Coherence of *Microcystis* species revealed through population genomics

Olga M. Pérez-Carrascal,[a#] Yves Terrat,[a] Alessandra Giani,[b] Nathalie Fortin,[c] Charles W. Greer,[c] Nicolas Tromas,[a#] B. Jesse Shapiro[a]

[a]Département de Sciences biologiques, Université de Montréal, Montréal, Québec, Canada

[b]Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[c]National Research Council of Canada, Montreal, Quebec, Canada

O.M.P.C. and Y.T. contributed equally to this work.

Address correspondence to Olga M. Pérez-Carrascal, olga.maria.perez-carrascal@umontreal.ca or Nicolas Tromas, nicolas.tromas@umontreal.ca

Abstract

Microcystis is a genus of freshwater cyanobacteria which causes harmful blooms in ecosystems worldwide. Some *Microcystis* strains produce harmful toxins such as microcystin, impacting drinking water quality. *Microcystis* colony morphology, rather than genetic similarity, is often used to classify *Microcystis* into morphospecies. However, colony morphology is a plastic trait which can change depending on environmental and laboratory culture conditions, and is thus an inadequate criterion for species delineation. Furthermore, *Microcystis* populations are thought to disperse globally and constitute a homogeneous gene pool. However, this assertion is based on relatively incomplete characterization of *Microcystis* genomic diversity. To better understand these issues, we performed a population genomic analysis of 33 newly sequenced genomes (of which 19 were resequenced to check for mutation in culture) mainly from Canada and Brazil. We identified eight *Microcystis* clusters of genomic similarity, only four of which correspond to named morphospecies and monophyletic groups. Notably, *M. aeruginosa* is paraphyletic, distributed across four genomic clusters, suggesting it is not a coherent species. Most monophyletic groups are specific to a unique geographic location, suggesting biogeographic structure over relatively short evolutionary time scales. Higher homologous recombination rates within than between clusters further suggest that monophyletic groups might adhere to a Biological Species-like concept, in which barriers to gene flow maintain species distinctness. However, certain genes – including some involved in microcystin and micropeptin biosynthesis – are recombined between monophyletic groups in the same geographic location, suggesting local adaptation. Together, our results show the importance of using genomic criteria for *Microcystis* species delimitation and suggest the existence of locally adapted lineages and genes.

Importance

The genus *Microcystis* is responsible for harmful and often toxic cyanobacterial blooms across the world, yet it is unclear how and if the genus should be divided into ecologically and genomically distinct species. To resolve the controversy and uncertainty surrounding *Microcystis* species, we performed a population genomic analysis of *Microcystis* genome from public databases, along with new isolates from Canada and Brazil. We inferred that significant genetic substructure exists within *Microcystis*, with several species being maintained by barriers to gene flow. Thus, *Microcystis* appears to be among a growing number of bacteria that adhere to a Biological Species-like Concept (BSC). Barriers to gene flow are permeable, however, and we find evidence for relatively frequent cross-species horizontal gene transfer (HGT) of genes that may be involved in local adaptation. Distinct clades of *Microcystis* (putative species) tend to have distinct profiles of toxin biosynthesis genes, and yet toxin genes are also subject to cross-species HGT and local adaptation. Our results thus pave the way for more informed classification, monitoring and understanding of harmful *Microcystis* blooms.

KEYWORDS: *Microcystis*, morphospecies, cyanotoxins, Horizontal Gene Transfer (HGT), homologous recombination, local adaptation, genomics, speciation, Biological Species Concept.

Introduction

Several categories of species concepts can be used as the basis to delimit and organize biological diversity, and two concepts in particular have been recently applied to bacteria and archaea. The Ecological Species Concept (ESC) proposes that speciation is driven by divergent natural selection between distinct ecological niches, while the Biological Species Concept (BSC) emphasizes gene flow (*e.g.* homologous recombination) as a cohesive force within species (1). The Stable Ecotype Model (SEM) is a version of the ESC tailored to bacteria, under the general assumption that adaptive mutations spread more rapidly by clonal expansion than by recombination (2). However, certain bacteria and archaea have relatively high rates of recombination, such that a BSC-like concept could apply - but strictly the BSC will never apply to bacteria that reproduce clonally and occasionally exchange genes across species boundaries. Several archaea and bacteria appear to fit a BSC-like model, showing higher recombination within species than between species (1, 3-5). Barriers to recombination could be maintained by natural selection or genetic incompatibilities, or due to physical separation (*i.e.* allopatry). Allopatric speciation is thought to be rare for globally dispersed bacteria, but does appear to occur among geographically separated hotspring archaea (6). Therefore, a plurality or spectrum of species concepts is probably necessary to fit the diverse lifestyles and recombination frequencies observed across microbes (1).

When two different species inhabit a common niche or geographic location, they may exchange genes beneficial to local adaptation. For example, *Vibrio cholerae* 'core' housekeeping genes are freely recombined among *V. cholerae* from both Bangladesh and the American coast, to the exclusion of the sister species *V. metecus* (7). However, in the integron (part of the genome

subject to particularly frequent recombination), *V. cholerae* undergoes more genetic exchange with *V. metecus* from the same geographic location (USA) than with *V. cholerae* from a different location (Bangladesh). This suggests that species cohesion is maintained across most of the core genome, while certain 'accessory' genes are exchanged across species boundaries to promote local adaptation. Identifying such locally adapted genes that cross species boundaries can provide insight into the genetic basis of adaptation to different environments (8, 9).

Here we consider the common bloom-forming cyanobacterium *Microcystis* as a model of speciation and local adaptation. *Microcystis* is a genus containing a great deal of genetic diversity (10) and capable of frequent recombination (11-13). Previous genetic and genomic studies have suggested that *Microcystis* is globally distributed, with little geographic structure (10, 11, 14, 15). Thus, it is plausible that *Microcystis* represents a single, globally distributed and homogeneous gene pool, adhering to a BSC-like model (1, 5, 11, 16). However, multiple attempts have been made to classify *Microcystis* into several species, based on various morphological and genetic criteria (17-19).

Microcystis is able to form colonies or cell aggregates covered by exopolysaccharide or mucilage (20). Several *Microcystis* strains are known to synthesize intracellular toxins, which are thought to be released to the environment primarily when cells die and lyse (21). Together, toxins and cell decomposition followed by oxygen depletion threaten the health of humans and animals. *Microcystis* colony morphology, cell size, and the structure of mucilage have been used for decades as taxonomic criteria to classify *Microcystis* in morphospecies or morpho-types (International Code of Botanical Nomenclature) (17, 22). However, morphospecies classifications are often inconsistent with genetic, genomic, and phylogenetic analyses (17, 23).

These inconsistencies may occur because *Microcystis* colonies can change morphology or become unicellular without any genotypic changes (17, 20, 24-26).

At least 51 morphospecies have been described within the *Microcystis* genus (http://www.algaebase.org/browse/taxonomy/?id=7066) (*e.g. M. aeruginosa*, *M. panniformis*, *M. viridis*, *M. wesenbergii*, *M. flos-aquae*, *M. novacekii*, and *M. ichthyoblabe*) (17, 27) One of the most studied and frequently reported is *M. aeruginosa*. Under laboratory conditions, it has been shown that colonies of *M. wesenbergii* morphospecies could become morphologically similar to colonies of *M. aeruginosa*, after just a few hours of culture (26). If a *Microcystis* strain undergoes colony morphology changes, it can then become over-classified into several morphospecies. Thus, the number of morphospecies may not reflect the number of *Microcystis* species based on other bacterial systematic approaches (17).

Because of these inconsistencies, several authors have attempted to reclassify *Microcystis* morphospecies using additional systematic approaches, like 16S rRNA gene sequence identity, DNA-DNA hybridization, phylogenetic analysis of conserved genes, and Average Nucleotide Identity (ANI) (10, 12, 14, 17, 19, 28, 29). For example, five *Microcystis* morphospecies (*M. aeruginosa*, *M. ichthyoblabe*, *M. novacekii*, *M. viridis*, and *M. wesenbergii*) were proposed to be reclassified as a single bacterial species (*M. aeruginosa*) (17). These five *Microcystis* strains showed 16S rRNA gene sequence identities higher than the usual cutoff value used to define bacterial species (>97%), DNA-DNA genome sequences hybridization values were also higher than the cutoff (>70%), and colony morphologies are generally similar (17, 26). Other studies showed that *M. aeruginosa* morphospecies together with other morphospecies are a single species complex with average nucleotide identity (ANI) values higher than 95%, which is

consistent with hybridization values higher than 70% – a standard rule in bacterial species delineation (10, 14, 30). Despite the high similarity in their core genomes, *Microcystis* are diverse in their gene content, resulting in large accessory genomes that can harbour genes related to the biosynthesis of harmful toxins or secondary metabolites (10, 14, 31). *Microcystis* classification based on accessory genes (*e.g*. toxins and polysaccharides) has also been proposed (14, 28).

In this study, we present a population genomics analysis using 33 newly sequenced genomes (of which 19 were resequenced after several years in laboratory culture) belonging to seven *Microcystis* morphospecies isolated mainly from Brazil and Canada over a 15-year period. We aimed to investigate the coherence of *Microcystis* morphospecies using phylogenomic and homologous recombination analyses. We identified four *Microcystis* clades corresponding to morphospecies, which are mainly restricted to particular geographical regions, with recombination rates higher within than between them, consistent with a BSC-like model. Meanwhile, *M. aeruginosa* morphospecies are paraphyletic and geographically unstructured, meaning that *M. aeruginosa* may in fact include multiples sub-species. In contrast with the general preference for recombination within clades, we also observed occasional HGT between clades. Many of these cross-species HGT events may be involved in local adaptation. Finally, we studied the profiles of genes related to the biosynthesis of secondary metabolites (such as microcystin) to determine if different *Microcystis* clades have a characteristic profile of biosynthetic genes.

Results

Phylogenetic coherence of named morphospecies

In order to assess the coherence of named *Microcystis* morphospecies, we sequenced 33 isolates of *Microcystis* (see Table S1 in the supplemental material). These genomes were initially classified into seven different morphospecies (*M. aeruginosa*, M. *flos-aquae*, *M. panniformis*, *M. wesenbergi*, *M. viridis*, *M*. sp. and *M. novacekii*). The size of the *Microcystis* assembled genomes ranged between 3.3 and 4.6 Mb with average GC content of 42.7%. The shared core genome consisted of 1,260 genes and the pangenome contained 16,928 genes. We supplemented these newly sequenced genomes with 26 additional *Microcystis* genomes downloaded from GenBank and compared them using three measures of genetic similarity: phylogeny, hierBAPS clustering and ANI (Average Nucleotide identity). Most of our genomes have pairwise ANI values > 95% with some exceptions: 199 out of 6241 pairwise comparisons have values between 93 and 94%, mostly involving comparison with Ma_AC_P_00000000_S299 (see Fig. S1, and Table S2 in the supplemental material). ANI values >95% are generally considered to include members of a single species. However, the *Microcystis* genomes do not constitute a single homogenous ANI cluster; rather, significant substructure is evident (Fig. S1).

To explore this substructure, we built a core genome phylogeny using 152 conserved genes also present in an outgroup (Fig. 1) and also clustered the aligned core genomes using hierBAPS (32). Of the 33 newly sequenced genomes, 19 were resequenced after several years in culture. The resequenced genomes differed from their ancestor by an average of 48 point mutations, and always clustered with their ancestor in the phylogenetic tree, suggesting that evolution in the laboratory had little impact on the structure of the phylogeny (Table S3). Additionally, we

178 isolated and sequenced a single colony from one of the batch cultures. This colony genome
179 (S217Col) clustered on the phylogeny with its parent culture (S217Cul) with a phylogenetic
180 distance of zero (Fig. 1), suggesting that a single colony is representative of the entire culture.
181
182 The hierBAPS analysis yielded eight groups (*Mpa*, *Mfl*, *Mvi*, *Mwe*, *Mae*1, *Mae*2, *Mae*3 and
183 *Mae*4) which are indicated alongside the phylogenetic tree (Fig. 1). In general, hierBAPS
184 clusters were consistent with the phylogeny, with some exceptions likely due to the susceptibility
185 of hierBAPS to long-branch attraction (32, 33). Four out of eight hierBAPS clusters
186 corresponded to monophyletic groups in the phylogeny and named morphospecies: *M.*
187 *panniformis* (*Mpa*), *M. wesenbergii* (*Mwe*), *M. viridis* (*Mvi*) and *M. flos-aquae* (*Mfl*), while the
188 other clusters contained multiple named morphospecies (Fig. 1). The *M. aeruginosa*
189 morphospecies was paraphyletic and distributed across four hierBAPS clusters (*Mae*1, *Mae*2,
190 *Mae*3 and *Mae*4), two of which (*Mae1* and *Mae2*) were monophyletic clades. Based on these
191 phylogenetic and population structure analyses, *Microcystis* appears to comprise six well-defined
192 monophyletic groups (*Mpa*, *Mfl*, *Mvi*, *Mwe*, *Mae*1, *Mae*2), four of which are congruent with the
193 morphospecies (*Mpa*, *Mfl*, *Mvi* and *Mwe*). We noted that these last four monophyletic
194 morphospecies tended to be fairly specific for a particular geographic location (either Canada or
195 Brazil), suggesting possible local adaptation, local and recent clonal expansion, or reduced
196 migration of these lineages. Even if *Microcystis* are generally closely related at >95% ANI, there
197 is clear and significant substructure within the genus.
198
199 We also investigated the pangenome content within each of eight hierBAPS clusters. The four
200 clusters corresponding to *M. panniformis*, *M. wesenbergii*, *M. viridis* and *M. flos-aquae* had
201 highly conserved core genomes (between 75% and 96% of genes shared by all members of the

cluster) while *M. aeruginosa* morphospecies had much smaller core genomes (between 53% and 64% of genes shared by all members; see Table S4 in the supplemental material). This is consistent with the four monophyletic morphospecies representing coherent clusters of genomic similarity, with the paraphyletic *M. aeruginosa* being an amalgam of high genetic diversity and variable gene content.

Higher homologous recombination rates within than between clusters supports a BSC-like concept

We next asked if homologous recombination could explain the cohesion of the monophyletic groups. To address this question, we estimated the ratio of homologous recombination to mutation rates (*r/m*) within and between eight hierBAPS groups. We found that the four monophyletic morphospecies (*M. panniformis*, *M. wesenbergii*, *M. novacekii* and *M. flos-aquae*) all have *r/m* ratios 2-3X higher within than between clades (Fig. 2). Recombination rates were generally low for *M. aeruginosa* both within and between clades (Fig. 2). When the resequenced genomes (from a second time point of the same culture) were excluded, the *r/m* tended to increase within clades (see Table S5 in the supplemental material), presumably because mutation but not recombination took place in the culture. Overall, these results suggest that the cohesion of four monophyletic groups could be driven or reinforced by preferential recombination within versus between groups, consistent with a BSC-like model of speciation. Conversely, the other four groups, consisting of *M. aeruginosa*, appeared to engage in relatively little recombination and thus defied delineation based on the BSC.

Frequent local horizontal gene transfer (HGT)

From the core genome phylogenetic tree, certain monophyletic clades showed strong geographic preferences. For example, *M. flos-aquae* was found almost uniquely in Canada, while *M. panniformis* was found only in Brazil (Fig. 1). In contrast, *M. wesenbergii* is found in both Canada and Brazil (Fig. 1), suggesting that morphospecies can be stable in ways that transcend geography boundaries. As previously observed in *Vibrio*, different species in the same geographic region may exchange genes, possibly leading to local adaptation (7). To identify potential locally adapted *Microcystis* genes, we screened gene trees for instances where two different named morphospecies (which formed distinct monophyletic groups in the species tree; Fig. 1) clustered together in the same monophyletic group (with bootstrap support >90%), consistent with cross-species HGT. As we were particularly interested in local HGT, we identified monophyletic groups of two distinct species, all isolated from the same region (*i.e.* Canada or Brazil, but not both). We screened a total of 25,157 core and accessory genes from 79 *Microcystis* genomes (53 reported here, including colonies and resequenced isolates, and 26 previously published). We considered 12,084 informative gene trees (that included 4 or more leafs). Of these trees, 590 (4.9% of the total) showed a pattern of non-local HGT (with Canadian and Brazilian isolates grouping together in the same well-supported clade), whereas slightly more (923 genes; 7.6% of the total) were consistent with local HGT. This suggests that geography, and possibly local adaptation, is an important factor in shaping rates of HGT. Local HGT events, on average, appear to be more recent than non-local events: in 80 out of 923 local HGTs, the phylogenetic distances within the recombined clade were equal to zero (Table 1), suggesting relatively recent HGT in local compared to non-local events (Fisher's exact test, Odds ratio = 1.90, P = 0.004). Local HGTs are also much more likely to be functionally annotated, compared to non-local HGTs which involve mostly hypothetical genes (Table 1; Fisher's exact test, Odds ratio = 2.99, P < 2.2e-16). While these differences could have many possible

explanations, we speculate that non-local HGT events are enriched in phages and other poorly annotated mobile or selfish genetic elements, while local HGTs involve metabolically or ecologically relevant genes, which are more likely to have been studied and annotated. Consistent with this explanation, the non-local pangenome is dominated by genes involved in DNA replication, recombination and repair (COG category L, Fig. S2; X^2 test, $P < 0.05$ after Bonferroni correction for multiple hypothesis testing), which is suggestive of self-replicating and recombining mobile elements (Fig. S2 and Data Set S1). Overall, these results suggest that local HGT events are relatively recent (and thus more frequently observed) and possibly more ecologically relevant (and less "selfish") than non-local HGTs.

Figure 3 illustrates a few noteworthy examples of local HGT events. For example, the phylogenetic trees of two neighboring genes encoding the *hicA-hicB* toxin-antitoxin system showed phylogenetic distances almost equal to zero (between 0 and 0.0008 substitutions per site), clustering Brazilian genomes of three different hierBAPS clusters into a single group (Fig. 3A and B), whereas these clusters are well-separated on the species tree (Fig. 3E). This suggests that the toxin-antitoxin system has been subjected to recent cross-species HGT in Brazil. The *hicAB* module is a mobile element that has been previously described in bacteria, archaea, plasmids and phages (34, 35) and at least 31 *hicB* antitoxins and 21 *hicA* toxins have been reported in *M. aeruginosa* (36). The *hicAB* module seems to act as a phage defense system, arresting cell growth in response to phage infection (36).

Another two genes encoding CRISPR *cas1* endonucleases also showed a signature of recent local HGT, with phylogenetic distances almost equal to zero (between 0 and 0.0001), clustering Canadian genomes in a single clade (Fig. 3C). These two genes are neighbors located on the

same contig, flanked by a hypothetical gene and a CRISPR-associated endoribonuclease (*cas2*). We also identified local HGT events involving other toxin-antitoxin genes, cyanotoxins (such as *mcyB*; Fig. 3D), endonucleases, and others (Data Set S1).

Clade-specific profiles of biosynthetic gene clusters

Having shown examples of cyanotoxin genes being involved in local HGT (Fig. 3D; Data Set S1), we sought to more broadly characterize the distribution of cyanotoxins and other biosynthetic gene clusters across *Microcystis* clades. Specifically, we asked whether *Microcystis* clades or morphospecies tended to have a characteristic profile of biosynthetic genes, despite potentially rapid gain and loss of these genes. The biosynthesis genes of secondary metabolites are usually found in gene clusters (37, 38). *Microcystis* species can synthesize a variable number of secondary metabolites, many of which are toxic to humans and other animals (39, 40).

We identified 34 known secondary metabolite gene clusters within all the *Microcystis* genomes using the software AntiSMASH (Fig. 4; Table S6). AntiSMASH identifies these genes based on a protein database and NRPS (Nonribosomal Peptide Synthetases) and PKS (Polyketide synthase) domain analysis (41). Eight out of 34 secondary metabolic gene clusters were present and complete in at least one *Microcystis* genome (Fig. 4).

We observed that *Microcystis* genomes lacking the microcystin cluster (*mcy*) usually contained another gene cluster instead. For example, *M. flos-aquae* lacked *mcy* but instead had genes related to the biosynthesis of anabaenopeptins (*apn*) (Fig. 4). However, other studies have found microcystin-producing strains of *M. flos-aquae* (42), suggesting that the genomes reported here likely undersample the diversity of biosynthetic gene clusters present in nature. Consistent with

297 relatively high diversity within clades, genomes within the same clade tended to have similar, but
298 non-identical profiles of gene clusters, with *M. aeruginosa* clades being among the most diverse.
299 *M. aeruginosa* clades also tended to have a high coding potential for toxins, including the
300 complete *mcy* and *mdn* (microviridin B) gene clusters. However, some genomes in *M.*
301 *aeruginosa* sub-clade *Mae1* lacked the *mcy* genes, consistent with loss or HGT. In contrast, *M.*
302 *wesenbergii* encoded relatively few biosynthetic gene clusters, consistent with previously
303 reported low toxin production and microcystin gene absence (42-45).
304
305 Viewed in aggregate, these biosynthetic gene clusters are part of the *Microcystis* accessory
306 genome. However, certain gene clusters are core to specific monophyletic groups. In *M. flos-*
307 *aquae* for example, *anp* genes were always present (core) and *mcy* genes were absent. These
308 group-specific core gene clusters could provide potential niche adaptations and ecological
309 distinctness. On the other hand, certain biosynthetic genes such as *mcyB* (Fig. 3D) are exchanged
310 across species boundaries. Thus, biosynthetic genes may contribute to both species-specific and
311 species-transcending adaptations (46).
312
313 **Discussion**
314 In this study, we investigated the correspondence among morphospecies and genome-informed
315 species definitions using dozens of *Microcystis* isolates from both Northern and Southern
316 hemispheres, primarily from Canada and Brazil. We assessed the genomic cohesion of
317 *Microcystis* clades within *Microcystis* species by measuring the genome similarities (phylogeny,
318 hierBAPS clustering and ANI values) and homologous recombination within and between
319 clades.
320

We found that *Microcystis* genomes used in this study together with the reference genomes fell into a single genomic complex (ANI values higher than 95%). Previous studies have suggested a universal cutoff of 95% ANI as adequate for species delineation. These studies described a genetic discontinuity or bimodal distribution with peaks higher than 95% (intra-species) and lower than 83% (inter-species), but the mechanism for this discontinuity is unclear, and it is difficult to exclude sampling bias as a reason for the discontinuity (47, 48). We observed that within a 95% ANI cluster of *Microcystis*, there is substantial genetic substructure, potentially containing distinct species or sub-species. Four of the eight sub-clusters we identified corresponded to named morphospecies, while the others were mostly composed of genomes classified as *M. aeruginosa* morphospecies. We concluded that *M. aeruginosa* is paraphyletic with a mixed geographical pattern, while the morphospecies *M. flos-aquae*, *M. panniformis*, *M. wesenbergii*, *M. viridis* consisted of well-defined clades within *Microcystis* species complex (Fig. 1).

What are the mechanisms that can explain the genetic structure within *Microcystis*? Ecological selection (the ESC), barriers to gene flow (BSC-like), and biogeography (allopatric divergence) could all play a role. Previous studies based on a smaller sample of *Microcystis* genomes (10), or marker genes (14, 15) suggested that there are few if any biogeographic barriers in *Microcystis* to migrate, leading to a globally mixed population (49, 50) – and that *Microcystis* should be defined as a single species (17, 18). Consistent with this, *M. aeruginosa* is globally distributed and paraphyletic in our core phylogeny. However, the four monophyletic morphospecies tend to be geographically restricted, possibly due in this case to limited migration (at least on recent time scales) and/or local adaptation. The four morphospecies could also represent short-lived clonal expansions, or biases due to incomplete sampling. However, to minimize bias, our sampling was

345 performed repeatedly over 15 years, with similar methods in both Brazil and Canada. We also
346 inferred more frequent homologous recombination within than between morphospecies (Fig. 2),
347 consistent with a BSC-like model maintaining genetic distinctness (51-53). While morphospecies
348 coherence may thus be maintained by barriers to gene flow, we suspect that morphospecies
349 divergence was initiated by selection for ecological distinctness (1). Although the precise
350 ecological differences between morphospecies are unknown (45, 54), we found that each of the
351 four monophyletic morphospecies had a distinct core genome and distinct profile of biosynthetic
352 gene clusters (Fig. 4) – both of which could provide potential ecological adaptations. Further
353 experimental study will be required to fully test the hypothesis of ecological distinctness among
354 morphospecies.
355
356 The BSC-like model requires more frequent recombination within than between species, but also
357 allows occasional recombination of "globally adaptive" genes across species boundaries. Similar
358 to previous observations in *Vibrio* (7), we inferred a significant proportion of cross-species HGT
359 events occurred within the same geographic location, suggesting local environmental adaptation.
360 Local HGT events tend to be phylogenetically more recent than non-local events, suggesting that
361 they occur at relatively higher frequency. Local HGTs are also enriched in genes of annotated
362 (non-hypothetical) function, including cyanotoxin genes such as microcystin (*mcyB*) and
363 cyanopeptolin (*mcnC*, *mcnF* and *mcnG*). Microcystin genes are likely of ancient origin in
364 cyanobacteria (55) and it has been suggested that they subsequently experienced significant
365 homologous recombination and positive selection (56, 57). Microcystin genes also show
366 biogeographic patterns. For example, the *mcyD* gene has distinctive alleles found in Japanese
367 *Microcystis* isolates but not elsewhere (58). Our inference of local, cross-species HGT of toxin
368 genes further supports the idea that they may be locally adapted.

Genes involved in phage defense systems were also involved in local HGT events, suggesting that local adaptation could be driven by local viruses. First, the *hicAB* operon appears to have been shared among at least three distinct hierBAPS clusters (*Mvi*, *Mpa*, and *Mae4*) in Brazil (Fig. 3), and has previously been suggested to be involved in phage defense and prone to HGT (34, 36). Second, *cas1*, which encodes the most conserved protein in the CRISPR–Cas defense system (59), appears to have been exchanged among three distinct clusters of *M. aeruginosa* in Canada (Fig. 3). This is consistent with previous evidence suggesting that CRISPR–Cas genes are subject to HGT and natural selection (59, 60). Thus, local HGT could promote adaptation to local phages.

Taken together, our results resolve some of the longstanding confusion surrounding *Microcystis* species and suggest new avenues for future research. While all *Microcystis* genomes sampled to date are monophyletic and closely related, there is significant genetic substructure suggesting the existence of several distinct species. The distinctiveness of these species appears to be maintained by barriers to gene flow, consistent with a BSC-like model (1, 5). Whether gene flow barriers are mainly geographic, genetic, or ecological is a subject for future investigation. While different *Microcystis* species appear to inhabit different niches, as evidence by geographic preferences and distinct profile of biosynthetic gene clusters, the nature of their ecological distinctiveness should also be a subject of future field and laboratory studies.

Materials and methods

Genome sequencing, assembly and binning

Over the past 15 years, we collected 30 *Microcystis* isolates from Brazil, Canada and the United States. The *Microcystis* strains were initially characterized as morphospecies based on their colony morphology, according to Komárek (27, 61, 62). Seven morphospecies were identified: *M. aeruginosa*, M. *flos-aquae*, *M. panniformis*, *M. wesenbergi*, *M. viridis*, *M.* sp., and *M. novacekii*. We performed the DNA extraction for these strains between 2006 and 2017 (see Table S1 in the supplemental material). The 30 *Microcystis* genomes were sequenced using the Illumina HiSeq 2500 platform with 125bp paired-end reads. The genomic Illumina libraries (with average fragment size 360bp) were prepared using the NEB (New England Biolabs®) low input protocol. We also sequenced the DNA of four single *Microcystis* colonies isolated manually under the microscope. Three of these colonies were new isolates recovered from Lake Champlain (Quebec, Canada) in 2017 without culture, while the fourth came from a culture that was also sequenced (in bulk) in this study, for a total of 33 new sequenced *Microcystis* genomes. Before DNA extraction and sequencing, each colony was washed 10-15 times with Z8 medium using a micropipette. DNA extraction was performed directly on each colony using the ChargeSwitch® gDNA Mini Bacteria Kit.

Of the 30 *Microcystis* isolates, 19 had been maintained in culture for several years until 2017. Thus, we extracted DNA and re-sequenced from these 19 cultures in 2011, 2016 and 2017 to check for contamination and mutations in the *Microcystis* genome over time and differences between culture and colony sequences (Table S1). Together, the *Microcystis* genomes, of which

14 were sequenced once, 18 twice, and one three times comprised a total of 53 genome sequences.

Sequences from cultures and colonies were assembled with the software IDBA-UD v1.1.3 (63), producing contigs belonging to both *Microcystis* and associated heterotrophic bacteria, which are naturally associated with *Microcystis* (64, 65). The software Anvi'o v3.0 was used to filter, cluster and select the contigs belonging to *Microcystis* (66). The associated bacterial genomes will be described in a forthcoming manuscript. To cluster contigs into genomes, we used tetra-nucleotide frequency, GC content, and taxonomic annotation based on the Centrifuge software, as implemented in Anvi'o (66, 67). The gene prediction and annotation were done using Prodigal v2.6.3 and Prokka v1.12 packages, respectively (68-70). The raw reads and the *Microcystis* genomes contigs are available in GenBank under Bioproject number PRJNA507251.

Phylogenomic analysis

A core genome of 152 single copy genes shared by 79 *Microcystis* genomes (53 and 26 genomes reported here and previously, respectively) and two outgroups (*Anabaena variabilis* ATCC 29413 and *Synechocystis* sp. PCC6803) was identified using the software Roary and blastn-all. First, a core genome for the 79 *Microcystis* genomes (minimum value of 90% amino acid identity) was identified using Roary. The outgroups were initially excluded due to their high divergence from *Microcystis*. To identify the homologous genes in the outgroups, blast-all was used (Blastp similarities higher than 60%). The common core genes between Roary and blast-all were selected and used to create a core gene alignment. Each homologous gene was aligned separately using muscle (71). The concatenated and degapped alignment of length 129,835 bp was used for building a phylogenetic tree in RAxML v8.2.4, using the GTRGAMMA model,

437 with 100 bootstraps (72). Using the same method, another core phylogenetic tree was inferred
438 without the previously published reference genomes. This concatenated core-alignment
439 comprised 222 genes (211,589 bp degapped) (see Fig. S3 in the supplemental material).
440
441 **Clustering analysis of *Microcystis* genomes**
442 A multiple genome alignment of the 79 genomes (53 and 26 genomes reported here and the
443 references, respectively) was performed using Mugsy 1.2.3 (73) (see Table S1 in the
444 supplemental material). The DNA core-alignment was extracted from the Mugsy output using a
445 python script (625,795 bp long). The core-alignment was used to perform genetic population
446 structure and cluster analysis with hierBAPS (The hierarchical Bayesian Analysis of Population
447 Structure) (32). We used as input parameters two clustering levels and an expected number of
448 cluster (k) equals to 10, 20 and 40. HierBAPS delineates the population using nested clustering.
449 In this method, rare genotypes (distantly related to better sampled clades) often cluster together
450 due to long-branch attraction (32). As a result, HierBAPS clusters could be incongruent with the
451 phylogeny inferred by maximum likelihood, which is less sensitive to long-branch attraction.
452
453 **Pairwise nucleotide identity between *Microcystis* genomes**
454 The 33 newly sequenced genomes and 26 previously published were compared using a python
455 module (average_nucleotide_identity.py) to estimate the Average Nucleotide Identity by
456 Mummer and by Blast (ANIm and ANIb values between genome pairs) (see Fig. S1 in the
457 supplemental material) (https://github.com/widdowquinn/pyani) (74). Bacterial genomes with
458 DNA–DNA hybridization (DDH) of at least 70%, are considered as the same species and usually
459 show values of ANI >95%. Hence, a cutoff higher than 95 for the ANI values between genome
460 pairs is used to identify genomes within the same genomic cluster or species (30, 47, 75). The

461 pairwise identities were plotted using the R package ggplot (http://ggplot2.org/) and the function
462 heatmap2 (76).
463
464 **Pangenome analysis in *Microcystis* genomic clusters**
465 Pangenomes were inferred for each *Microcystis* genomic cluster with more than 2 genomes (see
466 Table S2 in the supplemental material). A global pangenome estimation was generated using all
467 the *Microcystis* genomes excluding the 2 shorter genomes Ma_AC_P_00000000_S299 and
468 Ma_QC_C_20070823_S18). Those genomes were excluded of the pangenome analyses because
469 of their reduced size compared to the average size (10% and 30% for
470 Ma_AC_P_00000000_S299 and Ma_QC_C_20070823_S18, respectively).
471 Ma_QC_C_20070823_S18 showed the lowest coverage (28X) and Ma_AC_P_00000000_S299
472 appeared to be contaminated with another cyanobacterium (*Anabaena*).
473 The software Roary v3.12.0 was used to generate the pangenomes. Prior to the execution of
474 Roary the genomes were automatically annotated using Prokka v1.12 (70). The genomes in
475 GGF3 format generated with Prokka were used as input to Roary. Roary was executed using a
476 minimum percentage of amino acid identity of 90% for blastp, which was set up according to the
477 similarities in the genomes higher than 94% and Roary recommendations. MultiFasta alignment
478 of the core genes were created using PRANK v150803 implemented in Roary (77).
479 We also did a pangenome analysis to find homologous and accessory genes within 53
480 *Microcystis* genomes from Brazil, Canada, USA (including the shorter ones), and the 26
481 reference genomes from the NCBI database. Roary allowed us to identify 370 clusters of
482 homologues (shared in 79 genomes) and 1059 (shared in 76 up to 78 genomes). Roary also
483 identified 23,728 accessory genes or genes shared by less than 76 genomes.
484

SNPs and deletion identification between duplicates genomes

The calling of SNP variants (Single Nucleotide Polymorphism) was done using snippy v3.2 (https://github.com/tseemann/snippy) with default parameters. SNPs between duplicate genomes were identified using one of genomes as reference.

Homologous recombination rates across *Microcystis* genomic clusters

Using the 53 newly sequenced *Microcystis* genomes, we investigated rates of homologous recombination within and between genomic clusters or across the phylogenetic tree. To do this, we estimated the relative effect of recombination versus mutation (*r/m*) rates using ClonalFrameML v1.11-3 (78). Briefly, the degapped core genome alignment generated by Mugsy v2.2.1 (1,274,628 bp) (73) was split in several subalignments using pyfasta (https://pypi.org/project/pyfasta/). The subalignments corresponded to the genomic clusters defined using phylogenomic and the population structure analyses. In order to estimate *r/m* between clusters, we also created subalignments for pair of clusters. ClonalFrameML was executed using a bootstrap of 100 replicates (emsim=100). The input phylogenies given in ClonalFrameML were generated using RAXML with bootstrap of 100 replicates (72). The transition/transversion ratios also used as an inputs in ClonalFrameML were estimated using PHYML v3.0 under the model of nucleotides substitution HKY85 (79). ClonalFrameML analyses excluded the two smallest *Microcystis* genomes Ma_AC_P_00000000_S299 and Ma_QC_C_20070823_S18. ClonalFrameML analyses within clusters were also conducted after removing the replicate (resequenced) genomes.

Identification of horizontally transferred and locally adapted genes

To identify genes transferred across species boundaries, we screened gene trees for instances of two distinct species (hierBAPS clusters in the species tree) clustering together in the same monophyletic group, whereas they are normally distantly related in the species tree. As a signature of local adaptation, we additionally screened for such cross-species HGT events that occurred among two different species from the same country.

For local adaptation analysis, we worked with the clusters generated with Roary and using 53 *Microcystis* genomes from Brazil, Canada, USA, and the 26 reference genomes from the NCBI database. Once the gene clusters were identified with Roary, the alignments of the nucleotide sequences in each gene cluster (core genes present in >75 strains (>95%) and accessory genes) were generated using the MAFFT software v7.271 (80). Maximum likelihood phylogenetic trees for each alignment were inferred using FastTreeMP v2.1.8 and the generalized time-reversible model (GTR) for nucleotide substitution (81). The trees were visualized with graphlan v0.9.7 (82).

The phylogenetic trees that showed local (geographic) adaptation signatures were identified using a Perl script (https://figshare.com/articles/Monophy_screening_tree_files_for_the_detection_of_local_adaptations/7661009) to screen phylogenetic trees and identify monophyletic groups with a particular level of bootstrap support (in our case, 90%). The script also allows monophyletic groups including particular combination of isolates (*e.g.* from different morphospecies or geographic locations) to be identified, with a given minimal branch length (in our case, this parameter was set to 0) and number of isolates (in our case, 4) within the group. A phylogeny was considered as positive for non-local HGT if Canadian and Brazilian isolates were together in the same clade supported by a bootstrap value > 90%, while a phylogeny positive for the local-HGT showed Brazilian or Canadian isolates, but not both in the same well-supported clade. The phylogenies

532 with a signature of HGT were then manually curated to remove those consisting solely of HGT
533 within a single hierBAPS cluster. Genes in the accessory genome and core genome were
534 functionally annotated using the eggNOG database (83, 84). The full HGT gene set is reported in
535 Data Set S1.
536
537 **Inferring secondary metabolic pathways in *Microcystis* genomes**
538 We evaluated the metabolite profiles for individual *Microcystis* genomes using the package
539 antiSMASH v4.0.2. The annotated genomes using Prokka were used as input to AntiSMASH
540 (41). Two additional biosynthetic clusters absent in the AntiSMASH database (Aeruginosamide
541 (NCBI accessions numbers CCH92964- CCH92969) and Microginin (NCBI accessions numbers
542 CAQ48259-CAQ48262)) were added manually. Based on the antiSMASH results, we generated
543 a matrix of presence-absence of genes related to the biosynthesis of secondary metabolites. The
544 matrix was visualized using the R package ggplot and the function heatmap2 (76). All-against-all
545 BLASTP analysis was applied to find the best reciprocal hits between proteins in the database
546 and the proteins in the *Microcystis* genomes (85). The proteins with the best reciprocal hit were
547 extracted; to be considered as present in the database, the amino acid identity had to be > 60%,
548 and > 30 % of the length of the sequences had to be aligned, with an e-value of 10^{-5}.
549
550

551 **Figure legends**

552

553 **Figure 1. Phylogenetic tree of 53 Brazilian, Canadian and USA *Microcystis* genomes and 26**
554 **globally sampled reference genomes.** A core genome of 152 homologous genes shared by 79
555 *Microcystis* genomes and the outgroups (*Anabaena variabilis* ATCC 29413 and *Synechocystis*
556 sp. PCC6803) was identified using software Roary (77). The eight hierBAPS clusters are
557 highlighted in colored boxes. The genomes from the same isolate at a different time have letter
558 "D or T" at the end of their names and are indicated with a black asterisk. The genomes from a
559 bulk culture and a single colony from the same cultured are indicated with a hash. The three
560 genomes from uncultured colonies from Lake Champlain (Quebec, Canada) are indicated with a
561 red asterisk and "Col" at the end of their name. The font colors indicate the geographical origin
562 (Brazil: orange, Canada: blue, USA: purple, other: black). The abbreviated hierBAPS cluster
563 names *Mpn, Mwe, Mfl, Mvi* and *Mae* correspond to *M. panniformis*, *M. wesenbergii M. flos-*
564 *aquae*, *M. viridis* and *M. aeruginosa*, respectively. The tree bar scale indicates number of
565 nucleotide substitutions per site.

566

567 **Figure 2. Relative contribution of recombination/mutation (*r/m*) within (diagonal) and**
568 **among eight *Microcystis* hierBAPS clades**. The *r/m* estimation included the replicates genomes
569 (from a second time point of the same culture). The hierBAPS clusters are represented with the
570 same symbols and abbreviations used in Fig. 1. See Table S5 for the *r/m* values with and without
571 replicate genomes.

572

573 **Figure 3. Phylogenetic trees of accessory genes with an inferred history of local HGT.** a)
574 *hicA*, b) *hicB*, c) two Cas1 genes, d) *mcyB*, and e) the core genome phylogeny. The symbols

represent their corresponding hierBAPS clades. The font colors of the names indicate the geographical origin. The clades showing geographical signatures (local HGT) are highlighted in orange (Brazil) and blue (Canada). The bars below the trees indicate units of nucleotide substitutions per site.

Figure 4. Distribution of biosynthetic gene clusters across *Microcystis*. a. Phylogenetic tree of 53 Brazilian, Canadian and USA *Microcystis* genomes and 26 reference genomes. **b**. Presence and absence of the genes encoding secondary metabolites in each *Microcystis* genome shown as a heatmap. Rows and columns represent the genomes and genes, respectively. The presence and absence of genes are indicated in blue and white, respectively. The shade of blue increases with the amino acid similarity to the reference database. The ten biosynthetic clusters are enclosed by colored rectangles and their names appear at the bottom of the figure.

Figure S1. Heatmaps for the hierarchical clustering of 79 *Microcystis* genomes based on ANIm (A) and ANIb (B) (53 genomes sequenced in this work, plus 26 NCBI genomes published before). Rows and columns represent the genomes. The scale bar indicates the pairwise ANI values. The genomes from the same isolate at a different time point have a letter "D or T" at the end of their names and are indicated with black asterisks. The genomes from colonies and culture from the same isolate are indicated with a hash. The genomes names from colony end with the word Col and are indicated with a red asterisk.

Figure S2. Functional enrichment analysis in core, accessory, local and non-local HGT gene sets. Functional annotations were from eggNOG. a) Hierarchical clustering analysis based in the relative abundance for each functional category. b) Functional categories that showed

599 statistically significant ($P < 0.05$) differences in one or more comparison between groups (core,
600 accessory genes, local, or non-local HGT gene sets, defined as described in Methods). The chi-
601 square (X^2) test was conducted using the package STAMP (86). Raw p-values are indicated in
602 black, and after Bonferroni correction in blue. Asterisks denote significant ($p < 0.05$) differences
603 after correction.
604
605 **Figure S3. Phylogenetic tree of 53 Brazilian, Canadian and USA *Microcystis* genomes**. A
606 core genome of 222 homologous genes shared by 53 *Microcystis* genomes and the outgroups
607 (*Anabaena variabilis* ATCC 29413 and *Synechocystis sp*. PCC6803) was identified using the
608 software Roary (77). The concatenated core-alignment (222 genes - 211589 bp long) was used
609 for building a phylogenetic tree in the RAxML program, using the GTRGAMMA model, with
610 100 bootstrap replicates (72). The genomes from the same isolate at a different time have letter
611 "D or T" at the end of their names and are indicated with a black asterisk. The genomes from
612 colony and culture from the same isolate are indicated with a hash. The genomes names from
613 colonies end with the word Col and are indicated with a red asterisk.
614

615 **Supplementary table legends**

616

617 **Table S1. Genome characteristics of 33 *Microcystis* isolates from Brazil, Canada and USA.**
618 The genomes from the same isolate at a different time point are indicated with black asterisks.
619 The genomes from colonies and culture from the same isolate are indicated with hashes. The
620 genomes names from colonies end with "Col" are indicated with red asterisks.

621

622 **Table S2. ANIb and ANIm values corresponding to pairwise comparison between genomes.**

623

624 **Table S3**. **SNP variants between 19 duplicates (resequenced) genomes.** The duplicate
625 (resequenced) genomes are indicated with “D” at the end of their name.

626

627 **Table S4. Core genome composition for each hierBAPS clusters and all the genomes.** The
628 core genomes were estimated for each individual hierBAPS cluster using the software Roary.
629 The number of genomes per pan-genome and the ANIm values within clusters are indicated.

630

631 **Table S5. The relative effect of recombination versus mutation (*r/m*) within and between**
632 **hierBAPS clusters.** The *r/m* estimations after removing duplicate genomes are indicated within
633 parentheses and with *.

634

635 **Table S6.** Biosynthetic pathways.

636

637 **Data Set S1. Genes showing evidence of local and non-local HGT.** The list of genes with local
638 and non-local annotation, phylogenetic distances to the node, and functional annotation with
639 Prokka and eggNOG are indicated.
640
641 **Acknowledgments.** We are grateful to David Bird for assistance isolating strains and
642 maintaining cultures. This work was supported by the Genome Québec and Genome Canada-
643 funded ATRAPP Project (Algal blooms, Treatment, Risk Assessment, Prediction and
644 Prevention). **NT** was funded by a project from the European Union's Horizon 2020 research and
645 innovation program under the Marie Sklodowska-Curie grant agreement No 656647. Cultures
646 collections (Brazilian and some Canadian strains) were partially obtained and maintained thanks
647 to CNPq and FAPEMIG grants to **AG.** We also want to acknowledge the financial support of the
648 National Research Council. We declare that we have no financial conflicts of interest.
649

650 **References**

651

652 1. Shapiro BJ. 2018. What microbial population genomics has taught us about speciation.
653 In: Polz M., Rajora O. (eds) Population genomics: Microorganisms. Population
654 genomics. Springer, Cham. https://doiorg/101007/13836_2018_10.

655 2. Wiedenbeck J, Cohan FM. 2011. Origins of bacterial diversity through horizontal genetic
656 transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35:957-976.

657 3. Whitaker RJ, Grogan DW, Taylor JW. 2005. Recombination shapes the natural
658 population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. Mol Biol
659 Evol 22:2354-61.

660 4. Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, Krause DJ,
661 Whitaker RJ. 2012. Patterns of gene flow define species of thermophilic Archaea. PLoS
662 Biol 10:e1001265.

663 5. Bobay LM, Ochman H. 2017. Biological species are universal across Life's domains.
664 Genome Biol Evol doi:10.1093/gbe/evx026.

665 6. Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ. 2009. Biogeography of the
666 *Sulfolobus islandicus* pan-genome. Proc Natl Acad Sci U S A 106:8605-10.

667 7. Boucher Y, Cordero OX, Takemura A, Hunt DE, Schliep K, Bapteste E, Lopez P, Tarr
668 CL, Polz MF. 2011. Local mobile gene pools rapidly cross species boundaries to create
669 endemicity within global *Vibrio cholerae* populations. MBio 2:e00335–10.

670 8. van Elsas JD, Bailey MJ. 2002. The ecology of transfer of mobile genetic elements.
671 FEMS Microbiol Ecol 42:187-97.

672 9. Whittaker KA, Rynearson TA. 2017. Evidence for environmental and ecological
673 selection in a microbe with no geographic limits to gene flow. Proc Natl Acad Sci U S A
674 114:2651-2656.
675 10. Humbert JF, Barbe V, Latifi A, Gugger M, Calteau A, Coursin T, Lajus A, Castelli V,
676 Oztas S, Samson G, Longin C, Medigue C, de Marsac NT. 2013. A tribute to disorder in
677 the genome of the bloom-forming freshwater cyanobacterium *Microcystis aeruginosa*.
678 PLoS One 8:e70747.
679 11. Tanabe Y, Sano T, Kasai F, Watanabe MM. 2009. Recombination, cryptic clades and
680 neutral molecular divergence of the microcystin synthetase (mcy) genes of toxic
681 cyanobacterium *Microcystis aeruginosa*. BMC Evol Biol 9:115.
682 12. Tanabe Y, Kasai F, Watanabe MM. 2007. Multilocus sequence typing (MLST) reveals
683 high genetic diversity and clonal population structure of the toxic cyanobacterium
684 *Microcystis aeruginosa*. Microbiol 153:3695-703.
685 13. Tanabe Y, Kaya K, Watanabe MM. 2004. Evidence for recombination in the microcystin
686 synthetase (mcy) genes of toxic cyanobacteria *Microcystis* spp. J Mol Evol 58:633-41.
687 14. Harke MJ, Steffen MM, Gobler CJ, Otten TG, Wilhelm SW, Wood SA, Paerl HW. 2016.
688 A review of the global ecology, genomics, and biogeography of the toxic
689 cyanobacterium, *Microcystis* spp. Harmful Algae 54:4-20.
690 15. van Gremberghe I, Leliaert F, Mergeay J, Vanormelingen P, Van der Gucht K, Debeer
691 AE, Lacerot G, De Meester L, Vyverman W. 2011. Lack of phylogeographic structure in
692 the freshwater cyanobacterium *Microcystis aeruginosa* suggests global dispersal. PLoS
693 One 6:e19561.
694 16. Shapiro BJ, Leducq JB, Mallet J. 2016. What Is Speciation? PLoS Genet 12:e1005860.

17. Otsuka S, Suda S, Shibata S, Oyaizu H, Matsumoto S, Watanabe MM. 2001. A proposal for the unification of five species of the cyanobacterial genus *Microcystis* Kutzing ex Lemmermann 1907 under the rules of the Bacteriological Code. Int J Syst Evol Microbiol 51:873-9.

18. Otsuka S, Suda S, Li RH, Watanabe M, Oyaizu H, Matsumoto S, Watanabe MM. 1999. Phylogenetic relationships between toxic and non-toxic strains of the genus *Microcystis* based on 16S to 23S internal transcribed spacer sequence. FEMS Microbiol Lett 172:15-21.

19. Komárek J. 2010. Recent changes (2008) in cyanobacteria taxonomy based on a combination of molecular background with phenotype and ecological consequences (genus and species concept). Hydrobiologia:245–59.

20. Xiao M, Li M, Reynolds CS. 2018. Colony formation in the cyanobacterium *Microcystis*. Biol Rev Camb Philos Soc 93:1399-1420.

21. Philip T. Orr GJJ. 1998. Relationship between microcystin production and cell division rates in nitrogen - limited *Microcystis aeruginosa* cultures. Limnol Ocean 43:1604-1614.

22. Compère P. 2005. The nomenclature of Cyanophyta under the Botanical Code. Arch Hydrobiol/Algol Stud (Cyanobact Res 6) 117:31-37.

23. Zhang JY, Guan R, Zhang HJ, Li H, Xiao P, Yu GL, Du L, Cao DM, Zhu BC, Li RH, Lu ZH. 2016. Complete genome sequence and genomic characterization of *Microcystis panniformis* FACHB 1757 by third-generation sequencing. Stand Genomic Sci 11:11.

24. Ma J, Brookes JD, Qin B, Paerl HW, Gao G, Wu P, Zhang W, Deng J, Zhu G, Zhang Y, Xu H, Niu H. 2014. Environmental factors controlling colony formation in blooms of the cyanobacteria *Microcystis* spp. in Lake Taihu, China. Harmful Algae 31:136-142.

718 25. Wang WJ, Zhang YL, Shen H, Xie P, Yu J. 2015. Changes in the bacterial community
719 and extracellular compounds associated with the disaggregation of *Microcystis* colonies.
720 Biochem Syst Ecol 61:62-66.
721 26. Otsuka S, Suda S, Li RH, Matsumoto S, Watanabe MM. 2000. Morphological variability
722 of colonies of *Microcystis* morphospecies in culture. J Gen Appl Microbiol 46:39-50.
723 27. Komárek J. 1991. A review of water-bloom forming *Microcystis* species, with regard to
724 populations from Japan. Archiv Fur Hydrobiologie:115-127.
725 28. Xu S, Sun Q, Zhou X, Tan X, Xiao M, Zhu W, Li M. 2016. Polysaccharide biosynthesis-
726 related genes explain phenotype-genotype correlation of *Microcystis* colonies in Meiliang
727 Bay of Lake Taihu, China. Sci Rep 6:35551.
728 29. Castenholz RW, Norris TB. 2005. Revisionary concepts of species in the cyanobacteria
729 and their applications. Algol Stud:53–69.
730 30. Konstantinidis KT, Tiedje JM. 2005. Towards a genome-based taxonomy for
731 prokaryotes. J Bacteriol 187:6258-64.
732 31. Christiansen G, Molitor C, Philmus B, Kurmayer R. 2008. Nontoxic strains of
733 cyanobacteria are the result of major gene deletion events induced by a transposable
734 element. Mol Biol Evol 25:1695-704.
735 32. Cheng L, Connor TR, Siren J, Aanensen DM, Corander J. 2013. Hierarchical and
736 spatially explicit clustering of DNA sequences with BAPS software. Mol Biol Evol
737 30:1224-8.
738 33. Castillo-Ramirez S, Corander J, Marttinen P, Aldeljawi M, Hanage WP, Westh H, Boye
739 K, Gulay Z, Bentley SD, Parkhill J, Holden MT, Feil EJ. 2012. Phylogeographic
740 variation in recombination rates within a global clone of methicillin-resistant
741 *Staphylococcus aureus*. Genome Biol 13:R126.

742 34. Makarova KS, Grishin NV, Koonin EV. 2006. The HicAB cassette, a putative novel,
743 RNA-targeting toxin-antitoxin system in archaea and bacteria. Bioinformatics 22:2581-4.
744 35. Makarova KS, Wolf YI, Koonin EV. 2009. Comprehensive comparative-genomic
745 analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in
746 prokaryotes. Biol Direct 4:19.
747 36. Makarova KS, Wolf YI, Snir S, Koonin EV. 2011. Defense islands in bacterial and
748 archaeal genomes and prediction of novel defense systems. J Bacteriol 193:6039-6056.
749 37. Welker M, von Dohren H. 2006. Cyanobacterial peptides - nature's own combinatorial
750 biosynthesis. FEMS Microbiol Rev 30:530-63.
751 38. Sussmuth RD, Mainz A. 2017. Nonribosomal peptide synthesis-principles and prospects.
752 Angew Chem Int Ed Engl 56:3770-3821.
753 39. Tillett D, Parker DL, Neilan BA. 2001. Detection of toxigenicity by a probe for the
754 microcystin synthetase A gene (mcyA) of the cyanobacterial genus Microcystis:
755 comparison of toxicities with 16S rRNA and phycocyanin operon (Phycocyanin
756 Intergenic Spacer) phylogenies. Appl Environ Microbiol 67:2810-8.
757 40. Pearson LA, Dittmann E, Mazmouz R, Ongley SE, D'Agostino PM, Neilan BA. 2016.
758 The genetics, biosynthesis and regulation of toxic specialized metabolites of
759 cyanobacteria. Harmful Algae 54:98-111.
760 41. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, Suarez Duran HG, de
761 Los Santos ELC, Kim HU, Nave M, Dickschat JS, Mitchell DA, Shelest E, Breitling R,
762 Takano E, Lee SY, Weber T, Medema MH. 2017. antiSMASH 4.0-improvements in
763 chemistry prediction and gene cluster boundary identification. Nucleic Acids Res
764 45:W36-W41.

765 42. Via-Ordorika L, Fastner J, Kurmayer R, Hisbergues M, Dittmann E, Komarek J, Erhard
766 M, Chorus I. 2004. Distribution of microcystin-producing and non-microcystin-
767 producing *Microcystis* sp. in European freshwater bodies: detection of microcystins and
768 microcystin genes in individual colonies. Syst Appl Microbiol 27:592-602.
769 43. Yoshida M, Yoshida T, Satomi M, Takashima Y, Hosoda N, Hiroishi S. 2008. Intra-
770 specific phenotypic and genotypic variation in toxic cyanobacterial *Microcystis* strains. J
771 Appl Microbiol 105:407-15.
772 44. Xu Y, Wu Z, Yu B, Peng X, Yu G, Wei Z, Wang G, Li R. 2008. Non-microcystin
773 producing *Microcystis wesenbergii* (Komarek) Komarek (Cyanobacteria) representing a
774 main waterbloom-forming species in Chinese waters. Environ Pollut 156:162-7.
775 45. Otten TG, Paerl HW. 2011. Phylogenetic inference of colony isolates comprising
776 seasonal *Microcystis* blooms in Lake Taihu, China. Microb Ecol 62:907-18.
777 46. Majewski J, Cohan FM. 1999. Adapt globally, act locally: the effect of selective sweeps
778 on bacterial sequence diversity. Genetics 152:1459-74.
779 47. Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species
780 definition for prokaryotes. Proc Natl Acad Sci U S A 102:2567-2572.
781 48. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput
782 ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun
783 9:5114.
784 49. Moreira C, Spillane C, Fathalli A, Vasconcelos V, Antunes A. 2014. African origin and
785 europe-mediated global dispersal of the cyanobacterium *Microcystis aeruginosa*. Curr
786 Microbiol 69:628-33.

787 50. Marmen S, Aharonovich D, Grossowicz M, Blank L, Yacobi YZ, Sher DJ. 2016.
788 Distribution and habitat specificity of potentially-toxic *Microcystis* across climate, land,
789 and water use gradients. Front Microbiol 7:271.
790 51. Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial
791 speciation. Science 315:476-80.
792 52. Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria
793 and archaea. ISME J 3:199-208.
794 53. Shapiro BJ, Polz MF. 2014. Ordering microbial diversity into ecologically and
795 genetically cohesive units. Trends Microbiol 22:235-47.
796 54. Imai H, Chang KH, Kusaba M, Nakano S. 2009. Temperature-dependent dominance of
797 *Microcystis* (Cyanophyceae) species: *M. aeruginosa* and *M. wesenbergii*. J Plankton Res
798 31:171–178.
799 55. Rantala A, Fewer DP, Hisbergues M, Rouhiainen L, Vaitomaa J, Borner T, Sivonen K.
800 2004. Phylogenetic evidence for the early evolution of microcystin synthesis. Proc Natl
801 Acad Sci U S A 101:568-73.
802 56. Tooming-Klunderud A, Fewer DP, Rohrlack T, Jokela J, Rouhiainen L, Sivonen K,
803 Kristensen T, Jakobsen KS. 2008. Evidence for positive selection acting on microcystin
804 synthetase adenylation domains in three cyanobacterial genera. BMC Evol Biol 8:256.
805 57. Mikalsen B, Boison G, Skulberg OM, Fastner J, Davies W, Gabrielsen TM, Rudi K,
806 Jakobsen KS. 2003. Natural variation in the microcystin synthetase operon *mcyABC* and
807 impact on microcystin production in *Microcystis* strains. J Bacteriol 185:2774-85.
808 58. Moreira C, Vasconcelos V, Antunes A. 2013. Phylogeny of microcystins: evidence of a
809 biogeographical trend? Curr Microbiol 66:214-21.

59. Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJ, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der Oost J, Backofen R, Koonin EV. 2015. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol 13:722-36.

60. Godde JS, Bickerton A. 2006. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. J Mol Evol 62:718-29.

61. Komárek J, Komárková-Legnerová J, Santanna CL, Azevedo MTP, Senna PC. 2002. Two common *Microcystis* species (Chroococcales, Cyanobacteria) from tropical America, including *M. panniformis* sp. nov. Cryptogamie Algol 23:159-177.

62. Komárek J, Komárková J. 2002. Review of the European *Microcystis*-morphospecies (Cyanoprokaryotes) from nature. Czech Phycol 2:1-24.

63. Peng Y, Leung HC, Yiu SM, Chin FY. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420-8.

64. Meyer KA, Davis TW, Watson SB, Denef VJ, Berry MA, Dick GJ. 2017. 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.

65. Li Q, Lin F, Yang C, Wang J, Lin Y, Shen M, Park MS, Li T, Zhao J. 2018. A large-scale comparative metagenomic study reveals the functional interactions in six bloom-forming *Microcystis*-epibiont communities. Front Microbiol 9:746.

66. Eren AM, Esen OC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319.

834 67. Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and sensitive
835 classification of metagenomic sequences. Genome Res 26:1721-1729.
836 68. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal:
837 prokaryotic gene recognition and translation initiation site identification. BMC
838 Bioinformatics 11:119.
839 69. Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. 2012. Gene and translation initiation
840 site prediction in metagenomic sequences. Bioinformatics 28:2223-30.
841 70. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics
842 30:2068-9.
843 71. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time
844 and space complexity. BMC Bioinformatics 5:113.
845 72. Stamatakis A, Aberer AJ, Goll C, Smith SA, Berger SA, Izquierdo-Carrasco F. 2012.
846 RAxML-Light: a tool for computing terabyte phylogenies. Bioinformatics 28:2064-6.
847 73. Angiuoli SV, Salzberg SL. 2011. Mugsy: fast multiple alignment of closely related whole
848 genomes. Bioinformatics 27:334-42.
849 74. Richter M, Rossello-Mora R. 2009. Shifting the genomic gold standard for the
850 prokaryotic species definition. Proc Natl Acad Sci U S A 106:19126-31.
851 75. Konstantinidis KT, Ramette A, Tiedje JM. 2006. Toward a more robust assessment of
852 intraspecies diversity, using fewer genetic markers. Appl Environ Microbiol 72:7286-93.
853 76. Wickham H. 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag New
854 York.
855 77. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D,
856 Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis.
857 Bioinformatics 31:3691-3.

858 78. Didelot X, Wilson DJ. 2015. ClonalFrameML: efficient inference of recombination in
859 whole bacterial genomes. PLoS Comput Biol 11:e1004041.
860 79. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New
861 algorithms and methods to estimate maximum-likelihood phylogenies: assessing the
862 performance of PhyML 3.0. Syst Biol 59:307-21.
863 80. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
864 improvements in performance and usability. Mol Biol Evol 30:772-80.
865 81. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood
866 trees for large alignments. PLoS One 5:e9490.
867 82. Asnicar F, Weingart G, Tickle TL, Huttenhower C, Segata N. 2015. Compact graphical
868 representation of phylogenetic data and metadata with GraPhlAn. PeerJ 3:e1029.
869 83. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T,
870 Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P. 2016. eggNOG 4.5:
871 a hierarchical orthology framework with improved functional annotations for eukaryotic,
872 prokaryotic and viral sequences. Nucleic Acids Res 44:D286-93.
873 84. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P.
874 2017. Fast Genome-Wide Functional Annotation through Orthology Assignment by
875 eggNOG-Mapper. Mol Biol Evol 34:2115-2122.
876 85. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997.
877 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.
878 Nucleic Acids Res 25:3389-402.
879 86. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. 2014. STAMP: statistical analysis of
880 taxonomic and functional profiles. Bioinformatics 30:3123-4.
881

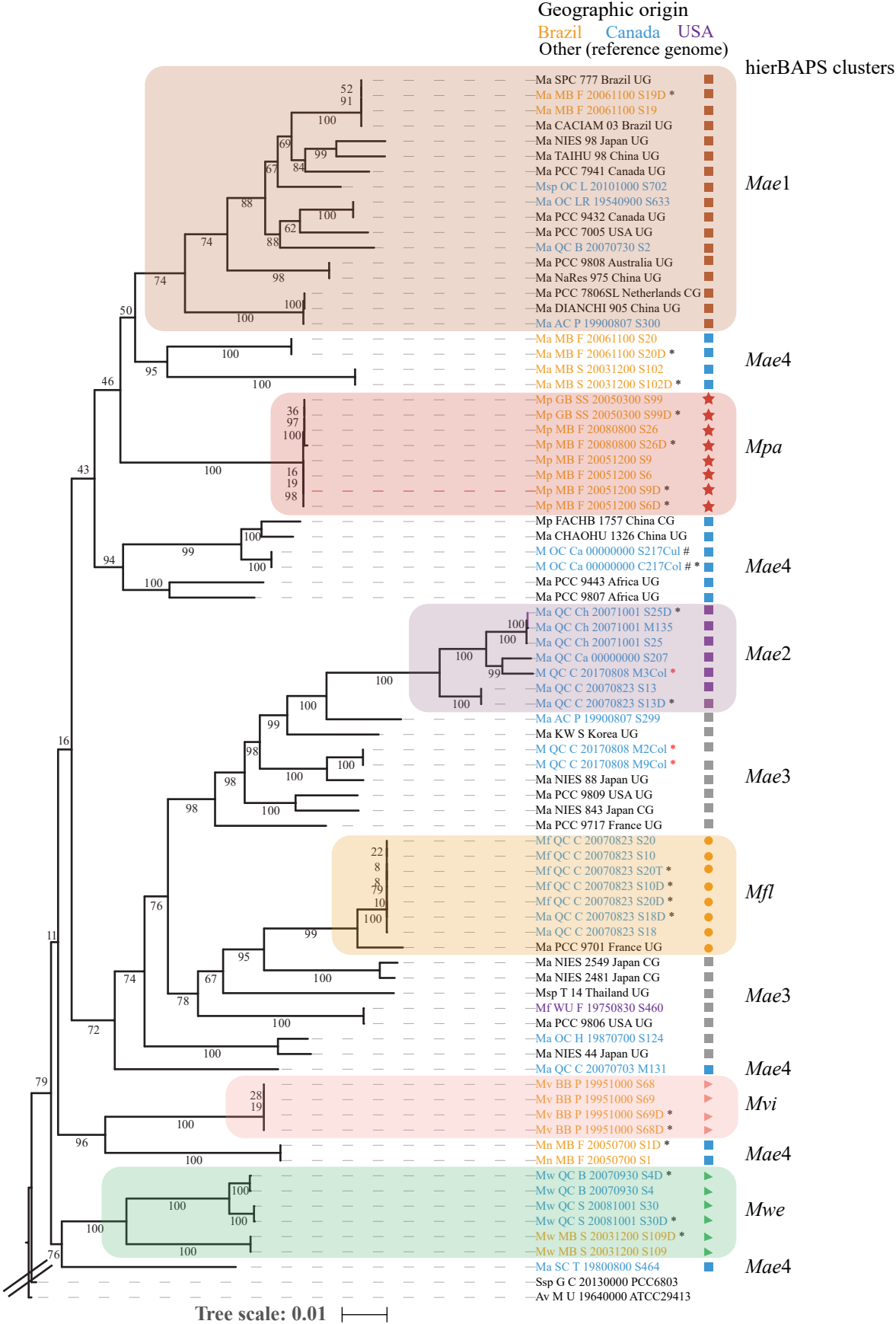
882 **Table 1.** Breakdown of horizontally transferred genes by geography.

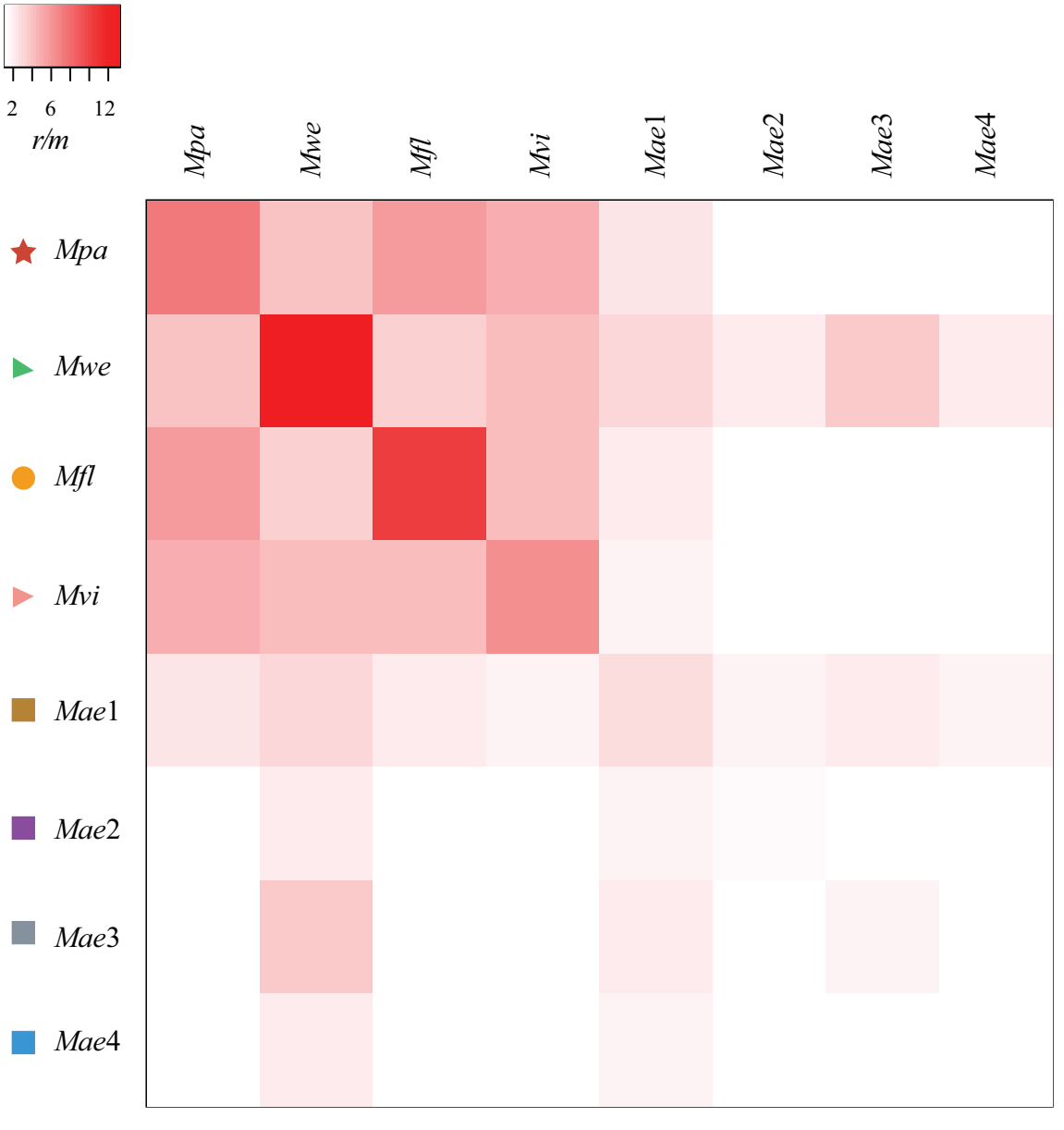
	Local HGT (Canada-only or Brazil-only clades)	Non-local HGT (Canada-Brazil clades)
Hypothetical	474	448
Known Function	449*	142
Phylogenetical distances equal to zero	80*	28
Total	923	590

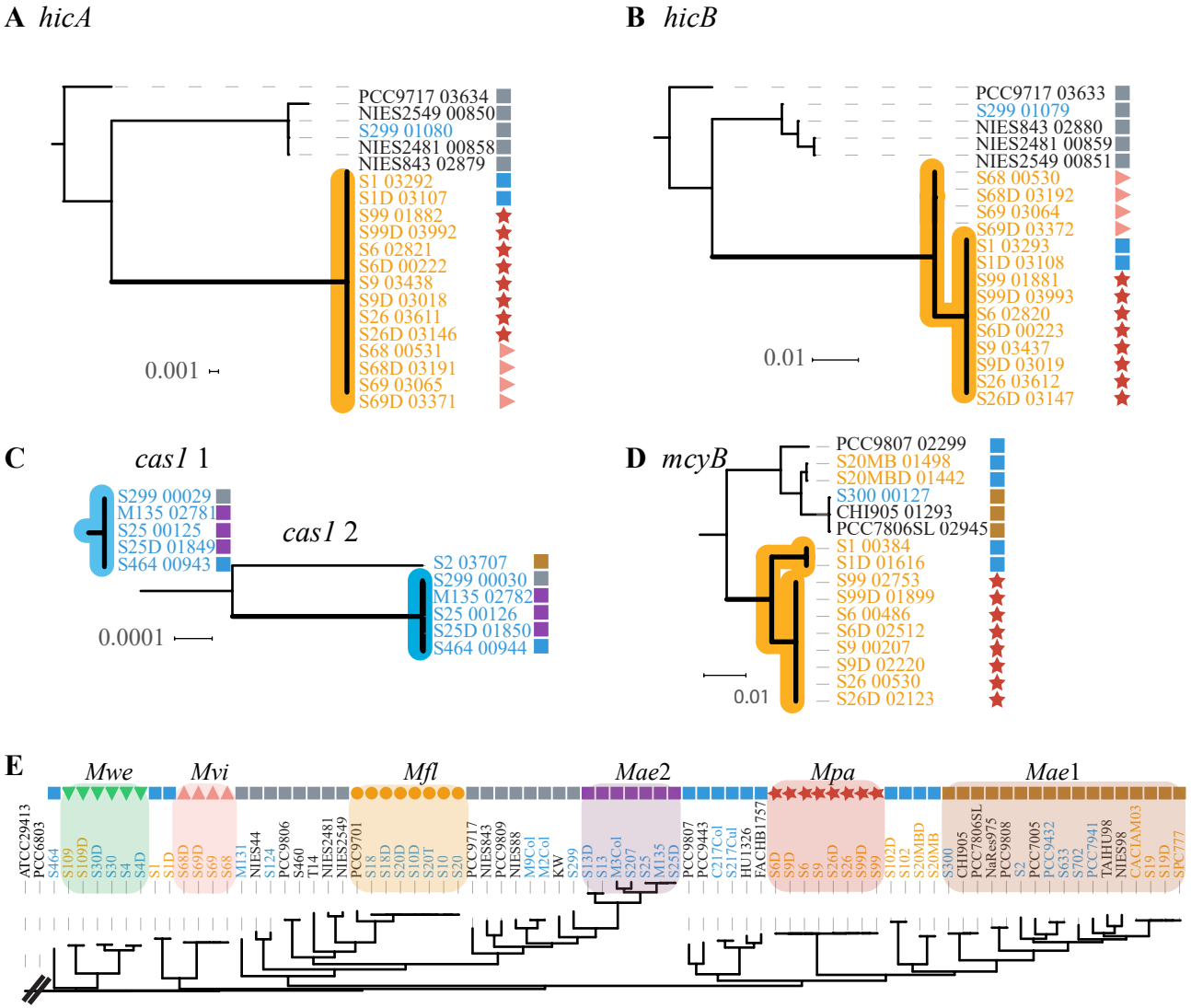
883 * denotes a significant enrichment among local vs. non-local HGT events (Fisher's exact test, $P <$

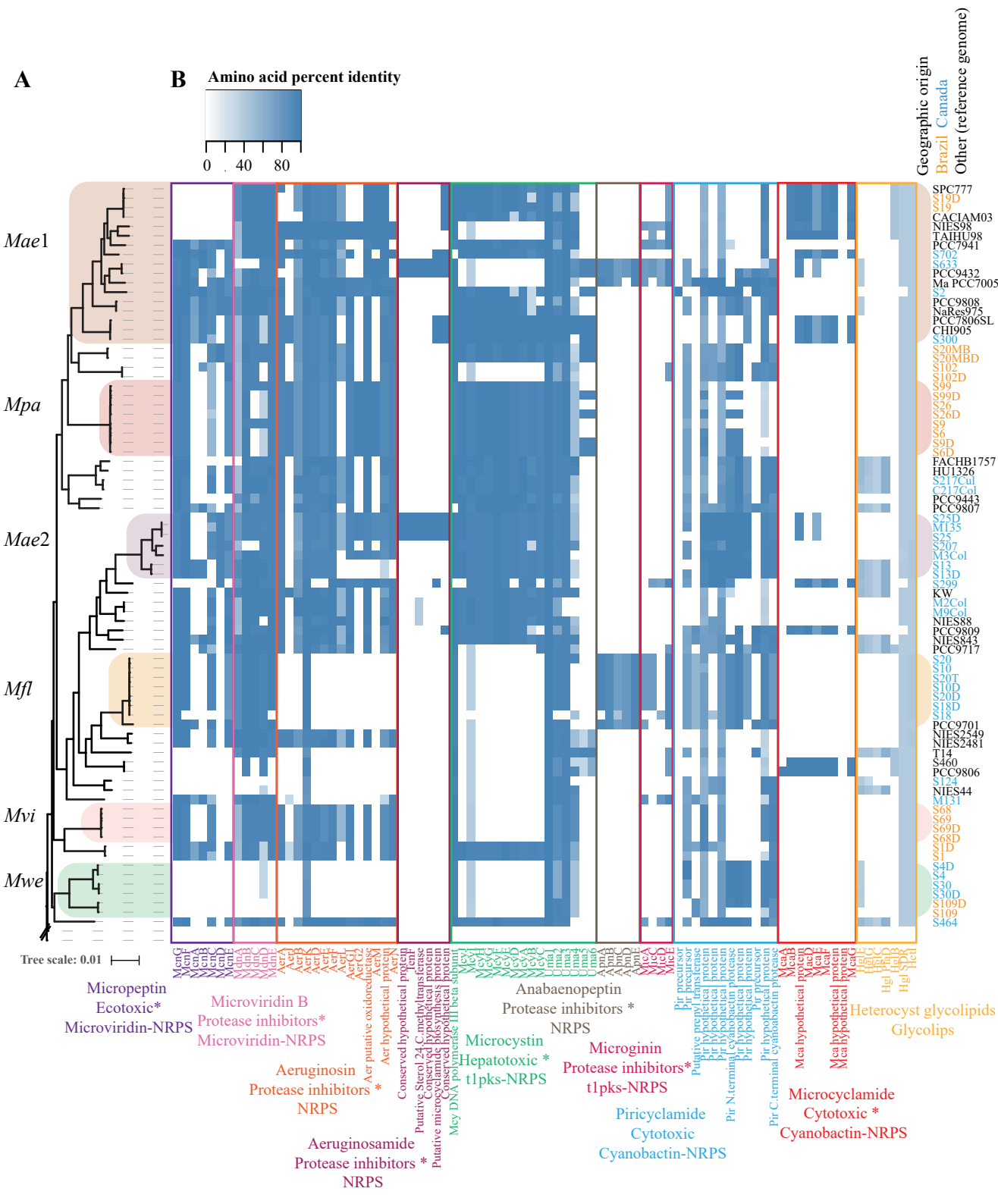
884 0.01).

885


Geographic origin
Brazil
Canada
USA
Other (reference genome)
hierBAPS clusters
Ma SPC 777 Brazil UG
Ma MB F 20061100 S19D *
Ma MB F 20061100 S19
Ma CACIAM 03 Brazil UG
Ma NIES 98 Japan UG
Ma TAIHU 98 China UG
Ma PCC 7941 Canada UG
Msp OC L 20101000 S702
Ma OC LR 19540900 S633
Ma PCC 9432 Canada UG
Ma PCC 7005 USA UG
Ma QC B 20070730 S2
Ma PCC 9808 Australia UG
Ma NaRes 975 China UG
Ma PCC 7806SL Netherlands CG
Ma DIANCHI 905 China UG
Ma AC P 19900807 S300
Mae1
Ma MB F 20061100 S20
Ma MB F 20061100 S20D *
Ma MB S 20031200 S102
Ma MB S 20031200 S102D *
Mae4
Mp GB SS 20050300 S99
Mp GB SS 20050300 S99D *
Mp MB F 20080800 S26
Mp MB F 20080800 S26D *
Mp MB F 20051200 S9
Mp MB F 20051200 S6
Mp MB F 20051200 S9D *
Mp MB F 20051200 S6D *
Mpa
Mp FACHB 1757 China CG
Ma CHAOHU 1326 China UG
M OC Ca 00000000 S217Cul #
M OC Ca 00000000 C217Col # *
Ma PCC 9443 Africa UG
Ma PCC 9807 Africa UG
Mae4
Ma QC Ch 20071001 S25D *
Ma QC Ch 20071001 M135
Ma QC Ch 20071001 S25
Ma QC Ca 00000000 S207
M QC C 20170808 M3Col *
Ma QC C 20070823 S13
Ma QC C 20070823 S13D *
Mae2
Ma AC P 19900807 S299
Ma KW S Korea UG
M QC C 20170808 M2Col *
M QC C 20170808 M9Col *
Ma NIES 88 Japan UG
Ma PCC 9809 USA UG
Ma NIES 843 Japan CG
Ma PCC 9717 France UG
Mae3
Mf QC C 20070823 S20
Mf QC C 20070823 S10
Mf QC C 20070823 S20T *
Mf QC C 20070823 S10D *
Mf QC C 20070823 S20D *
Ma QC C 20070823 S18D *
Ma QC C 20070823 S18
Ma PCC 9701 France UG
Mfl
Ma NIES 2549 Japan CG
Ma NIES 2481 Japan CG
Msp T 14 Thailand UG
Mf WU F 19750830 S460
Ma PCC 9806 USA UG
Mae3
Ma OC H 19870700 S124
Ma NIES 44 Japan UG
Ma QC C 20070703 M131
Mae4
Mv BB P 19951000 S68
Mv BB P 19951000 S69
Mv BB P 19951000 S69D *
Mv BB P 19951000 S68D *
Mvi
Mn MB F 20050700 S1D *
Mn MB F 20050700 S1
Mae4
Mw QC B 20070930 S4D *
Mw QC B 20070930 S4
Mw QC S 20081001 S30
Mw QC S 20081001 S30D *
Mw MB S 20031200 S109D *
Mw MB S 20031200 S109
Mwe
Ma SC T 19800800 S464
Mae4
Ssp G C 20130000 PCC6803
Av M U 19640000 ATCC29413
Tree scale: 0.01


2
6
12
r/m
Mpa
Mwe
Mfl
Mvi
Mae1
Mae2
Mae3
Mae4
Mpa
Mwe
Mfl
Mvi
Mae1
Mae2
Mae3
Mae4


A hicA
PCC9717 03634
NIES2549 00850
S299 01080
NIES2481 00858
NIES843 02879
S1 03292
S1D 03107
S99 01882
S99D 03992
S6 02821
S6D 00222
S9 03438
S9D 03018
S26 03611
S26D 03146
S68 00531
S68D 03191
S69 03065
S69D 03371
0.001
B hicB
PCC9717 03633
S299 01079
NIES843 02880
NIES2481 00859
NIES2549 00851
S68 00530
S68D 03192
S69 03064
S69D 03372
S1 03293
S1D 03108
S99 01881
S99D 03993
S6 02820
S6D 00223
S9 03437
S9D 03019
S26 03612
S26D 03147
0.01
C cas1 1
S299 00029
M135 02781
S25 00125
S25D 01849
S464 00943
cas1 2
S2 03707
S299 00030
M135 02782
S25 00126
S25D 01850
S464 00944
0.0001
D mcyB
PCC9807 02299
S20MB 01498
S20MBD 01442
S300 00127
CHI905 01293
PCC7806SL 02945
S1 00384
S1D 01616
S99 02753
S99D 01899
S6 00486
S6D 02512
S9 00207
S9D 02220
S26 00530
S26D 02123
0.01
E
Mwe
Mvi
Mfl
Mae2
Mpa
Mae1
ATCC29413
PCC6803
S464
S109
S109D
S30D
S30
S4
S4D
S1
S1D
S68D
S69D
S69
S68
M131
NIES44
S124
PCC9806
S460
T14
NIES2481
NIES2549
PCC9701
S18
S18D
S20D
S10D
S20T
S10
S20
PCC9717
NIES843
PCC9809
NIES88
M9Col
M2Col
KW
S299
S13D
S13
M3Col
S207
S25
M135
S25D
PCC9807
PCC9443
C217Col
S217Col
HU1326
FACHB1757
S6D
S9D
S6
S9
S26D
S26
S99D
S99
S102D
S102
S20MBD
S20MB
S300
CHI905
PCC7806SL
NaRes975
PCC9808
S2
PCC7005
PCC9432
S633
S702
PCC7941
TAIHU98
NIES98
CACIAM03
S19
S19D
SRC777
Tree scale: 0.01
Geographic origin
Canada Brazil Other (reference genome)


A
B
Amino acid percent identity
0
40
80
Geographic origin
Brazil
Canada
Other (reference genome)
Mae1
Mpa
Mae2
Mfl
Mvi
Mwe
Tree scale: 0.01
SPC777
S19D
S19
CACIAM03
NIES98
TAIHU98
PCC7941
S702
S633
PCC9432
Ma PCC7005
S2
PCC9808
NaRes975
PCC7806SL
CHI905
S300
S20MB
S20MBD
S102
S102D
S99
S99D
S26
S26D
S9
S6
S9D
S6D
FACHB1757
HU1326
S217Col
C217Col
PCC9443
PCC9807
S25D
M135
S25
S207
M3Col
S13
S13D
S299
KW
M2Col
M9Col
NIES88
PCC9809
NIES843
PCC9717
S20
S10
S20T
S10D
S20D
S18D
S18
PCC9701
NIES2549
NIES2481
T14
S460
PCC9806
S124
NIES44
M131
S68
S69
S69D
S68D
S1D
S1
S4D
S4
S30
S30D
S109D
S109
S464
MenG
MenF
MenA
MenB
MenC
MenE
MdnA
MdnB
MdnC
MdnD
MdnE
AerA
AerB
AerD
AerE
AerF
AerL
AerQ
AerO
AerM
AerN
Aer putative oxidoreductase
Aer hypothetical protein
Conserved hypothetical protein
TenI
Putative Sterol 24-C-methyltransferase
Conserved hypothetical protein
Putative microcyclamide biosynthesis protein McaA
Mcy DNA polymerase III beta subunit
McyJ
McyI
McyH
McyG
McyF
McyE
McyD
McyA
McyB
McyC
Uma2
Uma1
Uma3
Uma4
Uma5
Uma6
AmA
AmB
AmC
AmD
AmE
McA
McC
McD
McE
Pir precursor
Pir precursor
Putative prenyl transferase
Pir hypothetical protein
Pir hypothetical protein
Pir hypothetical protein
Pir hypothetical protein
Pir N-terminal cyanobactin precursor
Pir hypothetical protein
Pir hypothetical protein
Pir precursor
Pir hypothetical protein
Pir C-terminal cyanobactin precursor
McaA
McaB
Mca hypothetical protein
McaD
McaE
McaF
Mca hypothetical protein
Mca hypothetical protein
McaG
HglE
HglC
HglB
HglD
HglA
SDR
TrmT
Micropeptin
Ecotoxic*
Microviridin-NRPS
Microviridin B
Protease inhibitors*
Microviridin-NRPS
Aeruginosin
Protease inhibitors *
NRPS
Aeruginosamide
Protease inhibitors *
NRPS
Anabaenopeptin
Protease inhibitors *
NRPS
Microcystin
Hepatotoxic *
t1pks-NRPS
Microginin
Protease inhibitors*
t1pks-NRPS
Piricyclamide
Cytotoxic
Cyanobactin-NRPS
Microcyclamide
Cytotoxic *
Cyanobactin-NRPS
Heterocyst glycolipids
Glycolips