of 0.964 and 0.699. These threshold values of ANI and dDDH will allow genospecies classification of newly sequenced strains as the reconstruction of a core genome phylogenetic tree of all representatives is not practical each time a new genome becomes available. It is worth noting that our proposed ANI and dDDH thresholds may need to be reevaluated as more genomes are analyzed, especially with the addition of

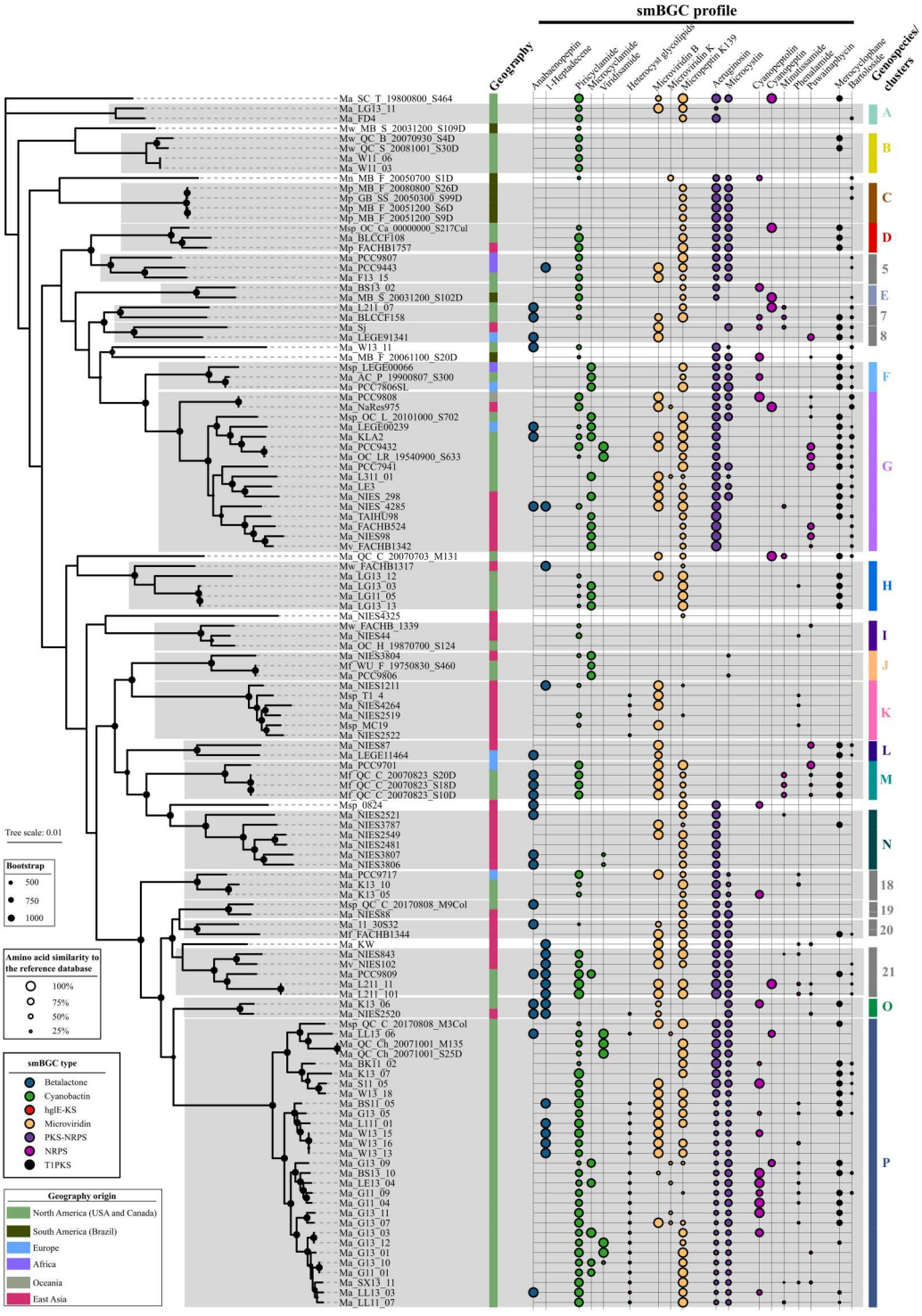

Fig. 3. Profile of smBGCs across *Microcystis*. Bootstrap support values were calculated from 1000 replicates; values above 500 are indicated. *M. aeruginosa* Ma_SC_T_19800800_S464 was used as an outgroup. From left to right: 1, geographic origin of isolates; 2, smBGC; 3, *Microcystis* genospecies or groups. Colored bars and letters on the right side of the figure indicate genospecies supported by ANI and dDDH clustering. Gray bars and numbers indicate groups that were not supported by ANI and dDDH clustering. The tree bar scale indicates the number of nucleotide substitutions per site. The area of circles indicates the amino acid similarity to the reference database.

Table 1. Coherence between proposed marker genes and genome-based similarity measures and genospecies and cluster assignment. Squared Pearson correlation coefficients (r^2) between marker gene similarities and ANI, digital dDDH values, and core gene similarity values, as well as cluster fidelity of the marker gene placement relative to the core gene tree (placement/expected), for both the 16 genospecies and 23 clusters. All comparisons are pairwise (16*S*, $N = 1830$; ITS, $N = 1830$; all others, $N = 7260$). Cluster fidelity was not possible for 16*S* rRNA or ITS due to the low number of available sequences ($N = 61$ for both). Bold correlations were significant at $P \leq 0.05$.

Marker gene	ANI	dDDH	Core gene similarity value	Genospecies assignment fidelity	Cluster assignment fidelity
Traditional					
16*S* rRNA	0.1167	0.1058	0.0999	–	–
ITS	0.2421	0.2373	0.2196	–	–
cpcBA-IGS	0.4289	0.4493	0.4106	90.4	86.7
Proposed					
acrB	**0.6846**	**0.7002**	**0.7333**	95.8	92.0
amtB	**0.7565**	**0.7442**	**0.7183**	100	95.6
cpoB	**0.7110**	**0.7211**	**0.7511**	100	91.2
dppA	**0.6502**	**0.6515**	**0.6632**	98.9	92.9
glnA	**0.5881**	**0.6091**	**0.6447**	100	97.4
helY	**0.7243**	**0.7185**	**0.7566**	97.8	95.6
murC	**0.7430**	**0.7191**	**0.6394**	100	93.8
recJ	**0.7013**	**0.6953**	**0.7104**	100	89.4
sbcC	**0.7641**	**0.7619**	**0.8308**	100	94.7
trpE	**0.6226**	**0.6358**	**0.6351**	97.8	92.0
yccS	**0.7575**	**0.7495**	**0.7754**	100	95.6

more closed (i.e., complete) genomes that will improve the identification of complete orthologous regions, as well as pan- and core genes. Ultimately, more complete genomes will increase reliability of ANI and dDDH thresholds across the phylogeny.

Pangenome analysis also revealed that less than half (35.6%) of each genome consisted of shared core genes, with 26% of pan genes being unique to a single strain. These results corroborate a highly plastic accessory *Microcystis* genome (*13*, *38*). By using both core sequence similarities and pan-gene presence-absence to construct phylogenetic trees, our proposed *Microcystis* classification considers the phylogenetic relationships among core genes that reflect evolution during adaptation and speciation, as well as phylogenetic relationships among noncore genes that reflect ecologically relevant species-environment interactions (*13*). Moreover, our proposed classification scheme supports Rosselló-Móra and Amann's (*21*) definition of a bacterial species, which is considered to represent "a category that circumscribes monophyletic and genomically and phenotypically coherent populations of individuals that can be clearly discriminated from other such entities by means of standardized parameters." A challenge ahead will be the phenotypic characterization of our proposed putative genospecies.

Although there is considerable debate regarding the roles of dispersal and environmental conditions in microbial biogeography (*5*, *50*, *51*), some strains within a genospecies seemed to reflect specific biogeography, with genospecies C all from South America, genospecies N from Southeast Asia, and genospecies P from North America. However, overall, we did not detect general geographic patterns in ANI, dDDH, or core gene phylogenetic distance metrics. Thus, it appears that geographic dispersal is not a major driver of *Microcystis* phylogenetic relationships. By contrast, our analysis of secondary metabolites corroborates the possibility for environmental filtering and ecologically relevant genomic variation between strains (*3*, *13*, *52*). While the specific functions of many secondary metabolites in bacteria are unknown, in general, they mediate response of species to other organisms or the environment. For example, secondary metabolites can serve as weapons against competitors or predators and regulators of symbiosis and nutrient acquisition and transport (*53*). Hence, secondary metabolites may provide some insight into certain ecotypes [sensu (*54*)] that may have arisen through environmental adaptation. For example, microcystin is thought to bind to enzymes as a means of protecting them from degradation by reactive oxygen species (ROS) produced during a bloom (*55*), whereas nonmicrocystin producing strains use enzymatic degradation of ROS (*4*). Hence, observed succession from toxic to nontoxic strains during the course of a bloom (*56*) may reflect a shift in ROS protection strategies from one based on production of high-nitrogen–containing microcystin to one based on enzymatic degradation of ROS when nitrogen becomes limiting (*57*, *58*). Our classification scheme could be used to differentiate between strains with the potential to produce microcystin (genospecies C, D, F, O, and P and groups 5, 18, 19, 20, and 21) from those without microcystin capabilities (genospecies A, B, E, H, I, K, L, M, and N and group 7), except for genospecies G and J and group 8, in which both putatively toxic and nontoxigenic strains are included. Whether all different smBGC profiles correspond to different *Microcystis* ecotypes remains unknown. However, recent analysis suggests that smBGC profiles may not correlate strongly with core genome phylogeny due to the high prevalence of HGT events (*59*). Cao *et al.* (*59*) recently determined that four transferred genes that could encode for microcystin biosynthesis in *M.*

panniformis FACHB-1757 (member of genospecies D in our classification) were likely derived from *Planktothrix* via HGT. Given the high nitrogen content of microcystins and their potential role in protection from ROS, a selective advantage is likely afforded to microcystin-capable *Microcystis* genotypes in nitrogen-rich systems.

While additional *Microcystis* genomes will strengthen the genomic classification of *Microcystis*, we are now better positioned to begin reassessing ecotypes and ecological niches of *Microcystis* [e.g., (*4*)] within the context of a proposed phylogeny. To aid in advancing the phenotypic characterization of *Microcystis* globally without the need for whole-genome sequencing and analysis, we propose 11 new marker genes that in select concatenated pairs, or even individually for first approximations, can replace the traditional marker gene approaches for identifying *Microcystis*. In particular, we suggest the pairing of *glnA* with either *acrB* or *helY* or *sbcC* with *helY* or *amtB* for identifying *Microcystis* genospecies. These proposed marker genes may be used to reduce the need for shotgun sequencing and analysis efforts to identify genospecies by enabling researchers to PCR amplify the single gene(s) and/or hypervariable regions with the primers proposed in table S5. Once amplified, either long- or short-read sequencing may be performed for full-length amplicons or hypervariable specific regions, respectively, and the resulting sequence compared to the existing sequences to identify the most closely related taxon and ultimately the genospecies classification. Once identified, characterization of nutrient uptake kinetics, light- and temperature-dependent growth rates, microbiomes, and other phenotypic traits [sensu (*3*, *60*, *61*–*63*)] can be defined and assigned for each genospecies, with the end result being a more concrete understanding of *Microcystis* diversity, ecology, and evolution.

Uncertainty in available metadata may limit the conclusions of our study. Namely, we cannot confirm that the provided morphospecies classifications assigned to these genomes in NCBI have been verified microscopically or using consistent criteria. These morphospecies identities may have been assigned via sequence similarity to previously deposited genomes with inaccurate classifications, thus propagating existing errors. In addition, identification may have been assigned with inappropriate methods (e.g., marker genes and universal OGRI). Hence, we must rely on comparing our genomic classification to the morphospecies name as assigned in NCBI in our analysis, although it may not be entirely accurate. We selected genomes from NCBI with the genus labeled as "*Microcystis*" (see Materials and Methods for further details). We did not find evidence that any *Microcystis* genome was misclassified at the genus level. Specifically, our use of closely related outgroups, such as *Aphanocapsa montana* BDHKU210001 (see fig. S1) to identify the root *Microcystis* genome indicated that all *Microcystis* genomes were more closely related to each other than to any closely related outgroup.

We offer a new whole-genome–based taxonomy, along with genetic markers that provide the beginnings of a new *Microcystis* classification system that will advance research characterizing the *Microcystis* niche. Our proposed *Microcystis* classification considers phylogenetic relationships among core genes (reflecting evolution during speciation), as well as phylogenetic relationships among noncore genes (reflecting species-environment interactions). This whole-genome–based classification approach overcomes historical limitations in *Microcystis* taxonomy, long troubled by morphological plasticity and cryptic species and ecotypes. This innovation will advance ecological and evolutionary research and management of this important HAB taxon.

MATERIALS AND METHODS

Study design

The current taxonomy of *Microcystis* is based on morphology. The problematic morphological taxonomy of *Microcystis* was initially challenged by the introduction of 16*S* rRNA genes. However, 16*S* rRNA genes are still insufficient for taxonomic resolution at the species level. At present, an increasing number of complete or draft genome sequences of *Microcysts* have been published, allowing for the application of whole-genome–based taxonomic tools including OGRI. However, as the OGRI threshold may not be applicable to *Microcystis* species, we needed to reevaluate the OGRI threshold. In this study, we first downloaded all *Microcystis* genomes from the NCBI database and then removed low-quality or duplicated genomes. We then identified monophyletic and genomically coherent clusters supported by both core gene– and pan-gene–based phylogenies. The clusters that were also supported jointly by ANI and dDDH clustering would be considered genospecies. We proposed new candidate marker genes that, when phylogenetically compared, produced similar groupings as the core gene– and pangenome-based phylogenies.

Dataset preparation

We downloaded all 173 complete and draft *Microcystis* genomes from the NCBI database (https://ncbi.nlm.nih.gov/genome/), on 8 August 2021. These genomes consisted of 6 complete and 114 draft genomes identified as *M. aeruginosa*, 11 draft genomes identified as *M. flos-aquae*, 10 draft genomes identified as *M. wesenbergii*, 1 complete and 8 draft genomes identified as *M. panniformis*, 1 complete and 5 draft genomes identified as *M. viridis*, 2 draft genomes identified as *Microcystis novacekii*, 1 draft genome identified as *M. elabens*, and 1 complete and 13 draft genomes with uncertain taxonomic status. We narrowed this group by selecting the highest-quality genome for strains that have been sequenced multiple times (N = 2 or 3) using "dereplicate" function of dRep v3.0.0 (*64*) with the default settings, leaving a total of 9 complete and 136 draft *Microcystis* genomes. We further filtered this dataset by removing poorly assembled genomes based on number of contigs (>1000), completeness (<97%), and contamination (>1%) using CheckM (version 1.1.3) using cyanobacteria-specific marker genes and default parameters (*65*). This quality filtering left 122 genomes that constitute the basis for further analysis (table S1). The estimated genome size was adjusted to account for its estimated completeness and contamination using the equation: estimated genome size = assembly genome size/estimated completeness/(1 + estimated contamination) (*66*).

Pangenome analysis

Identification of the pangenome requires first creating a rooted tree to determine taxonomic relationships and species evolution (*67*). However, introducing genomes of other genera for pangenome analysis would notably reduce the core gene number. To minimize this issue, we took a two-step approach to create the phylogenetic trees. Phylogenetic analyses were first performed with 122 *Microcystis* genomes along with one of several related species (based on ANI values), including *A. montana* BDHKU210001 (accession number:

NZ_JTJD00000000.1), *Cyanobium gracile* PCC-6307 (NC_019675.1), *Gloeocapsa* sp. PCC-7428 (GCA_000317555.1), and *Synechococcus elongatus* PCC-6301 (AP008231.1) (*5*). Trees were constructed using the up-to-date bacterial core gene pipeline, which infers a maximum likelihood phylogeny using the concatenated sequences of 92 universal single-copy core genes (*68*) identified across all *Microcystis* and each related, non-*Microcystis* species. All phylogenetic trees showed that *M. aeruginosa* Ma_SC_T_19800800_S464 was the basal taxon (fig. S1). We then eliminated all non-*Microcystis* strains and used *M. aeruginosa* Ma_SC_T_19800800_S464 as the outgroup in all subsequent phylogenetic analyses.

The entire gene pool of a bacterial clade, referred to as a pangenome, can be divided into a core genome (a set of genes shared by all genomes), accessory genes (a set of genes present in some genomes), and unique genes (a set of genes found in only one genome) (*69*). Genomes were annotated using Prokka (version 1.14.5, with default parameters) (*70*), followed by processing with PEPPAN (version 1.0.5), which provides consistent gene and pseudogene annotation and identification and exclusion of paralogs (*71*). A gene-presence tree was constructed using PEPPAN_parser. Single-copy core genes were aligned using MAFFT (version 7.471) (*72*) and concatenated using the FASconCAT-G software (version 1.04) (*73*). The maximum likelihood tree was inferred using RAxML-NG (version 1.0.3) (*74*) using the GTR + G +I model, which was selected as the best model by ModelTest-NG (version 0.1.7) and 1000 bootstrap replications (*75*). Trees were visualized using iTOL (http://itol.embl.de/) (*76*). Putative smBGCs encoded within the *Microcystis* genomes were identified and analyzed using the online server antiSMASH 6.0 (*77*) with detection strictness set to "relaxed" and all extra features activated.

Genospecies boundaries

We defined boundaries between *Microcystis* genospecies based on Chun *et al.*'s (*37*) OGRI, which relies on multiple sequence similarity approaches. For each pair of 122 genomes, we calculated ANI using BLASTN (*78*) and digital dDDH (*78*). The dDDH values between genome pairs were predicted using the GGDC 2.1 web server (*79*) available at http://ggdc.dsmz.de/distcalc2.php with the recommended Formula 2 and the alignment tool BLAST+. Heatmap and hierarchical cluster analyses were performed using the "pheatmap" package with default parameters in R (*80*). Average nucleotide identities were high among the 122 genomes. In pairwise genome comparisons, 70% of ANI values (range = 94.0 to 99.9%; median = 95.3%) were above the nominal prokaryote species threshold of 95%, indicating that existing ANI criteria may not be useful in delimiting *Microcystis*. In contrast, only 11% of dDDH values (range = 56.0 to 99.8%; median = 63.2%) were above the nominal prokaryote species threshold of 70%, suggesting that dDDH may be more useful. Using these OGRI results, we examined all pairwise ANI and dDDH values of the suspected monophyletic clusters from the core gene phylogeny. We evaluated all within-cluster pairwise ANI and dDDH values to find the lowest within-cluster value for each metric and then identified the highest pairwise ANI and dDDH values for any member of a cluster to all other genomes outside of that cluster (i.e., "between-cluster" comparisons). We identified new thresholds by identifying all clusters with at least three genomes in which the lowest within-cluster ANI and dDDH values were higher than the highest between-cluster ANI and dDDH values. We then found the lowest ANI and dDDH values from this subset of clusters meeting these criteria and set that as the new threshold for *Microcystis* genospecies. Clusters of two or more *Microcystis* genomes were defined as putative genospecies if they were (i) monophyletic and genomically coherent as identified by core and pan-gene trees and (ii) met our within-cluster and between-cluster OGRI thresholds (*37*).

Identification of marker genes for assigning *Microcystis* taxonomy

Given that use of the 16*S* rRNA genes for species-level resolution within *Microcystis* is limited [e.g., (*5*)], we explored the possibility of using single-copy core genes as taxonomic markers for *Microcystis* genospecies [sensu (*41*, *42*)]. Potential marker genes initially were identified as single-copy core genes with a length of >1400 bp (comparable to the size of the 16*S* rRNA gene). These were aligned within the 122 sequences for a single gene and used to build trees based on the maximum likelihood approach implemented in RAxML-NG 1.0.3 (*74*). The topological structures of the marker gene phylogenetic trees were compared with the phylogenetic tree generated from the core genome and genes with the best genospecies placement when compared to the core gene tree were considered marker genes. In addition, pairwise nucleotide sequence similarity values obtained for the marker genes were calculated and compared to dDDH, ANI, and pairwise core gene nucleotide sequence similarity values by correlation analysis. Pearson's correlation coefficient was used for the correlation analysis with linear regression.

Statistical analyses

All statistical analyses were completed in the R statistical environment v.4.0.3 (www.r-project.org). We first tested for associations between OGRIs and single-copy core gene phylogenetical distance. Pairwise genetic distance of concatenated core genes or marker genes between genomes was calculated in MEGA X using the Kimura 2-parameter model (*44*). The calculation of similarity values was performed as follows: similarity percentage = (1 − genetic distance) × 100%. Geographic distance between sites was calculated using the "rdist.earth" function in the R "fields" library v 13.3 using estimated GPS coordinates based on source lake name. Generalized linear models (GLMs) were used to assess the relationships between index measures (using default family = Gaussian, link = identity). Deviation explained by GLMs coupled with *P* values was used to assess the significance and strength of the relationships.

We then evaluated the pairwise correlation (Pearson's correlation) between dDDH, ANI, and pairwise core gene similarity values with the pairwise nucleotide sequence similarity values obtained for the 16*S* rRNA and other marker genes, individually. Values were considered statistically significant at $P < 0.05$.

Supplementary Materials

This PDF file includes:

Figs. S1 to S7
Tables S1 to S7
References

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. J. Komárek, J. Kaštovsky, J. Mareš, J. R. Johansen, Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014, using a polyphasic approach. *Preslia* **86**, 295–335 (2014).
2. I. van Gremberghe, F. Leliaert, J. Mergeay, P. Vanormelingen, K. Van der Gucht, A.-E. Debeer, G. Lacerot, L. De Meester, W. Vyverman, Lack of phylogeographic structure in the freshwater cyanobacterium *Microcystis aeruginosa* suggests global dispersal. *PLOS ONE* **6**, e19561 (2011).
3. K. V. Cook, C. Li, H. Cai, L. Krumholz, K. D. Hambright, H. W. Paerl, M. M. Steffen, A. E. Wilson, M. A. Burford, H.-P. Grossart, D. P. Hamilton, H. Jiang, A. Sukenik, D. Latour, E. I. Meyer, J. Padisák, B. Qin, R. M. Zamor, G. Zhu, The global *Microcystis* interactome. *Limnol. Oceanogr.* **65**, S194–S207 (2020).
4. G. J. Dick, M. B. Duhaime, J. T. Evans, R. M. Errera, C. M. Godwin, J. J. Kharbush, H. S. Nitschky, M. A. Powers, H. A. Vanderploeg, K. C. Schmidt, D. J. Smith, C. E. Yancey, C. C. Zwiers, V. J. Denef, The genetic and ecophysiological diversity of *Microcystis*. *Environ. Microbiol.* **23**, 7278–7313 (2021).
5. M. J. Harke, M. M. Steffen, C. J. Gobler, T. G. Otten, S. W. Wilhelm, S. A. Wood, H. W. Paerl, A review of the global ecology, genomics, and biogeography of the toxic cyanobacterium, *Microcystis* spp. *Harmful Algae* **54**, 4–20 (2016).
6. H. Shen, L. R. Song, Comparative studies on physiological responses to phosphorus in two phenotypes of bloom-forming *Microcystis*. *Hydrobiologia* **592**, 475–486 (2007).
7. X. Tan, H. H. Gu, Y. L. Ruan, J. J. Zhong, K. Parajuli, J. Y. Hu, Effects of nitrogen on interspecific competition between two cell-size cyanobacteria: *Microcystis aeruginosa* and *Synechococcus* sp. *Harmful Algae* **89**, 101661 (2019).
8. J. Komárek, J. Komárkova, Review of the European *Microcystis* morphospecies (Cyanoprokaryotes) from nature. *Czech Phycology* **2**, 1–24 (2002).
9. S. Otsuka, S. Suda, R. H. Li, S. Matsumoto, M. M. Watanabe, Morphological variability of colonies of *Microcystis* morphospecies in culture. *J. Gen. Appl. Microbiol.* **46**, 39–50 (2000).
10. E. Briand, N. Escoffier, C. Straub, M. Sabart, C. Quiblier, J.-F. Humbert, Spatiotemporal changes in the genetic diversity of a bloom-forming *Microcystis aeruginosa* (Cyanobacteria) population. *ISME J.* **3**, 419–429 (2009).
11. Y. Tanabe, M. M. Watanabe, Local expansion of a panmictic lineage of water bloom-forming cyanobacterium *Microcystis aeruginosa*. *PLOS ONE* **6**, e17085 (2011).
12. L. Frangeul, P. Quillardet, A.-M. Castets, J.-F. Humbert, H. C. Matthijs, D. Cortez, A. Tolonen, C.-C. Zhang, S. Gribaldo, J.-C. Kehr, Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics* **9**, 274 (2008).
13. K. A. Meyer, T. W. Davis, S. B. Watson, V. J. Denef, M. A. Berry, G. J. Dick, Genome sequences of lower Great Lakes *Microcystis* sp reveal strain-specific genes that are present and expressed in western Lake Erie blooms. *PLOS ONE* **12**, e0183859 (2017).
14. O. M. Pérez-Carrascal, Y. Terrat, A. Giani, N. Fortin, C. W. Greer, N. Tromas, B. J. Shapiro, Coherence of *Microcystis* species revealed through population genomics. *ISME J.* **13**, 2887–2900 (2019).
15. M. Xiao, M. Li, C. S. Reynolds, Colony formation in the cyanobacterium *Microcystis*. *Biol. Rev.* **93**, 1399–1420 (2018).
16. O. Clerck, M. D. Guiry, F. Leliaert, Y. Samyn, H. Verbruggen, Algal taxonomy: A road to nowhere? *J. Phycol.* **49**, 215V225 (2013).
17. T. H. Struck, J. L. Feder, M. Bendiksby, S. Birkeland, J. Cerca, V. I. Gusarov, S. Kistenich, K.-H. Larsson, L. H. Liow, M. D. Nowak, B. Stedje, L. Bachmann, D. Dimitrov, Finding evolutionary processes hidden in cryptic species. *Trends Ecol. Evol.* **33**, 153–163 (2018).
18. D. Bickford, D. J. Lohman, N. S. Sodhi, P. K. L. Ng, R. Meier, K. Winker, K. K. Ingram, I. Das, Cryptic species as a window on diversity and conservation. *Trends Ecol. Evol.* **22**, 148–155 (2007).
19. C. C. Thompson, L. Chimetto, R. A. Edwards, J. Swings, E. Stackebrandt, F. L. Thompson, Microbial genomic taxonomy. *BMC Genomics* **14**, 913 (2013).
20. K. de Queiroz, Species concepts and species delimitation. *Syst. Biol.* **56**, 879–886 (2007).
21. R. Rossello-Mora, R. Amann, Past and future species definitions for Bacteria and Archaea. *Syst. Appl. Microbiol.* **38**, 209–216 (2015).
22. B. A. Neilan, D. Jacobs, A. E. Goodman, Genetic diversity and phylogeny of toxic cyanobacteria determined by DNA polymorphisms within the phycocyanin locus. *Appl. Environ. Microbiol.* **61**, 3875–3883 (1995).
23. A. Wilmotte, S. Golubic, Morphological and genetic criteria in the taxonomy of Cyanophyta/Cyanobacteria. *Arch. Hydrobiol.* **64**, 1–24 (1991).
24. Y. Ouahid, G. Pérez-Silva, F. F. del Campo, Identification of potentially toxic environmental *Microcystis* by individual and multiple PCR amplification of specific microcystin synthetase gene regions. *Environ. Toxicol.* **20**, 235–242 (2005).
25. W. Tan, Y. Liu, Z. Wu, S. Lin, G. Yu, B. Yu, R. Li, *cpc*BA-IGS as an effective marker to characterize *Microcystis wesenbergii* (Komárek) Komarek in Kondrateva (cyanobacteria). *Harmful Algae* **9**, 607–612 (2010).
26. C. R. Woese, O. Kandler, M. L. Wheelis, Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci.* **87**, 4576–4579 (1990).
27. C. Lepère, A. Wilmotte, B. Meyer, Molecular diversity of *Microcystis* strains (Cyanophyceae, Chroococcales) based on 16S rDNA sequences. *Syst. Geogr. Plants* **70**, 275–283 (2000).
28. S.-J. Chun, Y. Cui, J. J. Lee, I.-C. Choi, H.-M. Oh, C.-Y. Ahn, Network analysis reveals succession of *Microcystis* genotypes accompanying distinctive microbial modules with recurrent patterns. *Water Res.* **170**, 115326 (2020).
29. D. Huo, Y. Chen, T. Zheng, X. Liu, X. Zhang, G. Yu, Z. Y. Qiao, R. Li, Characterization of *Microcystis* (Cyanobacteria) genotypes based on the internal transcribed spacer region of rRNA by next-generation sequencing. *Front. Microbiol.* **9**, 971 (2018).
30. I. Janse, W. E. A. Kardinaal, M. Meima, J. Fastner, P. M. Visser, G. Zwart, Toxic and nontoxic microcystis colonies in natural populations can be differentiated on the basis of rRNA gene internal transcribed spacer diversity. *Appl. Environ. Microbiol.* **70**, 3979–3987 (2004).
31. D. X. Guan, X. Y. Wang, H. C. Xu, L. Chen, P. F. Li, L. Q. Ma, Temporal and spatial distribution of *Microcystis* biomass and genotype in bloom areas of Lake Taihu. *Chemosphere* **209**, 730–738 (2018).
32. P. H. Moisander, M. Ochiai, A. Lincoff, Nutrient limitation of *Microcystis aeruginosa* in northern California Klamath River reservoirs. *Harmful Algae* **8**, 889–897 (2009).
33. Z. J. Wang, Y. Liu, Y. Xu, P. Xiao, R. H. Li, The divergence of *cpcBA*-IGS sequences between *Dolichospermum* and *Aphanizomenon* (Cyanobacteria) and the molecular detection of *Dolichospermum flos-aquae* in Taihu Lake, China. *Phycologia* **52**, 447–454 (2013).
34. Y. Tanabe, F. Kasai, M. M. Watanabe, Multilocus sequence typing (MLST) reveals high genetic diversity and clonal population structure of the toxic cyanobacterium Microcystis aeruginosa. *Microbiology* **153**, 3695–3703 (2007).
35. L. M. Rodriguez-R, J. C. Castro, N. C. Kyrpides, J. R. Cole, J. M. Tiedje, K. T. Konstantinidis, How much do rRNA gene surveys underestimate extant bacterial diversity? *Appl. Environ. Microbiol.* **84**, e00014-18 (2018).
36. Y. Okazaki, S. Fujinaga, M. M. Salcher, C. Callieri, A. Tanaka, A. Kohzu, H. Oyagi, H. Tamaki, S.-I. Nakano, Microdiversity and phylogeographic diversification of bacterioplankton in pelagic freshwater systems revealed through long-read amplicon sequencing. *Microbiome* **9**, 24–24 (2021).
37. J. Chun, A. Oren, A. Ventosa, H. Christensen, D. R. Arahal, M. S. da Costa, A. P. Rooney, H. Yi, X. W. Xu, S. De Meyer, M. E. Trujillo, Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int. J. Syst. Evol. Microbiol.* **68**, 461–466 (2018).
38. J.-F. Humbert, V. Barbe, A. Latifi, M. Gugger, A. Calteau, T. Coursin, A. Lajus, V. Castelli, S. Oztas, G. Samson, C. Longin, C. Medigue, N. Tandeau de Marsac, A tribute to disorder in the genome of the bloom-forming freshwater cyanobacterium *Microcystis aeruginosa*. *PLOS ONE* **8**, e70747 (2013).
39. A. Willis, J. N. Woodhouse, Defining cyanobacterial species: Diversity and description through genomics. *CRC Crit. Rev. Plant. Sci.* **39**, 101–124 (2020).
40. Y. Jin, J. L. Zhou, J. Zhou, M. D. Hu, Q. Zhang, N. Kong, H. G. Ren, L. Liang, J. J. Yue, Genome-based classification of *Burkholderia cepacia* complex provides new insight into its taxonomic status. *Biol. Direct* **15**, 6 (2020).
41. Y. C. Zhang, S. Qiu, Examining phylogenetic relationships of *Erwinia* and *Pantoea* species using whole genome sequence data. *Antonie Van Leeuwenhoek* **108**, 1037–1046 (2015).
42. D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P. A. Chaumeil, P. Hugenholtz, A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
43. A. Caputo, P. E. Fournier, D. Raoult, Genome and pan-genome analysis to classify emerging bacteria. *Biol. Direct* **14**, 5 (2019).
44. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
45. T. Thiergart, G. Landan, W. F. Martin, Concatenated alignments and the case of the disappearing tree. *BMC Evol. Biol.* **14**, 266 (2014).
46. S. Otsuka, S. Suda, S. Shibata, H. Oyaizu, S. Matsumoto, M. M. Watanabe, A proposal for the unification of five species of the cyanobacterial genus *Microcystis* Kützing ex Lemmermann 1907 under the Rules of the Bacteriological Code. *Int. J. Syst. Evol. Microbiol.* **51**, 873–879 (2001).
47. W. Zhang, I. Lou, W. Ung, Y. Kong, K. M. Mok, Analysis of cylindrospermopsin- and microcystin-producing genotypes and cyanotoxin concentrations in the Macau storage reservoir. *Hydrobiologia* **741**, 51–68 (2014).
48. A. B. F. Pacheco, I. A. Guedes, S. M. F. O. Azevedo, Is qPCR a reliable indicator of cyanotoxin risk in freshwater? *Toxins* **8**, 172 (2016).
49. L. J. Beversdorf, S. D. Chaston, T. R. Miller, K. D. McMahon, Microcystin mcyA and mcyE gene abundances are not appropriate indicators of Microcystin concentrations in lakes. *PLOS ONE* **10**, e0125353 (2015).
50. K. D. Hambright, J. E. Beyer, J. D. Easton, R. M. Zamor, A. C. Easton, T. C. Hallidayschult, The niche of an invasive marine microbe in a subtropical freshwater impoundment. *ISME J.* **9**, 256–264 (2015).

51. J. B. H. Martiny, B. J. M. Bohannan, J. H. Brown, R. K. Colwell, J. A. Fuhrman, J. L. Green, M. C. Horner-Devine, M. Kane, J. A. Krumins, C. R. Kuske, P. J. Morin, S. Naeem, L. Ovreas, A.-L. Reysenbach, V. H. Smith, J. T. Staley, Microbial biogeography: Putting microorganisms on the map. *Nat. Rev. Microbiol.* **4**, 102–112 (2006).
52. F. Dini-Andreote, J. C. Stegen, J. D. van Elsas, J. F. Salles, Disentangling mechanisms that mediate the balance between stochastic and deterministic processes in microbial succession. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E1326–E1332 (2015).
53. A. L. Demain, A. Fang, The natural functions of secondary metabolites. *Adv. Biochem. Eng. Biotechnol.* **69**, 1–39 (2000).
54. F. M. Cohan, Bacterial species and speciation. *Syst. Biol.* **50**, 513–524 (2001).
55. Y. Zilliges, J.-C. Kehr, S. Meissner, K. Ishida, S. Mikkat, M. Hagemann, A. Kaplan, T. Börner, E. Dittmann, The cyanobacterial hepatotoxin microcystin binds to proteins and increases the fitness of *Microcystis* under oxidative stress conditions. *PLOS ONE* **6**, e17615 (2011).
56. M. A. Berry, T. W. Davis, R. M. Cory, M. B. Duhaime, T. H. Johengen, G. W. Kling, J. A. Marino, P. A. Den Uyl, D. Gossiaux, G. J. Dick, V. J. Denef, Cyanobacterial harmful algal blooms are a biological disturbance to western Lake Erie bacterial communities. *Environ. Microbiol.* **19**, 1149–1162 (2017).
57. F. L. Hellweger, R. M. Martin, F. Eigemann, D. J. Smith, G. J. Dick, S. W. Wilhelm, Models predict planned phosphorus load reduction will make Lake Erie more toxic. *Science* **376**, 1001–1005 (2022).
58. C. E. Yancey, D. J. Smith, P. A. Den Uyl, O. G. Mohamed, F. Yu, S. A. Ruberg, J. D. Chaffin, K. D. Goodwin, A. Tripathi, D. H. Sherman, G. J. Dick, Metagenomic and metatranscriptomic insights into population diversity of *Microcystis* blooms: Spatial and temporal dynamics of mcy genotypes, including a partial operon that can be abundant and expressed. *Appl. Environ. Microbiol.* **88**, e0246421 (2022).
59. H. Cao, D. Xu, T. Zhang, Q. Ren, L. Xiang, C. Ning, Y. Zhang, R. Gao, Comprehensive and functional analyses reveal the genomic diversity and potential toxicity of *Microcystis*. *Harmful Algae* **113**, 102186 (2022).
60. H. Y. Cai, H. L. Jiang, L. R. Krumholz, Z. Yang, Bacterial community composition of size-fractioned aggregates within the phycosphere of cyanobacterial blooms in a eutrophic freshwater lake. *PLOS ONE* **9**, e102879 (2014).
61. C. Dziallas, H. P. Grossart, Temperature and biotic factors influence bacterial communities associated with the cyanobacterium *Microcystis* sp. *Environ. Microbiol.* **13**, 1632–1641 (2011).
62. H. W. Paerl, P. E. Kellar, Significance of bacterial-*Anabaena* (Cyanophyceae) associations with respect to N_2 fixation in freshwater. *J. Phycol.* **14**, 254–260 (1978).
63. H. Xu, D. Zhao, R. Huang, X. Cao, J. Zeng, Z. Yu, K. V. Hooker, K. D. Hambright, Q. Wu, Contrasting network features between free-living and particle-attached bacterial communities in Taihu Lake. *Microb. Ecol.* **76**, 303–313 (2018).
64. M. R. Olm, C. T. Brown, B. Brooks, J. F. Banfield, dRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
65. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
66. S. Nayfach, Z. J. Shi, R. Seshadri, K. S. Pollard, N. C. Kyrpides, New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).
67. S. W. Graham, R. G. Olmstead, S. C. H. Barrett, Rooting phylogenetic trees with distant outgroups: A case study from the commelinoid monocots. *Mol. Biol. Evol.* **19**, 1769–1781 (2002).
68. S. I. Na, Y. O. Kim, S. H. Yoon, S. M. Ha, I. Baek, J. Chun, UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction. *J. Microbiol.* **56**, 280–285 (2018).
69. J. O. McInerney, A. McNally, M. J. O'Connell, Why prokaryotes have pangenomes. *Nat. Microbiol.* **2**, 17040 (2017).
70. T. Seemann, Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
71. Z. M. Zhou, J. Charlesworth, M. Achtman, Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. *Genome Res.* **30**, 1667–1679 (2020).
72. K. Katoh, D. M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
73. P. Kück, G. C. Longo, FASconCAT-G: Extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Front. Zool.* **11**, 81 (2014).
74. A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
75. J. Felsenstein, Confidence-limits on phylogenies—An approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
76. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
77. K. Blin, S. Shaw, A. M. Kloosterman, Z. Charlop-Powers, G. P. van Wezel, M. H. Medema, T. Weber, antiSMASH 6.0: Improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).
78. A. F. Auch, M. von Jan, H. P. Klenk, M. Goker, Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand Genomic Sci.* **2**, 117–134 (2010).
79. J. P. Meier-Kolthoff, A. F. Auch, H. P. Klenk, M. Goker, Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **14**, 60 (2013).
80. R. Kolde, Pheatmap: Pretty heatmaps. R Package Version 1.0.12 (2019); https://CRAN.R-project.org/package=pheatmap..

Acknowledgments: We would like to thank the members of the University of Oklahoma Plankton Ecology and Limnology Laboratory for helpful discussions and statistical assistance and for commenting on earlier versions of this manuscript. In particular, we thank K. V. Cook for invaluable statistical and bioinformatic advice, helpful discussions, and critical editorial feedback. **Funding:** This work was supported by National Science Foundation (DEB 1831061) and U.S. Geological Survey (G21AP10181). **Author contributions:** Conceptualization: H.C. and K.D.H. Bioinformatics: H.C. and C.J.M. Interpretation: H.C., C.J.M., J.E.B., L.R.K., and K.D.H. Writing—original draft: H.C. Writing—review and editing: H.C., C.J.M., J.E.B., L.R.K., and K.D.H. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All genomic data used in this study were acquired from GenBank: https://ncbi.nlm.nih.gov/genome/. All accession numbers are provided in table S1.

Submitted 30 June 2022
Accepted 9 December 2022
Published 13 January 2023
10.1126/sciadv.add3783