



Article

# Microbial Diversity and Toxin Risk in Tropical Freshwater Reservoirs of Cape Verde

Ana P. Semedo-Aguiar <sup>1,2,3</sup> , Jose B. Pereira-Leal <sup>1</sup> and Ricardo B. Leite <sup>1,\*</sup> 

<sup>1</sup> Instituto Gulbenkian de Ciência, Rua da Quinta Grande n°6, 2780-156 Oeiras, Portugal; aaguiar@igc.gulbenkian.pt (A.P.S.-A.); jleal@igc.gulbenkian.pt (J.B.P.-L.)

<sup>2</sup> Programa de Pós-Graduação Ciência para o Desenvolvimento, Rua da Quinta Grande n°6, 2780-156 Oeiras, Portugal

<sup>3</sup> Universidade Jean Piaget de Cabo Verde, Campus da Praia, Caixa Postal 775, Palmarejo Grande, Praia, Cabo Verde

\* Correspondence: rleite@igc.gulbenkian.pt; Tel.: +351-21-440-7975

Received: 23 March 2018; Accepted: 3 May 2018; Published: 5 May 2018



**Abstract:** The Cape Verde islands are part of the African Sahelian arid belt that possesses an erratic rain pattern prompting the need for water reservoirs, which are now critical for the country's sustainability. Worldwide, freshwater cyanobacterial blooms are increasing in frequency due to global climate change and the 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. Faveta potentially toxic and bloom-forming *Microcystis* sp., dominated our samples, while a *Cryptomonas* green algae and *Gammaproteobacteria* dominated Saquinho and Poilão reservoirs. We reconstructed and assembled the *Microcystis* genome, extracted from the metagenome of bulk DNA from Faveta water. Phylogenetic analysis of *Microcystis* cf. *aeruginosa* CV01's genome revealed its close relationship with other *Microcystis* genomes, as well as clustering with other continental African strains, suggesting geographical coherency. In addition, it revealed several clusters of known toxin-producing genes. This survey reinforces the need to better understand the country's microbial ecology as a whole of water reservoirs on the rise.

**Keywords:** toxins; *Cyanobacteria*; NGS; *Microcystis*; MAGs; metagenome; freshwater; microbial communities; Cape Verde; reservoirs

**Key Contribution:** In this study, we identified the presence of genes and microbial species associated with toxin production. This knowledge allows for the prediction of potential public health issues and the control of the impacts of water impairment. Our work opens the way for other NGS based studies, targeting the biodiversity of others sites for a better understanding of the environment and public health issues of Cape Verde.

## 1. 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, which are able to produce toxins and foul odors and discolor the water [13,14]. The cyanobacterial phylum has many genera that produce toxins, also called cyanotoxins, and in freshwater bodies, toxic and non-toxic strains can co-exist and dominate at different times [15,16]. Poisoning with cyanotoxins occurs through consumption of contaminated food or water, or during aquatic recreational activities, causing many adverse symptoms like skin irritation, acute diarrhea, and liver and nervous tissue damage, leading to severe health problems, or death in humans, cattle, domestic animals, and wildlife [6,13,17–19]. Hence, the risk of toxin production occurring increases the need for monitoring plans to prevent toxin-related impairments and costs.

Molecular-based methods combined with sequencing offer the ability to not only identify possible toxin producers but also target species-specific toxins, validating the presence or absence of toxin-related pathways. Nevertheless, DNA-based molecular methods cannot predict if toxins are being produced and released to the environment.

The leveraging of molecular methods provided by Next Generation Sequencing (NGS) allows researchers to gain new insights into microbial community structure in environment samples, identify new community members, and discover alternative bioindicators of water quality.

Environmental DNA samples are often difficult to purify to axenic conditions, so *in vitro* assembly of genomes recovered from the sequencing reads of metagenomes, MAGs, have been accomplished on several microbial community studies and environmental surveys [20–24]. These genomics and metagenomic approaches enable, for instance, the reconstruction of genomes from unculturable samples or fast monitoring of cyanobacterial bloom formation [20].

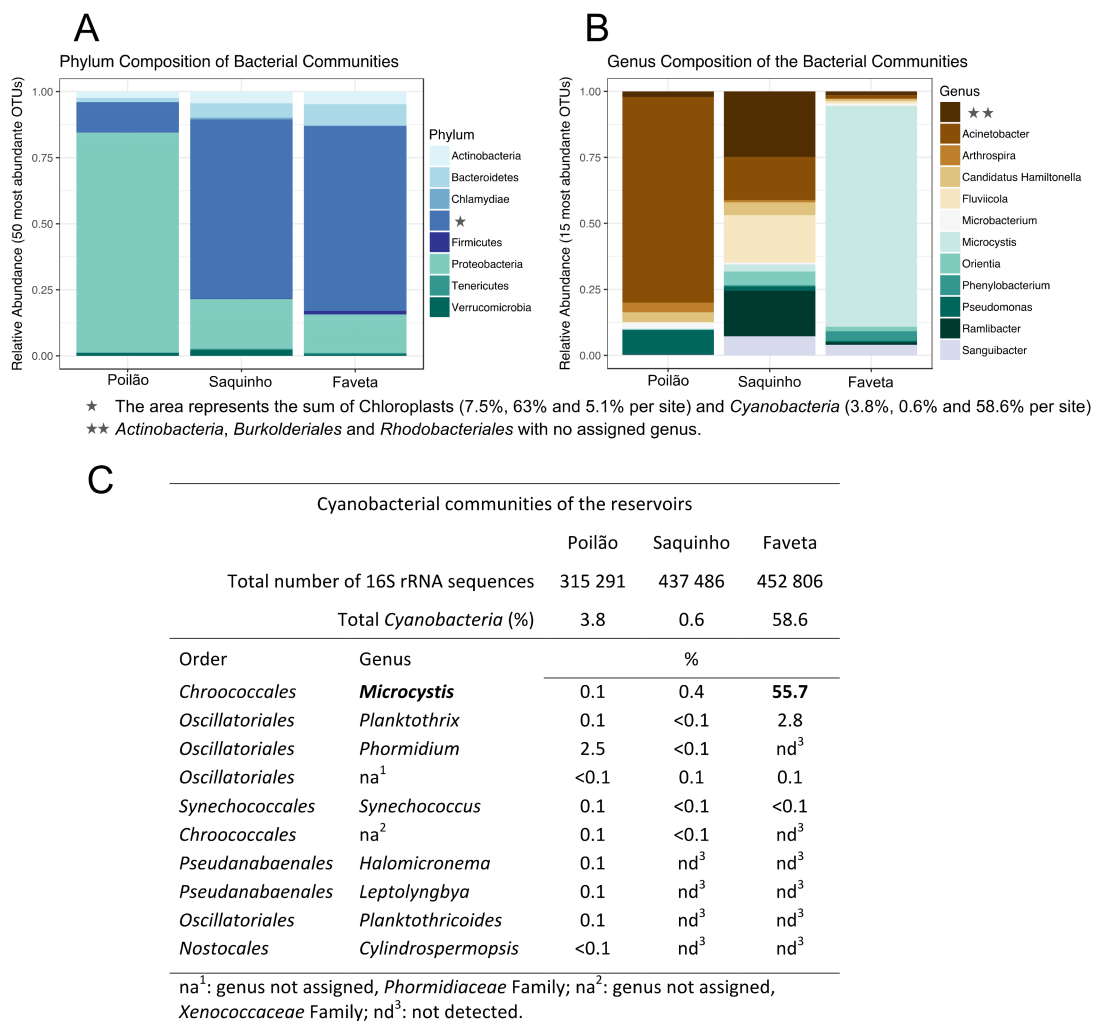
To determine the current potential of risk of toxin production in the Cape Verdean freshwater reservoirs, we performed NGS analysis of 16S rRNA gene amplicon sequences to identify microalgae and bacteria in the reservoirs. Our results show the presence of several cyanobacterial genera well known to produce toxins in all reservoirs. Furthermore, upon the detected dominance of one cyanobacterial strain in Faveta reservoir we decided to perform a metagenomic approach for identification of toxin related genes. We 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 a potential future threat in these reservoirs.

## 2. Results

### 2.1. Diversity of the Microbial Communities in the Reservoirs

We profiled the bacterial biodiversity in three reservoirs of the island of Santiago using 16S rRNA gene amplicon next generation sequencing. This marker gene is universal in prokaryotes, and it is also present in eukaryotic algae, which allowed us to detect both bacteria and eukaryotic phytoplankton. For Poilão reservoir, we obtained 804,043 reads (two replicates), for Saquinho 1,271,923 reads (three replicates), and for Faveta 1,354,273 reads (three replicates), which resulted in 315,291, 437,486 and 425,806 16S rRNA reads for Poilão, Saquinho, and Faveta, respectively.

Regarding taxonomic distribution, Poilão had the highest number of operational taxonomic units (OTUs) with 1356 OTUs, followed by Saquinho with 966 and Faveta with 950 different OTUs. The OTUs identified in Poilão reservoir represented 39 phyla and included 1.5% of unassigned sequences. *Proteobacteria* was the numerically most abundant phylum with 76% of all OTUs detected, of which 69.7% were *Gamma*proteobacteria, dominated by the genus *Acinetobacter* (51.6%). The second most abundant phylum was *Bacteroidetes*, representing 4.4% of the OTUs; *Betaproteobacteria* accounted for 3.0% and *Actinobacteria* 2.3%. *Cyanobacteria* in Poilão, which represented 3.5%, are mainly represented by the *Phormidium* genus accounting for 2.5% of the OTUs. In Poilão reservoir, sequences assigned to green algae (derived from chloroplasts) were 7.5% of the OTUs and approximately equal amounts of *Chlorophyta* and stramenopiles. Main results regarding the composition of microbial communities are shown in Figure 1.



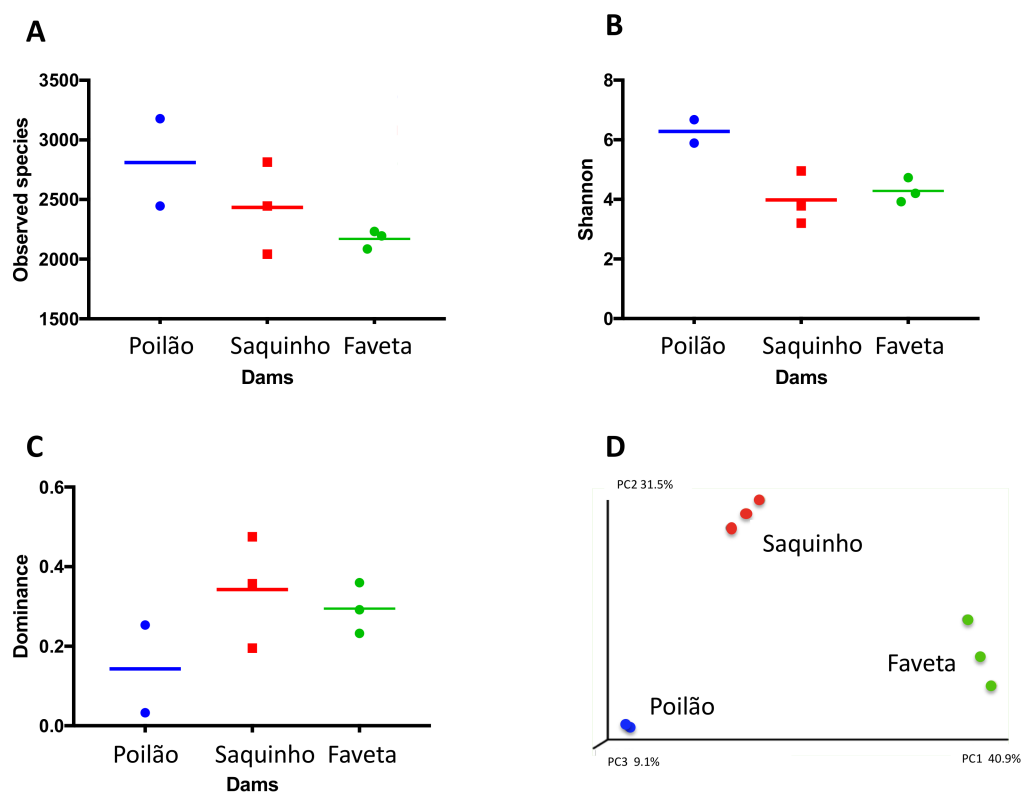
**Figure 1.** Composition of the microbe communities surveyed in Cape Verde. (A,B) Composition of the microbial communities detected in the three sampled Cape Verdean freshwater reservoirs; (C) Cyanobacteria communities detected in the surveyed reservoirs.

In the Saquinho reservoir, 37 phyla were detected along a small fraction (0.5%) of unassigned sequences. The *Proteobacteria* represented 20% of the OTUs, *Bacteroidetes* represented 5.4%, *Actinobacteria* accounted for 4.1%, and *Verrucomicrobia* represented 2.4%. *Cyanobacteria* in this reservoir represented only 0.6% of the OTUs, the lowest percentage of the three sampled reservoirs. This site had the highest number of OTUs associated with green algae (63%), distributed between the *Cryptophyta*

with 62% and stramenopiles with 1% of the identified OTUs (Figure 1A). Blastn of the cryptophyte 16S rRNA sequence revealed a hit for *Cryptomonas curvata* plastid partial ribosomal RNA operon (strain CCAC 0006) with 97% identity.

Faveta reservoir had the lowest number of phyla, accounting for 29 distinct phyla and 1.1% of the sequences classified as unknown. The most abundant phylum was *Cyanobacteria* with 58.6% of the OTUs and included two known genera to harbor toxic species: *Microcystis* accounted for 55.7% and *Planktothrix* 2.8% (Figure 1B,C). *Proteobacteria* was the second most abundant phylum with 18.6% assignments, *Bacteroidetes* was next with 7.0%, and *Actinobacteria* was third with 5.8%. The green algae at this site represented 5.1% of the identified OTUs.

In summary, our survey revealed distinct microbial community structures of the three sampled reservoirs, which is confirmed by the beta diversity analysis in Figure 2.



**Figure 2.** Biodiversity indices per replicate and average. (A) Number of total observed species, (B) Shannon, (C) dominance, and (D) principal coordinates analysis of Beta diversity by weighted Unifrac.

## 2.2. Toxin Producing Genera/Species

Putative toxin producers were identified in the three reservoirs. *Leptospira*, *Vibrio*, *Escherichia*, and *Salmonella* genera were detected in very low relative amounts, each representing less than 0.15% of the OTUs. The total frequency of these potentially pathogenic genera per reservoir was 0.22% in Poilão, 0.06% in Saquinho, and 0.16% of the OTUs in Faveta. We identified several cyanobacterial genera in the three sampled reservoirs. Representatives of these cyanobacteria are known from the literature to produce toxins. In Faveta reservoir, we identified the presence of cyanobacteria from the *Microcystis* and *Planktothrix* genera, reported to produce cyanotoxins, microcystin, and anatoxin-a, as well as other potentially toxic molecules like chlorinated and sulfated variants of aeruginosin, cyanopetolin, microginin, and microviridin [25–27]. Poilão reservoir had smaller relative quantities of cyanobacteria, but more genera were identified (Figure 1 and Figure S1), some of which were potentially toxic, such as *Microcystis*, *Phormidium*, *Planktothrix*, and *Cylindrospermopsis*. These cyanobacteria are known

to produce microcystin, homo and anatoxin-a, saxitoxin, and cylindrospermopsin, as well as other metabolites with different levels of toxicity. In Saquinho reservoir, the relative quantity of cyanobacteria was the lowest (0.6%), but even so the most represented genus was the potentially toxic *Microcystis*.

### 2.3. The Presence of a Dominant *Microcystis* Species in Faveta: Reconstruction of *M. cf. aeruginosa* CV01 Genome

Upon the detection of a dominant *Microcystis* species in the Faveta reservoir, we performed a metagenomic analysis to identify the strains and possible toxin related genes present in the pangenome. It is known that toxic *Microcystis* can co-exist alongside non-toxic strains, and that they can dominate at different times; therefore, the detection of *Microcystis* alone may not present a microcystin poisoning risk. With metagenomic sequencing, however, we were able to not only identify the organisms but also interrogate the presence of toxin-related genes in reconstructed genomes. Metagenomic binning using CONCOCT resulted in 28 bins, including a bin (#19) containing a 98.79% complete assembly of *M. aeruginosa* with a 4.7 Mbp genome (based on a CheckM assessment with a total of 75.6% of the reads mapping back to its contigs). All the other bins were also classified in terms of taxonomy, completeness, contamination, and heterogeneity (Table S1). Most of the remaining bins contained insignificant levels of completeness or contamination, but bin #16 revealed some *Microcystis* contigs with a low degree of completeness (1.72%). Bin #24 was mainly composed of *Actinobacteria* and bin #7 *Alphaproteobacteria*. We could not extract any additional information regarding these bins due to their level of incompleteness.

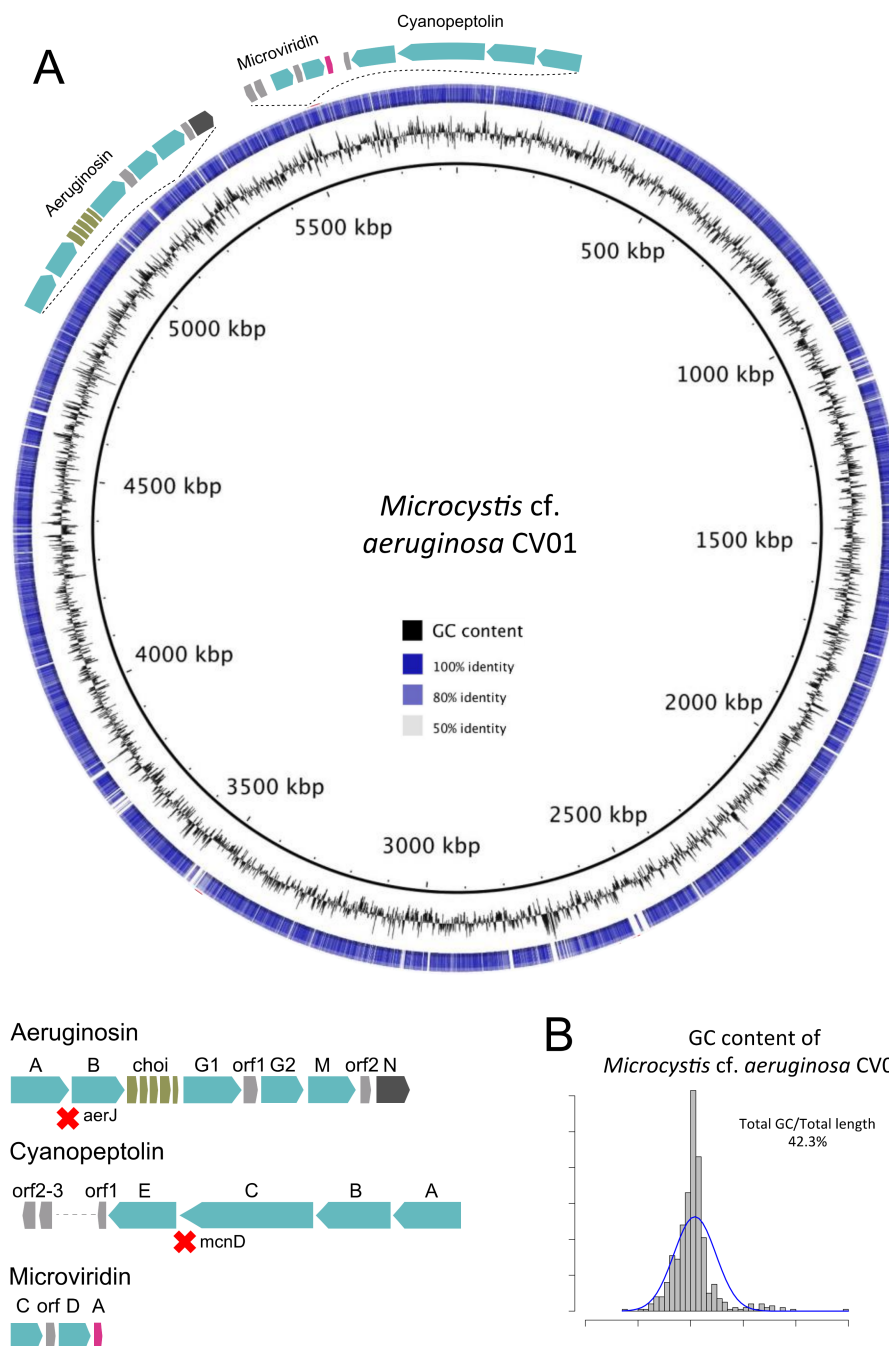
We decided to further refine binning by using the SPAdes coverage-based assembly. After SPAdes assembly refinement, we were able to generate a genome of 4.9 Mbp with a completeness level of 99.8% with 84% of the reads mapping back to contig reads from the sequencing effort (Figure S2 and Table S1). The SPAdes assembly improved the completeness level and total genome size, distributed over 262 contigs with lengths over 1000 bp and an average GC content of 42.3%. From a total of 5484 genes annotated, we identified 45 RNA genes, of which 42 were tRNA genes, one tmRNA, and one rRNA set (5S/16S/23S), as well as 5051 protein-coding genes and predicted functions for 91% of all genes (Blast2GO). Considering the importance of metabolite transport and transduction of signals through membranes in cyanobacteria, we also searched for and identified 112 genes with signal peptides and 3443 genes with transmembrane domains. Major statistical attributes of *M. cf. aeruginosa* CV01 genome are described in Table 1.

**Table 1.** 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	5484	100.0
Protein-coding genes	5051	92.1
RNA genes	45	0.8
Genes with function prediction	4996	91.1
Genes assigned to COGs	4166	76.0
Genes with Pfam domains	3407	62.1
Genes with signal peptides	112	2.0
Genes with transmembrane helices	3443	62.8
CRISPR repeats	9	-

*M. cf. aeruginosa* CV01 and *M. aeruginosa* NIES-843 share 3192 genes (plotted in a graphical representation in Figure 3). *M. cf. aeruginosa* CV01 holds 1334 genes that are absent from *M. aeruginosa* NIES-843, which in turn has 1493 genes that *M. cf. aeruginosa* CV01 does not have. The most significant absence is the microcystin synthesis cluster in *M. cf. aeruginosa* CV01, as well as *pilT* gene from type IV

pilus system. PilT protein is involved in cell functions like motility, cell adhesion, biofilm formation, and DNA uptake in bacteria, and is believed to be absent in *Microcystis* strains that do not synthesize microcystin [28,29]. Genes responsible for the transport of potassium like *kdpA*, *kdpB*, and *kdpC*, thought to be involved in salt tolerance in *Microcystis* [30], were also missing from the islander *M. cf. aeruginosa* CV01 genome.

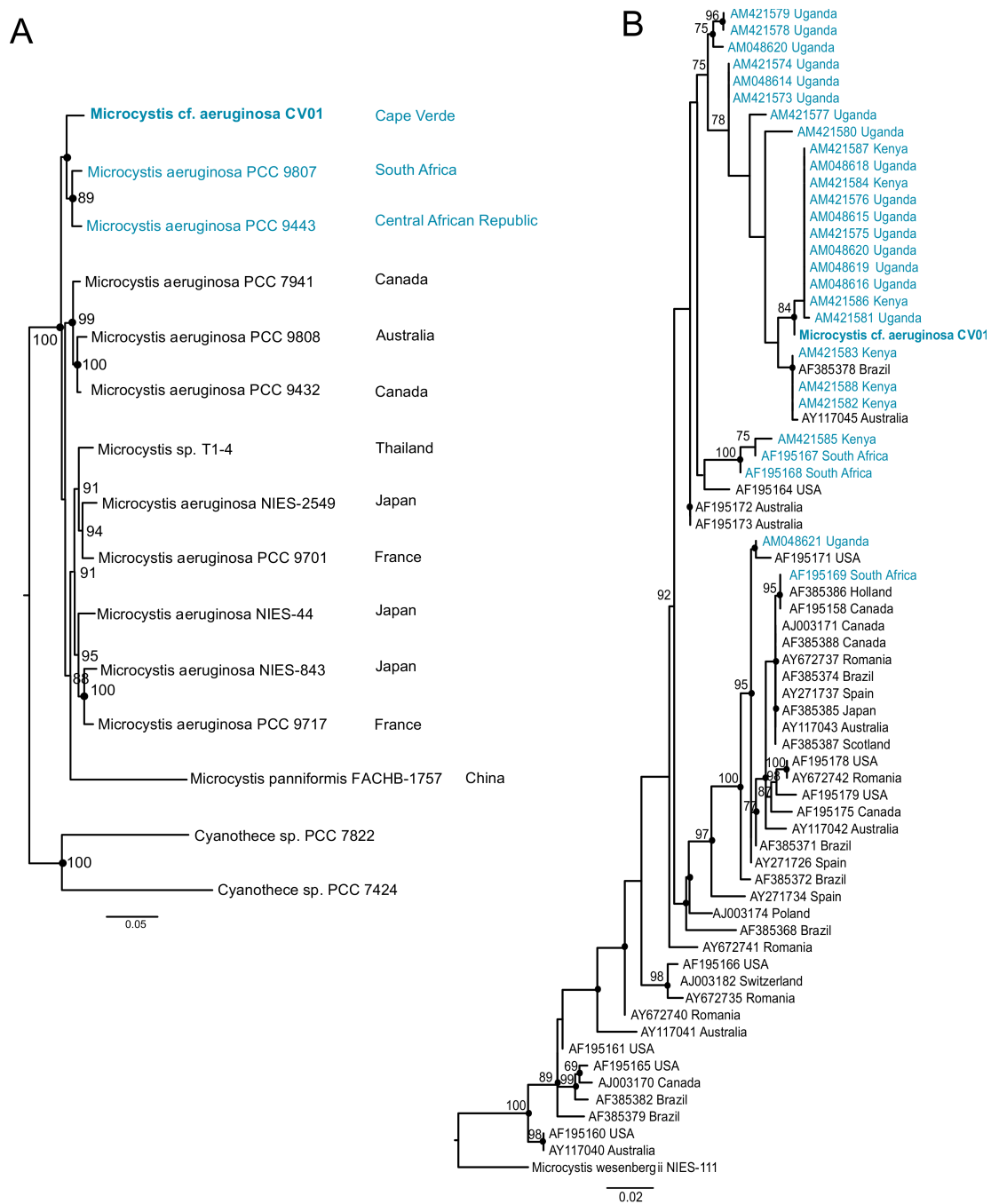


**Figure 3.** Representation of the genome of *M. cf. aeruginosa* CV01 (outer ring), sequenced from the environmental sample of Faveta reservoir. (A) The shades of blue in the circle are indicative of pairwise genomic sequence similarity according to blastn, while blank are non-corresponding regions, using as reference the genome of *M. aeruginosa* NIES-843. The relative location and structure of three gene clusters of potentially toxin-producing secondary metabolites are represented; (B) The histogram represents the GC content of the assembled *M. cf. aeruginosa* CV01 genome.

Four partial prophage regions were detected in *M. cf. aeruginosa* CV01. The first region, with GC content 45.3%, is 9.4 Kb, which encodes for 8 proteins and harbors protein-coding sequences (CDS) from a previously described infecting phage, P-TIM68, usually associated with *Prochlorococcus* Myoviridae virus that contains photosystem I gene sequences [31] and transposase sequences. A second partial phage region with 12 CDS and GC content of 43.0% spans over 11.8 Kb and contains phage sequences from the *Microcystis* infecting Myoviridae phage MaMv-DC [32]. Sequences of this phage and of phage Ma-LMM01 [33] are present in a third region with 7.1 Kb in length, 40.4% GC content, and 8 CDS. Finally, the fourth prophage region of 5.8 Kb in length, coding for 9 proteins and GC content of 40.6%, was detected, and CDS from a ssDNA marine virus reported to infect *Synechococcus* [34] were identified. Regarding CRISPR arrays, the defense mechanism of *Cyanobacteria* [35,36], we detected nine CRISPR repetitive units, varying from 0.3 Kb to 11.8 Kb in length, with direct repeat lengths from 35 to 38 bp. The CRISPR regions and the prophage did not overlap, but two clusters of CRISPR direct repeats (DR) were identical to the *Microcystis* phage Ma-LMM01 portion, a memory mechanism to their introduction into the CRISPR locus, providing immunity to further infection by that phage. Other characteristics of the *M. cf. aeruginosa* CV01 genome, such as KEGG orthologs and COG functional categories, are summarized in Tables S2 and S3.

#### 2.4. Phylogenetic Analysis

Our phylogenetic analysis using seven cyanobacterial genes identified the dominant cyanobacterium in Faveta reservoir as a *M. aeruginosa* strain, which we named *M. cf. aeruginosa* CV01 (Figures 3A and 4A). *M. cf. aeruginosa* CV01 is placed close to two strains collected in African water bodies: *M. aeruginosa* PCC 9443 collected from a fishpond in Landjica, Central African Republic; and *M. aeruginosa* PCC 9807, collected in Hartbeespoort Dam in Pretoria, South Africa (Pasteur Culture Collection of Cyanobacteria). Phylogenetic studies using the PC-IGS intergenic spacer region confirmed the clustering of *M. cf. aeruginosa* CV01 with other continental African strains, as shown in Figure 4B, in a branch that contains both microcystin producers and non-producers collected from Ugandan and Kenyan water bodies.



**Figure 4.** Phylogeny of the Cape Verdean strain *M. cf. aeruginosa* CV01. The African strains are represented in blue. (A) represents the location of *M. cf. aeruginosa* CV01 sequenced in this study among a set of fully sequenced genomes of *Microcystis*. The phylogenetic tree was inferred with MrBayes for 1 million generations and Maximum Likelihood (ML) was inferred with IQ-Tree for 1000 bootstraps, using GTR+I+G model for a set of 7 housekeeping genes (*ftsZ*, *glnA*, *gltX*, *gyrB*, *pgi*, *recA*, and *tpi*). Posterior probabilities higher than 85 are in the nodes. Dots represent ML bootstrap values higher than 65 (B) Phylogenetic tree of 69 *Microcystis* strains from all over the world locating the Cape Verdean strain among other continental African *M. aeruginosa* strains. ML phylogenies were inferred with IQ-Tree with 1000 bootstraps, and Bayesian phylogenies were inferred with MrBayes for 3 million generations, using GTR+I+G model for the PC-IGS region. Posterior probabilities higher than 75 are represented, and ML bootstrap values higher than 65 are represented by dots.



### 2.5. Toxin Genes and Toxic Species

In order to identify the risk of production of toxic cyanobacterial metabolites, we searched the *M. cf. aeruginosa* CV01 genome for non-ribosomal peptide synthases (NRPS) and NRPS/polyketide synthases (PKS) hybrid gene clusters. Genes that contribute to the synthesis of aeruginosin (NRPS/PKS/saccharide), cyanopeptolin (micropeptin), and microviridin molecules were detected. Halogenase genes *aerJ* and *mcnD* were not detected in aeruginosin, or in cyanopeptolin gene clusters or anywhere else in *M. cf. aeruginosa* CV01's genome. The absence of halogenase genes suggests the formation of non-halogenated variants of aeruginosin and cyanopeptolin. The search for other potentially toxin-related genes identified two adjacent sequences that showed some similarity to *mcyA* and *mcyB* (48 and 46% identity) in the genome. Nevertheless, other genes that are part of microcystin gene cluster were not identified, so there was no evidence for a complete microcystin gene cluster.

Putative genes for the synthesis of cyanopeptide metabolites like spumigin (aeruginosin), aeruginoside (aeruginosin), ambiguine (terpene-alkaloid), and piricyclamide (post-translational modified peptides) were also detected, but not their complete gene clusters. Finally, one unknown NRPS/PKS gene cluster was identified, showing the potential to produce peptide molecules that are yet unknown.

We took advantage of our metagenome assembly of Faveta's reservoir to search for genes related to toxin production using blast search, and we managed to identify a potential orthologous of *mcyI* gene from the microcystin cluster and another orthologous for the *anaI*, a unit from the anatoxin-a gene cluster.

### 3. Discussion

The studied reservoirs show three distinct microbial/microalgae community profiles, despite being located on the same island and in a radius of 15 kilometers from each other: in two sites, bacteria were dominant (*Proteobacteria* and *Cyanobacteria*), and in the other reservoir microalgae belonging to the cryptophytes were the most abundant taxa. The dominant species from one of the reservoirs was identified as a *M. aeruginosa* strain through phylogenetic studies, placing it closer to other strains collected in continental Africa. *Microcystis* spp. were detected in all three reservoirs, as well as other cyanobacteria known to bloom and produce cyanotoxins. Our analysis of the assembled *M. cf. aeruginosa* CV01 genome revealed that it can produce toxins, and therefore a potential risk of toxin production can exist in Cape Verdean reservoirs.

The absence of halogenase genes in aeruginosin and cyanopeptolin gene clusters is known to have no influence on the metabolite's next biosynthesis steps [37]; thus, it can be effectively synthesized by *M. cf. aeruginosa* CV01. It has been suggested that halogenase genes were present in an ancient form of cyanobacteria and were repeatedly lost in several lineages [38]. Regarding the relatively low sequence identity found in the sequences similar to *mcyA* and *mcyB* detected in the genome, it suggests that *M. cf. aeruginosa* CV01 gene degeneration implies loss of function of these genes. The discovery of large insertions and deletions affecting this cluster suggest this scenario [39]. The detection of selected microcystin and anatoxin genes in a metagenome survey does not imply by itself synthesis in Faveta reservoir, since it can belong to a pool of diverse organisms. Hence, the real toxic potential of these metabolites is difficult to determine, as the actual synthesis pathways are not fully known, and the diversity and plasticity of gene clusters in cyanobacteria allows for the production of a huge variety of analogous molecular structures and chemical activities. Therefore, further toxicological screenings need to be performed in order to fully understand these pathways.

Our microbial biodiversity survey revealed distinct community structures in the three reservoirs (Figure 2). In fact, *Proteobacteria* dominated Poilão reservoir, mainly bacteria from the *Acinetobacter* genus, while in Saquinho reservoir microalgae from the *Cryptomonas* genus were dominant, and in Faveta reservoir the cyanobacterial strain *M. cf. aeruginosa* CV01 was the most abundant. Despite being located on the same island and having common microbial groups, we found that each phyla's quantitative distribution varied substantially between

reservoirs (Figure 1A). Nevertheless, the profiles from the reservoirs identify groups that are common in other water bodies studied worldwide despite the differences in relative quantities as is the case of *Cyanobacteria*, *Cryptophyta*, *Actinobacteria*, *Bacteroidetes*, *Verrucomicrobia*, and the three clades of *Proteobacteria* (*Alpha*, *Beta*, and *Gamma*) [14,40]. Actually, the microbial profile of Poilão reservoir resembles those of lakes close to urban areas, where *Acinetobacter* is the dominant genus [41].

In each reservoir, one genus prevailed with a relative abundance above 50%: in Poilão it was *Acinetobacter*, in Saquinho it was *Cryptomonas*, and in Faveta the *Microcystis* genus.

The analysis of the local diversity indices of the replicates from each reservoir revealed consistency and reinforced the sites' distinct microbial communities (Figure 2A–C). The indices also showed that microbial communities presented different dominant genera in each of the reservoirs, as well as abundance of different taxa in all sites as typically found in freshwater bodies around the world [7–10,12].

Besides the operational starting date differences between the reservoirs and no physical communication between lakes, abiotic factors specific for each site might be involved in the dominance variations within the microbe communities.

The dominance of one cyanobacterial strain in Faveta allowed us to assemble and fully study its genome, and to identify genes, allowing the reconstruction of toxin pathways and assessing the toxin risk inherent in this specific strain. Exploration of the assembled genome also revealed genomic features in common with other *M. aeruginosa* genomes (Table 1). Some phage genes were found integrated in the genome of *M. cf. aeruginosa*. Myoviridae “photosynthetic” freshwater cyanophages (Ma-LMM01 and MaMV-DC) were also found. These genes are thought to play an important role during phage infection by supplementing the host with the production of photosynthesis proteins, a process that can be also beneficial to the host during the infection process, as suggested by some authors [31,42,43]. These horizontal gene transfer events are shaping the genome architecture of the *Microcystis* genus, providing a supplementary advantage that can be important during cyanobacteria blooms. A region containing chlorophyll *a* apoproteins A1 and A2 synthesis genes was also identified, but since these are single copy genes located near transposase sequences in this new genome, they were probably misidentified as having phage origin.

We identified four NRPS/PKS gene clusters that could synthesize potentially toxic metabolites: three well-known metabolites (aeruginosin, cyanopeptolin, and microviridin, represented in Figure 3A) and another metabolite from one yet unknown gene cluster.

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

Cyanobacterial blooms occur in freshwater reservoirs distributed worldwide where *M. aeruginosa* is one of the most frequently detected species. Actually, many long-term studies have reported toxic blooms in lakes and rivers from Kenya, Uganda, Senegal, Morocco, and South Africa [44–50], often dominated by *M. aeruginosa*, as we also detected on the island of Santiago. Moreover, a recent review on the occurrence of cyanobacterial blooms in Africa [51] shows that there is limited information from western African countries, including Cape Verde, exposing the need for further studies in countries where water quality is threatened and scarce. Therefore, our study increases the available information on cyanobacterial communities described for the western African region. The scarcity of renewable freshwater resources of archipelagic states like Cape Verde is aggravated by the terrain that favors torrential water flows and strong anthropogenic pressures on the environment leading to eutrophication of its freshwater bodies and increased risk of toxic algal blooms.

The main threat and concern from our analysis was the identification of a bacterial community dominated by *M. cf. aeruginosa* CV01, which signals the possibility of toxic blooms in Cape

Verdean reservoirs, since exponential growth is typical of this species. Alongside cyanobacteria, *Enterobacteriaceae* family members were also detected, although they are considered to be transient of the bacterial community in freshwaters, as they have anthropogenic and zoonotic origin [8].

The occurrence of blooms and toxin production are potential life-threatening risks to public health, so monitoring plans are very important. The costs involved in these control and containment strategies can be prohibitive, especially for low and middle-income countries. Hence, approaches like the one proposed in this work, which enabled the identification of potentially toxic cyanobacterial genera through 16S rRNA gene markers, could be an interesting alternative, without the time-consuming and expertise-dependent microscope identification of toxin-producing organisms or mass spectrometry-base identification of toxins. NGS is still not widely available, but DNA sampling kits are easy to use and can be sent to sequencing facilities at cost-effective prices. Other more sophisticated technics are possible such as lab on a chip, mass spectrometry, or even portable NGS devices, which can be adapted to use our workflow in the field, but if a simple molecular lab is available, a PCR assay could also be efficient at detecting the presence of specific putative cyanotoxin genes. These strategies can alert authorities and populations before bloom formation and toxin production.

Cyanobacteria are currently being developed and used for bio-production of metabolites and biomass in algal farms, for example, in the production of human dietary supplements, fertilizers, animal feed or biodiesel, to name a few uses [52–56]. Cape Verde has little land suitable for agriculture, but it has temperature and light conditions that are speculated to be suitable for simple bioreactors for biomass production using photosynthetic microorganisms like green algae or cyanobacteria. The observation of spontaneous blooms of cyanobacteria in freshwater reservoirs lends credibility to this hypothesis, and opens the way for a new productive industry in the archipelago.

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

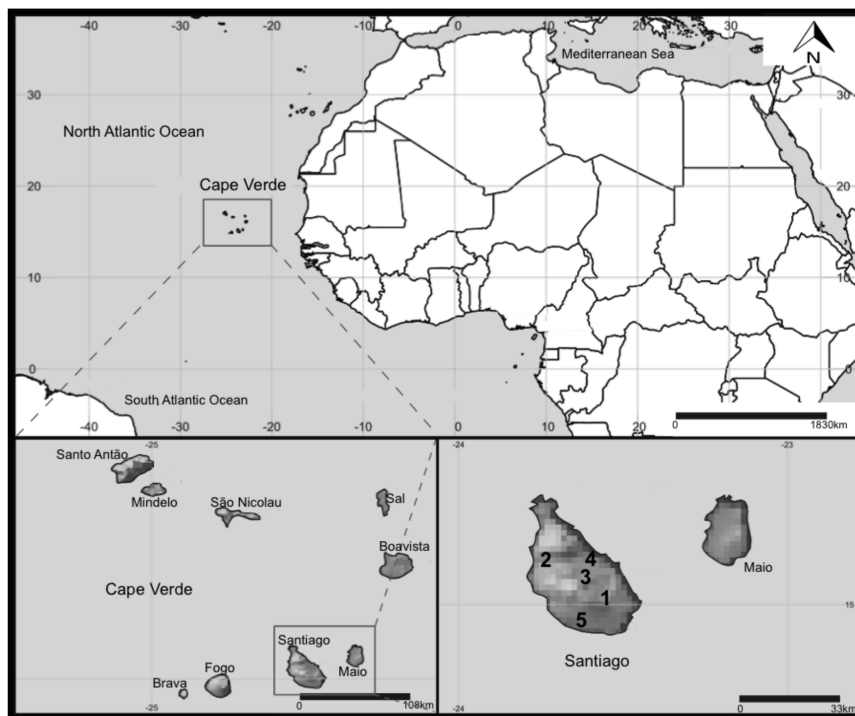
## 4. Materials and Methods

### 4.1. Study Sites and Sampling

This study was conducted on the Cape Verdean Island of Santiago (Figure 5). During sampling, in February 2014, only Poilão, Saquinho, and Faveta reservoirs had water, while Figueira Gorda and Salineiro were empty at the time of sampling. Other contextual information on Santiago Island's reservoirs is summarized in Table 2.

**Table 2.** Characteristics of the studied reservoirs in February 2014 (Government of Cape Verde).

Reservoir	Location Coordinates	Start of Impoundment	Theoretical Maximum Volume (m <sup>3</sup> )	At Full Capacity?
Poilão	15°04'25.0" N, 23°33'25.8" W	2006	1,200,000	yes
Saquinho	15°08'11.3" N, 23°42'27.7" W	2013	563,000	no
Faveta	15°05'54.5" N, 23°37'24.1" W	2013	536,565	yes
F. Gorda	15°07'5.8" N, 23°35'36.5" W	2014	1,455,272	empty
Salineiro	14°57'03.5" N, 23°38'00.4" W	2013	561,464	empty



**Figure 5.** Map of Cape Verde and location of the reservoirs in Santiago Island sampled in this study. 1—Poilão, 2—Saquinho, and 3—Faveta. 4—Figueira Gorda and 5—Salineiro reservoirs where empty at the time of sampling.

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  $\mu$ L Sterivex-GV filters (Millipore, Billerica, MA, USA) by vacuum filtration. Filters were kept on ice and in the dark while being transported to the lab for DNA extraction.

#### 4.2. DNA Extraction

Genomic DNA was extracted from the individual filters using PowerWater Sterivex DNA Isolation Kit (MO BIO, Carlsbad, CA, 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.

#### 4.3. 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 (Section 4.2) 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 [57].

The purified DNA was sequenced on an Illumina MiSeq platform using a 250 base pair paired-end DNA library (Illumina, San Diego, CA, USA) 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 (Oeiras, Portugal).

#### 4.4. Taxonomic Composition and Identification of Potentially Toxic Genus/Species

We used the QIIME v1.9.1 workflow [58] 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 [59], 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 compared with Greengenes v13\_8 [60], using Uclust algorithm [61] with default parameters into Operational Taxonomic Units—OTUs, with identity threshold at 97%. Taxonomy assignments were done with RDP Classifier v2.2 [62]. To assess the consistency of the sampling and the biodiversity present in each sampling site, diversity indices were calculated: number of Observed species, Shannon's index, and Simpson's Dominance index [63]. Archaea sequences represented less than 0.1% in each reservoir and were not considered to the analysis.

#### 4.5. Graphical Representation of the Microbial Communities' Structure Detected in the Reservoirs

Phyloseq package [64] in RStudio 1.1.419 [65] was used to perform a graphical representation of the structure of the microbial community in each of the sampled sites. A complete assignment of the more abundant bacterial genera in our samples was generated by Kraken [66] running the full Bacterial Genbank Reference sequences, excluding the sequences assigned previously as chloroplasts by QIIME.

#### 4.6. Metagenomic Sequencing and Assembly of the *Microcystis cf. aeruginosa* CV01 Genome

The same initial DNA extracted for the amplicon libraries was subsequently used to perform a whole metagenome sequencing of Faveta reservoir using Illumina MiSeq v3 kit, with a 200× coverage. The DNA libraries were created using Nextera DNA Library Preparation Kit, and sequencing was executed by IGC's Gene Expression Facility.

The sequenced reads were assembled with SPAdes v3.6.0 [67] and k-mers of 21, 33, 55, 77, 99, and 127. The GC content of the assembled metagenome from Faveta reservoir revealed two distinct contributions: one clearly in the range of the GC content typical of *Microcystis aeruginosa* (42–43%) [68] and the other peak at approximately 63% for other contributors present in the sample, such as several GC rich strains of *Gamma*, *Alpha*, and *Betaproteobacteria* (Figure S3). The peak with lower GC was in the range of our target species and most of its reads have high coverage numbers, since they belong to the numerically most abundant species. Two different approaches were used to isolate this genomic contribution: one using CONCOCT v0.41 [69], a binning tool that operates based on sequence coverage and composition, and another exploring the high coverage percentage of a single genomic contribution by assembling reads from the raw sequencing data that had covered above 60×. In both methods, short sequences were filtered out (length < 1000 bp). Taxonomy of CONCOCT bins was then determined by Kraken [66] using the default Minikraken database. BWA 0.7.16 [70] and Samtools 1.3 [71] were used to map the reads back to isolated genomes, while CheckM [72] was used to get metrics on the quality of the isolation method.

#### 4.7. Genome Annotation and General Features

We used QUAST [73] to determine the statistics of the assembly of the *M. cf. aeruginosa* CV01 genome. Annotation of the genome was done with PROKKA v1.11 [74] and RAST v2.0 [75] using default parameters to identify putative genes (coding and non-coding sequences). Protein function prediction and annotation of the predicted genes was done against KEGG orthologs (KOs) [76] and clusters of orthologs of proteins with Blast2GO [77] and eggNOG v4.5 [78]. Identification of CRISPR repeats, typical in *Cyanobacteria* and in the *Microcystis* genus, was performed with the web server CRISPRFinder [79] and Recognition Tool CRT v1.1 [80], considering a minimum of 3 repeat units. The prediction of transmembrane topology and signal peptide sites was done using Phobius [81] and SignalP v4.1 [82]. Prediction of transmembrane helices in proteins was accomplished using TMHMM v2.0 [83], prediction of ribosomal RNA subunits was done with RNAmmer v1.2 [84], while tRNA and tmRNA genes prediction was done using ARAGORN v1.2.36 [85]. To evaluate the

presence and possible origin of prophage sequences, identification and annotation of these sequences were performed using PHASTER [86].

#### 4.8. Detection of Toxin Genes in *M. cf. aeruginosa* CV01

To identify the presence of toxin genes in *M. cf. aeruginosa* CV01, we searched the genome to identify protein domains common in toxin gene clusters with Pfam v29.0 [87]. Next, we compared these sequences in our genome with putative toxin synthase genes like microcystin, nodularin, cylindrospermopsin, anatoxin-a, saxitoxin, microviridin, aeruginosin, and micropeptin, which are known to be present in cyanobacteria. We used the online tool antiSMASH v3.0 [88] to detect the presence of non-ribosomal peptide synthase (NRPS) and/or polyketide synthase (PKS) gene clusters and other domains of natural products typical of cyanobacterial metabolites. Next, by manually curating the sequences that we identified as being part of toxin synthesis process, we reconstructed the gene clusters ensuring the overlap of contig ends.

#### 4.9. *M. cf. aeruginosa* CV01 Phylogenetic Analysis

To reconstruct the phylogenetic relationships of *Microcystis* strains, we inferred a species tree using a set of 12 publicly available, fully sequenced genomes. The species tree was constructed from a concatenation of DNA sequences from a set of seven single-copy housekeeping genes present in all genomes (*ftsZ*, *glnA*, *gltX*, *gyrB*, *pgi*, *recA*, and *tpi*), using maximum likelihood phylogenies to infer the genetic variation of the 12 genomes. Gene sequences were separately aligned with MAFFT v.7.220 [89] and trimmed using Gblocks v0.91b [90,91]. To infer ML phylogenies, we used IQ-Tree v1.5.5 [92] to compute and support the trees by calculating 1000 bootstraps using GAMMA distribution and GTR+I model. Bayesian phylogenies were inferred with MrBayes v3.2 [93] for 1 million generations using the same model and a discarded burn-in rate of 25% of the initial generations. To investigate the relationship between *M. cf. aeruginosa* CV01 and other continental African, tropical, and temperate climate strains, we used the phycocyanin alpha subunit and phycocyanin beta subunit (*cpcA-cpcB*) intergenic space region of the phycocyanin gene cluster (PC-IGS) for a set of publicly available nucleic acid sequences from *Microcystis* strains with worldwide distribution. Trees were computed using IQ-Tree with 1000 bootstraps. Bayesian phylogenies were inferred with MrBayes for 3 million generations. Phylogenies were inferred with IQ-Tree under a GAMMA distribution and GTR+I model.

#### 4.10. Nucleotide Sequence Accession Numbers

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

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2072-6651/10/5/186/s1>, Figure S1: Rarefaction curves per number of reads for each sampled reservoir, Figure S2: Faveta reservoir, Figure S3: the GC content of the raw reads and the final assembled metagenome from Faveta reservoir, Table S1: Degree of completeness and contamination and percentage of mapped reads, Table S2: KEGG functional annotation of the predicted transcriptome of *Microcystis cf. aeruginosa* CV01, Table S3: COG functional categories of the predicted transcriptome of *Microcystis cf. aeruginosa* CV01.

**Author Contributions:** Conceived and designed the experiments: A.P.S.-A., J.B.P.-L., and R.B.L. Performed the experiments: A.P.S.-A. and R.B.L. Analyzed the data: A.P.S.-A. and R.B.L. Contributed reagents/materials/analysis tools: J.B.P.-L. and R.B.L. Wrote the paper: A.P.S.-A., J.B.P.-L., and R.B.L.

**Acknowledgments:** We wish to thank W. Szymaniak for help in accessing the sample sites, and Linda Amaral-Zettler for valuable insights and review of the text and the Gene Expression Facility from Instituto Gulbenkian de Ciência for all the technical support. The work was supported by the project PTDC/MAR-BIO/4132/2014 from the Portuguese Science and Technology Foundation (FCT). Ana P. Semedo-Aguiar was a recipient of fellowship SFRH/BD/113752/2015 from FCT, under the auspices of the Programa de Pós-Graduação de Ciência para o Desenvolvimento (PGCD), and RBL was a recipient of fellowship SFRH/BPD/91518/2012 from FCT.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Lobo de Pina, A. Hidroquímica e Qualidade das Águas Subterrâneas da Ilha de Santiago—Cabo Verde. Ph.D. Thesis, Aveiro University, Aveiro, Portugal, 2009.
2. Dodds, W.K.; Bouska, W.W.; Eitzmann, J.L.; Pilger, T.J.; Pitts, K.L.; Riley, A.J.; Schloesser, J.T.; Thornbrugh, D.J. Eutrophication of U.S. freshwaters: Analysis of potential economic damages. *Environ. Sci. Technol.* **2009**, *43*, 12–19. [[CrossRef](#)] [[PubMed](#)]
3. Heisler, J.; Glibert, P.M.; Burkholder, J.M.; Anderson, D.M.; Cochlan, W.; Dennison, W.C.; Dortch, Q.; Gobler, C.J.; Heil, C.A.; Humphries, E.; et al. Eutrophication and harmful algal blooms: A scientific consensus. *Harmful Algae* **2008**, *8*, 3–13. [[CrossRef](#)] [[PubMed](#)]
4. Paerl, H.W.; Paul, V.J. Climate change: Links to global expansion of harmful cyanobacteria. *Water Res.* **2012**, *46*, 1349–1363. [[CrossRef](#)] [[PubMed](#)]
5. Paerl, H.W.; Scott, J.T. Throwing fuel on the fire: Synergistic effects of excessive nitrogen inputs and global warming on harmful algal blooms. *Environ. Sci. Technol.* **2010**, *44*, 7756–7758. [[CrossRef](#)] [[PubMed](#)]
6. O’Neil, J.M.; Davis, T.W.; Burford, M.A.; Gobler, C.J. The rise of harmful cyanobacteria blooms: The potential roles of eutrophication and climate change. *Harmful Algae* **2012**, *14*, 313–334. [[CrossRef](#)]
7. Jones, S.E.; Newton, R.J.; McMahon, K.D. Evidence for structuring of bacterial community composition by organic carbon source in temperate lakes. *Environ. Microbiol.* **2009**, *11*, 2463–2472. [[CrossRef](#)] [[PubMed](#)]
8. Newton, R.J.; Jones, S.E.; Eiler, A.; McMahon, K.D.; Bertilsson, S. A Guide to the Natural History of Freshwater Lake Bacteria. *Microbiol. Mol. Biol. Rev.* **2011**, *75*, 14–49. [[CrossRef](#)] [[PubMed](#)]
9. Oh, S.; Caro-Quintero, A.; Tsementzi, D.; DeLeon-Rodriguez, N.; Luo, C.; Poretsky, R.; Konstantinidis, K.T. Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl. Environ. Microbiol.* **2011**, *77*, 6000–6011. [[CrossRef](#)] [[PubMed](#)]
10. Eiler, A.; Drakare, S.; Bertilsson, S.; Pernthaler, J.; Peura, S.; Rofner, C.; Simek, K.; Yang, Y.; Znachor, P.; Lindström, E.S. Unveiling Distribution Patterns of Freshwater Phytoplankton by a Next Generation Sequencing Based Approach. *PLoS ONE* **2013**, *8*, e53516. [[CrossRef](#)] [[PubMed](#)]
11. Yan, Q.; Bi, Y.; Deng, Y.; He, Z.; Wu, L.; Van Nostrand, J.D.; Shi, Z.; Li, J.; Wang, X.; Hu, Z.; et al. Impacts of the Three Gorges Dam on microbial structure and potential function. *Nat. Sci. Rep.* **2015**, *5*, 8605. [[CrossRef](#)] [[PubMed](#)]
12. Eiler, A.; Bertilsson, S. Flavobacteria blooms in four eutrophic lakes: Linking population dynamics of freshwater bacterioplankton to resource availability. *Appl. Environ. Microbiol.* **2007**, *73*, 3511–3518. [[CrossRef](#)] [[PubMed](#)]
13. Kosten, S.; Huszar, V.L.M.; Bécares, E.; Costa, L.S.; van Donk, E.; Hansson, L.A.; Jeppesen, E.; Kruk, C.; Lacerot, G.; Mazzeo, N.; et al. Warmer climates boost cyanobacterial dominance in shallow lakes. *Glob. Chang. Biol.* **2012**, *18*, 118–126. [[CrossRef](#)]
14. Eiler, A.; Bertilsson, S. Composition of freshwater bacterial communities associated with cyanobacterial blooms in four Swedish lakes. *Environ. Microbiol.* **2004**, *6*, 1228–1243. [[CrossRef](#)] [[PubMed](#)]
15. Conradie, K.R.; Barnard, S. The dynamics of toxic *Microcystis* strains and microcystin production in two hypertrophic South African reservoirs. *Harmful Algae* **2012**, *20*, 1–10. [[CrossRef](#)]
16. Davis, T.W.; Berry, D.L.; Boyer, G.L.; Gobler, C.J. The effects of temperature and nutrients on the growth and dynamics of toxic and non-toxic strains of *Microcystis* during cyanobacteria blooms. *Harmful Algae* **2009**, *8*, 715–725. [[CrossRef](#)]
17. Chorus, I.; Falconer, I.R.; Salas, H.J.; Bartram, J. Health risks caused by freshwater cyanobacteria in recreational waters. *J. Toxicol. Environ. Heal. Part B* **2000**, *3*, 323–347. [[CrossRef](#)]
18. Paerl, H.W.; Hall, N.S.; Calandrino, E.S. Controlling harmful cyanobacterial blooms in a world experiencing anthropogenic and climatic-induced change. *Sci. Total Environ.* **2011**, *409*, 1739–1745. [[CrossRef](#)] [[PubMed](#)]
19. Paerl, H.W.; Huisman, J. Blooms Like It Hot. *Science* **2008**, *320*, 57–58. [[CrossRef](#)] [[PubMed](#)]
20. Alvarenga, D.O.; Fiore, M.F.; Varani, A.M. A metagenomic approach to cyanobacterial genomics. *Front. Microbiol.* **2017**, *8*, 809. [[CrossRef](#)] [[PubMed](#)]
21. Cabello-Yeves, P.J.; Zemskay, T.I.; Rosselli, R.; Coutinho, F.H.; Zakharenko, A.S.; Blinov, V.V.; Rodriguez-Valera, F. Genomes of novel microbial lineages assembled from the sub-ice waters of Lake Baikal. *Appl. Environ. Microbiol.* **2018**, *84*. [[CrossRef](#)] [[PubMed](#)]

22. He, S.; Stevens, S.L.R.; Chan, L.-K.; Bertilsson, S.; Glavina del Rio, T.; Tringe, S.G.; Malmstrom, R.R.; McMahon, K.D. Ecophysiology of freshwater Verrucomicrobia inferred from Metagenome-Assembled Genomes. *mSphere* **2017**, *2*, e00277-17. [[CrossRef](#)] [[PubMed](#)]
23. Pinto, F.; Tett, A.; Armanini, F.; Asnicar, F.; Boscaini, A.; Pasolli, E.; Zolfo, M.; Donati, C.; Salmaso, N.; Segata, N. Draft genome sequences of novel Pseudomonas, Flavobacterium, and Sediminibacterium strains from a freshwater ecosystem. *Genome Announc.* **2018**, *6*, e00009-18. [[CrossRef](#)] [[PubMed](#)]
24. Mehrshad, M.; Amoozegar, M.A.; Ghai, R.; Shahzadeh Fazeli, S.A.; Rodriguez-Valera, F. Genome reconstruction from metagenomic data sets reveals novel microbes in the brackish waters of the Caspian Sea. *Appl. Environ. Microbiol.* **2016**, *82*, 1599–1612. [[CrossRef](#)] [[PubMed](#)]
25. Kohler, E.; Grundler, V.; Haussinger, D.; Kurmayer, R.; Gademann, K.; Pernthaler, J.; Blom, J.F. The toxicity and enzyme activity of a chlorine and sulfate containing aeruginosin isolated from a non-microcystin-producing *Planktothrix* strain. *Harmful Algae* **2014**, *39*, 154–160. [[CrossRef](#)] [[PubMed](#)]
26. Quiblier, C.; Wood, S.; Echenique Subiabre, I.; Heath, M.; Villeneuve, A.; Humbert, J.-F. A review of current knowledge on toxic benthic freshwater cyanobacteria—Ecology, toxin production and risk management. *Water Res.* **2013**, *47*, 5464–5479. [[CrossRef](#)]
27. Rounge, T.B.; Rohrlack, T.; Nederbragt, A.J.; Kristensen, T.; Jakobsen, K.S. A genome-wide analysis of nonribosomal peptide synthetase gene clusters and their peptides in a *Planktothrix rubescens* strain. *BMC Genom.* **2009**, *10*, 396. [[CrossRef](#)] [[PubMed](#)]
28. Schuergers, N.; Wilde, A. Appendages of the cyanobacterial cell. *Life* **2015**, 700–715. [[CrossRef](#)] [[PubMed](#)]
29. Nakasugi, K.; Alexova, R.; Svenson, C.J.; Neilan, B.A. Functional analysis of PilT from the toxic Cyanobacterium *Microcystis aeruginosa* PCC 7806. *J. Bacteriol.* **2007**, *189*, 1689–1697. [[CrossRef](#)] [[PubMed](#)]
30. Sandrini, G.; Huisman, J.; Matthijs, H.C.P. Potassium sensitivity differs among strains of the harmful cyanobacterium *Microcystis* and correlates with the presence of salt tolerance genes. *FEMS Microbiol. Lett.* **2015**, *362*, fnv121. [[CrossRef](#)] [[PubMed](#)]
31. Sharon, I.; Alperovitch, A.; Rohwer, F.; Haynes, M.; Glaser, F.; Atamna-Ismaeel, N.; Pinter, R.Y.; Partensky, F.; Koonin, E.V.; Wolf, Y.I.; et al. Photosystem I gene cassettes are present in marine virus genomes. *Nature* **2009**, *461*, 258–262. [[CrossRef](#)] [[PubMed](#)]
32. Ou, T.; Gao, X.C.; Li, S.H.; Zhang, Q.Y. Genome analysis and gene nblA identification of *Microcystis aeruginosa* myovirus (MaMV-DC) reveal the evidence for horizontal gene transfer events between cyanomyovirus and host. *J. Gen. Virol.* **2015**, *96*, 3681–3697. [[CrossRef](#)] [[PubMed](#)]
33. Yoshida, T.; Nagasaki, K.; Takashima, Y.; Shirai, Y.; Tomaru, Y.; Takao, Y.; Sakamoto, S.; Hiroishi, S.; Ogata, H. Ma-LMM01 infecting toxic *Microcystis aeruginosa* illuminates diverse cyanophage genome strategies. *J. Bacteriol.* **2008**, *190*, 1762–1772. [[CrossRef](#)] [[PubMed](#)]
34. Huang, S.; Wang, K.; Jiao, N.; Chen, F. Genome sequences of siphoviruses infecting marine *Synechococcus* unveil a diverse cyanophage group and extensive phage–host genetic exchanges. *Environ. Microbiol.* **2012**, *14*, 540–558. [[CrossRef](#)] [[PubMed](#)]
35. Cai, F.; Axen, S.D.; Kerfeld, C.A. Evidence for the widespread distribution of CRISPR-Cas system in the Phylum Cyanobacteria. *RNA Biol.* **2013**, *10*, 687–693. [[CrossRef](#)] [[PubMed](#)]
36. Yang, C.; Lin, F.; Li, Q.; Li, T.; Zhao, J. Comparative genomics reveals diversified CRISPR-Cas systems of globally distributed *Microcystis aeruginosa*, a freshwater bloom-forming cyanobacterium. *Front. Microbiol.* **2015**, *6*. [[CrossRef](#)] [[PubMed](#)]
37. Cadel-Six, S.; Dauga, C.; Castets, A.M.; Rippka, R.; Bouchier, C.; Tandeau de Marsac, N.; Welker, M. Halogenase genes in nonribosomal peptide synthetase gene clusters of *Microcystis* (Cyanobacteria): Sporadic distribution and evolution. *Mol. Biol. Evol.* **2008**, *25*, 2031–2041. [[CrossRef](#)] [[PubMed](#)]
38. Nishizawa, T.; Ueda, A.; Nakano, T.; Nishizawa, A.; Miura, T.; Asayama, M.; Fujii, K.; Harada, K.; Shirai, M. Characterization of the locus of genes encoding enzymes producing heptadepsipeptide micropeptin in the unicellular cyanobacterium *Microcystis*. *J. Biochem.* **2011**, *149*, 475–485. [[CrossRef](#)] [[PubMed](#)]
39. Christiansen, G.; Molitor, C.; Philmus, B.; Kurmayer, R. Nontoxic strains of cyanobacteria are the result of major gene deletion events induced by a transposable element. *Mol. Biol. Evol.* **2008**, *25*, 1695–1704. [[CrossRef](#)] [[PubMed](#)]
40. Humbert, J.-F.; Dorigo, U.; Cecchi, P.; Le Berre, B.; Debroyas, D.; Bouvy, M. Comparison of the structure and composition of bacterial communities from temperate and tropical freshwater ecosystems. *Environ. Microbiol.* **2009**, *11*, 2339–2350. [[CrossRef](#)] [[PubMed](#)]



41. McLellan, S.L.; Fisher, J.C.; Newton, R.J. The microbiome of urban waters. *Int. Microbiol.* **2015**, *18*, 141–149. [[CrossRef](#)] [[PubMed](#)]
42. Lindell, D.; Jaffe, J.D.; Johnson, Z.I.; Church, G.M.; Chisholm, S.W. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **2005**, *438*, 86–89. [[CrossRef](#)] [[PubMed](#)]
43. Yoshida-Takashima, Y.; Yoshida, M.; Ogata, H.; Nagasaki, K.; Hiroishi, S.; Yoshida, T. Cyanophage infection in the bloom-forming cyanobacteria *Microcystis aeruginosa* in surface freshwater. *Microbes Environ.* **2012**, *27*, 350–355. [[CrossRef](#)] [[PubMed](#)]
44. Okello, W.; Portmann, C.; Erhard, M.; Gademann, K.; Kurmayer, R. Occurrence of microcystin-producing cyanobacteria in Ugandan freshwater habitats. *Environ. Toxicol.* **2010**, *25*, 367–380. [[CrossRef](#)] [[PubMed](#)]
45. Thomazeau, S.; Houdan-Fourmont, A.; Couté, A.; Duval, C.; Couloux, A.; Rousseau, F.; Bernard, C. The contribution of Sub-Saharan African strains to the phylogeny of cyanobacteria: Focusing on the nostocaceae (nostocales, cyanobacteria). *J. Phycol.* **2010**, *46*, 564–579. [[CrossRef](#)]
46. Haande, S.; Rohrlack, T.; Ballot, A.; Røberg, K.; Skulberg, R.; Beck, M.; Wiedner, C. Genetic characterisation of *Cylindrospermopsis raciborskii* (Nostocales, Cyanobacteria) isolates from Africa and Europe. *Harmful Algae* **2008**, *7*, 692–701. [[CrossRef](#)]
47. Berger, C.; Ba, N.; Gugger, M.; Bouvy, M.; Rusconi, F.; Couté, A.; Troussellier, M.; Bernard, C. Seasonal dynamics and toxicity of *Cylindrospermopsis raciborskii* in Lake Guiers (Senegal, West Africa). *FEMS Microbiol. Ecol.* **2006**, *57*, 355–366. [[CrossRef](#)] [[PubMed](#)]
48. Douma, M.; Ouahid, Y.; Campo, F.F.; Del Loudiki, M.; Mouhri, K.; Oudra, B. Identification and quantification of cyanobacterial toxins (microcystins) in two Moroccan drinking-water reservoirs (Mansour Eddahbi, Almassira). *Environ. Monit. Assess.* **2010**, *160*, 439–450. [[CrossRef](#)] [[PubMed](#)]
49. Matthews, M.; Bernard, S. Eutrophication and cyanobacteria in South Africa’s standing water bodies: A view from space. *S. Afr. J. Sci.* **2015**, *111*, 1–8. [[CrossRef](#)]
50. Haande, S.; Ballot, A.; Rohrlack, T.; Fastner, J.; Wiedner, C.; Edvardsen, B. Diversity of *Microcystis aeruginosa* isolates (Chroococcales, Cyanobacteria) from East-African water bodies. *Arch. Microbiol.* **2007**, *188*, 15–25. [[CrossRef](#)] [[PubMed](#)]
51. Ndlela, L.L.; Oberholster, P.J.; Van Wyk, J.H.; Cheng, P.H. An overview of cyanobacterial bloom occurrences and research in Africa over the last decade. *Harmful Algae* **2016**, *60*, 11–26. [[CrossRef](#)] [[PubMed](#)]
52. Englund, E.; Pattanaik, B.; Ubhayasekera, S.J.K.; Stensjö, K.; Bergquist, J.; Lindberg, P. Production of Squalene in *Synechocystis* sp. PCC 6803. *PLoS ONE* **2014**, *9*, e90270. [[CrossRef](#)] [[PubMed](#)]
53. Zanchett, G.; Oliveira-Filho, E. Cyanobacteria and Cyanotoxins: From Impacts on Aquatic Ecosystems and Human Health to Anticarcinogenic Effects. *Toxins (Basel)* **2013**, *5*, 1896–1917. [[CrossRef](#)] [[PubMed](#)]
54. Enzing, C.; Ploeg, M.; Barbosa, M.; Lolke, S. *Microalgae-Based Products for the Food and Feed Sector: An Outlook for Europe*; European Union: Luxembourg, 2014.
55. Simas-Rodrigues, C.; Villela, H.D.M.; Martins, A.P.; Marques, L.G.; Colepicolo, P.; Tonon, A.P. Microalgae for economic applications: Advantages and perspectives for bioethanol. *J. Exp. Bot.* **2015**, *66*, 4097–4108. [[CrossRef](#)] [[PubMed](#)]
56. Brennan, L.; Owende, P. Biofuels from microalgae—A review of technologies for production, processing, and extractions of biofuels and co-products. *Renew. Sustain. Energy Rev.* **2010**, *14*, 557–577. [[CrossRef](#)]
57. Klindworth, A.; Pruesse, E.; Schweer, T.; Peplies, J.; Quast, C.; Horn, M.; Glockner, F.O. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **2013**, *41*, e1. [[CrossRef](#)] [[PubMed](#)]
58. Caporaso, J.G.; Kuczynski, J.; Stombaugh, J.; Bittinger, K.; Bushman, F.D.; Costello, E.K.; Fierer, N.; Peña, A.G.; Goodrich, J.K.; Gordon, J.I.; et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **2010**, *7*, 335–336. [[CrossRef](#)] [[PubMed](#)]
59. Bokulich, N.A.; Subramanian, S.; Faith, J.J.; Gevers, D.; Gordon, I.; Knight, R.; Mills, D.A.; Caporaso, J.G. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* **2013**, *10*, 57–59. [[CrossRef](#)] [[PubMed](#)]
60. McDonald, D.; Price, M.N.; Goodrich, J.; Nawrocki, E.P.; DeSantis, T.Z.; Probst, A.; Andersen, G.L.; Knight, R.; Hugenholtz, P. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **2012**, *6*, 610–618. [[CrossRef](#)] [[PubMed](#)]
61. Edgar, R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **2010**, *26*, 2460–2461. [[CrossRef](#)] [[PubMed](#)]

62. Wang, Q.; Garrity, G.M.; Tiedje, J.M.; Cole, J.R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **2007**, *73*, 5261–5267. [[CrossRef](#)] [[PubMed](#)]
63. Morris, E.K.; Caruso, T.; Fischer, M.; Hancock, C.; Obermaier, E.; Prati, D.; Maier, T.S.; Meiners, T.; Caroline, M.; Wubet, T.; et al. Choosing and using diversity indices: Insights for ecological applications from the German Biodiversity Exploratories. *Ecol. Evol.* **2014**, *4*, 3514–3524. [[CrossRef](#)] [[PubMed](#)]
64. McMurdie, P.J.; Holmes, S. Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* **2013**, *8*, e61217. [[CrossRef](#)] [[PubMed](#)]
65. Racine, J.S. RStudio: A platform-independent IDE for R and Sweave. *J. Appl. Econom.* **2012**, *27*, 167–172. [[CrossRef](#)]
66. Wood, D.E.; Salzberg, S.L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **2014**, *15*, R46. [[CrossRef](#)] [[PubMed](#)]
67. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Pribelski, A.D.; et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [[CrossRef](#)] [[PubMed](#)]
68. Humbert, J.-F.; Barbe, V.; Latifi, A.; Gugger, M.; Calteau, A.; Coursin, T.; Lajus, A.; Castelli, V.; Oztas, S.; Samson, G.; et al. A tribute to disorder in the genome of the bloom-forming freshwater cyanobacterium *Microcystis aeruginosa*. *PLoS ONE* **2013**, *8*, e70747. [[CrossRef](#)] [[PubMed](#)]
69. Alneberg, J.; Bjarnason, B.S.; de Bruijn, I.; Schirmer, M.; Quick, J.; Ijaz, U.Z.; Lahti, L.; Loman, N.J.; Andersson, A.F.; Quince, C. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **2014**, *11*, 1144–1146. [[CrossRef](#)] [[PubMed](#)]
70. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [[CrossRef](#)] [[PubMed](#)]
71. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [[CrossRef](#)] [[PubMed](#)]
72. 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.* **2015**, *25*, 1043–1055. [[CrossRef](#)] [[PubMed](#)]
73. Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **2013**, *29*, 1072–1075. [[CrossRef](#)] [[PubMed](#)]
74. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **2014**, *30*, 2068–2069. [[CrossRef](#)] [[PubMed](#)]
75. Aziz, R.K.; Bartels, D.; Best, A.A.; DeJongh, M.; Disz, T.; Edwards, R.A.; Formsma, K.; Gerdes, S.; Glass, E.M.; Kubal, M.; et al. The RAST Server: Rapid annotations using subsystems technology. *BMC Genom.* **2008**, *9*, 75. [[CrossRef](#)] [[PubMed](#)]
76. Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **2016**, *44*, D457–D462. [[CrossRef](#)] [[PubMed](#)]
77. Conesa, A.; Goetz, S.; García-Gómez, J.M.; Terol, J.; Talón, M.; Robles, M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674–3676. [[CrossRef](#)] [[PubMed](#)]
78. Huerta-Cepas, J.; Szklarczyk, D.; Forslund, K.; Cook, H.; Heller, D.; Walter, M.C.; Rattei, T.; Mende, D.R.; Sunagawa, S.; Kuhn, M.; et al. EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **2016**, *44*, D286–D293. [[CrossRef](#)] [[PubMed](#)]
79. Grissa, I.; Vergnaud, G.; Pourcel, C. CRISPRfinder: A web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* **2007**, *35*, W52–W57. [[CrossRef](#)] [[PubMed](#)]
80. Bland, C.; Ramsey, T.L.; Sabree, F.; Lowe, M.; Brown, K.; Kyrpides, N.C.; Hugenholtz, P. CRISPR Recognition Tool (CRT): A tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinform.* **2007**, *8*, 209. [[CrossRef](#)] [[PubMed](#)]
81. Kall, L.; Krogh, A.; Sonnhammer, E.L.L. Advantages of combined transmembrane topology and signal peptide prediction—The Phobius web server. *Nucleic Acids Res.* **2007**, *35*, W429–W432. [[CrossRef](#)] [[PubMed](#)]

82. Petersen, T.N.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nat. Methods* **2011**, *8*, 785–786. [[CrossRef](#)] [[PubMed](#)]
83. Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes<sup>11</sup>Edited by F. Cohen. *J. Mol. Biol.* **2001**, *305*, 567–580. [[CrossRef](#)] [[PubMed](#)]
84. Lagesen, K.; Hallin, P.; Rødland, E.A.; Stærfeldt, H.H.; Rognes, T.; Ussery, D.W. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **2007**, *35*, 3100–3108. [[CrossRef](#)] [[PubMed](#)]
85. Laslett, D.; Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **2004**, *32*, 11–16. [[CrossRef](#)] [[PubMed](#)]
86. Arndt, D.; Grant, J.R.; Marcu, A.; Sajed, T.; Pon, A.; Liang, Y.; Wishart, D.S. PHASTER: A better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **2016**, *44*, W16–W21. [[CrossRef](#)] [[PubMed](#)]
87. Finn, R.D.; Coghill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-vegas, A.; et al. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **2016**, *44*, D279–D285. [[CrossRef](#)] [[PubMed](#)]
88. Weber, T.; Blin, K.; Duddela, S.; Krug, D.; Kim, H.U.; Brucoleri, R.; Lee, S.Y.; Fischbach, M.A.; Müller, R.; Wohlleben, W.; et al. AntiSMASH 3.0—A comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.* **2015**, *43*, W237–W243. [[CrossRef](#)] [[PubMed](#)]
89. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [[CrossRef](#)] [[PubMed](#)]
90. Castresana, J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Mol. Biol. Evol.* **2000**, *17*, 540–552. [[CrossRef](#)] [[PubMed](#)]
91. Talavera, G.; Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **2007**, *56*, 564–577. [[CrossRef](#)] [[PubMed](#)]
92. Nguyen, L.T.; Schmidt, H.A.; Von Haeseler, A.; Minh, B.Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274. [[CrossRef](#)] [[PubMed](#)]
93. Ronquist, F.; Teslenko, M.; Van Der Mark, P.; Ayres, D.L.; Darling, A.; Höhna, S.; Larget, B.; Liu, L.; Suchard, M.A.; Huelsenbeck, J.P. MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **2012**, *61*, 539–542. [[CrossRef](#)] [[PubMed](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).