Microbial diversity and toxin risk in tropical freshwater reservoirs of Cape Verde

Ana P. Semedo-Aguiar,[1,2,3] José B. Pereira-Leal,[1] Ricardo B. Leite[1*]

[1] Instituto Gulbenkian de Ciência, Rua da Quinta Grande nº6, 2780-156 Oeiras, Portugal

[2] Programa de Pós-Graduação Ciência para o Desenvolvimento (PGCD)

[3] Universidade Jean-Piaget de Cabo Verde

Running Head: Toxin risk survey in Cape Verdean reservoirs

* Corresponding author

E-mail: rleite@igc.gulbenkian.pt (RBL)

ORCID iD: 0000-0002-9622-3895

Abstract

The Cape Verde islands are part of the African Sahelian arid belt that possesses an irregular rainy season between August and October. This erratic rain pattern has prompted the need for water reservoirs, now critical for the country's sustainability. Worldwide, freshwater cyanobacterial blooms are increasing in frequency due to global climate change and eutrophication of water bodies, particularly in reservoirs. To date there have been no risk assessments of cyanobacterial toxin production in these man-made structures. We evaluated this potential risk using 16S rRNA gene amplicon sequencing and full metagenome sequencing in freshwater reservoirs of Cape Verde.

Our analysis revealed the presence of several potentially toxic cyanobacterial genera in all sampled reservoirs (Poilão, Saquinho and Faveta). In Faveta *Microcystis* sp., a genus well known for toxin production and bloom-formation, dominated our samples, while a green algae of the genus *Cryptomonas* and *Gammaproteobacteria* dominated Saquinho and Poilão.

Taking advantage of the dominance of *Microcystis* in the Faveta reservoir, we were able to reconstruct and assemble its genome, extracted from a metagenome of bulk DNA from Faveta water. We named it *Microcystis* cf. *aeruginosa* CV01, for which a phylogenetic analysis revealed to have a close relationship with other genomes from those taxa, as well as other continental African strains, suggesting geographical coherency. In addition, it revealed several clusters of known toxin-producing genes.

This assessment of Cape Verdean freshwater microbial diversity and potential for toxin production reinforces the need to better understand the microbial ecology as a whole of water reservoirs on the rise.

Introduction

The available freshwater in the African archipelago of Cape Verde (DMS coordinates 15°07'12.51" N, 23°36'18.62" W) does not cover its needs. In addition, overexploitation and saline intrusion can impair the quality of groundwater [1]. Several water storage structures are being planned around the country in order to collect storm water needed to increase irrigated areas and modernize agriculture. However, these waters can carry nutrients of natural and anthropogenic origin, creating conditions for eutrophication and exponential growth of microalgae. These algal blooms have deleterious impacts on public health, water quality, and environmental issues, as well as economic costs due to bottom anoxia, release of noxious products, and toxic metabolites [2,3]. Actually, these events are occurring more frequently worldwide, and it is thought that global climatic changes are a major contributor to this problem [4–6].

To evaluate the risk of occurrence of algal blooms in freshwater bodies it is important to characterize their microbial composition. In non-eutrophic freshwater systems, the most commonly abundant bacterial groups are, in order of decreasing relative abundance, *Actinobacteria*, *Bacteroidetes*, *Proteobacteria* (Alpha, Beta and Gamma clades), *Verrucomicrobia* and *Cyanobacteria* [7–12]. In contrast, eutrophic waters contain microbial communities that include large numbers of cyanobacteria,

60 which are able to produce toxins and foul odors and discolor the water [13,14]. The
61 cyanobacterial phylum has many genera that produce toxins, also called cyanotoxins,
62 and in freshwater bodies, toxic and non-toxic strains can co-exist and dominate at
63 different times [15,16]. Poisoning with cyanotoxins occurs through consumption of
64 contaminated food or water, or during aquatic recreational activities, causing many
65 adverse symptoms like skin irritation, acute diarrhea, and liver and nervous tissue
66 damage, leading to severe health problems, or death in humans, cattle, domestic
67 animals and wildlife [6,13,17–19]. Hence, the risk of having toxin production
68 increases the need for monitoring plans to prevent toxin-related impairments and
69 costs.
70 Molecular-based methods combined with sequencing offer the ability to not only
71 identify possible toxin producers but also target species-specific toxins, validating the
72 presence or absence of toxin-related pathways. Nevertheless, DNA-based molecular
73 methods cannot predict if toxins are being produced and released to the
74 environment.
75 The leveraging of molecular methods provided by Next Generation Sequencing (NGS)
76 allows researchers to gain new insights into microbial community structure in
77 environment samples, identify new community members and discover potential
78 bioindicators of water quality.
79 To determine the current risk of toxin production in the Cape Verdean freshwater
80 reservoirs, we performed NGS analysis of 16S rRNA gene amplicon sequences to
81 identify microalgae and bacteria in the reservoirs. Our results show the presence of
82 several cyanobacterial genera well known to produce toxins in all reservoirs. We
83 were able to reconstruct, for the first time, the full genome of a potentially toxic

cyanobacterium from Cape Verde, based on the full metagenome sequencing data of Faveta reservoir. Analysis of this genome revealed the presence of genetic machinery used to synthesize cyanotoxins. The results of our biodiversity survey, phylogenetic analysis, and genome reconstruction, lead us to conclude that toxin risk is a reality and potential future threat in these reservoirs.

Materials and methods

Study sites and Sampling

This study was conducted on the Cape Verdean Island of Santiago (Fig 1). During sampling, in February 2014, only Poilão (B1), Saquinho (B2) and Faveta (B3) reservoirs had water while Figueira Gorda and Salineiro were empty at the tie of sampling. Other contextual information on Santiago Island's reservoirs is summarized in Table 1.

Fig 1. Map of Cape Verde and location of the reservoirs in Santiago Island sampled in this study.

DMS coordinates of the reservoirs: 1-Poilão (15˚04'25.0''N, 23˚33'25.8''W), 2-Saquinho (15˚08'11.3''N, 23˚42'27.7''W) and 3-Faveta (15˚05'54.5''N, 23˚37'24.1''W). 4-Figueira Gorda (15˚07'5.8''N, 23˚35'36.5''W) and 5-Salineiro (14˚57'03.5''N, 23˚38'00.4''W) reservoirs where empty at the time of sampling.

Table 1. Characteristics of the studied reservoirs in February 2014 (Government of Cape Verde)

Reservoir	Location	Start of impoundment	Theoretical maximum volume (m^3)	At full capacity?
Poilão	São Lourenço dos Orgãos	2006	1,200,000	yes
Saquinho	Santa Catarina	2013	563,000	no
Faveta	São Salvador do Mundo	2013	536,565	yes
Figueira Gorda	Santa Cruz	2014	1,455,272	empty
Salineiro	Cidade Velha	2013	561,464	empty

Triplicate water samples were collected from 0.5 to 1 m depth at each study site in the same sampling area, using a five-liter bucket. Microbial cells were concentrated by filtering the five liters of water through 0.22 µL Sterivex-GV filters (Millipore, Billerica, MA, USA) by vacuum filtration. Filters where kept on ice and in the dark while transported to the lab for DNA extraction.

DNA extraction

Genomic DNA was extracted from the individual filters using PowerWater Sterivex DNA Isolation Kit (MO BIO, USA) following the manufacturer's protocol. The amount of the DNA extracted was later quantified using NanoDrop 1000 spectrophotometer (Thermo-Fisher Scientific, Wilmington, DE, USA) measuring the UV absorption at 260 nm and 280 nm wavelengths.

Bacterial diversity: 16S rRNA gene amplicon sequencing

To assess the bacterial diversity in the Cape Verdean reservoirs, we used the 16S rRNA gene as a marker for biodiversity.

The extracted environmental DNA was amplified using primers targeting the V3 and V4 hypervariable regions of the 16S rRNA gene. PCR amplification of the 464 bp fragments was performed with the general bacterial primer pair 341F/785R [20].

The purified DNA was sequenced on an Illumina MiSeq platform using a 250 bp paired-end DNA library to generate at least 100,000 reads per sample. 16S ribosomal RNA gene amplicon sequencing was performed by Instituto Gulbenkian de Ciência (IGC) Gene Expression Facility.

Taxonomic composition and identification of potentially toxic genus/species

We used the QIIME v1.9.1 workflow [21] for demultiplexing, removing barcodes, quality filtering, clustering, and taxonomy assignment of the reads. Briefly, merging of the paired-end reads was done with a defined minimum overlap of 8 bp and a maximum difference of 20%, followed by quality filtering of the reads using default parameters [22], for each reservoir. Sequences were then clustered at a 97% identity threshold and all clusters with less than 4 sequences were removed thus reducing noisy reads. For each cluster a reference sequence was chosen (the first) and by comparison with Greengenes v13_8 [23], using UCLUST algorithm [24] with default parameters into Operational Taxonomic Units - OTUs. Taxonomy assignments were

145 done with RDP Classifier v2.2 [25]. To assess the consistency of the sampling and the
146 biodiversity present in each sampling site, diversity indices where calculated:
147 number of Observed species, Shannon's index and Simpson's Dominance index [26].
148

149 Graphical representation of the microbial communities' 150 structure detected in the reservoirs

151 Phyloseq package [27] in RStudio 1.1.419 [28] was used to perform a graphical
152 representation of the structure of the microbial community in each of the sampled
153 sites. A complete assignment of the more abundant bacterial genera in our samples,
154 was generated by Kraken running the full Bacterial Genbank Reference sequences
155 [29], excluding the sequences assigned previously as chloroplasts by QIIME.
156

157 Sequencing and assembly of the *Microcystis* cf. *aeruginosa* 158 CV01 genome

159 A whole metagenome sequencing of Faveta reservoir was performed using Illumina
160 MiSeq v3 kit with a 200x coverage. The DNA libraries were created using Nextera
161 DNA Library Preparation Kit, and sequencing was executed by IGC's Gene Expression
162 Facility.
163 The sequenced reads were assembled with SPAdes v3.6.0 [30] using single-cell
164 mode, and k-mers of 21, 33, 55, 77, 99 and 127. The GC content of the assembled
165 metagenome from Faveta reservoir revealed two distinct contributions: one clearly
166 in the range of the GC content typical of *Microcystis aeruginosa* (42-43%) [31] and

167 the other peak, at approximately 63% for other contributors present in the sample,
168 such as several GC rich strains of *Gamma, Alpha,* and *Betaproteobacteria* (S1 Fig).
169 The peak with lower GC was in the range of our target species and most of its reads
170 have high coverage numbers since they belong to the numerically most abundant
171 species. Two different approaches were used to isolate this genomic contribution:
172 one using CONCOCT v0.41 [32] a binning tool that operates based on sequence
173 coverage and composition, and another exploring the high coverage percentage of a
174 single genomic contribution by assembling reads from the raw sequencing data that
175 had coverage above 60x. In both methods, short sequences were filtered out (length
176 < 1,000 bp). Taxonomy of CONCOCT bins was then determined by Kraken [29] using
177 the default Minikraken database. BWA 0.7.16 (33) and Samtools 1.3 [34] were used
178 to map the reads back to isolated genomes while CheckM [35] was used to get
179 metrics on the quality of the isolation method.
180

181 Genome annotation and general features

182 We used QUAST [36] to determine the statistics of the assembly of the *M.* cf.
183 *aeruginosa* CV01 genome. Annotation of the genome was done with PROKKA v1.11
184 [37] and RAST v2.0 [38] using default parameters to identify putative genes (coding
185 and non-coding sequences). Protein function prediction and annotation of the
186 predicted genes was done against KEGG orthologs (KOs) [39] and clusters of
187 orthologs of proteins with Blast2GO [40] and eggNOG v4.5 [41]. Identification of
188 CRISPR repeats, typical in *Cyanobacteria* and in the *Microcystis* genus, was
189 performed with the web server CRISPRFinder [42] and Recognition Tool CRT v1.1

190 [43], considering a minimum of 3 repeat units. The prediction of transmembrane
191 topology and signal peptide sites was done using Phobius [44] and SignalP v4.1 [45].
192 Prediction of transmembrane helices in proteins was accomplished using TMHMM
193 v2.0 [46], prediction of ribosomal RNA subunits was done with RNAmmer v1.2 [47],
194 while tRNA and tmRNA genes prediction was done using ARAGORN v1.2.36 [48]. To
195 evaluate the presence and possible origin of prophage sequences, identification and
196 annotation of these sequences were performed using PHASTER [49].
197

198 Detection of toxin genes in *M.* cf. *aeruginosa* CV01

199 To identify the presence of toxin genes in *M.* cf. *aeruginosa* CV01, we searched the
200 genome to identify protein domains common in toxin gene clusters with Pfam v29.0
201 [50]. Next, we compared these sequences in our genome with putative toxin
202 synthase genes like microcystin, nodularin, cylindrospermopsin, anatoxin-a,
203 saxitoxin, microviridin, aeruginosin and micropeptin, which are known to be present
204 in cyanobacteria. We used the online tool antiSMASH v3.0 [51] to detect the
205 presence of non-ribosomal peptide synthase (NRPS) and/or polyketide synthase
206 (PKS) gene clusters and other domains of natural products typical of cyanobacterial
207 metabolites. Next, by manually curating the sequences that we identified as being
208 part of toxin synthesis process, we reconstructed the gene clusters ensuring the
209 overlap of contig ends.
210

211 *M*. cf. *aeruginosa* CV01 phylogenetic analysis

212 To reconstruct the phylogenetic relationships of *Microcystis* strains we inferred a
213 species tree using a set of 12 publicly available fully sequenced genomes. The species
214 tree was constructed from a concatenation of DNA sequences from a set of seven
215 single copy housekeeping genes present in all genomes (*ftsZ*, *glnA*, *gltX*, *gyrB*, *pgi*,
216 *recA* and *tpi*), using maximum likelihood phylogenies to infer the genetic variation of
217 the 12 genomes. Gene sequences were separately aligned with MAFFT v.7.220 [52]
218 and trimmed using Gblocks v0.91b [53,54]. To infer ML phylogenies we used RaxML
219 v8.1.20 [55], to compute and support the trees, by calculating 1000 non-parametric
220 bootstraps using GAMMA distribution and GTR+I+G model, model that was
221 previously estimated using Prottest [56]. Bayesian phylogenies were inferred with
222 MrBayes v3.2 [57] for 1 million generations using the same model and a discarded
223 burn-in rate of 25% of the initial generations. To investigate the relationship
224 between *M.* cf. *aeruginosa* CV01 and other continental African, tropical, and
225 temperate climate strains, we used the phycocyanin alpha subunit and phycocyanin
226 beta subunit (cpcA-cpcB) intergenic space region of the phycocyanin gene cluster
227 (PC-IGS) for a set of publicly available nucleic sequences from *Microcystis* strains
228 with worldwide distribution. Trees were computed using RaxML (1,000 non-
229 parametric bootstraps). Bayesian phylogenies were inferred with MrBayes for 3
230 million generations. Phylogenies were inferred under a GAMMA distribution and
231 GTR model.
232

233 # Results

234

235 ## Diversity of the microbial communities in the reservoirs

236 We profiled the bacterial biodiversity in three reservoirs of the island of Santiago
237 using 16S rRNA gene amplicon next generation sequencing. This marker gene is
238 universal in bacteria and it is also present in chloroplasts of eukaryotes, which
239 allowed us to detect both bacteria and eukaryotic phytoplankton. For Poilão
240 reservoir we obtained 804,043 reads (two replicates), for Saquinho 1,271,923 reads
241 (three replicates) and Faveta 1,354,273 reads (three replicates), which resulted in
242 315,291, 437,486 and 425,806 16S rRNA reads for Poilão, Saquinho, and Faveta,
243 respectively.
244 Regarding taxonomic distribution, Poilão had the highest number of operational
245 taxonomic units (OTUs) with 1,356 OTUs, followed by Saquinho with 966 and Faveta
246 with 950 different OTUs. The OTUs identified in Poilão reservoir represented 39
247 phyla and included 1.5 % of unassigned sequences. *Proteobacteria* was the
248 numerically most abundant phylum with 76% of all OTUs detected, of which 69.7%
249 were *Gammaproteobacteria,* dominated by the genus *Acinetobacter* (51.6%). The
250 second most abundant phylum was *Bacteroidetes*, representing 4.4% of the OTUs,
251 *Betaproteobacteria* accounted for 3.0% and *Actinobacteria* 2.3%. Cyanobacteria in
252 Poilão, which represented 3.5%, are mainly represented by the *Phormidium* genus
253 accounting for 2.5% of the OTUs. In Poilão reservoir, sequences assigned to green
254 algae (derived from chloroplasts) were 7.5% of the OTUs and approximately equal

255 amounts of *Chlorophyta* and stramenopiles. Main results regarding the composition
256 of microbial communities are shown in Fig 2.
257
258
259 **Fig 2. Composition of the microbe communities surveyed in Cape Verde.** (A-B)
260 Composition of the microbial communities detected in the three sampled Cape
261 Verdean freshwater reservoirs: Poilão-B1, Saquinho-B2 and Faveta-B3. (C)
262 Cyanobacteria communities detected in the surveyed reservoirs.
263
264 In the Saquinho reservoir, 37 phyla were detected and 0.5% of unassigned
265 sequences. The *Proteobacteria* represented 20% of the OTUs, *Bacteroidetes* 5.4%,
266 *Actinobacteria* accounted for 4.1%, and *Verrucomicrobia* 2.4%. *Cyanobacteria* in this
267 reservoir represented only 0.6% of the OTUs, the lowest percentage of the three
268 sampled reservoirs. This site had the highest number of OTUs associated with green
269 algae (63%), distributed between the *Cryptophyta* with 62% and stramenopiles with
270 1% of the identified OTUs (Fig 2A). Blastn of the cryptophyte 16S rRNA sequence
271 revealed a hit for *Cryptomonas curvata* plastid partial ribosomal RNA operon (strain
272 CCAC 0006) with 97% identity.
273 Faveta reservoir had the lowest number of phyla, accounting for 29 distinct phyla
274 and 1.1% of the sequences classified as unknown. The most abundant phylum was
275 *Cyanobacteria* with 58.6% of the OTUs and included two genera known to harbor
276 toxic species: *Microcystis* accounted for 55.7% and *Planktothrix* 2.8% (Fig 2B-C).
277 *Proteobacteria* was the second most abundant phylum with 18.6% assignments;

278 *Bacteroidetes* was next with 7.0% and *Actinobacteria* third with 5.8%: The green
279 algae at this site represented 5.1% of the identified OTUs.
280 In summary, our survey revealed distinct microbial community structures of the
281 three sampled reservoirs, which is confirmed by the beta diversity analysis in Fig 3D.
282
283
284 **Fig 3. – Biodiversity indices per replicate and average.** (A) Number of total observed
285 species, (B) Shannon, (C) Dominance and (D) Principal Coordinates Analysis of Beta
286 diversity by weighted Unifrac. Levels of significance: *$p < 0.05$. Poilão (D1); Saquinho
287 (D2); Faveta (D3).
288
289

290 Toxin producing genera/species

291 We identified several cyanobacterial genera in the three sampled reservoirs, some of
292 which are known from the literature to produce toxins. In Faveta reservoir we
293 identified the presence of cyanobacteria from the *Microcystis* and *Planktothrix*
294 genera, reported to produce cyanotoxins, microcystin and anatoxin-a, as well as
295 other potentially toxic molecules like chlorinated and sulfated variants of
296 aeruginosin, cyanopetolin, microginin, and microviridin [58–60]. Poilão reservoir had
297 smaller relative quantities of cyanobacteria, but more genera were identified (Fig 2
298 and S2 Fig) some of which were potentially toxic like, *Microcystis*, *Phormidium,*
299 *Planktothrix* and *Cylindrospermopsis.* These cyanobacteria are known to produce
300 microcystin, homo and anatoxin-a, saxitoxin, and cylindrospermopsin, as well as

301 other metabolites with different levels of toxicity. In Saquinho reservoir the relative
302 quantity of cyanobacteria was the lowest (0.6%), but even so the most represented
303 genus was the potentially toxic *Microcystis*.
304

305 *M.* cf. *aeruginosa* CV01 genome

306 CONCOCT produced 23 bins, including a bin containing an assembly of *M. aeruginosa*
307 with a 4.7 Mbp genome. According to CheckM with a completeness level of 98.79%
308 with a total of 75.6% of the reads mapping back to its contigs. The SPADES
309 assembled genome was 4.9 Mbp with a completeness level of 99.8% with 84% of the
310 reads mapping back to contig reads from the sequencing effort (S3 Fig and S1 Table).
311 The SPADES assembly was chosen due to its estimated completeness level and total
312 genome size (plotted in a graphical representation in Fig 4), distributed over 262
313 contigs with length over 1,000 bp and an average GC content of 42.3%. From a total
314 of 5,484 genes annotated, we identified 45 RNA genes, of which 42 were tRNA
315 genes, one tmRNA, and one rRNA set (5S/16S/23S), as well as 5051 protein-coding
316 genes and predicted functions for 91% of all genes (Blast2GO). Considering the
317 importance of metabolite transport and transduction of signals through membranes
318 in cyanobacteria, we also searched for and identified 112 genes with signal peptides
319 and 3,443 genes with transmembrane domains. Major statistical attributes of *M.* cf.
320 *aeruginosa* CV01 genome are described in Table 2.
321
322

323 **Fig 4. Representation of the genome of *M.* cf. *aeruginosa* CV01, sequenced from**
324 **the environmental sample of Faveta reservoir.** (A) The shades of blue in the circle
325 are indicative of pairwise genomic sequence similarity according to blastn, using as
326 reference the genome of *M. aeruginosa* NIES-843. The location and structure of
327 three gene clusters of potentially toxin-producing secondary metabolites are
328 represented. (B) The histogram represents the GC content of the assembled *M*. cf.
329 *aeruginosa* CV01 genome.
330
331 **Table 2. Genome statistics for *M.* cf. *aeruginosa* CV01.**

Attribute	Value	% of total
Genome size (bp)	4,918,369	100.0
DNA coding (bp)	3,900,077	79.3
DNA G+C (bp)	2,080,470	42.3
DNA scaffolds	262	-
Total genes	5,484	100.0
Protein coding genes	5,051	92.1
RNA genes	45	0.8
Genes with function prediction	4,996	91.1
Genes assigned to COGs	4,166	76.0
Genes with Pfam domains	3,407	62.1
Genes with signal peptides	112	2.0
Genes with transmembrane helices	3,443	62.8
CRISPR repeats	9	-

332

333

334 Four partial prophage regions were detected in *M.* cf. *aeruginosa* CV01. The first
335 region, with GC content 45.3%, is 9.4 Kb, encodes for 8 proteins and harbors protein
336 coding sequences (CDS) from a previously described infecting phage, P-TIM68,
337 usually associated with *Prochlorococcus* Myoviridae virus that contains photosystem
338 I gene sequences [61] and transposase sequences. A second partial phage region
339 with 12 CDS and GC content of 43.0% spans over 11.8 Kb and contains phage
340 sequences from the *Microcystis* infecting Myoviridae phage MaMv-DC [62].
341 Sequences of this phage and of phage Ma-LMM01 [63] are present in a third region
342 with 7.1 Kb in length, 40.4% GC content and 8 CDS. Finally, the forth prophage region
343 of 5.8 Kb in length, coding for 9 proteins and GC content of 40.6% was detected, and
344 CDS from a ssDNA marine virus reported to infect *Synechococcus* [64] were
345 identified. Regarding CRISPR arrays, the defense mechanism of *Cyanobacteria*
346 [65,66] we detected eight CRISPR repetitive units, varying from 0.3Kb and 11.8Kb in
347 length, with direct repeat lengths from 35 to 38 bp. The CRISPR regions and the
348 prophage were not overlapping but two clusters of CRISPR direct repeats (DR) were
349 identical to the *Microcystis* phage Ma-LMM01 portion, a memory mechanism to
350 their introduction into the CRISPR locus, providing immunity to further infection by
351 that phage. Other characteristics of the *M.* cf. *aeruginosa* CV01 genome, such as
352 KEGG orthologs and COG functional categories, are summarized in S2-S3 Tables.

353

354 **Phylogenetic analysis**

355 Our phylogenetic analysis using seven cyanobacterial genes, as described previously
356 in Materials and methods, identified the dominant cyanobacterium in Faveta
357 reservoir as a *M. aeruginosa* strain, which we named *M.* cf. *aeruginosa* CV01 (Fig
358 5A). *M.* cf. *aeruginosa* CV01 is placed close to two strains collected in African water
359 bodies: *M. aeruginosa* PCC 9443 collected from a fishpond in Landjica, Central
360 African Republic and *M. aeruginosa* PCC 9807 collected in Hartbeespoort Dam in
361 Pretoria, South Africa (Pasteur Culture Collection of Cyanobacteria). Phylogenetic
362 studies using the PC-IGS intergenic spacer region confirmed the clustering of *M.* cf.
363 *aeruginosa* CV01 with other continental African strains, as shown in Fig 5B, in a
364 branch that contains both microcystin producers and non-producers collected from
365 Ugandan and Kenyan water bodies [67].
366
367
368 **Fig 5. Phylogeny of the Cape Verdean strain *M.* cf. *aeruginosa* CV01.** (A) represents
369 the location of *M.* cf. *aeruginosa* CV01 sequenced in this study among a set of fully
370 sequenced genomes of *Microcystis*. The phylogenetic tree was inferred with
371 MrBayes for 1 million generations, using a GAMMA distribution and GTR substitution
372 matrix for a set of 7 housekeeping genes (*ftsZ*, *glnA*, *gltX*, *gyrB*, *pgi*, *recA* and *tpi*).
373 Dots represent posterior probability values higher than 85. (B) A phylogenetic tree of
374 69 *Microcystis* strains from all over the world locating the Cape Verdean strain
375 among other continental African *M. aeruginosa* strains. ML phylogenies were
376 inferred with RaxML with 1,000 bootstraps and Bayesian phylogenies were inferred
377 with MrBayes for 3 million generations, using a GAMMA distribution and GTR
378 substitution matrix for the intergenic spacer region of the phycocyanin cluster PC-IGS

379 region of 69 globally distributed *Microcystis* strains. Dots represent posterior
380 probability values higher than 75. The African strains are represented in blue in (A)
381 and (B).
382

383 Toxin genes and toxic species

384 In order to identify the risk of production of toxic cyanobacterial metabolites, we
385 searched the *M.* cf. *aeruginosa* CV01 genome for non-ribosomal peptide synthases
386 (NRPS) and NRPS/polyketyde synthases (PKS) hybrid gene clusters.
387 Genes that contribute to the synthesis of aeruginosin (NRPS/PKS/saccharide),
388 cyanopeptolin (micropeptin) and microviridin molecules were detected. Aeruginosin
389 and cyanopeptolin gene clusters were present and were non-halogenated (*aerJ* and
390 *mcnD* genes where not detected) (Fig 4). Regarding microcystin gene clusters, some
391 genes were present, nevertheless the full gene cluster was not represented.
392 Putative genes for the synthesis of cyanopetide metabolites like spumigin
393 (aeruginosin), aeruginoside (aeruginosin), ambiguine (terpene-alkaloid), and
394 piricyclamide (post-translational modified peptides) were also detected, but not
395 their complete gene clusters. Finally, one unknown NRPS/PKS gene cluster was
396 identified, showing the potential to produce peptide molecules that are yet
397 unknown.
398

399 Nucleotide sequence accession numbers

400 The nucleotide sequence data are available at DDBJ/EMBL/GenBank under the
401 accession number SUB2733390.

402

403 # Discussion

404

405 The studied reservoirs show three distinct microbial/microalgae community profiles,
406 despite being located on the same island and in a radius of 15 kilometers from each
407 other: in two sites, bacteria were dominant (*Proteobacteria* and *Cyanobacteria*) and
408 in the other reservoir microalgae belonging to the cryptophytes were the most
409 abundant taxa. The dominant species from one of the reservoirs was identified as a
410 *M. aeruginosa* strain through phylogenetic studies, placing it closer to other strains
411 collected in continental Africa. *Microcystis* spp. were detected in all three reservoirs,
412 as well as other cyanobacteria known to bloom and produce cyanotoxins. Our
413 analysis of the assembled *M.* cf. *aeruginosa* CV01 genome revealed that it can
414 produce toxins and therefore, a potential risk of toxin production can exist in Cape
415 Verdean reservoirs.
416 Our microbial biodiversity survey revealed distinct community structures in the three
417 reservoirs (Fig 3D). In fact, *Proteobacteria* dominated Poilão reservoir, mainly
418 bacteria from the *Acinetobacter* genus, while in Saquinho reservoir microalgae from
419 the *Cryptomonas* genus were dominant, and in Faveta reservoir the cyanobacterial
420 strain *M.* cf. *aeruginosa* CV01 was the most abundant. Despite being located on the
421 same island and having common microbial groups, we found that each phyla's
422 quantitative distribution varied substantially between reservoirs (Fig 2A).
423 Nevertheless, the profiles from the reservoirs identify groups that are common in
424 other water bodies studied worldwide despite the differences in relative quantities

425 as is the case of *Cyanobacteria*, *Cryptophyta*, *Actinobacteria*, *Bacteroidetes*,
426 *Verrucomicrobia*, and the three clades of *Proteobacteria* (Alpha, Beta and Gamma)
427 [14,68]. Actually the microbial profile of Poilão reservoir resembles those of lakes
428 close to urban areas, where *Acinetobacter* is the dominant genus [69].
429 In each reservoir, one genus prevailed with a relative abundance above 50%: in
430 Poilão was this was *Acinetobacter*, in Saquinho it was *Cryptomonas,* and in Faveta
431 the *Microcystis* genus.
432 The analysis of the local diversity indices of the replicates from each reservoir
433 revealed consistency and reinforced the sites' distinct microbial communities (Fig
434 3A-C). The indices also showed that microbial communities presented different
435 dominant genera in each of the reservoirs, as well as abundance of different taxa in
436 all sites as typically found in freshwater bodies around the world [7–10,12].
437 Besides the operational starting date differences between the reservoirs (Table 1)
438 and no physical communication between lakes, abiotic factors specific for each site
439 might be involved in the dominance variations within the microbe communities.
440 The dominance of one cyanobacterial strain in Faveta, allowed us to assemble and
441 fully study its genome, and to identify genes, allowing the reconstruction of toxin
442 pathways and assessing the toxin risk inherent in this specific strain. Exploration of
443 the assembled genome also revealed genomic features in common with other *M.*
444 *aeruginosa* genomes (Table 2). Some phage genes were found integrated in the
445 genome of *M. cf. aeruginosa*. Myoviridae "photosynthetic" freshwater cyanophages
446 (Ma-LMM01 and MaMV-DC) were also found. These genes are thought to play an
447 important role during phage infection by supplementing the host with the
448 production of photosynthesis proteins, a process that can be also beneficial to the

449 host during the infection process as suggested by some authors [61,70,71]. These
450 horizontal gene transfer events are shaping the genome architecture of the
451 *Microcystis* genus, providing a supplementary advantage that can be important
452 during cyanobacteria blooms. A region containing chlorophyll a apoproteins A1 and
453 A2 synthesis genes was also identified, but since these are single copy genes located
454 near transposase sequences in this new genome, they were probably misidentified
455 as having phage origin.

456 We identified four NRPS/PKS gene clusters that could synthesize potentially toxic
457 metabolites: three well-known metabolites (aeruginosin, cyanopeptolin, and
458 microviridin, represented in Fig 4A) and another metabolite from one yet unknown
459 gene cluster. The real toxic potential of these metabolites is difficult to determine, as
460 the actual synthesis pathways are not fully known, so further toxicological screenings
461 need to be performed.

462 Phylogenetic markers placed the Cape Verdean strain among others from freshwater
463 bodies from Africa, albeit Cape Verde being a distant archipelago from the
464 continental Africa. The identification in all reservoirs of other cyanobacterial genera
465 known to be toxin producers like *Phormidium, Planktothrix, and Cylindrospermopsis*,
466 increases the potential risk of toxin production. Other studies in African water bodies
467 have identified these and other potentially cyanotoxin producers, raising the
468 possibility of future occurrence of other cyanobacterial genera in Cape Verdean
469 freshwater reservoirs.

470 Cyanobacterial blooms occur in freshwater reservoirs distributed worldwide where
471 *M. aeruginosa* is one of the most frequently detected species. Actually, many long-
472 term studies have reported toxic blooms in lakes and rivers from Kenya, Uganda,

473 Senegal, Morocco, and South Africa [67,72–77], often dominated by *M. aeruginosa,*
474 as we also detected in the island of Santiago. Moreover a recent review on the
475 occurrence of cyanobacterial blooms in Africa [78] shows that there is limited
476 information from western African countries, including Cape Verde, exposing the
477 need for further studies in countries were water quality is threatened and scarce.
478 Therefore, our study increases the available information on cyanobacterial
479 communities described for the western African region. The scarcity of renewable
480 freshwater resources of archipelagic states like Cape Verde is aggravated by the
481 terrain that favors torrential water flows and strong anthropogenic pressures on the
482 environment leading to eutrophication of its freshwater bodies and increased risk of
483 toxic algal blooms.
484 The main threat and concern from our analysis, was the identification of a bacterial
485 community dominated by *M.* cf. *aeruginosa* CV01, which signals the possibility of
486 toxic blooms in Cape Verdean reservoirs, since exponential growth is typical of this
487 species.
488 The occurrence of blooms and toxin production are potential life-threatening risks to
489 public health, so monitoring plans are very important. The costs involved in these
490 control and containment strategies can be prohibitive especially for low and middle-
491 income countries. Hence, approaches like the one proposed in this work, which
492 enabled the identification of potentially toxic cyanobacterial genera through 16S
493 rRNA gene markers could be an interesting alternative, without the time-consuming
494 and expertise-dependent microscope identification of toxin-producing organisms or
495 mass spectrometry-base identification of toxins. NGS is still not widely available but
496 DNA sampling kits are easy to use and can be sent to sequencing facilities at cost-

497 effective prices. Other more sophisticated technics are possible such as lab on a chip,
498 mass spectrometry or even portable NGS devices, that can be adapted to use our
499 workflow in the field, but if a simple molecular lab is available, a PCR assay could also
500 be efficient to detect the presence of specific putative cyanotoxin genes. These
501 strategies can alert authorities and populations before bloom formation and toxin
502 production.

503 Cyanobacteria are currently being developed and used for bio-production of
504 metabolites and biomass in algal farms, for example in the production of human
505 dietary supplements, fertilizers, animal feed or biodiesel, to name a few uses [79–
506 83]. Cape Verde has little land suitable for agriculture, but it has temperature and
507 light conditions that are speculated to be suitable for simple bioreactors for biomass
508 production using photosynthetic microorganisms like green algae or cyanobacteria.
509 The observation of spontaneous blooms of cyanobacteria in freshwater reservoirs
510 lends credibility to this hypothesis, and opens the way for a new productive industry
511 in the archipelago.

512 The present work identifies the existence of real risk for cyanotoxin production in
513 Cape Verdean freshwater reservoirs. Similar structures are planned, which will also
514 need to be studied and monitored. Future work should include studies on the
515 dynamics of the local microbial communities, as well as characterize how
516 environmental factors are affecting their organization, in order to predict and
517 control the impacts of water impairment and toxin production on public health and
518 on the economy. In this study we made use of many freely available open source
519 tools, which represent, to our knowledge, innovative research strategies in Cape
520 Verde. This study will open the way for further research on microbial biodiversity

521 and other genomic studies in the archipelago, and raise questions relevant for
522 different areas of research and application.
523

524 Aknowledgements

525 We wish to thank Dr. W. Szymaniak for help in accessing the sample sites and the
526 Gene Expression Facility from Instituto Gulbenkian de Ciência for all the technical
527 support. Ana P. Semedo-Aguiar was recipient of fellowship
528 SFRH/BD/113752/2015 from the Portuguese Science and Technology Foundation
529 (FCT), under the auspices of the Programa de Pós-Graduação Ciência para o
530 Desenvolvimento (PGCD) and RBL was recipient of fellowship
531 SFRH/BPD/91518/2012 from FCT.
532

533 Author Contributions

534

535 Conceived and designed the experiments: APSA JBPL RBL. Performed the
536 experiments: APSA RBL. Analyzed the data: APSA RBL. Contributed
537 reagents/materials/analysis tools: JBPL RBL. Wrote the paper: APSA JBPL RBL.
538
539

References

1. Lobo de Pina A. Hidroquímica e qualidade das águas subterrâneas da ilha de Santiago - Cabo Verde. Universidade de Aveiro;2009.
2. Dodds WK, Bouska WW, Eitzmann JL, Pilger TJ, Pitts KL, Riley AJ, et al. Eutrophication of U.S. freshwaters: analysis of potential economic damages. Environ Sci Technol. 2009;43(1):12–9.
3. Heisler J, Glibert PM, Burkholder JM, Anderson DM, Cochlan W, Dennison WC, et al. Eutrophication and harmful algal blooms: A scientific consensus. Harmful Algae. 2008;8(1):3–13.
4. Paerl HW, Paul VJ. Climate change: Links to global expansion of harmful cyanobacteria. Water Res. 2012;46(5):1349–63.
5. Paerl HW, Scott JT. Throwing fuel on the fire: Synergistic effects of excessive nitrogen inputs and global warming on harmful algal blooms. Environ Sci Technol. 2010;44(20):7756–8.
6. O'Neil JM, Davis TW, Burford MA, Gobler CJ. The rise of harmful cyanobacteria blooms: The potential roles of eutrophication and climate change. Harmful Algae. 2012;14:313–34.
7. Jones SE, Newton RJ, McMahon KD. Evidence for structuring of bacterial community composition by organic carbon source in temperate lakes. Environ Microbiol. 2009;11(9):2463–72.
8. Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. A Guide to the Natural History of Freshwater Lake Bacteria. Microbiol Mol Biol Rev. 2011;75(1):14–49.

564 9. Oh S, Caro-Quintero A, Tsementzi D, DeLeon-Rodriguez N, Luo C, Poretsky R,
565 et al. Metagenomic insights into the evolution, function, and complexity of the
566 planktonic microbial community of Lake Lanier, a temperate freshwater
567 ecosystem. Appl Environ Microbiol. 2011;77(17):6000–11.
568 10. Eiler A, Drakare S, Bertilsson S, Pernthaler J, Peura S, Rofner C, et al. Unveiling
569 Distribution Patterns of Freshwater Phytoplankton by a Next Generation
570 Sequencing Based Approach. PLoS One. 2013;8(1):1–10.
571 11. Yan Q, Bi Y, Deng Y, He Z, Wu L, Van Nostrand JD, et al. Impacts of the Three
572 Gorges Dam on microbial structure and potential function. Nat Sci Reports.
573 2015;5(8605):1–9.
574 12. Eiler A, Bertilsson S. Flavobacteria blooms in four eutrophic lakes: Linking
575 population dynamics of freshwater bacterioplankton to resource availability.
576 Appl Environ Microbiol. 2007;73(11):3511–8.
577 13. Kosten S, Huszar VLM, Bécares E, Costa LS, van Donk E, Hansson LA, et al.
578 Warmer climates boost cyanobacterial dominance in shallow lakes. Glob
579 Chang Biol. 2012;18(1):118–26.
580 14. Eiler A, Bertilsson S. Composition of freshwater bacterial communities
581 associated with cyanobacterial blooms in four Swedish lakes. Environ
582 Microbiol. 2004;6(12):1228–43.
583 15. Conradie KR, Barnard S. The dynamics of toxic Microcystis strains and
584 microcystin production in two hypertrofic South African reservoirs. Harmful
585 Algae. 2012;20:1–10.
586 16. Davis TW, Berry DL, Boyer GL, Gobler CJ. The effects of temperature and
587 nutrients on the growth and dynamics of toxic and non-toxic strains of

588 Microcystis during cyanobacteria blooms. Harmful Algae. 2009;8(5):715–25.
589 17. Chorus I, Falconer IR, Salas HJ, Bartram J. Health risks caused by freshwater
590 cyanobacteria in recreational waters. J Toxicol Environ Heal Part B.
591 2000;3(4):323–47.
592 18. Paerl HW, Hall NS, Calandrino ES. Controlling harmful cyanobacterial blooms
593 in a world experiencing anthropogenic and climatic-induced change. Sci Total
594 Environ. 2011;409(10):1739–45.
595 19. Paerl HW, Huisman J. Blooms Like It Hot. Science. 2008;320(5872):57–8. doi:
596 10.1126/science.1155398 PMID: 18388279
597 20. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al.
598 Evaluation of general 16S ribosomal RNA gene PCR primers for classical and
599 next-generation sequencing-based diversity studies. Nucleic Acids Res.
600 2013;41(1):e1–e1.
601 21. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK,
602 et al. QIIME allows analysis of high- throughput community sequencing data.
603 Nat Methods. 2010;7(5):335–6.
604 22. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon I, Knight R, et al.
605 Quality-filtering vastly improves diversity estimates from Illumina amplicon
606 sequencing. Nat Methods. 2013;10(1):57–9.
607 23. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al.
608 An improved Greengenes taxonomy with explicit ranks for ecological and
609 evolutionary analyses of bacteria and archaea. ISME J. 2012;6(3):610–8.
610 24. Edgar RC. Search and clustering orders of magnitude faster than BLAST.
611 Bioinformatics. 2010;26(19):2460–1.

612 25. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid
613 assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ
614 Microbiol. 2007;73(16):5261–7.
615 26. Morris EK, Caruso T, Fischer M, Hancock C, Obermaier E, Prati D, et al.
616 Choosing and using diversity indices: insights for ecological applications from
617 the German Biodiversity Exploratories. Ecol Evol. 2014;4(18):3514–24.
618 27. McMurdie PJ, Holmes S. Phyloseq: An R Package for Reproducible Interactive
619 Analysis and Graphics of Microbiome Census Data. PLoS One. 2013;8(4).
620 28. Team Rs. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA;
621 2016.
622 29. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification
623 using exact alignments. Genome Biol. 2014;15(3):R46.
624 30. Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov AS, et al.
625 SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell
626 Sequencing. J Comput Biol. 2012;19(5):455–77.
627 31. Humbert J-F, Barbe V, Latifi A, Gugger M, Calteau A, Coursin T, et al. A tribute
628 to disorder in the genome of the bloom-forming freshwater cyanobacterium
629 Microcystis aeruginosa. PLoS One. 2013;8(8):e70747.
630 32. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al.
631 Binning metagenomic contigs by coverage and composition. Nat Methods.
632 2014;11(11):1144–46.
633 33. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler
634 transform. Bioinformatics. 2009;25(14):1754–60.
635 34. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence

636 Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
637 35. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM :
638 assessing the quality of microbial genomes recovered from isolates , single
639 cells , and metagenomes. 2015;1043–55.
640 36. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for
641 genome assemblies. Bioinformatics. 2013;29(8):1072–5.
642 37. Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics.
643 2014;30(14):2068–9.
644 38. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards R a, et al. The RAST
645 Server: rapid annotations using subsystems technology. BMC Genomics.
646 2008;9:75.
647 39. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a
648 reference resource for gene and protein annotation. Nucleic Acids Res.
649 2016;44(D1):D457–62.
650 40. Conesa A, Goetz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: A
651 universal tool for annotation, visualization and analysis in functional genomics
652 research. Bioinformatics. 2005;21(18):3674–6.
653 41. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al.
654 EGGNOG 4.5: A hierarchical orthology framework with improved functional
655 annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res.
656 2016;44(D1):D286–93.
657 42. Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered
658 regularly interspaced short palindromic repeats. Nucleic Acids Res.
659 2007;35(Web Server):W52–7.

660 43. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR
661 Recognition Tool (CRT): a tool for automatic detection of clustered regularly
662 interspaced palindromic repeats. BMC Bioinformatics. 2007;8(1):209.
663 44. Kall L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane
664 topology and signal peptide prediction--the Phobius web server. Nucleic Acids
665 Res. 2007;35(Web Server):W429–32.
666 45. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating
667 signal peptides from transmembrane regions. Nat Methods. 2011;8(10):785–
668 6.
669 46. Krogh a, Larsson B, von Heijne G, Sonnhammer E. Predicting transmembrane
670 protein topology with a hidden Markov model: application to complete
671 genomes. J Mol Biol. 2001;305(3):567–80.
672 47. Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW.
673 RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic
674 Acids Res. 2007;35(9):3100–8.
675 48. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA
676 genes in nucleotide sequences. Nucleic Acids Res. 2004;32(1):11–6.
677 49. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better,
678 faster version of the PHAST phage search tool. Nucleic Acids Res.
679 2016;44(W1):W16–21.
680 50. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam
681 protein families database: towards a more sustainable future. Nucleic Acids
682 Res. 2016;44(D1):D279–85.
683 51. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, et al. antiSMASH 3.0-

684 -a comprehensive resource for the genome mining of biosynthetic gene
685 clusters. Nucleic Acids Res. 2015;43(Web server issue):W237–43.
686 52. Katoh K, Standley DM. MAFFT multiple sequence alignment software version
687 7: Improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–
688 80.
689 53. Castresana J. Selection of Conserved Blocks from Multiple Alignments for
690 Their Use in Phylogenetic Analysis. Mol Biol Evol. 2000;17(4):540–52.
691 54. Talavera G, Castresana J. Improvement of phylogenies after removing
692 divergent and ambiguously aligned blocks from protein sequence alignments.
693 Syst Biol. 2007;56(4):564–77.
694 55. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-
695 analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
696 56. Darriba D, Taboada GL, Posada D. ProtTest 3 : fast selection of best-fit models
697 of protein evolution. Bioinformatics. 2011;28(8):1164–5.
698 57. Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, et al.
699 Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice
700 across a large model space. Syst Biol. 2012;61(3):539–42.
701 58. Kohler E, Grundler V, Haussinger D, Kurmayer R, Gademann K, Pernthaler J, et
702 al. The toxicity and enzyme activity of a chlorine and sulfate containing
703 aeruginosin isolated from a non-microcystin-producing Planktothrix strain.
704 Harmful Algae. 2014;39:154–60.
705 59. Quiblier C, Wood S, Echenique Subiabre I, Heath M, Villeneuve A, Humbert J-F.
706 A review of current knowledge on toxic benthic freshwater cyanobacteria -
707 Ecology, toxin production and risk management. Water Res.

708 2013;47(15):5464–79.
709 60. Rounge TB, Rohrlack T, Nederbragt AJ, Kristensen T, Jakobsen KS. A genome-
710 wide analysis of nonribosomal peptide synthetase gene clusters and their
711 peptides in a Planktothrix rubescens strain. BMC Genomics. 2009;10:396.
712 61. Sharon I, Alperovitch A, Rohwer F, Haynes M, Glaser F, Atamna-Ismaeel N, et
713 al. Photosystem I gene cassettes are present in marine virus genomes. Nature.
714 2009;461(7261):258–62.
715 62. Ou T, Gao XC, Li SH, Zhang QY. Genome analysis and gene nblA identification
716 of microcystis aeruginosa myovirus (MaMV-DC) reveal the evidence for
717 horizontal gene transfer events between cyanomyovirus and host. J Gen Virol.
718 2015;96(12):3681–97.
719 63. Yoshida T, Nagasaki K, Takashima Y, Shirai Y, Tomaru Y, Takao Y, et al. Ma-
720 LMM01 infecting toxic Microcystis aeruginosa illuminates diverse cyanophage
721 genome strategies. J Bacteriol. 2008;190(5):1762–72.
722 64. Huang S, Wang K, Jiao N, Chen F. Genome sequences of siphoviruses infecting
723 marine Synechococcus unveil a diverse cyanophage group and extensive
724 phage–host genetic exchanges. Environ Microbiol. 2012;14(2):540–58.
725 65. Cai F, Axen SD, Kerfeld C a. Evidence for the widespread distribution of
726 CRISPR-Cas system in the Phylum Cyanobacteria. RNA Biol. 2013;10(5):687–
727 93.
728 66. Yang C, Lin F, Li Q, Li T, Zhao J. Comparative genomics reveals diversified
729 CRISPR-Cas systems of globally distributed Microcystis aeruginosa, a
730 freshwater bloom-forming cyanobacterium. Front Microbiol. 2015;6(394). doi:
731 10.3389/fmicb.2015.00394 PMID: 26029174

732 67. Haande S, Ballot A, Rohrlack T, Fastner J, Wiedner C, Edvardsen B. Diversity of
733 Microcystis aeruginosa isolates (Chroococcales, Cyanobacteria) from East-
734 African water bodies. Arch Microbiol. 2007;188(1):15–25.
735 68. Humbert J-F, Dorigo U, Cecchi P, Le Berre B, Debroas D, Bouvy M. Comparison
736 of the structure and composition of bacterial communities from temperate
737 and tropical freshwater ecosystems. Environ Microbiol. 2009;11(9):2339–50.
738 69. McLellan SL, Fisher JC, Newton RJ. The microbiome of urban waters. Int
739 Microbiol. 2015;18(3):141–9.
740 70. Lindell D, Jaffe JD, Johnson ZI, Church GM, Chisholm SW. Photosynthesis genes
741 in marine viruses yield proteins during host infection. Nature.
742 2005;438(7064):86–9.
743 71. Yoshida-Takashima Y, Yoshida M, Ogata H, Nagasaki K, Hiroishi S, Yoshida T.
744 Cyanophage Infection in the Bloom-Forming Cyanobacteria Microcystis
745 aeruginosa in Surface Freshwater. Microbes Environ. 2012;27(4):350–5.
746 72. Okello W, Portmann C, Erhard M, Gademann K, Kurmayer R. Occurrence of
747 microcystin-producing cyanobacteria in Ugandan freshwater habitats. Environ
748 Toxicol. 2010;25(4):367–80.
749 73. Thomazeau S, Houdan-Fourmont A, Couté A, Duval C, Couloux A, Rousseau F,
750 et al. The contribution of sub-saharan african strains to the phylogeny of
751 cyanobacteria: Focusing on the nostocaceae (nostocales, cyanobacteria). J
752 Phycol. 2010;46(3):564–79.
753 74. Haande S, Rohrlack T, Ballot A, Røberg K, Skulberg R, Beck M, et al. Genetic
754 characterisation of Cylindrospermopsis raciborskii (Nostocales, Cyanobacteria)
755 isolates from Africa and Europe. Harmful Algae. 2008;7(5):692–701.

756 75. Berger C, Ba N, Gugger M, Bouvy M, Rusconi F, Couté A, et al. Seasonal
757 dynamics and toxicity of Cylindrospermopsis raciborskii in Lake Guiers
758 (Senegal, West Africa). FEMS Microbiol Ecol. 2006;57(3):355–66.
759 76. Douma M, Ouahid Y, Campo FF Del, Loudiki M, Mouhri K, Oudra B.
760 Identification and quantification of cyanobacterial toxins (microcystins) in two
761 Moroccan drinking-water reservoirs (Mansour Eddahbi, Almassira). Environ
762 Monit Assess. 2010;160(1–4):439–50.
763 77. Matthews M, Bernard S. Eutrophication and cyanobacteria in South Africa's
764 standing water bodies: A view from space. S Afr J Sci. 2015 May
765 28;111(5/6):1–8.
766 78. Ndlela LL, Oberholster PJ, Van Wyk JH, Cheng PH. An overview of
767 cyanobacterial bloom occurrences and research in Africa over the last decade.
768 Harmful Algae. 2016;60:11–26. doi: 10.1016/j.hal.2016.10.001
769 79. Englund E, Pattanaik B, Ubhayasekera SJK, Stensjö K, Bergquist J, Lindberg P.
770 Production of Squalene in Synechocystis sp. PCC 6803. Neilan B, editor. PLoS
771 One. 2014;9(3):e90270.
772 80. Zanchett G, Oliveira-Filho E. Cyanobacteria and Cyanotoxins: From Impacts on
773 Aquatic Ecosystems and Human Health to Anticarcinogenic Effects. Toxins
774 (Basel). 2013;5(10):1896–917.
775 81. Enzing C, Ploeg M, Barbosa M. Microalgae-based products for the food and
776 feed sector: an outlook for Europe. Luxembourg; 2014. doi: 10.2791/3339
777 82. Simas-Rodrigues C, Villela HDM, Martins AP, Marques LG, Colepicolo P, Tonon
778 AP. Microalgae for economic applications: advantages and perspectives for
779 bioethanol. J Exp Bot. 2015;66(14):4097–108.

780 83. Brennan L, Owende P. Biofuels from microalgae—A review of technologies for
781 production, processing, and extractions of biofuels and co-products. Renew
782 Sustain Energy Rev. 2010;14(2):557–77.
783

784 # Supporting Information

785

786 **S1 Fig. - The GC content of the raw reads and the final assembled metagenome**
787 **from Faveta reservoir.** GC content of the raw reads and of the assembled genomes
788 in the histograms of panels A and B, respectively.

789

790 **S2 Fig. – Rarefaction curves per number of reads for each reservoir sampled.**
791 Representation based in Chao1, Observed OTUs and Phylogenetic distance - PD
792 whole tree of the 16sRNA gene amplicons of Illumina sequencing. D1-Poilão; D2-
793 Saquinho; D3-Faveta.

794

795 **S3 Fig. – Faveta reservoir.** Degree of completeness and contamination of the *M.*
796 *aeruginosa* bin detected using CONCOCT and assembly produced by SPADES (panel
797 A), Graphical representation of a PCA of clusters based on composition and coverage
798 (bin 19 in yellow). Left Panel C - Histogram of GC distribution of sequences within bin
799 19, Right Panel C - Deviation of the average of the entire assembly, x-axis represents
800 GC and y-axis the size. Left Panel D - Histogram of Manhattan distance between
801 tetranucleotide signature of bin 19 and the entire assembly, Right Panel D -
802 Deviation of the average of the entire assembly, x axis represents GC and y axis the
803 size.

804

805 **S1 Online Fig. – Kraken analyses of taxon distribution in whole genome reads and**
806 **16S rRNA from Faveta reservoir.**

807

808 # Supplementary tables legends

809 **S1 Table. Degree of completeness and contamination and percentage of mapped**

810 **reads**

811

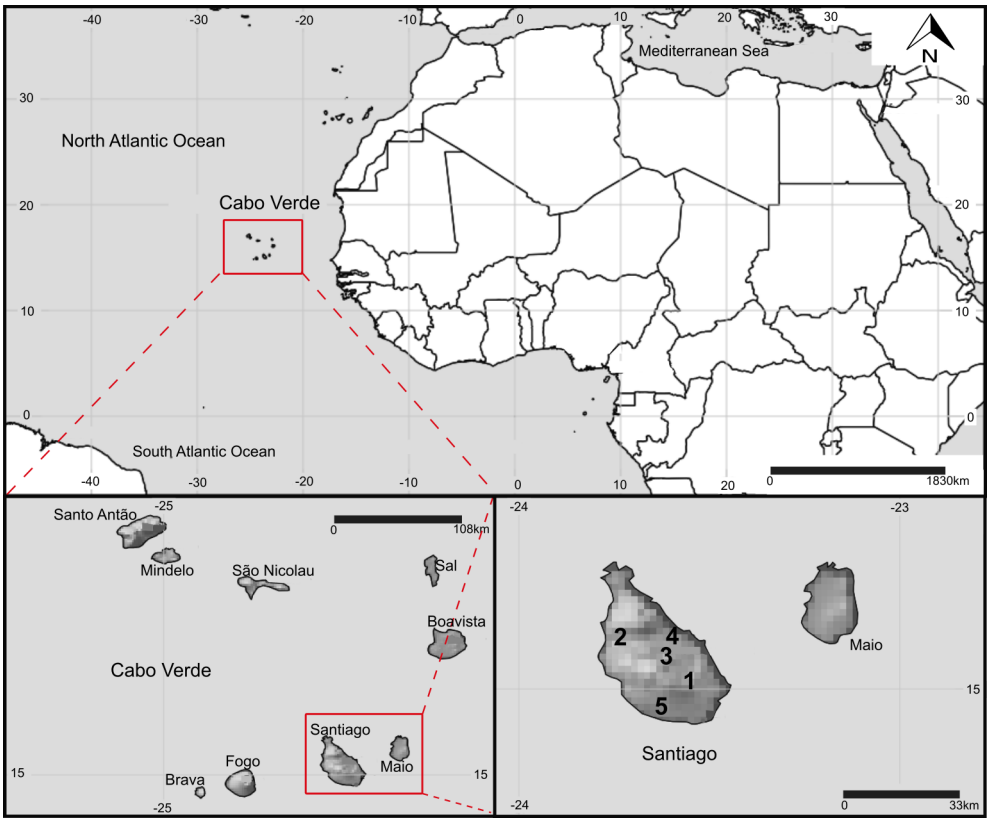
812 **S2 Table. KEGG functional annotation of the predicted transcriptome of *Microcystis***

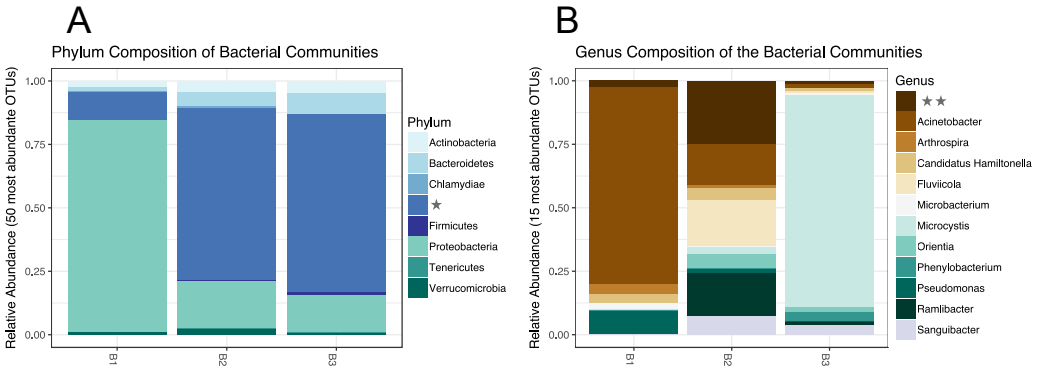
813 **cf. *aeruginosa* CV01**

814

815 **S3 Table. COG functional categories of the predicted transcriptome of *Microcystis***

816 **cf. *aeruginosa* CV01**


North Atlantic Ocean
Cabo Verde
South Atlantic Ocean
Mediterranean Sea
N
1830km
Santo Antão
Mindelo
São Nicolau
Sal
Boavista
Cabo Verde
Santiago
Maio
Fogo
Brava
108km
Santiago
Maio
1
2
3
4
5
33km



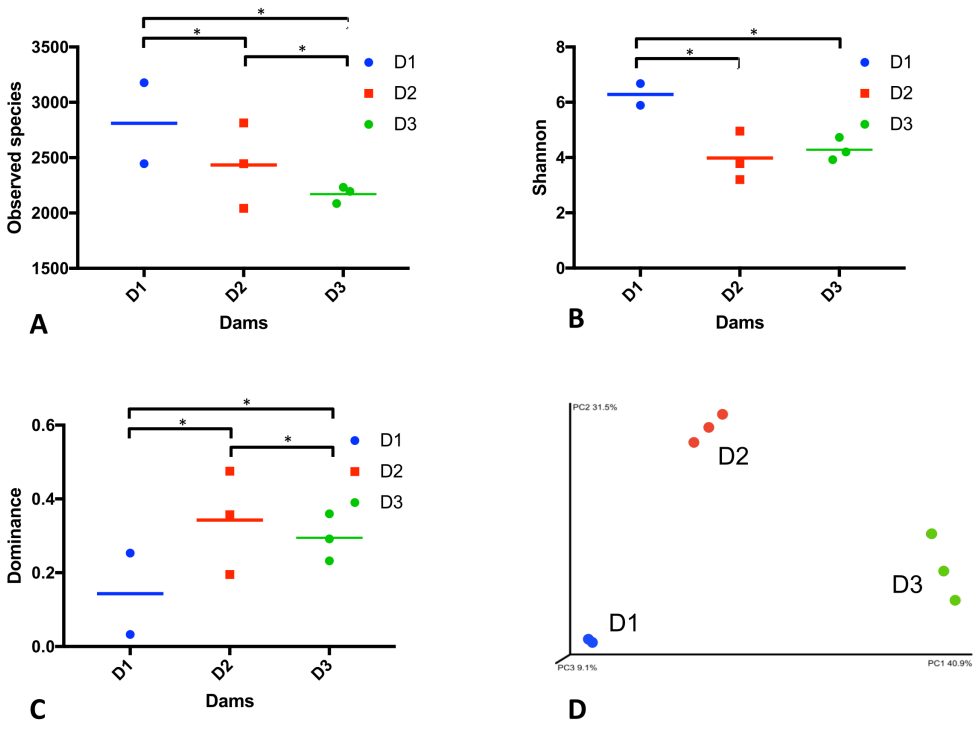
★ The area represents the sum of Chloroplasts (7.5%, 63% and 5.1% per site) and *Cyanobacteria* (3.8%, 0.6% and 58.6% per site)
★★ *Actinobacteria*, *Burkolderiales* and *Rhodobacteriales* with no assigned genus.

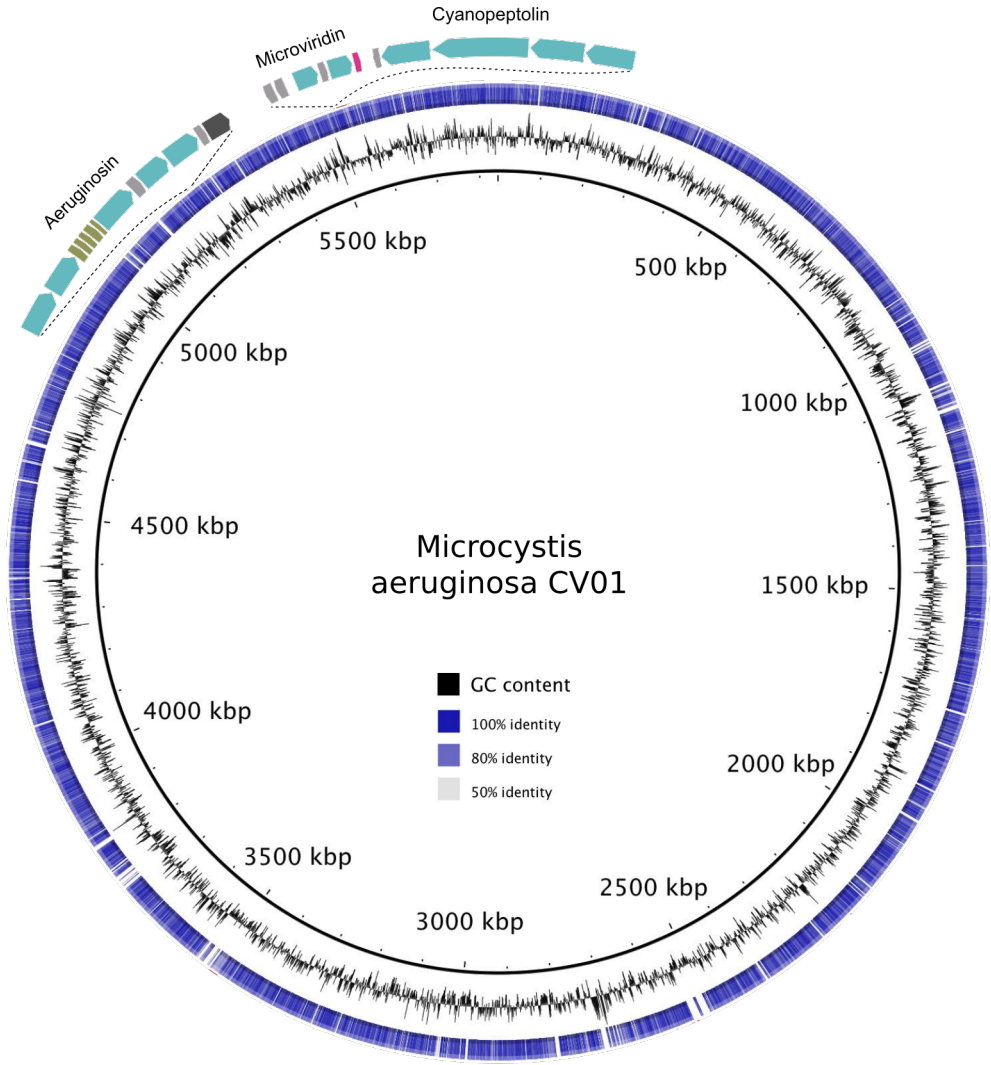
C

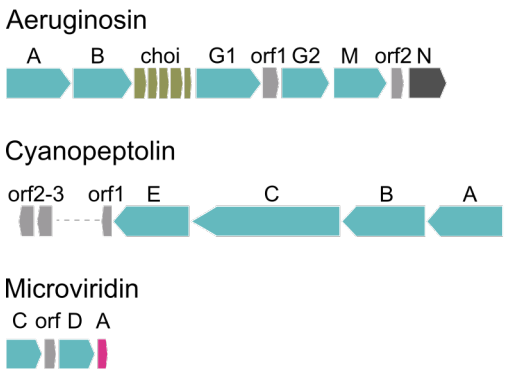
Cyanobacterial communities of the reservoirs

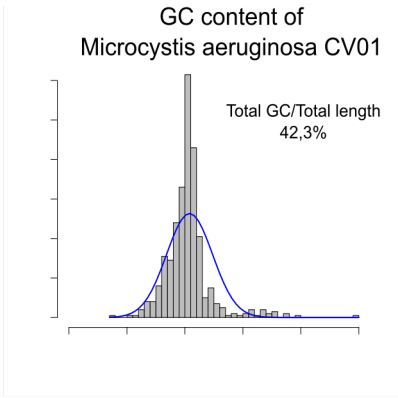
		Poilão	Saquinho	Faveta
Total number of 16S rRNA sequences		315 291	437 486	452 806
Total *Cyanobacteria* (%)		3.8	0.6	58.6
Order	Genus		%	
Chroococcales	***Microcystis***	0.1	0.4	**55.7**
Oscillatoriales	*Planktothrix*	0.1	<0.1	2.8
Oscillatoriales	*Phormidium*	2.5	<0.1	nd
Oscillatoriales	na[1]	<0.1	0.1	0.1
Synechococcales	*Synechococcus*	0.1	<0.1	<0.1
Chroococcales	na[2]	0.1	<0.1	nd
Pseudanabaenales	*Halomicronema*	0.1	nd	nd
Pseudanabaenales	*Leptolyngbya*	0.1	nd	nd
Oscillatoriales	*Planktothricoides*	0.1	nd	nd
Nostocales	*Cylindrospermopsis*	<0.1	nd	nd

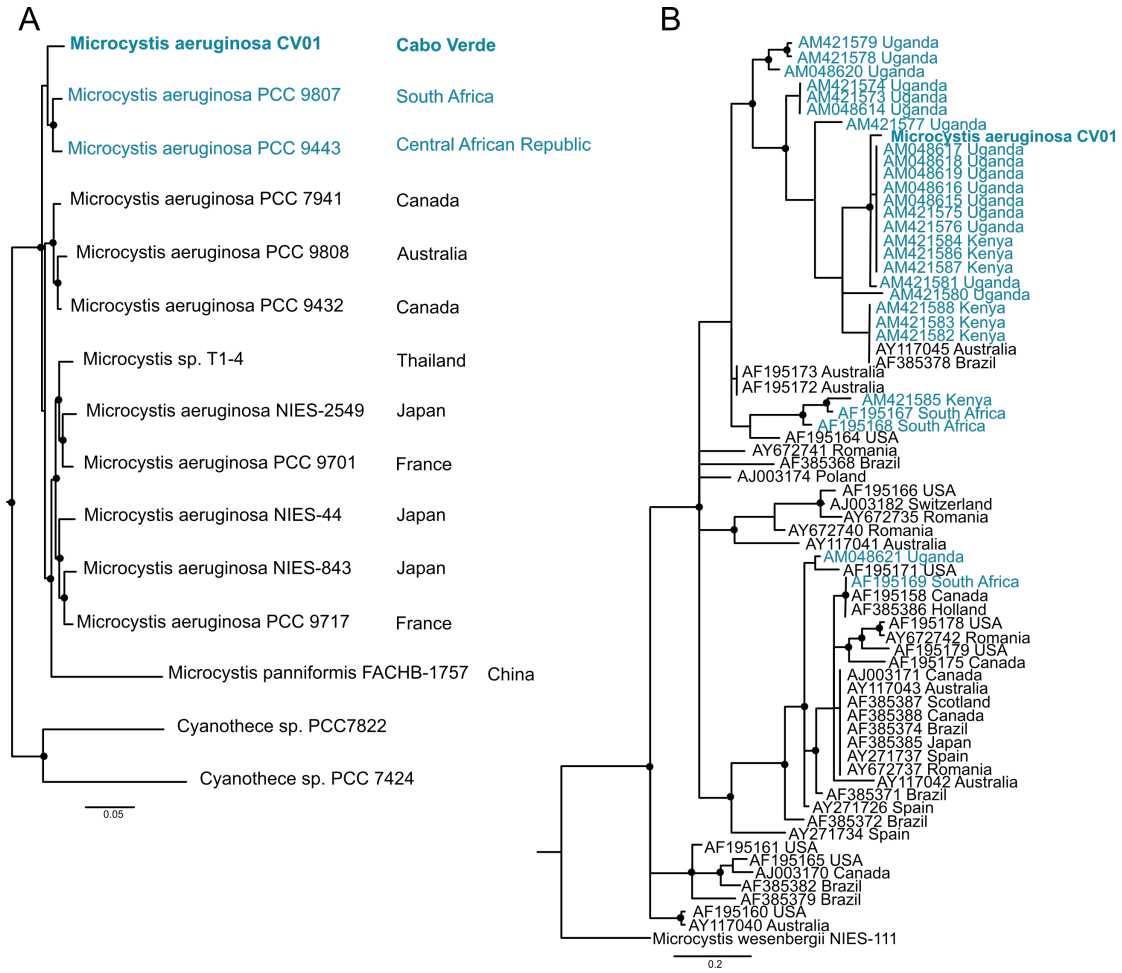
nd: not detected; na[1]: genus not assigned, *Phormidiaceae* Family; na[2]: genus not assigned, *Xenococcaceae* Family.


Observed species
3500
3000
2500
2000
1500
D1
D2
D3
Dams
A
Shannon
8
6
4
2
0
B
Dominance
0.6
0.4
0.2
0.0
C
PC2 31.5%
PC3 9.1%
PC1 40.9%
D


Cyanopeptolin
Microviridin
Aeruginosin
5500 kbp
500 kbp
5000 kbp
1000 kbp
4500 kbp
Microcystis
aeruginosa CV01
1500 kbp
GC content
100% identity
80% identity
50% identity
4000 kbp
2000 kbp
3500 kbp
2500 kbp
3000 kbp


Aeruginosin
A B choi G1 orf1 G2 M orf2 N
Cyanopeptolin
orf2-3 orf1 E C B A
Microviridin
C orf D A


GC content of
Microcystis aeruginosa CV01
Total GC/Total length
42,3%


A
Microcystis aeruginosa CV01
Cabo Verde
Microcystis aeruginosa PCC 9807
South Africa
Microcystis aeruginosa PCC 9443
Central African Republic
Microcystis aeruginosa PCC 7941
Canada
Microcystis aeruginosa PCC 9808
Australia
Microcystis aeruginosa PCC 9432
Canada
Microcystis sp. T1-4
Thailand
Microcystis aeruginosa NIES-2549
Japan
Microcystis aeruginosa PCC 9701
France
Microcystis aeruginosa NIES-44
Japan
Microcystis aeruginosa NIES-843
Japan
Microcystis aeruginosa PCC 9717
France
Microcystis panniformis FACHB-1757
China
Cyanothece sp. PCC7822
Cyanothece sp. PCC 7424
0.05
B
AM421579 Uganda
AM421578 Uganda
AM048620 Uganda
AM421574 Uganda
AM421573 Uganda
AM048614 Uganda
AM421577 Uganda
Microcystis aeruginosa CV01
AM048617 Uganda
AM048618 Uganda
AM048619 Uganda
AM048616 Uganda
AM048615 Uganda
AM421575 Uganda
AM421576 Uganda
AM421584 Kenya
AM421586 Kenya
AM421587 Kenya
AM421581 Uganda
AM421580 Uganda
AM421588 Kenya
AM421583 Kenya
AM421582 Kenya
AY117045 Australia
AF385378 Brazil
AF195173 Australia
AF195172 Australia
AM421585 Kenya
AF195167 South Africa
AF195168 South Africa
AF195164 USA
AY672741 Romania
AF385368 Brazil
AJ003174 Poland
AF195166 USA
AJ003182 Switzerland
AY672735 Romania
AY672740 Romania
AY117041 Australia
AM048621 Uganda
AF195171 USA
AF195169 South Africa
AF195158 Canada
AF385386 Holland
AF195178 USA
AY672742 Romania
AF195179 USA
AF195175 Canada
AJ003171 Canada
AY117043 Australia
AF385387 Scotland
AF385388 Canada
AF385374 Brazil
AF385385 Japan
AY271737 Spain
AY672737 Romania
AY117042 Australia
AF385371 Brazil
AY271726 Spain
AF385372 Brazil
AY271734 Spain
AF195161 USA
AF195165 USA
AJ003170 Canada
AF385382 Brazil
AF385379 Brazil
AF195160 USA
AY117040 Australia
Microcystis wesenbergii NIES-111
0.2