Unexpected Micro-Spatial Scale Genomic Diversity of the Bloom-Forming Cyanobacterium Aphanizomenon Gracile and its Phycosphere

Sebastien Halary (✉ sebastien.halary@mnhn.fr)
MNHN: Museum National d'Histoire Naturelle  https://orcid.org/0000-0002-7089-9459

Sébastien Duperron
MNHN: Museum National d'Histoire Naturelle

Sandra Kim Tiam
MNHN: Museum National d'Histoire Naturelle

Charlotte Duval
MNHN: Museum National d'Histoire Naturelle

Eloïse Dhénain
MNHN: Museum National d'Histoire Naturelle

Cécile Bernard
MNHN: Museum National d'Histoire Naturelle

Benjamin Marie
MNHN: Museum National d'Histoire Naturelle

Research

Keywords: cyanobacteria, heterotrophic bacteria, gracile, BCGs

Posted Date: June 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-617160/v1

License: ☇ ① This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Unexpected micro-spatial scale genomic diversity of the bloom-forming cyanobacterium *Aphanizomenon gracile* and its phycosphere

Running Title: Comparative omics of *Aphanizomenon gracile* and its phycosphere

Sébastien Halary¹*, Sébastien Duperron¹, Sandra Kim Tiam¹, Charlotte Duval¹, Eloïse Dhénain¹, Cécile Bernard¹, Benjamin Marie¹*

¹Muséum National d’Histoire Naturelle, CNRS, UMR7245 Mécanismes de Communication et Adaptation des Micro-organismes, 12 rue Buffon, 75005 Paris, France.

*Co-corresponding authors: sebastien.halary@mnhn.fr, benjamin.marie@mnhn.fr

Abstract

**Background:** Genotypic diversity within cyanobacteria populations, partly driven by horizontal gene transfers, is a key factor of their colonization success, allowing them to cope with spatio-temporal fluctuations of the environmental conditions. By providing complementary functions, such as oxidative stress protection, heterotrophic bacteria composing phycospheres play also an essential role in cyanobacteria adaptation. *Aphanizomenon gracile* is a species of toxinogen cyanobacteria blooming worldwide with severe consequences for fresh and brackish water ecosystems. While marker heterogeneity surveys have shown that harmful cyanobacteria blooms were not clonal populations, the real extent of genetic and functional diversity within a population of freshwater cyanobacteria, including *A. gracile*, and their associated phycospheres remains unclear.

**Results:** Here, comparative omics of four monoclonal strains of *A. gracile* isolated from a single drop of water reveals extensive heterogeneity of chemotypes and gene contents, despite constrained genome size and high similarity indices. These variations are remarkably associated with horizontal gene transfers (HGT) of biosynthetic gene clusters (BCG), and a novel siphophage infecting *A. gracile* displaying characteristics of temperate phages appears to participate to this genotypic diversification. In spite of high variability in heterotrophic taxa relative abundances, *A. gracile* phycospheres displayed an apparent functional redundancy implying biosynthesis of public goods.

**Conclusions:** Altogether, these results suggest that a bloom would constitute a hot-spot for *A. gracile* genotype diversification driven by cyanophages, where losses and gains of BCGs compels
cyanobacteria individuals to cooperate together and with heterotrophic bacteria in a black queen hypothesis compatible way.

**Background**

*Aphanizomenon* is a genus of filamentous harmful bloom-forming cyanobacteria (cyanoHAB) occurring worldwide in brackish and freshwater ecosystems\(^1\). *Aphanizomenon* can dominate phytoplanktonic communities in an exclusive way\(^2\), leading to blooms with multiple deleterious effects, including trophic network perturbations, microbial loop alterations, and release of a particularly broad range of toxic metabolites (*e.g.* microcystins, saxitoxins, cylindrospermopsins)\(^1,3\). Furthermore, its ability to fix inorganic nitrogen makes *Aphanizomenon* responsible for the subsequent growth of other non-nitrogen-fixing cyanoHAB such as *Microcystis*\(^4\). Despite the ubiquity and the high number of described strains of *Aphanizomenon*, little is known about the genetic diversity of the genus and genomic data are only available for few strains belonging to the species *A. flos-aquae*\(^3\).

The ecological success of cyanobacteria partly relies on the genotypic diversity existing within a single population\(^5,7\), allowing populations to cope with spatial and temporal environmental variations. In freshwater ecosystems, these population studies remain mainly focused on 16S rRNA-encoding gene fragment or internal transcribed spacer (ITS) sequencing and the presence/absence of cyanotoxin-biosynthesis genes in cyanobacterial strains from eutrophic lakes (*e.g.* *Microcystis aeruginosa*, *Planktothrix agardhii*, *Raphidiopsis* (anc. *Cylindrospermopsis* *raciborskii*)\(^5,8,9\) because of their environmental and health impacts. However, infra-specific genetic diversity in many freshwater cyanobacteria is obviously not restricted to these markers, and whole genome sequencing highlighted that populations often consist of diverse conspecific strain that display the same morphotype, yet sometimes only share a limited fraction of their genes, as low as 54% for example for some clones of *M. aeruginosa*\(^10\). Each strain thus contains only a subsample of the global population-level pangenome. The same can be observed for the metabolites, which display a high strain-specific profile, extending far beyond more cyanotoxin production capabilities\(^11\). Such infra-specific diversity is suspected to result in increased capabilities to cope with different selective pressures such as temperature, light intensity or nutrients availability, and improved exploitation of the water column and environmental micro-niches. However, little data is currently available documenting the extent of genome and chemotype diversity at a population level.

Furthermore, cyanobacteria are integrated in a complex network of interactions with co-occurring microbial populations. The most studied examples so far are predation by zooplanktonic grazers or competition with photosynthetic eukaryotes\(^2\). Other key players that participate in
cyanobacteria population regulation have been overlooked. Albeit known for a long time, the beneficial interactions established between cyanobacteria and associated heterotrophic bacteria comprising the so-called phycosphere (or cyanosphere) have recently enjoyed renewed interest. These heterotrophs have been shown to exploit metabolites released by cyanobacteria, which on their side benefit from reduced concentrations of reactive oxygen species in vitro, otherwise deleterious for photosynthesis, thanks to cyanosphere-produced peroxiredoxins and catalases. The broad range of metabolic capabilities carried out by cyanobacteria and the distantly-related bacterial phyla composing the cyanosphere may also promote a higher diversity of interactions. Another important yet neglected group are the cyanophages. Indeed, temperate cyanophages, mostly studied in marine environments, are thought to drive the genomic plasticity associated with genotypic diversity of cyanobacteria, and viral lysis to trigger bloom senescence.

In this study, we evaluate how representative one single individual can be of the potential pangenome of a blooming population of *A. gracile*, and attempt to decipher the interactions that may contribute to the adaptative success of this cyanobacterial species. For this, the inter-individual genomic and metabolite diversity of *A. gracile* are described from four monoclonal strains isolated from a single drop of water collected in an eutrophic freshwater peri-urban lake. Using Metagenome-Assembled Genomes (MAG) approach, we also identify members and features of the microbial community composing the *A. gracile* cyanosphere, which include heterotrophic bacteria and a novel cyanophage that may be involved in *A. gracile* population dynamics.

**Methods**

**Site and sampling conditions**

The lake of Champs-sur-Marne (48°51’50” N, 2°35’53” E) is located in the outer eastern suburb of Paris (France). This reservoir is a small (10.3 ha) shallow (mean depth of 2.4 m) water body isolated from the river system, used for recreational activities, and surveyed weekly for cyanobacteria and toxins presences since 2005. Raw water was collected from the sub-surface level in July 2010 and brought back refrigerated to the lab.

**Cyanobacterial isolation and culture conditions**

*A. gracile* strains isolations from one droplet were conducted by repeated manual transfers of single filaments on solid or liquid Z8 media (with or without nitrogen), at least 3 times, under an inverted microscope (Nikon ECLIPSE TS100). Growing clones were later cultured in 25 cm² culture flasks (Nunc, Roskilde, Denmark) containing 10 mL of Z8 medium. Cyanobacteria strains were maintained in the Paris Museum Collection (PMC) at 25°C, using daylight fluorescent tubes providing an irradiance level of 16 μmol photons.cm⁻².s⁻¹, with a photoperiod of 16h:8h light:dark.
Strains and cultures were monoclonal and non-axenic, deposited into the PMC collection under identifiers PMC 627.10, PMC 638.10, PMC 644.10 and PMC 649.10. For simplification purposes, these strains will be referred in the rest of the manuscript as PMC627, PMC638, PMC644 and PMC649, respectively.

**DNA extraction and sequencing**

DNA was extracted using a ZymoBIOMICS DNA mini kit (Zymo Research, CA) from cyanobacterial strains grown in flasks. Mechanical lysis was carried out using an ultrasonic cell probe Vibra Sonic (Granuloshop) for 30 s at a range of 100 % to 32.5 W. Total DNA was sequenced using both ILLUMINA HiSeq 2x250bp and SMRT cell PacBio RS2 platforms (Genoscreen, France). In addition, amplicon libraries of a ~500 bp fragment of the 16S rRNA-encoding gene corresponding to the V4-V5 variable region of *Escherichia coli* were constructed using primers 515F and 926R and sequenced on an Illumina MiSeq platform (2 x 250 bp, paired-end sequencing, Genoscreen, France).

**Metagenome assemblies, annotation and curation**

Scaffolds were assembled from MiSeq and Pacbio reads using SPAdes-based Unicycler hybrid-assembler, with default parameters, for the metagenome of each individual strain culture. For each metagenome, nodes from assembly graphs were clustered using MyCC (k-mer size=4, minimal sequence size=1000) and taxonomically annotated using Contig Annotation Tool. 16S rRNA-encoding genes were also extracted from these nodes using Metaxa and annotated using ACT. All assemblies were pairwise-aligned using Megablast (E-value ≤1E-10) and all sequences sharing at least 98 % similarity over the shortest sequence were considered as originating from the same genome. Congruent data between these diverse methodologies allowed to cluster together sequences from the different genomes constituting each metagenome (Metagenome Assembled Genomes, MAGs). In order to improve bacterial genome assemblies by simplifying assembly paths (i.e. to obtain longer nodes), nodes belonging to a same MAG detected in several metagenomes were aligned using CSAR. Resulting scaffolds were then aligned to assembly paths which were then curated using Bandage, following the software recommendations. MAG assembly completeness was assessed using CheckM. Paired-reads of each metagenome were mapped on each MAG using bowtie2, and coverages calculated using samtools. GC statistics were calculated using custom python scripts. Open Reading Frames (ORF) were predicted using Prodigal. Functional annotations including KEGG were obtained using eggnog-mapper with default parameters. Genome maps were generated using Circos.
A. gracile comparative genomics

Genes encoding rRNA and tRNA were detected using Barrnap\textsuperscript{36} and Aragorn\textsuperscript{37}, respectively. ITS sequences were manually extracted using 16S and 23S rRNA-encoding genes coordinates. Sequences from all A. gracile genomes were pairwise aligned using Blastn\textsuperscript{27}. Clusters of Orthologous genes (COG) were defined using OrthoFinder (E-value≤1E\textsuperscript{-05}, identity≥70\%)\textsuperscript{38}. OG frequency was calculated as the proportion of A. gracile genome possessing at least one member of a COG. Synteny index for each genome position was then calculated based on COG triplets using custom perl scripts. Relationship between OG frequency and synteny index was studied using a Pearson’s Chi-square test (H\textsubscript{0}: independence hypothesis, p-value<0.01). A. gracile core genome was defined as constituted by all COGs with an OG frequency of 100\%, all other COGs and singletons being considered as belonging to flexible genes set. Functional comparison between these both fractions was performed using Student tests (H\textsubscript{0}: no difference, p-value<0.01).

Cyanobacteria gene content comparisons

A database was constructed with the four A. gracile MAGs and 268 unique complete genomes of cyanobacteria downloaded from NCBI in addition to information about sampling date and location. ANI was calculated by Mummer between each pair of genomes belonging to a same species using pyani (https://github.com/widdowquinn/pyani). For all members of pairs displaying an ANI greater than or equal to the minimal ANI between our four A. gracile strains genomes (i.e. 99.08\%), clusters of orthologous genes were retrieved as described above. The gene content dissimilarity in all couples were assessed by the Bray-Curtis distance calculated from the matrix of COG presence/absence in each genome using R vegan package\textsuperscript{39}. For clarity purposes, only species gathering at least 3 couples were displayed on the graph.

Biosynthetic gene clusters were retrieved from genomes using AntiSmash v5.1.2 tool\textsuperscript{40}, except for saxitoxin gene clusters that have been additionally searched and confirmed after specific Blast search with sxtA using the MicroScope platform\textsuperscript{41}.

Bacteriophage MAG characterization and analysis

A circular MAG assembled from PMC627 metagenome was aligned against \textit{nr} using Megablast. Given the high similarity with its single significant hit (namely Cyanophage vB_AphaS-CL131, see Results), this MAG was then considered as a cyanophage genome and named Ag627 phage. Alignment between both sequences was visualized using Kablammo\textsuperscript{42}, and Average Nucleotide Identity determined using ANI calculator\textsuperscript{43}. In order to assess its replication strategy, an analysis based on Mavrich \textit{et al.}\textsuperscript{44} was performed. A bacteriophage genomes dataset was constructed composed by Ag627 and vB_AphaS-CL131 genomes as well as all “microbial virus” genomes.
available in NCBI RefSeq (2,528). Jaccard distances between each pair of genomes, proportional to the nucleotide distance, was calculated using MASH (k=16, s=10,000)\textsuperscript{45}. In order to reduce the time of the following procedures, only genomes displaying a Jaccard distance \( \leq 0.15 \) with at least one other genome were kept. ORFs were then predicted for the remaining 1,871 genomes, and groups of homologous genes clusterized using Orthofinder with default parameters\textsuperscript{38}. Gene content dissimilarity between every genome pair was then assessed by the Bray-Curtis distance calculated from cluster of homologous genes presence/absence in all genome matrix\textsuperscript{39}.

**Metabolite content characterization by high-resolution mass spectrometry**

Metabolite content was extracted from each strain cellular biomass using 0.75% methanol (100 \( \mu \)L for 1 mg dry mass) and further analysed on high resolution mass spectrometer, as previously described\textsuperscript{46}. Shortly, Ultra High Performance Liquid Chromatography (UHPLC) was performed on 2 \( \mu \)L of each of the metabolite extracts using a Polar Advances II 2.5-\( \mu \)m pore C\textsubscript{18} column (Thermo\textsuperscript{®}) at a 300 \( \mu \)L.min\textsuperscript{-1} flow rate with a linear gradient of acetonitrile in 0.1% formic acid (5 to 90% in 21 min). The metabolite contents were analysed in triplicates using an electrospray ionization hybrid quadrupole time-of-flight (ESI-QqTOF) high resolution mass spectrometer (Maxis II ETD, Bruker) on positive autoMSMS mode with information dependent acquisition (IDA), on the 50-1500 m/z range at 2 Hz or between 2-8 Hz speed, for MS and MS/MS respectively, according to relative intensity of parent ions, in consecutive cycle times of 2.5 s, with an active exclusion of previously analysed parents. The data were analysed with MetaboScape 4.0 software for internal recalibration (<0.5 ppm), molecular feature search and MGF export. Peak lists were generated from MS/MS spectra between 1 and 15 min, with a filtering noise threshold at 0.1% maximal intensity and combining various charge states and related isotopic forms. A molecular network then was created using the online workflow at Global Natural Products Social molecular networking (GNPS) (http://gnps.ucsd.edu)\textsuperscript{47} as previously described\textsuperscript{11}. The clustered spectra of the network were annotated by comparing monoisotopic mass to our in-house cyanobacteria metabolite databases according to MS and MS/MS fragmentation pattern matches. Molecular networks were visualized using Cytoscape 3.2.1.

**RESULTS**

**Metagenome-assembled genomes acquisition**

After suitable quality trimming and filtering of reads from both technologies, hybrid assemblies generated between 294 (PMC644) and 568 (PMC638) scaffolds, displaying coverage median values of from 18.15X (PMC638) to 26.18X (PMC644). All sequencing and assembly statistics are summarized in Table S1. Synergistic scaffold binning approach allowed to characterize between three (PMC644) and 11 (PMC649) MAGs in each culture sample. Then, scaffold pairwise alignments
clustered these data into 25 unique MAGs (see Table 1). Four MAGs correspond to the four divergent *A. gracile* genomes, 20 MAGs correspond to heterotrophic bacteria genomes including six circularized bacterial genomes, and one corresponds to a complete viral genome.

### Table 1. Genomic features of *A. gracile* strains

| Strain  | Size (Mb) | GC%         | Completeness% | #16S loci | #tRNA | #CDS   | #COGs | #Singletons |
|---------|-----------|-------------|---------------|-----------|-------|--------|-------|-------------|
| PMC627  | 5.32      | 38.39±4.18  | 98.22         | 6         | 43    | 4959   | 4507  | 306         |
| PMC638  | 5.38      | 38.39±4.26  | 94.89         | 5         | 46    | 4924   | 4609  | 212         |
| PMC644  | 5.39      | 38.29±4.18  | 99.55         | 6         | 41    | 4921   | 4612  | 215         |
| PMC649  | 5.24      | 38.35±4.17  | 96.89         | 3         | 38    | 4836   | 4534  | 221         |

**A. gracile genome features**

*A. gracile* genomes assembly graphs were highly entangled, with between 73 (PMC627) and 316 (PMC638) short and highly connected nodes (data not shown), indicating a substantial amount of repeated sequences. The synergistic approach used to improve assembly allowed to merge PMC627 sequences into 71 scaffolds, including 15 longer than 5 kb altogether representing 99.34% of the total genome length (Figure 1a). *A. gracile* MAGs displayed sizes comprised between 5.24 and 5.39 Mb, and a GC percent (GC%) average of 38.35±4.17% (Table 1, Figure 1.a). Estimations of assembly completeness varied between 94.89% (PMC638) to 99.55% (PMC644) and were not proportional to the genome lengths or to the number of CDSs. Divergence in number of transfer RNA genes was observed among strain genomes, from 38 (PMC649) to 46 (PMC638) in total (Table 1). On 21 tRNAs, only 10 displayed the same copy number in all strains (see Table S2). Comparable differences were observed for the 16S-ITS-23S rRNA operon copy number between genomes, from three (PMC649) to six copies (PMC627).

**A. gracile Clusters of Orthologous Genes**

In total 19,640 Open Reading Frames (ORFs) were predicted onto these genomes, with a number by strain comprised between 4,836 (PMC638) and 4,959 (PMC627) (Table 1). 93.26% (18,211) of these ORFs were gathered into 4,971 clusters of orthologous groups (COGs). The core genome, composed by genes shared by all four genomes, was represented by 3,832 COGs (77.08%), corresponding to 79.31% of PMC627 ORFs (Figure 1a-b). 13.56% and 8.99% of the COGs were shared exclusively between 3 and 2 genomes, respectively. Small numbers of COGs (between 1 to 8 by strain) were only composed by strain specific in-paralogs, but genomes displayed a higher number of strain specific single copy genes (singletons), between 212 to 306, representing up to 6.17% of the PMC627 gene content for instance. Given COGs and singletons, the current *A. gracile* pangenome was composed in total of 5,925 coding genes.
Figure 1. Comparative genomics between the four *A. gracile* strains. a) PMC627 *A. gracile* genome, with CDS positions (with NRPS/PKS highlighted in black), corresponding OG frequencies in four genomes, and synteny indices of genomic regions. b) Venn diagram of OG distribution between the four strains, with COGs (black labels), strain specific in-paralog clusters (white labels), and strain specific singletons. c) Percentages of PMC627 genome length represented by the different syntenic regions. d) Similarities of 16S-ITS-23S loci between strains and e) their structures.

**A. gracile** genome synteny

Considering gene triplet frames, the average synteny index was $80.55\pm1.31\%$ for one genome. The regions displaying a synteny index of 100% (conserved in all strains) represented $68.22\pm0.49\%$
of the genome lengths (see Figure 1.c for the PMC627 example). Genomic regions with synteny values of 66 and 33% represented 15.37 ±2.73% and 6.61±0.68% respectively of the genome length, and regions without synteny 9.79±2.20% (Figure 1.a, c). Genomic regions with synteny index <100% were evenly distributed along the genomes, and may gather up to 40 consecutive genes in PMC627. This distribution seems to share a pattern similar with the distribution along the genome of COGs frequency, *i.e.* the percentage of *A. gracile* genomes carrying them, what was confirmed by a Pearson’s Chi-square test (*H₀* independence hypothesis rejected with p-value<0.001, Figure 1.a).

**SSU ribosomal RNA encoding gene locus diversity**

Within a genome, no dissimilarity was observed among the different 16S rRNA or 23S rRNA copies. 16S rRNA sequences, with a conserved size of 1,485 bp, were identical between PMC638 and PMC649, but displayed a similarity of 99.86% in the other cases, corresponding to substitutions of 2 nucleotides (Figure 1.d). 23S rRNA sequences were strain specific, with a similarity varying between 99.36% and 99.78% between two strains, and a size of 2,715 bp for PMC627 and 2,824 bp for the three other genomes (Figure 1.d). Two types of internal transcribed spacers (ITS) were characterized in PMC627, 638 and 644: a short ITS sequence of 270 bp and a long one of 523 bp. This difference is related to the insertion in the long ITS type of a 240-bp region containing one tRNA-Ile and one tRNA-Ala supernumerary coding sequences. Both ITS types shared 97.24 % of similarity on their aligned regions, and were identical among strains. Only the short ITS sequence was detected in the PMC 649 genome.
Functional differences between core and flexible gene fractions

In total, 18,237 genes (92.85%) were associated to an orthologous group in EggNog bacterial database (4559.25±55.82 by genome), and 10,432 (53.11%) obtained a functional annotation. Proportions of orphans differed between core and flexible gene fractions, with 4.66±0.05% and 16.77±1.53 respectively (p-value<0.01, see Figure 2.a). Statistically significant differences were also detected for several functional categories between both fractions (p-values<0.01, see Figure 2.a).

Genes related to Energy production and conversion (C), Amino acid transport and metabolism (E), Carbohydrate transport and metabolism (G), Coenzyme transport and metabolism (H), Translation/ribosomal structure and biogenesis (J), Post-translational modifications (O) and Inorganic ion transport and metabolism (P), were over-represented in core-genome compared to flexible fraction. On the other hand, genes implied in replication/recombination and repair (L) and secondary metabolites biosynthesis, transport and catabolism (Q) represented a significantly higher proportion.
of genes in the flexible fraction, compared to the core-genome (Figure 2.a). Higher percentages of L-related genes in flexible fractions were associated with significantly higher proportions of prokaryotic defense system genes and transposases (Figure 2.b) while housekeeping functions related to DNA repair and recombination, messenger RNA biosynthesis and ribosome biogenesis were underrepresented compared to the core-genome (Figure 2.a).

Remarkably, the repertoires of biosynthetic gene clusters (BGC) specifically involved in secondary metabolite production were unique to each A. gracile strain, but always broader than those detected in the genomes of their close relatives (Figure 2.b). The A. gracile BGCs are notably involved in the production of saxitoxin (in PMC627 and 638, Figure 2.c), puwainaphycins (in PMC638, 644 and 649), anabaenopeptolins (in PMC627, 638 and 649), anacyclamides (in all strains), among others secondary metabolites. Beyond the occurrence of whole clusters in a single genome, divergences among strains also resulted to differences in their composition, as illustrated by the variability of COG frequencies within NRPS/PKS BGCs (Figure 1.a). The presence of transposases, and bacteriophage-related genes within clusters as exemplified by the saxitoxin BGC (Figure 2.c) could be involved in within BGCs recombination events.

Metabolomic profiles characterization

The high potential of secondary metabolite production was confirmed by direct metabolomic analysis of the biomass extracts by high resolution mass spectrometry (HRMS) and visualization with molecular networks (Figure 3.a,b), allowing for instance the detection of high amounts of saxitoxin and neosaxitoxin in both PMC627 and PMC638 strains (Figure 3.a). The global molecular network performed by GNPS (Figure 3.b) further highlights the production of a large set of metabolites by the four strains, including fatty acids, amino acids, nucleic acids, glycolipids, phospholipids, saccharides and carotenoids. It also reveals the production of various yet uncharacterized variants of puwainaphycins, in the three strains that specifically present puwainaphycin gene cluster. In total, in the 1400 detected analytes only 64 (4.57%) were synthesized in all cultures. The great majority of them (75.14%) were specific to a single strain, illustrating the large molecular diversity retrieved in the metabolome of the four strains beyond their genomic similarities and differences.
Figure 3. Metabolomic analysis. a) The saxitoxin and neosaxitoxin production in PMC627 and PMC638 illustrated by molecular network and MS/MS spectra observation is in agreement with sxt gene cluster observed in their respective genomes. b) The global molecular network of the four Aphanizomenon strains reveals the presence of various metabolite families and also illustrates the molecular diversity within these strains. c) Venn diagram of analytes detected in all A. gracile strains.

Gene content dissimilarity between closely related cyanobacteria strains

Among surveyed cyanobacterial genomes, 106 shared an ANI ≥ 99%, the minimal ANI observed among A. gracile strains, with at least one other genome (Figure 4). Within species variations can be limited in some species like Microcystis panniformis, M. flos-aquae or P. marinus, or extensive as for M. aeruginosa or more particularly for taxa like Fischerella thermalis or Aphanizomenon flos-aquae (both belonging to the Nostocales). Average gene content dissimilarity varied between 0.004 for Prochlorococcus marinus and 0.07 for Aphanizomenon gracile. Remarkably, this highest average value was observed among A. gracile genotypes that were sampled together from a single aquatic ecosystem and sampling point (and actually, drop of water).
Figure 4. Gene content dissimilarities between couples of genomes sharing ANI ≥99%, by cyanobacterial species. Red dots correspond to mean values. Values are displayed as white x when both compared cyanobacteria were sampled at the same location at the same time.

Genomic characterization of a novel cyanophage

Cyanophage MAG consisted in a single circular sequence of 120.7 kb coding for 165 proteins and displaying by far the greatest coverage of PMC627 metagenome (917.21X) from which it was assembled. Mapping of PMC627 metagenome reads on this MAG indicated a globally even coverage all along its sequence, with a mean value of 3.85 folds the PMC627 A. gracile MAG coverage (Figure 5.a). Four small genomic regions presented a greater coverage of ~5 to 8 folds the A. gracile genome mean coverage. These regions coded for transposases and one protein without annotation also present in several copies into the different PMC A. gracile genomes with a sequence similarity of 100% (Figure 5.a). Except for such regions, the coverage of the cyanophage MAG was globally null in the three other metagenomes, meaning that the cyanophage was undetectable in PMC638, PMC644 and PMC647.

Aphanizomenon phage vB_AphaS-CL131 genome (MG209611.1) was found to be the closest homolog (best hit) of this MAG, with an identity of 90.75% on 52% of the query length (dc-megablast, E-value=0.0, Figure 5.b). Both sequences shared an Average Nucleotide Identity (ANI) of 87.52%, and almost identical GC contents (39.6±0.05 % vs 39.7±0.05 %). At structural level, both displayed a depletion in CDS strand switching (Figure 5.a) characteristic of bacteriophage genomes,
and a well conserved synteny (Figure 5.b). Finally, 91 CDS were found to be homologous between the MAG and vB_AphaS-CL131 genome, representing 55.15% and 68.93% of Ag627 phage and (132 CDS on 112.7 kb) coding genes, respectively. Only 52 CDS of the MAG were affiliated to an orthologous group in the eggNOG database. As described for vB_AphaS-CL131, phage-specific functions (phage regulatory protein, terminase, tape measure, cell adhesion, virion morphogenesis, AntA/AntB antirepressor), as well as a CRISPR-associated protein (Cas_csx3) were detected by this approach, but no functions classically associated in relation to lysogeny were found (e.g. integrases, excisionases). Complementary phylogenetic analysis showed that both phages constitute a robustly supported clade, closely related to other Siphoviridae infecting Cyanobacteria (see Figure S1). These results led us to henceforth consider the MAG as a genome of a siphophage infecting *Aphanizomenon gracile* called vB_AphaS_Ag627.

![Figure 5.](image)

**Figure 5.** a) Genome of vB_AphaS_Ag627 cyanophage with CDS positions (with transposase coding sequences in red), genome coverage in four the metagenomes (divided by the respective mean coverage of cyanobacterial genome), and regions with 100% identity with the four *A. gracile* genomes. b) Megablast alignment between vB_AphaS_Ag627 and its closest relative. Links represent hits with an E-value<1E-05, color intensity depending of the hit length. c) Nucleotide distance and gene content dissimilarity between couples of bacteriophages genomes (from 1,834 genomes including 45 cyanophages and Ag627 phage). Dotted lines represent the two evolutionary modes described for bacteriophage genomes in\(^4\), High Gene Content Flux (HGCF) and Low Gene Content Flux (LGCF).

**The evolutionary mode of phage vB_AphaS_Ag627**

Comparison of nucleotide distance and gene content dissimilarity between available bacteriophage genomes resulted in a plot where couples of genomes were distributed following two modalities (Figure 5.c). One group of genome pairs were dispersed along a straight line, meaning that gene content dissimilarity between members was proportional to their nucleotide distance. This distribution is typical for genomes having essentially diverged from one another by mutation events
and was called Low Gene Content Flux evolutionary mode (LGCF)\textsuperscript{44}. All genome pairs of the 74 cyanophage genomes, lytic for the great majority (97.3%), showed a LGCF distribution. The second group of genome pairs formed a hyperbole, indicating a disproportionately high gene content dissimilarity compared to their nucleotidic distances. This is typical of a genome divergence enhanced by acquisition and/or loss of genes by Horizontal Gene Transfer (HGT). This evolutionary mode called High Gene Content Flux (HGCF)\textsuperscript{44} is specific to lysogenic replicative bacteriophages. vB\_AphaS\_Ag627 and CL131 genomes showed a nucleotidic distance of 0.10 and a gene content dissimilarity of 0.40, which corresponds to the HGCF evolutionary mode. Both these genomes displayed high nucleotide distances (> 0.45) with all other surveyed bacteriophage genomes, out of the range of confident values for such analysis\textsuperscript{44}.

**A. gracile** cyanosphere metagenome-based community structure

### Table 2. Cyanosphere associated MAGs statistics and taxonomic annotations

| MAG     | Size (Mb) | Completeness CheckM (%) | TAXON         |
|---------|-----------|--------------------------|---------------|
| Ac-Ac-01 | 2.39      | 100                      | Actinobacteria | Actinomycetales |
| Al-Ca-02 | 4.44      | 100                      | Alpha-proteobacteria | Caulobacteriales |
| Al-Rb-03 | 4.14      | 100                      | Alpha-proteobacteria | Rhodobacteriales |
| Al-Rb-04 | 3.45      | 100                      | Alpha-proteobacteria | Rhodobacteriales |
| Al-Rb-05 | 2.02      | 40.28                    | Alpha-proteobacteria | Rhodobacteriales |
| Al-Rb-06 | 3.60      | 100                      | Alpha-proteobacteria | Rhodobacteriales |
| Al-Rs-07 | 3.92      | 100                      | Alpha-proteobacteria | Rhodobacteriales |
| Al-Sm-08 | 2.71      | 95.83                    | Alpha-proteobacteria | Rhodobacteriales |
| Al-Sm-09 | 3.25      | 91.67                    | Alpha-proteobacteria | Rhodobacteriales |
| Al-Sm-10 | 3.13      | 100                      | Alpha-proteobacteria | Rhodobacteriales |
| Al-Sm-11 | 1.86      | 75                       | Alpha-proteobacteria | Rhodobacteriales |
| Ba-Ch-12 | 5.21      | 100                      | Bacteroidetes   | Chitinophagales  |
| Ba-Cy-13 | 5.35      | 100                      | Bacteroidetes   | Chitinophagales  |
| Ba-Cy-14 | 3.45      | 95.83                    | Bacteroidetes   | Chitinophagales  |
| Ba-Fb-15 | 0.71      | 38.02                    | Bacteroidetes   | Flavobacteriales |
| Ba-Sa-16 | 4.43      | 100                      | Bacteroidetes   | Flavobacteriales |
| Ba-Sb-17 | 0.60      | 25                       | Bacteroidetes   | Flavobacteriales |
| Ba-Un-18 | 0.11      | 0                        | Bacteroidetes   | Flavobacteriales |
| Be-Bu-19 | 4.12      | 100                      | Beta-proteobacteria | Burkholderiales |
| Ge-Ge-20 | 3.74      | 59.33                    | Gemmatimonadetes | Gemmatimonadales |

Twenty unique MAGs, including 11 complete and three nearly complete non-phototrophic bacterial genomes were characterized (Table 2). The related bacteria belonged to five taxa, namely the Actinobacteria, Bacteroidetes, Alpha- and Betaproteobacteria and Gemmatimonadetes. Except for Ba-Un-18, for which no related reads have been found in PMC627 and PMC644, all other MAGs
were identified in all metagenomes, with mean coverages greater than one (Table S2). Despite the similar MAG richness, α-diversity varied among the four cyanospheres, with Shannon indices comprised between 0.90 for PMC644 and 2.11 for PMC649, related to variations in MAG relative abundances among cultures (Table S2). Indeed, Actinobacteria (Actinomycetales), Betaproteobacteria and Gemmatimonadetes (Gemmatimonadales), each represented by a unique MAG (Table 2, Figure 6.a), were only abundant in PMC649 metagenome. Alphaproteobacteria and Bacteroidetes taxa, which gathered 10 and 8 MAGs, respectively (Table 2), displayed high relative abundances in all cyanosphere communities, despite variations (Figure 6.a). Alphaproteobacterial communities were highly divergent, and dominated by one to four MAGs with reads proportion greater than 10% (Figure 6.b). Only one MAG belonging to the Sphingomonadales order presented proportions greater than 10% in three of the four metagenomes. Bacteroidetes community structures displayed similar variability, with only three MAGs (belonging to orders Cytophagales, Flavobacterales and Saprospirales), representing more than 10% of Bacteroidetes in more than two metagenome communities. Mean MAG coverage ratio between non-phototrophic bacteria and the A. gracile genome, used as a crude proxy for cell densities on cyanobacterial filaments, also showed high variability among cultures. PMC627 displayed the lowest ratio for all heterotrophic taxa (Figure 6.c) with 1.24 Alphaproteobacteria and 0.40 Bacteroidetes genomes per cyanobacterial genome. Despite similar community structures in terms of taxa relative abundance between PMC638 and PMC644, these ratios were much greater in the latter for Alphaproteobacteria (4.16) and Bacteroidetes (7.76) compared to the former (2.78 and 3.45, respectively). PMC649 showed the most atypical profile. It was the only sample displaying a noticeable ratio of Actinobacteria (0.87), Betaproteobacteria (1.63) and Gemmatimonadales (0.53). PMC649 also showed the highest ratio of Alphaproteobacteria, with 6.63 genomes per A. gracile genome.
Figure 6. Taxonomic and functional study of the cyanosphere. a) Relative reads abundance of each taxa composing the cyanosphere. b) Heatmap illustrating the relative reads abundance of each MAG in Alphaproteobacteria and Bacteroidetes. Black squares correspond to not detected MAG. c) Ratio between taxa genomes coverages and *A. gracile* genome coverage by sample. d) Principal Coordinate Analysis based on KO presence/absence in metagenomic sequences, classified following their taxonomical annotations. Dotted ellipses correspond to confidence intervals for each taxa. e) Functional analysis of Alphaproteobacteria, Bacteroidetes and Cyanobacteria pangenomes. Heatmap represent KEGG module completeness in the three taxa for all functional categories (outer black circle). Cells length is proportional to the number of KO composing a module. Z-scores are calculated by column. Significantly greater completion in one of the three taxa is indicated by taxon-colored cell. (pairwise Student test with Bonferroni correction, p-value<0.01). Functional categories abbreviations: Nucl. – Nucleotide; Glyc – Glycans; Cofactors/Vit. –
Functional comparative metagenomics

In total, genetic communities were constituted by 12,520 to 28,626 genes (19,527±7,405) depending of the *A. gracile* strains, including 40.0 to 46.6% that received a KO affiliation. Genes from *Aphanizomenon*, Alphaproteobacteria and Bacteroidetes, represented 99.6 to 100% of the function annotated gene set except for PMC649 culture with where they only reached 61.8%. Considering KEGG orthologs (KO), 2,045 were from cyanobacterial genomes, 3,704 from Alphaproteobacteria and 2,621 from Bacteroidetes. In terms of frequencies, 90.31% of cyanobacterial KO were common to all *A. gracile* cultures, versus 50.45% for Alpha-proteobacteria and 42.69% for Bacteroidetes (data not shown). Principal Coordinates Analysis (PCoA) based on KO presence/absence displayed the taxon-related fractions of all metagenomes in distinct well-defined clusters (Figure 6.c). These clusters clearly separated Alphaproteobacteria, Bacteroidetes and Cyanobacteria groups, indicating divergences of their functional gene contents. Interestingly, the four points corresponding to the Cyanobacteria were more tightly packed together than those corresponding to Alphaproteobacteria and Bacteroidetes, suggesting less variation within the *Aphanizomenon* strains cluster than within the different associated Alphaproteobacteria or Bacteroidetes. An analysis of KEGG modules completion was performed, focused on these three taxa for statistical and exhaustiveness reasons because they were all present in all metagenomes and gathered the great majority of KO annotated genes (Figure 6.d). In total, KO were distributed into 289 KEGG modules, among which 106 showed significant differences in module completion between the three taxa (ANOVA, p-value<0.01, data not shown), confirming divergence observed in PCoA at the KO level.

Among all 289 functional modules retrieved in the three taxa abundant in all samples, namely Cyanobacteria, Alphaproteobacteria and Bacteroidetes, *Aphanizomenon* showed the lowest number of complete or nearly complete modules (Z-score closed to 1, Figure 6.d). Cyanobacterial protein functions have been associated to 164 modules, in which only 12 displayed significantly greater completion in Cyanobacteria compared to other taxa (pairwise Student test with Bonferroni correction, p-value<0.01). Half of them were related to “Energy Metabolism” and involved in nitrogen fixation, photosystems and ATP synthesis (Figure 6.d, Table S3). The other modules were associated with “Lipid Metabolism” (lactosylceramide biosynthesis), “Cofactors and Vitamins” (ubiquinone and tocopherol/tocorterol biosynthesis), “Terpenoids/Polyketides” (beta-carotene biosynthesis), “Other secondary metabolites biosynthesis” (staphyloferrin B synthesis) and “Gene set” functional categories (vancomycin resistance). Bacteroidetes KO were assigned to a larger
repertoire than Cyanobacteria, with a number of 227 modules, but their completion levels were generally low, especially in “Energy metabolism”, “Biosynthesis of other secondary metabolites”, “Xenobiotics biodegradation”, and “Gene set” functional categories (Figure 6.d). Only three modules showed a significant greater completion in Bacteroidetes, involved in citrate cycle (“Carbohydrates metabolism”), coenzyme M biosynthesis (“Energy metabolism”) and tetrahydrofolate biosynthesis (“Metabolism of cofactor and vitamins”, Figure 6.d, Table S3). Lastly, Alphaproteobacteria showed the broadest range of functional capabilities with KO associated with 271 functional modules equally distributed into 11 functional categories, displaying the highest Z-scores (Figure 5.d). Significant greater completions were found for 27 functional Alphaproteobacteria modules, most of them being related to “Energy metabolism”, with five modules involved in ATP synthesis and three in carbon fixation pathways. As observed in Cyanobacteria, Alphaproteobacteria displayed two modules with highest completion involved in “Biosynthesis of cofactors and vitamins”, namely fumarate and ubiquinone biosynthesis. These bacteria also showed the highest number of modules with significant greater completion in “Biosynthesis of terpenoids and polyketides” (enediyne, Type II polyketide backbone), “Biosynthesis of other secondary metabolites” (rebeccamycin and prodigiosin), all of them referred to be involved in potent antibacterial activities, as well as in “Xenobiotics biodegradation” (benzoate degradation), and gene set functional categories (Type 3 secretion system and effectors, cagA pathogenicity island signature, vancomycin and multidrug resistances) (Figure 6.d, Table S3).

DISCUSSION

Evaluation of micro-scale biodiversity within a single cyanobacterial population is a cornerstone for the establishment of a holistic theoretical framework of cyanoHAB bloom formation. In this study, we investigated the micro-scale genomic heterogeneity within Aphanizomenon gracile by comparing strains isolated from a single drop of freshwater during a bloom event, and the taxonomic and functional diversity of associated cyanospheres. We present the first sequenced genomes of Aphanizomenon gracile, and we show that strains isolated from a single drop of water display unexpectedly high genetic diversity, despite a constrained genome size (5.33±0.06 Mb) and the absence of self-evident differentiation at the small subunit 16S rRNA gene sequence and ANI levels.

A striking representation of the micro-scale A. gracile genomic heterogeneity is given by the low synteny scores, only slightly higher than those reported among 12 Microcystis aeruginosa strains sampled from distant geographical locations (80.55±1.31% vs 76±4%, respectively). Remarkably, low synteny in an A. gracile is closely associated with the presence of genes specific to this strain throughout its genome, rather than intra-genomic rearrangements, suggesting a strong impact of HGT – and potential gene loss, in the Aphanizomenon genomic plasticity, as already depicted in M.
The resulting gene content dissimilarities appear to be the highest when compared to those observed between pairs of closely-related cyanobacterial strains isolated from a single site and sampling time and could reflect the high frequency of HGT events occurring locally in a single blooming population, which is largely overlooked in comparative genomics studies. If genetic drift and/or selection under culture conditions may contribute to genomic variability among strains, it has probably a limited effect at the genome scale. Indeed, Perez-Carrascal et al.\textsuperscript{10} recently claimed that \textit{M. aeruginosa} strains re-sequenced after years in culture only differed from older genomes by a few point mutations. Furthermore, given the broad extent of the \textit{A. gracile} pangenome compared to the constrained genome size, it appears very unlikely that the genomic divergence observed among strains simply results from gene losses from a common ancestor.

Our analysis confirms that an \textit{Aphanizomenon} genome can carry two types of rRNA operons (\textit{rrn} operon)\textsuperscript{52,53}, mainly diverging by the insertion of two tRNA coding sequences in the longest ITS sequence. Intragenomic variability between ITS sequences is well known in Bacteria and is mediated by homologous recombination of sequence blocks\textsuperscript{54–57}. More interestingly, the number of total \textit{rrn} operon can vary among \textit{A. gracile} strains. This observation should lead to caution studies of \textit{Aphanizomenon}-associated community structures based on 16S or ITS amplicon-based approach alone, and questions the feasibility of a gene copy number catalogue for cyanobacteria, required for the improvement of gene copy normalization methods developed for meta-taxonomic surveys\textsuperscript{58}. The number of \textit{rrn} operon copies has been positively correlated with bacterial fitness, resource availability and utilization\textsuperscript{59}. Eutrophication of water bodies could thus promote the multiple occurrences of \textit{rrn} operons in \textit{A. gracile} genomes, which would allow cyanobacteria for a faster nutrient metabolization. Together with gene content dissimilarity, these results suggest that strains co-occurring within a single \textit{A. gracile} bloom could employ different metabolic strategies, illustrated by individual metabolic profiles, possibly linked to distinct ecological adaptations to stochastic environmental variations.

On the other hand, these divergences, mostly due to different secondary metabolites biosynthesis capabilities associated with the presence/absence of biosynthetic gene clusters as well as recombination events within BCGs, could reflect a cyanobacteria population ecology compatible with the black queen hypothesis\textsuperscript{60}. Indeed, although suspected to be involved in essential physiological and ecological traits (e.g. saxitoxins would play a role in Na\textsuperscript{+} homeostasis)\textsuperscript{61} biosynthetic clusters are costly in terms of genomic maintenance and resources for metabolite production. Dynamic losses and gains of BCGs, already known to be mediated by HGTs in cyanobacteria\textsuperscript{10,62}, would maintain an uneven distribution of these clusters among the members of the population, resulting to a decrease of resources requirements at the individual level, whereas the global population still benefits from all synthesized secondary metabolites. The fact that local HGTs in \textit{A. gracile} population seem to occur
more frequently than in less intensively blooming taxa, in the light of our results, supports that a bloom would constitute a hot-spot for cyanobacterial inter-genomic recombination and genotypic diversification, potentially driven by bacteriophages.

Indeed, except for secondary metabolites biosynthetic genes, flexible gene fractions are essentially enriched in genes coding for “Orphans, transposases and defense mechanisms”, all of these categories being usually associated with selfish genetic elements such as prophages. Disseminated virus-originating genes are common in cyanobacterial genomes, leading to consider cyanophages as key drivers of cyanobacteria genomic plasticity. The phage vB_AphaS-Ag627 identified in this study is clearly involved in this process, given the several identical transposase-coding genes it shares with the four *A. gracile* genomes, although its real impact remains elusive. Phage vB_AphaS-Ag627 is phylogenetically related to the vB_AphaS_CL131 siphophage, infecting *A. flos-aquae* in the Baltic sea. Together, they show a pattern of divergence typical of temperate bacteriophages. The lack of clues for a prophage integration into the *Aphanizomenon* genome could then reflect the maintenance of the viral genome under an episomic form into the infected cell, as it is actually observed for a growing number of lysogenic bacteriophages. This hypothesis is supported by the presence of AntA/AntB antirepressor, involved in temperate phages in the decision mechanism between lysis and lysogeny. Thus, it is likely that vB_AphaS_Ag627 lysogeny provides adaptative traits to infected *A. gracile* cells, but the low performance of functional annotation step do not allow to characterize them. At least, the presence of CRISPR associated protein suggests that vB_AphaS_Ag627 could play a role in the resistance against viral over-infections. vB_AphaS_Ag627 is probably able to replicate under culture conditions, since its genome coverage was found to be ~3.8 greater than *A. gracile* PMC627 one, although the cell culture never collapsed. But the absence of this cyanophage in all other cultures questions the levels of sensitivity and resistance of each individual strain composing the native population, and complementary work is required to better understand the consequences of differential cyanophage infection of distinct strains on natural *A. gracile* population dynamics.

Heterotrophic bacteria characterized in *A. gracile* cultures belong to taxa already reported to be commonly associated with cyanobacteria in culture or in the environment. Among them, Actinomycetales, Cytophagales and Rhodobacterales are reported to gather members able to lyse some cyanobacteria and are suspected to play a major role in bloom termination. However, their lysis capability is depending of the cyanobacterial species and experiments are required to test their ability to predate *Aphanizomenon*. Furthermore, other interactions with *A. gracile* occurring during its growth could also sustain these associations within the phycosphere. Some others, as Saprospirales and Sphingomonadales, have been shown to grow on cyanobacterial exudates, including secondary metabolites, or to rapidly increase in abundance after bloom collapse in lakes.
Sphingobacterales\textsuperscript{72}. This suggests that such bacteria are recruited in the phycosphere by cyanobacteria supply of metabolites. If all but one of these bacteria are detectable in every of the four cya
nospheres, suggesting host fidelity, their relative abundances fluctuate greatly from one to another. These variations could be at least partially related to the different metabolic profiles displayed by \textit{A. gracile} strains. Indeed, public good exchanges, partly mediated by extracellular vesicles\textsuperscript{77}, are thought to structure the association between cyanobacteria and their cya
nospheres\textsuperscript{12,14,15}. If the community structures differ among cya
nospheres, they seem to be largely functionally redundant. Functional module completeness analysis suggests a reduced metabolic capability range in \textit{Aphanizomenon}, compared to Bacteroidetes and mostly Alphaproteobacteria, but this striking difference could simply reflect the bias inherent to functional databases which are based on well-known model micro-organisms highly divergent from cyanobacteria. However, in addition to products from photosynthesis and nitrogen fixation, specifically performed by \textit{A. gracile}, which are already thought to be shared with the cyanosphere\textsuperscript{14}, many secondary metabolites, vitamins (B2, B3 and B6 or E) and cofactors, mostly synthesized by Cyanobacteria and Alphaproteobacteria could also constitute public goods structuring the community. In contrast, only one complete module responsible for the biosynthesis of tetrahydrofolate, a cofactor involved in nucleotide metabolism, is found in Bacteroidetes.

CONCLUSIONS

In conclusion, exploration of four strains of \textit{A. gracile} co-existing during a bloom event revealed levels of inter-individual genomic variability higher than previously recorded among conspecific strains of cyanobacteria. Although limited by the number of studied metagenomes, this study suggests that an even larger size of the \textit{A. gracile} pangenome may exist at a single bloom scale, and that a strong spatial heterogeneity of cyanobacterial genotypes and associated cya
nospheres may also occur, questioning the cyanobacterial population dynamics and numerous bacterial interactions leading to such events. A consequence is that genome sequencing of many strains is mandatory to gain proper overview of the potential of a whole population. This genotypic diversification is likely mostly mediated by temperate phages, including vB\textunderscore AphaS\textunderscore Ag627, the one newly identified during this study. Beyond the diversity statement, further experimental setups should further explore biotic interactions occurring on eutrophic lakes to decipher the mechanisms that regulate cyanoHAB populations until bloom collapse.

REFERENCES
1. Cires, S. & Ballot, A. A review of the phylogeny, ecology and toxin production of bloom-forming Aphanizomenon spp. and related species within the Nostocales (cyanobacteria). *Harmful Algae* **54**, 21–43 (2016).

2. Escalas, A. *et al.* Drivers and ecological consequences of dominance in periurban phytoplankton communities using networks approaches. *Water Res* **163**, 114893 (2019).

3. Bernard, C. *et al.* Cyanobacteria associated with the production of cyanotoxins. in *Handbook of Cyanobacterial Monitoring and Cyanotoxin Analysis* (ed. Wiley edition) (2017).

4. Moisander, P. H., Paerl, H. W. & Zehr, J. P. Effects of inorganic nitrogen on taxa-specific cyanobacterial growth and nifH expression in a subtropical estuary. *Limnol. Oceanogr.* **53**, 2519–2532 (2008).

5. Yepremian, C. *et al.* Microcystin ecotypes in a perennial Planktothrix agardhii bloom. *Water Res* **41**, 4446–56 (2007).

6. Kardinaal, W. E. A. *et al.* Competition for light between toxic and nontoxic strains of the harmful cyanobacterium Microcystis. *Appl. Environ. Microbiol.* **73**, 2939–2946 (2007).

7. Sabart, M. *et al.* Genetic diversity along the life cycle of the cyanobacterium Microcystis: highlight on the complexity of benthic and planktonic interactions. *Environ. Microbiol.* **17**, 901–911 (2015).

8. Kashtan, N. *et al.* Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science* **344**, 416–20 (2014).

9. Jackrel, S. L. *et al.* Genome evolution and host-microbiome shifts correspond with intraspecific niche divergence within harmful algal bloom-forming Microcystis aeruginosa. *Mol. Ecol.* **28**, 3994–4011 (2019).

10. Pérez-Carrascal, O. M. *et al.* Coherence of Microcystis species revealed through population genomics. *ISME J.* **13**, 2887–2900 (2019).

11. Kim Tiam, S. *et al.* Insights into the Diversity of Secondary Metabolites of Planktothrix Using a Biphasic Approach Combining Global Genomics and Metabolomics. *Toxins* **11**, (2019).
12. Paerl, H. W. & Kellar, P. E. Specific associations of blue-green algae Anabaena and Aphanizomenon with bacteria in freshwater blooms. *Journal of Phycology* - Wiley Online
   https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1529-8817.1976.tb02867.x (1976).
13. Morris, J. J., Johnson, Z. I., Szul, M. J., Keller, M. & Zinser, E. R. Dependence of the cyanobacterium Prochlorococcus on hydrogen peroxide scavenging microbes for growth at the ocean’s surface. *PLoS One 6*, e16805 (2011).
14. Cook, K. V. *et al.* The global Microcystis interactome. *Limnol. Oceanogr.* 65, S194–S207 (2020).
15. Morris, J. J., Kirkegaard, R., Szul, M. J., Johnson, Z. I. & Zinser, E. R. Facilitation of robust growth of Prochlorococcus colonies and dilute liquid cultures by ‘helper’ heterotrophic bacteria. *Appl Env. Microbiol* 74, 4530–4 (2008).
16. Ahlgren, N. A., Perelman, J. N., Yeh, Y.-C. & Fuhrman, J. A. Multi-year dynamics of fine-scale marine cyanobacterial populations are more strongly explained by phage interactions than abiotic, bottom-up factors. *Environ. Microbiol.* 21, 2948–2963 (2019).
17. Cairns, J., Coloma, S., Sivonen, K. & Hiltunen, T. Evolving interactions between diazotrophic cyanobacterium and phage mediate nitrogen release and host competitive ability. *R. Soc. Open Sci.* 3, 160839.
18. Ledreux, A. *et al.* Evidence for saxitoxins production by the cyanobacterium *Aphanizomenon gracile* in a French recreational water body. *Harmful Algae* 10, 88–97 (2010).
19. Rippka, R. Isolation and purification of cyanobacteria. *Methods Enzym.* 167, 3–27 (1988).
20. Parada, A. E., Needham, D. M. & Fuhrman, J. A. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Env. Microbiol* 18, 1403–14 (2016).
21. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19, 455–77 (2012).
22. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* **13**, e1005595 (2017).

23. Lin, H. H. & Liao, Y. C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci Rep* **6**, 24175 (2016).

24. von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol* **20**, 217 (2019).

25. Bengtsson-Palme, J. et al. METAXA2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol Ecol Resour* **15**, 1403–14 (2015).

26. Pruesse, E., Peplies, J. & Glockner, F. O. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–9 (2012).

27. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–10 (1990).

28. Chen, K. T. et al. CSAR: a contig scaffolding tool using algebraic rearrangements. *Bioinformatics* **34**, 109–111 (2018).

29. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–2 (2015).

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

31. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–9 (2012).

32. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).

33. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
34. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309–D314 (2019).

35. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639–45 (2009).

36. Seemann, T. *Bacterial ribosomal RNA predictor*. (2020).

37. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).

38. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019).

39. Oksanen, J. F. et al. vegan: Community Ecology Package. R package. (2019).

40. Blin, K. et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).

41. Médigue, C. et al. MicroScope—an integrated resource for community expertise of gene functions and comparative analysis of microbial genomic and metabolic data. *Brief. Bioinform.* **20**, 1071–1084 (2017).

42. Wintersinger, J. A. & Wasmuth, J. D. Kablammo: an interactive, web-based BLAST results visualizer. *Bioinformatics* **31**, 1305–6 (2015).

43. Yoon, S. H., Ha, S. M., Lim, J., Kwon, S. & Chun, J. A large-scale evaluation of algorithms to calculate average nucleotide identity. *Antonie Van Leeuwenhoek* **110**, 1281–1286 (2017).

44. Mavrich, T. N. & Hatfull, G. F. Bacteriophage evolution differs by host, lifestyle and genome. *Nat Microbiol* **2**, 17112 (2017).

45. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132 (2016).
46. Le Manach, S. et al. Global Metabolomic Characterizations of Microcystis spp. Highlights Clonal Diversity in Natural Bloom-Forming Populations and Expands Metabolite Structural Diversity. Front. Microbiol. 10, 791 (2019).

47. Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat. Biotechnol. 34, 828–837 (2016).

48. Kim, M., Oh, H.-S., Park, S.-C. & Chun, J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int. J. Syst. Evol. Microbiol. 64, 346–351 (2014).

49. Humbert, J.-F. et al. A tribute to disorder in the genome of the bloom-forming freshwater cyanobacterium Microcystis aeruginosa. PloS One 8, e70747 (2013).

50. Kettler, G. C. et al. Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. PLoS Genet. 3, e231 (2007).

51. Coleman, M. L. et al. Genomic islands and the ecology and evolution of Prochlorococcus. Science 311, 1768–1770 (2006).

52. Laamanen, M. J., Forsström, L. & Sivonen, K. Diversity of Aphanizomenon flos-aquae (cyanobacterium) populations along a Baltic Sea salinity gradient. Appl. Environ. Microbiol. 68, 5296–5303 (2002).

53. Park, H.-K. et al. Molecular Verification of Bloom-forming Aphanizomenon flos-aquae and Their Secondary Metabolites in the Nakdong River. Int. J. Environ. Res. Public. Health 15, (2018).

54. Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V. & Polz, M. F. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J. Bacteriol. 186, 2629–2635 (2004).

55. Lan, R. & Reeves, P. R. Recombination between rRNA operons created most of the ribotype variation observed in the seventh pandemic clone of Vibrio cholerae. Microbiol. Read. Engl. 144 (Pt 5), 1213–1221 (1998).
56. Tambong, J. T., Xu, R. & Bromfield, E. S. P. Intercistronic heterogeneity of the 16S-23S rRNA spacer region among Pseudomonas strains isolated from subterranean seeds of hog peanut (Amphicarpa bracteata). *Microbiol. Read. Engl.* **155**, 2630–2640 (2009).

57. Gürtler, V. The role of recombination and mutation in 16S-23S rDNA spacer rearrangements. *Gene* **238**, 241–252 (1999).

58. Starke, R., Pylro, V. S. & Morais, D. K. 16S rRNA Gene Copy Number Normalization Does Not Provide More Reliable Conclusions in Metataxonomic Surveys. *Microb. Ecol.* (2020) doi:10.1007/s00248-020-01586-7.

59. Klappenbach, J. A., Dunbar, J. M. & Schmidt, T. M. rRNA operon copy number reflects ecological strategies of bacteria. *Appl. Environ. Microbiol.* **66**, 1328–1333 (2000).

60. Morris, J. J., Lenski, R. E. & Zinser, E. R. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *mBio* **3**, (2012).

61. Holland, A. & Kinnear, S. Interpreting the Possible Ecological Role(s) of Cyanotoxins: Compounds for Competitive Advantage and/or Physiological Aide? *Mar. Drugs* **11**, 2239–2258 (2013).

62. Calteau, A. *et al.* Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. *BMC Genomics* **15**, (2014).

63. Broecker, F. & Moelling, K. Evolution of Immune Systems From Viruses and Transposable Elements. *Front. Microbiol.* **10**, 51 (2019).

64. Daubin, V. & Ochman, H. Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. *Genome Res.* **14**, 1036–1042 (2004).

65. Cortez, D., Forterre, P. & Gribaldo, S. A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaean and bacterial genomes. *Genome Biol.* **10**, R65 (2009).
66. Šulčius, S. et al. Genomic Characterization of Cyanophage vB_AphaS-CL131 Infecting Filamentous Diazotrophic Cyanobacterium Aphanizomenon flos-aquae Reveals Novel Insights into Virus-Bacterium Interactions. *Appl. Environ. Microbiol.* **85**, (2019).

67. Brolund, A., Franzén, O., Melefors, O., Tegmark-Wisell, K. & Sandegren, L. Plasmidome-analysis of ESBL-producing escherichia coli using conventional typing and high-throughput sequencing. *PloS One* **8**, e65793 (2013).

68. Xue, H. et al. Eco-Evolutionary Dynamics of Episomes among Ecologically Cohesive Bacterial Populations. *mBio* **6**, e00552-00515 (2015).

69. Ravin, N. V., Svarchevsky, A. N. & Dehò, G. The anti-immunity system of phage-plasmid N15: identification of the antirepressor gene and its control by a small processed RNA. *Mol. Microbiol.* **34**, 980–994 (1999).

70. Driscoll, C. B., Otten, T. G., Brown, N. M. & Dreher, T. W. Towards long-read metagenomics: complete assembly of three novel genomes from bacteria dependent on a diazotrophic cyanobacterium in a freshwater lake co-culture. *Stand. Genomic Sci.* **12**, (2017).

71. Parveen, B. et al. Bacterial communities associated with Microcystis colonies differ from free-living communities living in the same ecosystem. *Environ. Microbiol. Rep.* **5**, 716–724 (2013).

72. Guedes, I. A. et al. Close Link Between Harmful Cyanobacterial Dominance and Associated Bacterioplankton in a Tropical Eutrophic Reservoir. *Front. Microbiol.* **9**, (2018).

73. Zhu, L. et al. Bacterial Communities Associated with Four Cyanobacterial Genera Display Structural and Functional Differences: Evidence from an Experimental Approach. *Front. Microbiol.* **7**, (2016).

74. Chen, H. et al. Induction and resuscitation of the viable but nonculturable state in a cyanobacteria-lysing bacterium isolated from cyanobacterial bloom. *Microb. Ecol.* **63**, 64–73 (2012).

75. Rashidan, K. K. & Bird, D. F. Role of Predatory Bacteria in the Termination of a Cyanobacterial Bloom. *Microb. Ecol.* **41**, 97–105 (2001).
76. Ishii, H., Nishijima, M. & Abe, T. Characterization of degradation process of cyanobacterial hepatotoxins by a gram-negative aerobic bacterium. *Water Res.* **38**, 2667–2676 (2004).

77. Biller, S. J. *et al.* Bacterial Vesicles in Marine Ecosystems. *Science* **343**, 183–186 (2014).

**DECLARATIONS**

**Ethics approval and consent to participate**
Not applicable

**Consent for publication**
Not applicable

**Availability of data and material**
PMC strains are available upon request (www.mnhn.fr/en/collections/collection-groups/biological-resources-living-and-cryopreserved-cells). Raw reads were deposited into the GENBANK Sequence Read Archive (SRA) database under the BioProject PRJNA693796 (samples SAMN17438945 to SAMN17438948).

**Competing interests**
The authors declare that they have no competing interests

**Funding**
This work was supported by the ATM grant “ECOL-Cyanomique” from the AVIV department of the MNHN, Paris. SKT was funded by an ATER grant from the MNHN.

**Authors’ contributions**
CD cultured strains and extracted DNA. BM and SH designed the project. SH, SD, SKT, ED, CB and BM carried out the analyses. SH drafted the manuscript.

**Acknowledgements**
SH would like to thank Claude Yéprémian, for sharing his extensive knowledge about cyanobacteria. The authors thank the “Cyanobacteria and Eukaryotic Microalgae Collections” and the “Plateau technique de Spectrométrie de Masse Bio-organique” (UMR 7245 MCAM) of the MNHN, Paris, France.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.pdf