UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
Departamento de Botânica
Programa de Pós-Graduação em Biologia Vegetal

PPGBV
UFMG

UFMG

MARCELE LAUX

AVALIAÇÃO DA DIVERSIDADE GENÉTICA DA COMUNIDADE PLANCTÔNICA DE RESERVATÓRIOS DE ÁGUA DOCE TROPICAIS, ATRAVÉS DE ANÁLISE METAGENÔMICA

Tese apresentada ao Programa de Pós-Graduação em Biologia Vegetal do Departamento de Botânica do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Biologia Vegetal.

Área de Concentração Ecofisiologia

BELO HORIZONTE – MG

2016

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
Departamento de Botânica
Programa de Pós-Graduação em Biologia Vegetal

PPGBV
U F M G

UFMG

MARCELE LAUX

AVALIAÇÃO DA DIVERSIDADE GENÉTICA DA COMUNIDADE PLANCTÔNICA DE RESERVATÓRIOS DE ÁGUA DOCE TROPICAIS, ATRAVÉS DE ANÁLISE METAGENÔMICA

Tese apresentada ao Programa de Pós-Graduação em Biologia Vegetal do Departamento de Botânica do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Biologia Vegetal.

Área de Concentração Ecofisiologia

Orientador: Prof. Dr. Alessandra Giani
Universidade Federal de Minas Gerais

Coorientador: Prof. Dr. José Miguel Ortega
Universidade Federal de Minas Gerais

BELO HORIZONTE – MG

2016

Laux, Marcele.

Avaliação da diversidade genética da comunidade planctônica de reservatórios de água doce tropicais, através de análise metagenômica. [manuscrito] / Marcele Laux. – 2016.

198 f. : il. ; 29,5 cm.

Orientadora: Profa. Dra. Alessandra Giani. Coorientador: Prof. Dr. José Miguel Ortega.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas.

1. Bioma tropical. 2. Reservatórios - Teses. 3. Cianobactéria - Teses. 4. Metagenômica - Teses. 5. Botânica - Teses. I. Giani, Alessandra. II. Ortega, José Miguel. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 581.6

UFMG

Programa de Pós-Graduação em Biologia Vegetal
Universidade Federal de Minas Gerais
ICB - Departamento de Botânica

Tese defendida e aprovada em 31 de março de 2016, pela Banca Examinadora constituída pelos professores:

Dra. Alessandra Giani (UFMG)

Dra. Marli de Fátima Fiore (USP- Piracicaba – CENA)

Dr. Fernando Dini Andreote (USP- Piracicaba –CENA)

Dr. Ricardo Motta Pinto Coelho (UFMG)

Dr. Álvaro Cantini Nunes (UFMG)

Agradecimentos

Agradeço à professora Alessandra Giani pela orientação e o apoio que demonstrou mesmo antes de eu iniciar o doutorado, acreditando no nosso projeto e acreditando na minha capacidade de realizá-lo.

Agradeço ao Departamento de Botânica, em especial à professora Denise Maria Trombert de Oliveira, ao professor Cleber Cunha Figueredo, ao professor João Renato Stehmann e à professora Rosy Mary dos Santos Isaias.

Agradeço à equipe do Laboratório Biodados, em especial professor Miguel José Ortega, que disponibilizou seus servidores para que eu pudesse realizar meu trabalho, e ao colega Ricardo Vialle, por me ajudar nos primeiros passos da bioinformática.

Agradeço aos colegas do Laboratório de Ficologia, em especial Valquíria e Elenice pelo apoio técnico.

Agradeço à Selminha pelos cafés.

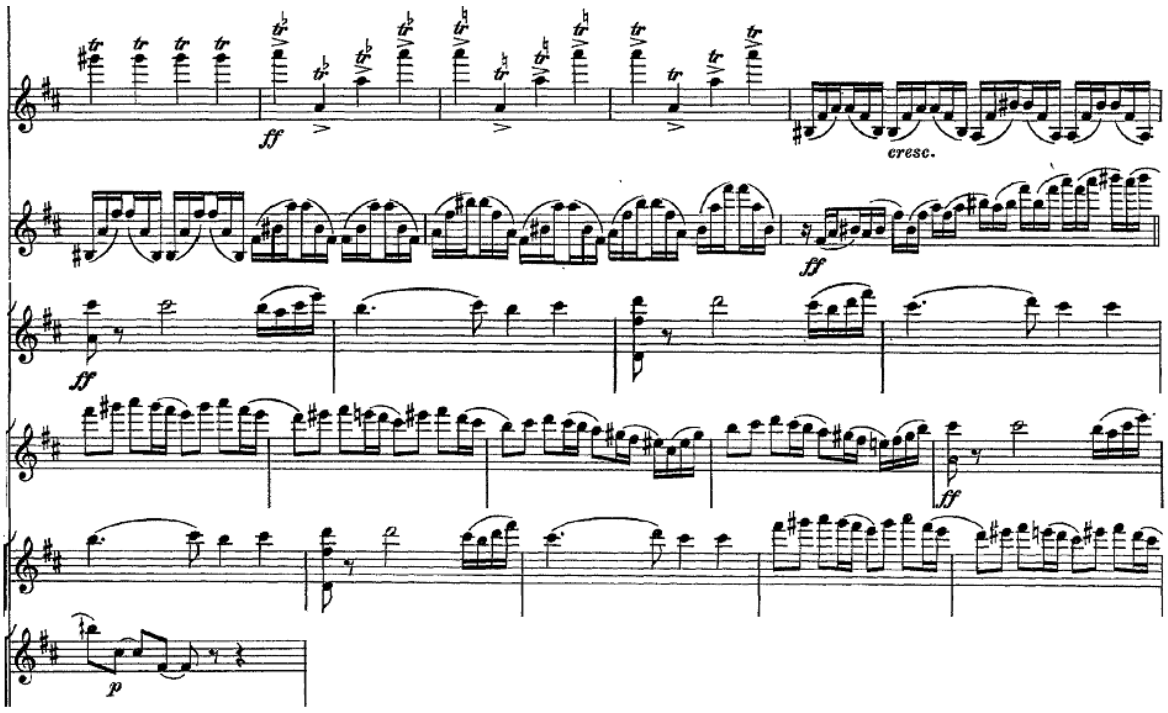
Agradeço às escolas públicas pelas quais passei, agradeço à UFRGS e à UFMG, agradeço ao CNPq, Faufrgs, Fundep e CAPES, cujas bolsas foram essenciais para eu seguir esse caminho.

Agradeço a minha família pelo apoio e estímulo.

Agradeço a meus amigos pela paciência, especialmente à Dávia Talgatti.

Agradeço à Cláudia, que me deu um lar e a segurança que eu precisava para ter coragem de continuar.

Por fim, meu agradecimento especial é para meu país, que me disponibilizou educação plena desde a pré-escola e um contexto social e cultural que permitiu que eu escolhesse e seguisse meu objetivo.



Concerto para Violino em D maior, Op. 35, I. Allegro moderato

Compasso 119-141, solo e tema

Tchaikovsky P. I., (1881)

LISTA DE ABREVIATURAS

WGS - “whole-genome shotgun”, estratégia de fragmentação randômica do DNA.
BLAST - “Basic Local Alignment Search Tool”, ferramenta de alinhamento local.
NCBI - “National Center for Biotechnology Information”, Centro Nacional para Informação de Biotecnologia.
LCA - “Lowest Common Ancestor”, Menor Ancestral Comum.
MEGAN - “Metagenome Analyzer”
MG-RAST - “Metagenomics RAST Server”
Read - fragmento de DNA da amostra ambiental.
SNVs - “single nucleotide variants”, detecção de variantes de nucleotídeos.
NGS - “Next Generation Sequencing”, sequenciamento de última geração.
NCD - “Neighbor Cell Distance”, distância da célula vizinha.
GLM - “linear generalized model”, modelo linear generalizado.
OTU - “operational taxonomic unit”, unidade taxonômica operacional.
SNV - “single nucleotide variants”, variantes (polimorfismos) de base única.
CMI - “Candidate Metagenomic Islands”, ilhas metagenômicas candidatas.
KEGG - “Kyoto Encyclopedia of Genes and Genomes”
COG - “Phylogenetic classification of proteins encoded in complete genomes”

RESUMO GERAL

No primeiro capítulo é apresentada a diversidade de picocianobactérias do reservatório de Volta Grande, através de taxonomia clássica, análises estatísticas complementares e abordagem metagenômica. A riqueza de espécies descritas morfologicamente apresentou concordância com a microdiversidade observada para os gêneros dominantes anotados de acordo com os bancos de referências NCBI (nt) e SILVA (ssu rRNA). No segundo capítulo apresentamos o perfil metagenômico do reservatório de Marimbondo em escala temporal e espacial. As variáveis ambientais apresentaram um gradiente de acordo com o aumento da biomassa da população de Microcystis. O perfil taxonômico da comunidade foi distinto entre as amostras com baixa e alta biomassa de Microcystis, inclusive correlacionado ao gradiente ambiental, porém o perfil funcional foi estável entre as amostras, indicando uma redundância funcional. Dessa forma, as distintas anotações taxonômicas podem refletir, na verdade, a dinâmica de diversificação/adaptação do conjunto gênico de uma população. O terceiro capítulo aborda a microdiversidade da população de M. aeruginosa do reservatório de Marimbondo, uma vez que um grande número de cepas foi previamente observado. As análises de recrutamento e detecção de variantes de base única (SNVs) foram empregadas para explorar os padrões de alinhamento e diferenciação entre os membros da população. A microdiversidade observada pode ser relacionada à estrutura genômica em mosaico, na qual o conjunto gênico alterna em função das necessidades funcionais da população. O capítulo quatro explora a variabilidade genética das populações de dois gêneros da família Aphanizomenonaceae da lagoa da Pampulha, um reservatório artificial urbano hipereutrófico impactado por constantes florações de cianobactérias. Através de diversas abordagens, a complexidade de gêneros e cepas estreitamente relacionadas foi explorada. O alinhamento múltiplo, incluindo gêneros e cepas da América do Sul, indicou maior similaridade da população local aos genomas de Raphidiopsis brookii D9 e Cylindrospermopsis raciborskii T3. Diversos fatores indicam que estejamos lidando com uma população ainda desconhecida, entre eles a falta de alinhamento completo com R. brookii D9, a similaridade a diversas sequências sul-americanas e a alta diversidade intraespecífica observada pela ampla distribuição de SNVs, em especial àqueles detectados nos operons de assimilação e conversão de nitrito/nitrato.

Palavras-chave: Cianobactérias, Metagenômica, Reservatórios, Bioma tropical

ABSTRACT

On the first chapter we present the diversity of picocyanobacteria in the Volta Grande reservoir, through classic taxonomy, complemmentary statistical analyses and metagenomics approach. The richness in morphologically described species was in line with the microdiversity observed in the dominant genera annotated, according to reference databases: NCBI (nt) e SILVA (ssu rRNA). On the second chapter se see the metagenomic profile of the Marimbondo reservoir, both in spacial and temporal scale. The environmental variables showed a gradient consistent with the increase of biomass in the Microcystis population. The taxonomic profile of the community was different between samples with low or high Microcystis biomass, as well as correlated with the environmental gradient. Meanwhile, the functional profile was stable among samples, indicating functional redundancy. Therefore, the distinct taxonomic annotations could actually reflect the diversification/adaptation dynamics of the gene pool in a given population. The third chapter is concerned with the microdiversity of the M. aeruginosa population in the the Marimbondo reservoir, due to number of strains previously observed. The recruitment analyses and variant calling (SNV) were used with the intent of exploring alignment patterns and differentiation among population members. The microdiversity we observed could be related to the mosaic-like genomic structure in which the gene pool alternates according to the functional needs of the population. The fourth chapter explores the genetic variability of the two genera of the Aphanizomenonaceae Family in Pampulha lake, an urban artificial hypereutrophic reservoir suffering the impact of constant cyanobacteria blooms. Through different approaches, we explored the complexity in closely related strains and genera. The alignment of south american strains and genera indicated greater similarity of the local population with the genomes of Raphidiopsis brookii D9 and Cylindrospermopsis raciborskii T3. Many factors might indicate that we are dealing with a still unknown population, among which is the lack of full alignment with R. brookii D9, the similarity with several south american sequences and the high intraspecific diversity. Said diversity was observed through the wide distribution in SNVs, especially those detected in assimilation and conversion of nitrite/nitrate operons.

Key Words: Cyanobacteria, Metagenomics, Freshwater Reservoir, Tropical Biome

Sumário

LISTA DE FIGURAS

Capítulo 1

Capítulo 2

Capítulo 3

Capítulo 4

Fig. 11 - Gráfico de cobertura da amostra P1M mapeada ao cluster SXT de *C. raciborskii* T3.
Fig. 12 - Distribuição de polimorfismos gênicos de alto impacto em todos os contigs de *C. raciborskii* CS505 e *R. brookii* D9 na amostra P1M.

LISTA DE TABELAS

Capítulo 1

Capítulo 2

Capítulo 3

Tab. 6 - Ocorrência (1 or 0) de alguns genes do metabolism central em cada genoma de referência de *M. aeruginosa* analisado, considerando contigs e genes com contribuição relativa acima de 75%.

Capítulo 4

Tab. 1 - Resumo dos atributos do metagenoma das amostras P1M e P3M após a etapa de pré-processamento, conteúdo gênico e identificação de regiões ribossomais, de acordo com a plataforma MG-RAST.
Tab. 2 - Anotação intraespecífica para a família Nostocaceae, de acordo com a plataforma MG-RAST.
Tab. 3 - Resumo de alinhamento e cobertura das amostras P1M e P3M de acordo com o alinhador BWA-MEM e funções BEDtools Coveragebed.
Tab. 4 - Profundidade e amplitude de cobertura de reads alinhados acima de 97% de similaridade a outros gêneros e espécies de Nostocaceae.
Tab. 5 - Sequências parciais brasileiras, uruguaias e australianas de *C. raciborskii* mapeadas para ambos os conjuntos de dados da lagoa da Pampulha.
Tab. 6 - Resumo da detecção de variantes (SNVs) para os genomas de *C. raciborskii* CS-505 e *R. brookii* D9.
Tab. 7 - Contigs dos genomas de referência que apresentaram maior proporção de SNVs, taxa de variação, número de variantes de alto impacto e amplitude de cobertura para cada amostra.
Tab. 8 - Genes mapeados em relação ao genoma de *C. raciborskii* CS-505 com maior número de variantes de alto impacto, mostrando a profundidade de cobertura (aDP), o efeito detectado (Effect), a profundidade de alelos (AD), a profundidade de cobertura filtrada (DP) e a frequência de alelos polimórficos (Freq).
Tab. 9 - Genes mapeados em relação ao genoma de *R. brookii* D9 com maior número de variantes de alto impacto, mostrando a profundidade de cobertura (aDP), o efeito detectado (Effect), a profundidade de alelos (AD), a profundidade de cobertura filtrada (DP) e a frequência de alelos polimórficos (Freq).
Tab. 10 - Resumo da detecção de variantes em relação ao cluster de biossíntese de saxitoxina de *C. raciborskii* T3, mostrando os atributos dos alelos e de alinhamento.

INTRODUÇÃO GERAL

As cianobactérias estão entre os principais organismos atuantes na dinâmica de ecossistemas global, integrando sistemas aquáticos e atmosfera através da liberação de oxigênio, sequestro de dióxido de carbono, fixação de nitrogênio e produção primária de biomassa (Beck et al., 2012). Apesar de serem organismos procarióticos, suas células apresentam um aparato fotossintético semelhante ao das plantas, que inclui clorofila-*a* e ambos os fotossistemas I e II (Komárek, 2006). Durante as últimas décadas, características ecológicas, ultraestruturais e evidências moleculares têm substancialmente influenciado e aumentado o conhecimento desse grupo (Sheath & Wehr, 2003; Whitton, 2012), sendo que a abordagem polifásica, que inclui caracteres morfológicos, citológicos, bioquímicos, ecológicos e moleculares, tem sido adotada para o sistema de classificação atual (Komárek et al., 2011, Komárek, 2016). Devido à longa história evolutiva desse grupo e à ampla capacidade de adaptação ao ambiente, diversas populações apresentam caracteres fenotípicos semelhantes, dificultando a delimitação morfológica e reforçando a importância da investigação da diversidade genética (Otsuka et al., 2000; Nguyen et al., 2012; Komárek et al., 2014). Para diversas espécies, sejam elas procariontes ou eucariontes, o conhecimento em termos de diversidade filogenética tem aumentado progressivamente, acarretando mudanças nos sistemas de classificação taxonômica, tradicionalmente ancorados em caracteres morfológicos. Para cianobactérias não seria diferente, de fato as relações filogenéticas apontam para apenas dois agrupamentos monofiléticos, à nível de ordem (Gloeobacterales e Nostocales), enquanto a vastidão de membros cocóides e filamentosos restantes apresentam características polifiléticas (Komárek, 2016).

Ambientes de água doce geralmente apresentam flutuações rápidas das características físico-químicas e biológicas, como por exemplo variações diárias de luminosidade, circulação da água relacionada ao vento e variações sazonais de níveis de nutrientes relacionadas à pluviosidade. Tal instabilidade exige que as comunidades microbianas sejam capazes de responder às diferentes condições ambientais, favorecendo, assim, a diferenciação genética intraespecífica, ou seja, de membros de uma mesma população (Moore et al., 1998; Van Elsas & Bailey, 2002; Ulrich et al., 2005; Coleman et al., 2006; Galhardo et al., 2007; Piccini et al., 2011; Cordero et al., 2012; Shapiro et al., 2012; Humbert et al., 2013; Cordero & Polz, 2014). Assim como os demais

organismos procarióticos, o genoma das cianobactérias é capaz de responder às variáveis ambientais através de diversos mecanismos (Baulina, 2012), contribuindo para a ampla diversidade genotípica intra e interespecífica (Hess, 2011; Humbert et al., 2013; Simm et al., 2015) e permitindo a ocupação de diversos nichos ecológicos através de diferentes estratégias adaptativas[6] (Castenholz & Norris, 2005; Komárek, 2006; 2011; Palinska et al., 2011). Além disso, diversas espécies de cianobactérias podem apresentar várias cópias do genoma (oligoplóides e poliploides) e o número de cópias depende do organismo, da fase de desenvolvimento e das condições ambientais (Whitton & Potts, 2007; Kurmayer & Kutzenberger, 2003; Griese et al., 2011; Pecoraro et al., 2011; Stucken et al., 2013).

Muitos impactos ambientais decorrentes das atividades humanas são diretamente relacionados ao processo de eutrofização de sistemas aquáticos, como despejo de esgoto doméstico, efluentes industriais e lixiviação de nutrientes, especialmente nitrogênio e fósforo, de áreas agrícolas. Tal aumento dos níveis de nutrientes disponíveis afeta a biodiversidade local, desde a comunidade microbiana até os níveis tróficos mais elevados, causando desequilíbrio entre as populações. Cianobactérias têm a habilidade de aumentar rapidamente a biomassa em resposta às variáveis ambientais, podendo estabelecer uma condição denominada 'floração' (Pope & Patel, 2008; Paerl & Paul, 2012; Paerl & Otten, 2013), na qual o crescimento celular de uma ou poucas espécies é intenso em reposta, principalmente, à disponibilidade de fósforo e nitrogênio (Downing et al., 2001; Giani et al., 2005; Makarewicz & Lewis, 2015; Hu et al., 2015). Florações geralmente envolvem uma população dominante de cianobactérias associada a um conjunto diverso de outros microrganismos (Litchman et al., 2010; Zarraonaindia et al., 2013; Pimentel & Giani, 2013; Penn et al., 2014), especialmente a comunidade heterotrófica, a qual exerce um papel fundamental no desenvolvimento da biomassa de cianobactérias (Saxton et al., 2011; Shen et al., 2011; Steffen et al., 2012) e está relacionada ao padrão de dominância de diferentes espécies do fitoplâncton (Bagatini et al., 2014; Cordero &Polz, 2014; Louati et al., 2015).

Com o progresso das metodologias de sequenciamento de DNA, hoje é possível abordar não apenas a composição de toda a comunidade planctônica, através de perfis taxonômicos, como também o perfil funcional e as vias de interação, sejam entre indivíduos ou com o ambiente (Oulas et al., 2015). Metagenômica consiste no

sequenciamento do DNA extraído diretamente do ambiente (Chen & Pachter, 2005), permitindo, desta forma, a análise do perfil genético à nível de comunidade, sem restrições quanto ao tipo de ambiente ou à estrutura do genoma (Tringe et al., 2005; Gilbert & Dupont, 2011; Oh et al., 2011; Wiedenbeck & Cohan, 2011; Escalas et al., 2013; Bendall et al., 2016) e potencialmente integrando atributos ecológicos, fenotípicos e metabólicos das populações predominantes (Rodriguez-Valera et al., 2009; Thomas et al., 2012; Escalas et al., 2013). Recentemente o monitoramento da qualidade da água de alguns sistemas de água doce tem sido realizado através de técnicas de metagenômica, aprimorando com isso o perfil de biodiversidade e permitindo explorar genes de interesse, inclusive com atributo quantitativo (Tan et al., 2015). Dentre as diversas aplicações, destacamos a avaliação de comunidades microbianas associadas ao impacto antrópico (Kisand et al., 2012), a análise temporal da variabilidade microbiana de sistemas lóticos e bacia de inundação (Van Rossum et al., 2015) e avaliação de processos de tratamento de água (Gomez-Alvarez et al., 2012; Chao et al., 2013). A técnica "*whole-genome shotgun*" (WGS) fornece a visualização global da comunidade através de fragmentos randômicos de DNA sequenciados de amostras ambientais, sem uso de marcadores genéticos (Venter et al., 2004). Através dessa técnica é possível avaliar os níveis de diversidade filogenética e polimorfismo intraespecífico global (Oh et al., 2011; Bendall et al., 2016) e explorar o perfil metabólico de uma população em seu próprio ambiente (Mitra et al., 2011; Tan et al., 2015). Desta forma, a técnica WGS possibilita tanto a busca por atributos que agrupam organismos dentro de uma população, quanto atributos que os distinguem dentro de uma comunidade (Whitaker & Banfield, 2006, Oulas et al., 2015).

Com um conjunto de amostras sequenciado através de WGS, uma ampla variedade de análises e estratégias podem ser adotadas, de acordo com o objetivo do trabalho. Uma das estratégias amplamente utilizadas para explorar os perfis de biodiversidade e funcional de comunidades microbianas é o alinhamento das sequências de DNA utilizando BLAST (Ferramenta de busca de alinhamento local básico), cuja classificação taxonômica pode corresponder a diversos bancos de dados, especialmente NCBI (Centro Nacional de Informação em Biotecnologia). Com o conjunto de sequências alinhadas ao banco de dados escolhido, a anotação taxonômica e funcional pode ser realizada utilizando o algoritmo 'Menor Ancetral Comum' ('*Lowest Common Ancestor*' ou LCA), amplamente difundido através do *software* MEGAN '*Metagenome Analyzer*' (Huson et al., 2011) e utilizado por plataformas automatizadas de análise como MG-

RAST (*Metagenomics RAST Server*) (Meyer et al., 2008). O algoritmo LCA atribui cada fragmento de DNA (*read*), previamente alinhado, ao nó hierárquico mais abrangente para aquele táxon, de acordo com parâmetros pré-definidos, ou seja, um *read* só alcançará o nível específico se ele apresentar os atributos necessários, como por exemplo escore de alinhamento alto, número de fragmentos que devem apresentar a mesma anotação ou a proporção de diferença permitida entre os escores dos diferentes alinhamentos e o escore mais alto para aquele táxon (Huson et al., 2011; Huson & Mitra, 2012). Exemplificando, se o melhor alinhamento para um *read* do conjunto de dados apresentar um escore muito alto, tal *read* será provavelmente posicionado à nível de espécie ou mesmo cepa. Se o melhor alinhamento apresentar um escore mais baixo, porém todas os *reads* com escore acima de 80% desse escore corresponderem a um mesmo gênero, por exemplo, esse grupo de *reads* será atribuído a esse gênero. Se o melhor escore for muito baixo, e entre os escores acima de 80% desse escore estiverem *reads* de um mesmo gênero, eles serão posicionados nos nós mais abrangentes. O perfil funcional pode ser gerado da mesma maneira, porém através do alinhamento de acordo com bancos de dados de proteínas previamente classificadas funcionalmente (Mitra et al. 2011).

A análise pode ser um pouco mais dirigida, desde que estejam disponíveis os genomas de referência de interesse. A técnica de recrutamento do genoma consiste no alinhamento de todas os *reads* do ambiente contra um genoma de referência, geralmente da espécie dominante, e permite determinar tanto aquelas regiões genômicas da espécie que estão preservadas no ambiente, com alinhamentos cuja identidade foi acima de 99%, quanto potenciais regiões hipervariáveis (Pašić et al., 2009). Os *reads* 'recrutados' geralmente apresentam uma contribuição maior de genes relacionados ao metabolismo central (Mira et al., 2010; Knoop et al., 2013; Cordero & Polz, 2014). As regiões hipervariáveis são caracterizadas por uma cobertura abaixo da média (número de *reads* alinhados para cada posição do genoma de referência) em relação ao alinhamento global. Essa baixa cobertura é decorrente da diferença entre a sequência de DNA do genoma de referência e os *reads* que 'quase' alinharam a essa região, ou seja, pode indicar a ausência daqueles genes do genoma de referência no ambiente, ou pode estar relacionado a genes compartilhados entre a população, que apresentam polimorfismos em relação à sequência de referência, cuja ocorrência e abundância variam de acordo com as necessidades funcionais (Tettelin et al., 2008; Mira et al., 2010; Shapiro et al., 2012; Cordero & Polz 2014).

Uma vez que as populações dominantes no ambiente de estudo foram definidas e a avaliação dos atributos de alinhamento e cobertura do genoma apresentou robustez suficiente para análises mais detalhadas, o mapeamento do metagenoma contra os genomas de referência fornece ainda mais ferramentas para avaliação da diversidade das populações predominantes. Uma das estratégias amplamente difundidas para avaliação da diversidade intraespecífica é a detecção de variantes de base única (SNVs), a qual tem sido recomendada para análises de genômica comparativa de comunidades microbianas, em especial utilizando sequências extraídas diretamente do ambiente, uma vez que as variações genéticas pontuais, à nível de nucleotídeos, podem ser detectadas mesmo entre *reads* de diferentes populações (Nielsen et al., 2011; Olson et al., 2015), evidenciando os genes compartilhados na dinâmica de populações e comunidade. Esta abordagem adiciona um poder de análise sem precedentes ao estudo de comunidades planctônicas.

Alguns conceitos fundamentais para a análise das comunidades microbianas de sistemas aquáticos abrangem grande parte da discussão desse trabalho e serão brevemente descritos. O 'pan-genoma' engloba todo o conjunto gênico correspondente a uma unidade taxonômica (Tettelin et al., 2008; Mira et al., 2010), ou seja, uma espécie que apresenta diversas populações, cujos membros apresentam alta diversidade genotípica para adaptarem-se a diversas pressões ambientais (como é o caso de populações de *Microcystis aeruginosa*), apresentará um amplo pan-genoma, cujo repertório gênico aumenta à medida que novas populações são estudadas (Van Elsas & Bailey, 2002; Ulrich et al., 2005; Coleman et al., 2006; Frangeul et al., 2008; Beck et al., 2012; Xie et al., 2012; Humbert et al., 2013; Cordero & Polz, 2014; Simm et al., 2015).

Ao explorar a composição taxonômica e/ou funcional de uma população microbiana, é provável que uma grande diversidade intraespecífica seja observada. Essa população apresentará um amplo conjunto gênico, que constituirá o pan-genoma, porém apresentará sub-conjuntos constituídos de membros relacionados, estes que constituem a chamada 'microdiversidade' (Acinas et al., 2004; Fuhrman, 2009; Kashtan et al., 2014). Uma população que apresenta uma elevada microdiversidade apresentará uma grande dversidade genotípica, o que tem sido considerado uma estratégia evolutiva que possibilita a persistência de uma população em ambientes instáveis, como sistemas de água doce (Komárek, 2006; 2011; Palinska et al., 2011; Beck et al., 2012; Cordero et al., 2012; Humbert et al., 2013; Cordero & Polz, 2014; Rosen et al., 2015). Mesmo entre

cepas muito próximas é possível observar um conteúdo gênico divergente (Hess, 2011; Boon et al., 2014; Kashtan et al., 2014).

Quando aplicamos o conceito de ‘pan-genoma’ ao conjunto gênico de uma comunidade, ao invés de um nível taxonômico específico, podemos adotar uma abordagem centrada no potencial funcional do ambiente e considerar o conjunto gênico como uma ‘comunidade de genes’ (Boon et al., 2014), na qual as interações entre os membros de populações distintas são diretamente relacionadas a demandas específicas do ambiente. Esses dois conceitos podem ser ainda relacionados a ‘metacomunidades’ de sistemas microbianos (Leibold et al., 2004), nas quais populações potencialmente interagem, mesmo entre múltiplas espécies.

APRESENTAÇÃO

Esta tese é composta de quatro capítulos, todos com importantes tabelas suplementares, que podem ser solicitadas no seguinte email: *marcelelaux@gmail.com*

No primeiro capítulo desta tese apresentamos a diversidade de espécies de picocianobactérias do reservatório de Volta Grande, através de técnicas da taxonomia clássica, somada a análises complementares utilizando uma abordagem estatística para delimitação de espécies e a abordagem metagenômica. A diversidade morfológica da população de *Microcystis aeruginosa* também é apresentada. A riqueza de espécies descritas morfologicamente apresentou concordância com a microdiversidade observada para os gêneros dominantes anotados de acordo com os bancos de referências NCBI (nt) e SILVA (ssu rRNA) (Quast et al., 2013). Provavelmente devido à condição oligo-mesotrófica desse reservatório, a quantidade de DNA extraído e *reads* sequenciados foi abaixo do esperado. Um conjunto de dados de WGS com número insuficiente de fragmentos de DNA acarreta baixa cobertura do metagenoma, esta que inviabiliza a avaliação concreta de perfis de biodiversidade e funcional representativos do sistema. A baixa cobertura costuma ser um obstáculo enfrentado por diversos pesquisadores, porém os dados podem fornecer resultados interessantes quando analisados de forma compatível ao seu potencial de gerar informação.

No segundo capítulo apresentamos o perfil taxonômico e funcional do reservatório de Marimbondo. As amostragens foram realizadas ao longo de um ano, seguindo um gradual aumento de biomassa de cianobactérias, e espacialmente, em quatro estações de coleta, no período de floração. As variáveis ambientais apresentaram um gradiente de acordo com o aumento da biomassa da população de *Microcystis*. O perfil taxonômico da comunidade foi distinto entre as amostras com baixa e alta biomassa de *Microcystis*, inclusive correlacionado ao gradiente ambiental, porém o perfil funcional não apresentou diferenças significativas. Esse perfil funcional estável pode indicar uma redundância funcional através das populações compondo a comunidade ao longo do período de amostragem, ou seja, as mudanças de contribuição dos membros da comunidade não acarretam em mudanças no perfil funcional do sistema. Dessa forma, as distintas anotações taxonômicas podem refletir, na verdade, a dinâmica de diversificação/adaptação do conjunto gênico de uma população (Steffen et al., 2014; Biller et al., 2015).

O terceiro capítulo aborda a microdiversidade da população de *Microcystis aeruginosa* do reservatório de Marimbondo, uma vez que um grande número de cepas foi previamente observado e a variabilidade genética de comunidades em ambiente natural a nível intraespecífico tem sido uma prática recentemente recomendada (Luo et al., 2015). As análises de recrutamento e detecção de variantes de base única (SNVs) são consideradas indicadoras de diversidade intra e interespecífica (Bendall et al., 2016) e foram empregadas para explorar os padrões de alinhamento e diferenciação entre os membros da população. Através de análise de recrutamento, baseada no genoma de referência de *M. aeruginosa* NIES-843, sugerimos 13 potenciais regiões hipervariáveis, cujos genes podem estar relacionados à dinâmica de diversificação genética. As sequências metagenômicas foram então alinhadas contra outros 14 genomas de *M. aeruginosa*, gerando um amplo mapeamento, com ampla cobertura a todos os genomas testados. A microdiversidade observada pode ser relacionada à estrutura genômica em mosaico, na qual o conjunto gênico alterna em função das necessidades funcionais da população (Carmichael & Gorham, 1981; Steffen et al., 2014; Desai & Walczak, 2015). Diversos genes relacionados ao metabolismo central e a regiões hipervariáveis apresentaram relevantes variações gênicas, reiterando o alto grau de plasticidade e microdiversidade apresentadas por essa espécie (Kaneko et al., 2007; Frangeul et al., 2008; Humbert et al., 2013) em resposta às condições ambientais.

O capítulo quatro explora a variabilidade genética das populações de dois gêneros da família Aphanizomenonaceae da lagoa da Pampulha, um reservatório artificial urbano hipereutrófico impactado por constantes florações de cianobactérias. A variação sazonal da contribuição de *Microcystis* e *Cylindrospermopsis* foi previamente observada nesse sistema, porém a contribuição das populações relacionadas à família Aphanizomenonaceae é alta ao longo de todo o ciclo anual. Através de diversas abordagens, a complexidade de gêneros e cepas estreitamente relacionadas foi explorada, especialmente a população de sequências relacionadas a *Cylindrospermopsis raciborskii* CS505 e *Raphidiopsis brookii* D9 (Stucken et al., 2010). As sequências metagenômicas mapeadas em relação aos genomas completos de C. *raciborskii* CS505 e *R. brookii* D9 e a diversas sequências parciais, principalmente da América do Sul, indicaram maior similaridade da população local à *R. brookii* D9 e *C. raciborskii* T3, devido, por exemplo, à completa ausência do cluster de biossíntese de cilindospermopsina, porém presença do cluster de biossíntese de saxitoxina. Diversos autores observaram um padrão

biogeográfico para esses gêneros, inclusive propondo que diferentes ecótipos de *C. raciborskii* podem ter diferentes habilidades, de acordo com as pressões ambientais locais (Piccini et al., 2011; Piccini et al., 2013; Hoff-Risseti et al., 2013). Diversos fatores nos levam a considerar que estejamos lidando com uma população ainda desconhecida, entre eles a falta de alinhamento completo com *R. brookii* D9, a similaridade a diversas sequências sul-americanas e a alta diversidade intraespecífica observada pela ampla distribuição de SNVs, em especial àqueles detectados nos operons de assimilação e conversão de nitrito/nitrato. A análise comparativa de populações à nível de família e à nível intraespecífico, descritas no presente trabalho, fornece valiosas informações sobre uma possível população sul-americana de Nostocaceae. Futuramente, a análise de recrutamento do metagenoma do reservatório da Pampulha em relação ao genoma completo de *C. raciborskii* T3 e subsequente detecção de SNVs global, incrementará o perfil genético e funcional dessa população.

OBJETIVOS

OBJETIVO GERAL

O objetivo geral do trabalho foi explorar questões sobre a diversidade de cianobactérias e da comunidade microbiana associada em reservatórios de água doce, através da estratégia metagenômica, a fim de relacionar a estrutura da comunidade às condições ambientais e explorar possíveis mecanismos de interação e diversificação.

OBJETIVOS ESPECÍFICOS

Capítulo 1: Elaborar o levantamento taxonômico da comunidade de picocianobactérias do reservatório de Volta Grande e relacionar o perfil morfológico com o perfil de biodiversidade gerado pela análise metagenômica.

Capítulo 2: Descrever o perfil de biodiversidade e o perfil funcional do reservatório de Marimbondo ao longo de um gradiente ambiental, relacionando os resultados obtidos pela metagenômica com as variáveis ambientais.

Capítulo 3: Explorar a diversidade genética à nível intraespecífico da população de *Microcystis* do reservatório de Marimbondo, destacando a estrutura da população e as características que podem ser relacionadas à dinâmica de interação com a comunidade microbiana e com o ambiente.

Capítulo 4: Analisar a população de cianobactérias dominante do reservatório da Pampulha, afim de determinar sua estrutura intraespecífica e atributos relacionados a sua capacidade de diversificação e adaptação ao ambiente.

REFERÊNCIAS

Acinas S. G., Klepac-Ceraj V., Hunt D. E., Pharino C., Ceraj I., Distel D. L., Polz M. F. (2004) Fine-scale phylogenetic architecture of a complex bacterial community. *Nature*, 430:551-554

Bagatini I. L., Eiler A., Bertilsson S., Klaveness D., Tessarolli L. P., Vieira A. A. H. (2014) Host-Specificity and Dynamics in Bacterial Communities Associated with Bloom-Forming Freshwater Phytoplankton. *PLoS ONE*, 9(1): e85950

Baulina, O. I. (2012) Ultrastructural plasticity of cyanobacteria. Springer Science & Business Media, 204p.

Beck C., Knoop H., Axmann I. M., Steuer R. (2012) The diversity of cyanobacterial metabolism: genome analysis of multiple phototrophic microorganisms. *BMC Genomics*, 13:56

Bendall M. L., Stevens S. LR., Chan LK., Malfatti S., Schwientek P., Tremblay J., Schackwitz W., Martin J., Pati A., Bushnell B., Froula J., Kang D., Tringe S. G., Bertilsson S., Moran M. A., Shade A., Newton R. J., McMahon K. D., Malmstrom R. R. (2016) Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *The ISME Journal*, 1–13

Biller S. J., Berube P. M., Lindell D., Chisholm S. W. (2015) *Prochlorococcus*: the structure and function of collective diversity. *Nat. Rev. Microbiol.*, 13:13-27

Boon E., Meehan C. J., Whidden C., Wong D. H. J., Langille M. G. I., Beiko R. G. (2014) Interactions in the microbiome: communities of organisms and communities of genes. *FEMS Microbiol Rev*, 38:90-118

Burke C., Steinberg P., Rusche D., Kjelleberg S., Thomas T. (2011) Bacterial community assembly based on functional genes rather than species. *PNAS*, 108(34):14288-14293

Carmichael, W.W., Gorham, P.R. (1981) The mosaic nature of toxic blooms of cyanobacteria. In The Water Environment (pp. 161-172). Springer US.

Castenholz R. W., Norris T. B. (2005) Revisionary concepts of species in the Cyanobacteria and their applications. *Arch. Hydrobiol. Algol. Stud.*, 117:53-69

Chao Y., Ma L., Yang Y., Ju F., Zhang X.X., Wu WW., Zhang T. (2013) Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment. *Nat. Scient. Rep*., 3:3550

Chen K., Pachter L. (2005) Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities. *PLoS Comput Biol*, 1(2):106-111

Coleman M. L., Sullivan M. B., Martiny A. C., Steglich C., Barry K., DeLong E. F., Chisholm S. W. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science*, 311:1768–1770

Cordero O. X., Ventouras L. A., DeLong E. F., Polz M. F. (2012) Public good dynamics drive evolution of iron aquisition strategies in natural bacterioplankton populations. *Proc. Natl. Acad. Sci. USA*, 109:20059–20064

Cordero O. X., Polz M. F. (2014) Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Rev. Microbiol.*, 12:263-273

Desai M. M., Walczak A. M. (2015) Flexible gene pools - Rapid genetic exchange leads to mosaic genomes in cyanobacterial populations. *Science*, 348:977-978

Downing J. A., Watson S. B., McCauley E. (2001) Predicting Cyanobacteria dominance in lakes. *Can. J. Fish. Aquat. Sci.*, 58:1905-1908

Drummond, G.M., Martins, C.S., Greco, M.B., Vieira, F. (2009) Biota Minas: Diagnostico do conhecimento sobre a biodiversidade no Estado de Minas Gerais-subsidio ao Programa Biota Minas. In Biota Minas: Diagnostico do conhecimento sobre a biodiversidade no Estado de Minas Gerais-subsidio ao Programa Biota Minas. Fundação Biodiversitas.

Escalas A., Bouvier T., Mouchet M. A., Leprieur F., Bouvier C., Troussellier M., Mouillot D. (2013) A unifying quantitative framework for exploring the multiple facets of microbial biodiversity across diverse scales. *Environmental Microbiology*, 15:2642-2657

Faith, D.P., Baker, A.M. (2006) Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. *Evolutionary bioinformatics*, 2.

Frangeul L., Quillardet, P., Castets A. M., Humbert J. F., Matthijs H. CP., Cortez D., Tolonen A., Zhang CC., Gribaldo S., Kehr J. C., Zilliges Y., Ziemert N., Becker S., Talla E., Latifi A., Billault A., Lepelletier A., Dittmann E., Bouchier C., Tandeau de Marsac N. (2008) Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics*, 9:274

Fuhrman J. A. (2009) Microbial community structure and its functional implications. *Nature*, 459:193-199

Galhardo R. S., Hastings P. J., Rosenberg S. M. (2007) Mutation as a stress response and the regulation of evolvability. *Crit. Rev. Biochem. Mol. Biol*., 42:399–435

Giani A., Bird D., Prairie Y., Lawrence J. (2005) Empirical study of cyanobacterial toxicity along a trophic gradient of lakes. *Can. J. Fish. Aq. Sc.*, 62:1-10

Gilbert J. A., Dupont C. L. (2011) Microbial Metagenomics: Beyond the Genome. *Annu. Rev. Mar. Sci.*, 3:347-71

Gomez-Alvarez V., Revetta R. P., Santo Domingo J. W. (2012) Metagenomic Analyses of Drinking Water Receiving Different Disinfection Treatments. *Appl. Envir. Microbiol*., 78(17):6095-6102

Griese, M., Lange, C., Soppa, J. (2011) Ploidy in cyanobacteria. *FEMS microbial. letters*, 323(2): 124-131

Grünwald, N.J., Goodwin, S.B., Milgroom, M.G., Fry, W.E. (2003) Analysis of genotypic diversity data for populations of microorganisms. *Phytopathology*, 93(6):738-746

Hess W. R. (2011) Cyanobacterial genomics for ecology and biotechnology. *Curr. Opin. Microbiol.*, 14:608-614

Hoff-Risseti C., Dörr F. A., Schaker P. D. C., Pinto E., Werner V. R., Fiore M. F. (2013) Cylindrospermopsin and Saxitoxin Synthetase Genes in *Cylindrospermopsis raciborskii* Strains from Brazilian Freshwater. *PLoS ONE*, 8(8): e74238

Hu C., Rea C., Yu Z., Lee J. (2015) Relative importance of Microcystis abundance and diversity in determining microcystin dynamics in Lake Erie coastal wetland and downstream beach water. *J. Appl. Microbiol.*, 120:138-151

Humbert JF., Barbe V., Latifi A., Gugger M., Calteau A., Coursin T., Lajus A., Castelli V., Oztas S., Samson G., Longin C., Medigue C., Tandeau de Marsac N. (2013) A Tribute to Disorder in the Genome of the Bloom-Forming Freshwater Cyanobacterium *Microcystis aeruginosa*. *PLoS ONE*, 8(8): e70747

Huson D. H., Mitra S., Ruscheweyh H. J., Weber N., Schuster S.C. (2011) Integrative analysis of environmental sequences using MEGAN 4. *Genome Res.*, 21:1552-1560

Huson D. H., Mitra S. (2012) Introduction to the Analysis of Environmental Sequences: Metagenomics with MEGAN. *Evolut. Genom*., series Methods in Molecular Biology, 856:415-429

Janoir, C., Denève, C., Bouttier, S., Barbut, F., Hoys, S., Caleechum, L., ... Monot, M. (2013) Adaptive strategies and pathogenesis of *Clostridium difficile* from in vivo transcriptomics. *Infection and immunity*, 81(10):3757-3769

Kaneko T., Nakajima N., Okamoto S., Suzuki I., Tanabe Y., Tamaoki M., Nakamura Y., Kasai F., Watanabe A., Kawashima K., Kishida Y., Ono A., Shimizu Y., Takahashi C., Minami C., Fujishiro T., Kohara M., Katoh M., Nakazaki N., Nakayama S., Yamada M., Tabata S., Watanabe M. M. (2007) Complete Genomic Structure of the Bloom-forming Toxic Cyanobacterium *Microcystis aeruginosa* NIES-843. *DNA Research*, 14:247-256

Kashtan N., Roggensack S. E., Rodrigue S., Thompson J. W., Biller S. J., Coe A., Ding H., Marttinen P., Malmstrom R. R., Stocker R., Follows M. J., Stepanauskas R., Chisholm S. W. (2014) Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild *Prochlorococcus*. Science, 344:416-420

Kisand V., Valente A., Lahm A., Tanet G., Lettieri T. (2012) Phylogenetic and Functional Metagenomic Profiling for Assessing Microbial Biodiversity in Environmental Monitoring. *PLoS ONE*, 7(8): e43630

Knoop H., Gründel M., Zilliges Y., Lehmann R., Hoffmann S., Lockau W., Steuer R. (2013) Flux Balance Analysis of Cyanobacterial Metabolism: The Metabolic Network of *Synechocystis* sp. PCC 6803. *PLoS Comput Biol*, 9(6): e1003081

Komárek, J. (2006) Cyanobacterial Taxonomy: Current Problems and Prospects for the Integration of Traditional and Molecular Approaches. *Algae*, 21(4):349-375

Komárek J. (2011) Introduction to the 18(th) IAC Symposium in Ceske Budejovice 2010, Czech Republic - Some current problems of modern cyanobacterial taxonomy. *Fottea*, 11: 1-7

Komárek, J., Kaštovský, J., Jezberová, J. (2011) Phylogenetic and taxonomic delimitation of the cyanobacterial genus *Aphanothece* and description of *Anathece* gen. nov. *Eur. J. Phycol.*, 46(3): 315-326

Komárek, J., Kastovsky, J., Mares, J., Johansen, J. R. (2014) Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014, using a polyphasic approach. *Preslia*, 86(4): 295-335

Komárek, J. (2016) Review of the cyanobacterial genera implying planktic species after recent taxonomic revisions according to polyphasic methods: state as of 2014. *Hydrobiologia*, 764(1): 259-270

Kurmayer, R., Kutzenberger, T. (2003) Application of real-time PCR for quantification of microcystin genotypes in a population of the toxic cyanobacterium *Microcystis* sp. *Appl. Envir. Microbiol.*, 69(11): 6723-6730

Leibold M. A., Holyoak M., Mouquet N., Amarasekare P., Chase J. M., Hoopes M. F., Holt R. D., Shurin J. B., Law R., Tilman D., Loreau M., Gonzalez A. (2004) The metacommunity concept: a frameworkfor multi-scale community ecology. *Ecol. Let*., 7: 601-613

Litchman E., Tezanos Pinto P., Klausmeier C. A., Thomas M. K., Yoshiyama K. (2010) Linking traits to species diversity and community structure in phytoplankton. *Hydrobiologia*, 653:15-28

Louati I., Pascault N., Debroas D., Bernard C., Humbert, J. F., Leloup J. (2015) Structural Diversity of Bacterial Communities Associated with Bloom-Forming Freshwater Cyanobacteria Differs According to the Cyanobacterial Genus. *PLoS ONE*, 10(11): e0140614

Luo C., Knight R., Siljander H., Knip M., Xavier R.J., Gevers D. (2015) Constrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.*, 33(10):1045-1052

Makarewicz J. C., Lewis T. W. (2015) Exploring spatial trends and causes in Lake Ontario coastal chemistry: Nutrients and pigments. *J. Great Lakes Res*., 41(3):794-800

Meyer F., Paarmann D., D'Souza M., Olson R., Glass E.M., Kubal M., Paczian T., Rodriguez A., Stevens R., Wilke A., Wilkening J., Edwards R.A. (2008) The Metagenomics RAST server — A public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics, 9:386

Mira A., Martín-Cuadrado A. B., D'Auria G., Rodríguez-Valera F. (2010) The bacterial pan-genome: a new paradigm in microbiology. *Int. Microbiol*., 13: 45-57

Mitra S., Rupek P., Richter D. C., Urich T., Gilbert J. A., Meyer F., Wilke A., Huson D.H. (2011) Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics*, 12 (Suppl. 1): S21

Moore L. R., Rocap G., Chisholm S. W. (1998) Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature*, 393:464

Morris, J. J., Lenski, R.E., Zinser, E.R. (2012) The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio*, 3(2):e00036-12

Neilan B. A., Dittmann E., Rouhiainen L., Bass R. A., Schaub V., Sivonen K., Borner T. (1999) Nonribosomal Peptide Synthesis and Toxigenicity of Cyanobacteria. *J. Bacteriol.*, 181(13):4089-4097

Nguyen V. L. A., Tanabe Y., Matsuura H., Kaya K., Watanabe M. M. (2012) Morphological, biochemical and phylogenetic assessments of water-bloom-forming tropical morphospecies of *Microcystis* (Chroococcales, Cyanobacteria). *Phycol. Res*., 60: 208-222

Nielsen R., Paul J. S., Albrechtsen A., Song Y. S. (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Gen.*, 12:443-451

Oh S., Caro-Quintero A., Tsementzi D., DeLeon-Rodriguez N., Luo C., Poretsky R., Konstantinidis K.T. (2011) Metagenomic Insights into Evolution, Function, and Complexity of the Planktonic Microbial Community of Lake Lanier, a Temperate Freshwater Ecosystem. *Appl. Environm. Microbiol.*, 77(17):6000-6011

Olson N. D., Lund S. P., Colman R. E., Foster J. T., Sahl J. W., Schupp J. M., Keim P., Morrow J. B., Salit M. L., Zook J. M. (2015) Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front. Genet*., 6:235

Otsuka S., Suda S., Li R., Matsumoto S., Watanabe M. M. (2000) Morphological variability of colonies of Microcystis morphospecies in culture. *J. Gen. Appl. Microbiol*., 46(1):39-50

Oulas A., Pavloudi C., Polymenakou P., Pavlopoulos G.A., Papanikolaou N., Kotoulas G., Arvanitidis C., Iliopoulos, I. (2015) Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinform. Biol. Insights*, 9:75

Paerl H. W., Otten T. G. (2013) Harmful Cyanobacterial Blooms: Causes, Consequences, and Controls.

Environmental Microbiology, DOI 10.1007/s00248-012-0159-y

Paerl H. W., Paul V. J. (2012) Climate change: Links to global expansion of harmful cyanobacteria. *Water Res*., 46(5):1349-1363

Palinska K. A., Deventer B., Hariri K., Lotocka M. (2011) A taxonomic study on Phormidium-group (cyanobacteria) based on morphology, pigments, RAPD molecular markers and RFLP analysis of the 16S rRNA gene fragment. *Fottea*, 11:41-55

Pašić L, Rodriguez-Mueller B., Martin-Cuadrado A., Mira A., Rohwer F., Rodriguez-Valera, F. (2009) Metagenomic islands of hyperhalophiles: the case of *Salinibacter ruber*. *BMC Genomics*, 10:570

Pecoraro, V., Zerulla, K., Lange, C., Soppa, J. (2011) Quantification of ploidy in proteobacteria revealed the existence of monoploid, (mero-) oligoploid and polyploid species. *PLoS One*, 6(1): e16392

Penn K., Wang J., Fernando S. C., Thompson J. R. (2014) Secondary metabolite gene expression and interplay of bacterial functions in a tropical freshwater cyanobacterial bloom. *The ISME Journal*, 8:1866–1878

Piccini C., Aubriot L., Fabre A., Amaral V., González-Piana M., Giani A., Figueredo C. C., Vidal L., Kruk C., Bonilla S. (2011) Genetic and eco-physiological differences of South American *Cylindrospermopsis raciborskii* isolates support the hypothesis of multiple ecotypes. *Harmful Algae*, 10:644-653

Piccini C., Aubriot L., D'Alessandro B., Martigani F., Bonilla S. (2013) Revealing Toxin Signatures in Cyanobacteria: Report Genes Involved in Cylindrospermopsin Synthesis from Saxitoxin-Producing *Cylindrospermopsis raciborskii*. *Adv. Microbiol.*, 3:289-296

Pimentel J. S. M., Giani A. (2013) Estimating toxic cyanobacteria in a Brazilian reservoir by quantitative real-time PCR, based on the microcystin synthetase D gene. *J. Appl. Phycol.*, 25: 1545-1554

Pope P. B., Patel B. K. C. (2008) Metagenomic analysis of a freshwater toxic cyanobacteria bloom. *FEMS Microbiol. Ecol.*, 64:9-27

Whitton B.A., Potts, M. (Eds.) (2007) The ecology of cyanobacteria: their diversity in time and space. Springer Science & Business Media. 669p.

Quast C., Pruesse E., Yilmaz P., Gerken J., Schweer T., Yarza P., Peplies J., Glöckner F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl. Acids Res.*, 41 (D1):D590-D596

Reynolds C. S. (1984) The Ecology of Freshwater Phytoplankton. Cambridge University Press, Cambridge. 396 p.

Rodriguez-Valera F., Martin-Cuadrado AB., Rodriguez-Brito B., Pašić L., Thingstad T. F., Rohwer F., Mira A. (2009) Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol*., 7:828-836

Rosen M. J., Davison M., Bhaya D., Fisher D. S. (2015) Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science*, 348: 1019-1023

Saxton M. A., Morrow E. A., Bourbonniere R. A., Wilhelm S. W. (2011) Glyphosate influence on phytoplankton community structure in Lake Erie. *J. Great Lakes Res.*, 37(4): 683-690

Shapiro B. J., Friedman J., Cordero O. X., Preheim S. P., Timberlake S. C., Szabó G., Polz M. F., Alm E.

J. (2012) Population genomics of early events in the ecological differentiation of bacteria. *Science*, 336:48–51

Sheath R. G., Wehr J. D. (2003) Freshwater Algae of North America. Elsevier Science, San Diego, California. 918 p.

Shen H., Niu Y., Xie P., Tao M., Yang X. (2011) Morphological and physiological changes in *Microcystis aeruginosa* as a result of interactions with heterotrophic bacteria. *Freshwater Biology*, 56: 1065–1080

Simm S., Keller M., Selymesi M., Schleiff E. (2015) The composition of the global and feature specific cyanobacterial core-genomes. *Front. Microbiol*., 6:219

Steffen M. M., Li Z., Effler T. C., Hauser L. J., Boyer G. L., Wilhelm S. W. (2012) Comparative Metagenomics of Toxic Freshwater Cyanobacteria Bloom Communities on Two Continents. *PLoS ONE*, 7(8): e44002

Steffen M. M., Dearth S. P., Dill B. D., Li Z., Larsen K. M., Campagna S. R., Wilhelm S. W. (2014) Nutrients drive transcriptional changes that maintain metabolic homeostasis but alter genome architecture in Microcystis. *The ISME Journal*, 8:2080-2092

Stucken K., John U., Cembella A., Murillo A. A., Soto-Liebe K., Fuentes-Valdés J. J., Friedel M., Plominsky A. M., Vásquez M., Glöckner G. (2010) The Smallest Known Genomes of Multicellular and Toxic Cyanobacteria: Comparison, Minimal Gene Sets for Linked Traits and the Evolutionary Implications. *PLoS ONE*, 5(2):e9235

Stucken, K., Koch, R., Dagan, T. (2013) Cyanobacterial defense mechanisms against foreign DNA transfer and their impact on genetic engineering. *Biol. Res.*, 46(4):373-382

Tan BF., Ng C., Nshimyimana J. P., Loh L. L., Gin K. Y. H., Thompson J. R. (2015) Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities. *Front. in Microbiol*., 6:1027

Tettelin H., Riley D., Cattuto C., Medini D. (2008) Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.*, 11(5):472-7.

Thomas T., Gilbert J., Meyer F. (2012) Metagenomics - a guide from sampling to data Analysis. *Microbial Informatics and Experimentation*, 2:3

Tringe S. G., Mering C., Kobayashi A., Salamov A. A., Chen K., Chang H. W., Podar M., Short J. M., Mathur E. J., Detter J. C., Bork P., Hugenholtz P., Rubin E. M. (2005) Comparative Metagenomics of Microbial Communities. *Science*, 308:554-557

Ulrich L. E., Koonin E. V., Zhulin I. B. (2005) One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol.*, 13:52–56

Van Elsas J. D., Bailey M. J. (2002) The ecology of transfer of mobile genetic elements. *FEMS Microbiol. Ecol.*, 42:187–197

Van Rossum T., Peabody M. A., Uyaguari-Diaz M. I., Cronin K. I., Chan M., Slobodan J. R., Nesbitt M. J., Suttle C. A., Hsiao W. W. L., Tang P. K. C., Prystajecky N. A., Brinkman F. S. L. (2015) Year-Long Metagenomic Study of River Microbiomes Across Land Use and Water Quality. *Front. Microbiol*. 6:1405

Venter J. C., Remington K., Heidelberg J. F., Halpern A. L., Rusch D., Eisen J. A., Wu D., Paulsen I., Nelson K. E., Nelson W., Fouts D. E., Levy S., Knap A. H., Lomas M. W., Nealson K., White O. (2004)

Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304:66-74

Watson, J.D., Baker, T.A., Bell, S.P., Gann, A., Levine, M., Losicke, R. (2015) Biologia molecular do gene. 7 ed. Artmed Editora. 916p.

Whitaker R. J., Grogan, D. W., Taylor J. W. (2005) Recombination Shapes the Natural Population Structure of the Hyperthermophilic Archaeon *Sulfolobus islandicus*. *Mol. Biol. Evol*., 22(12):2354-2361

Whitaker R. J., Banfield J. F. (2006) Population genomics in natural microbial communities. *TRENDS Ecol. Evol*., 21(9):508-516

Whitton B. A. (2012) Ecology of Cyanobacteria II: Their Diversity in Space and Time. Springer Science & Business Media, 760 p.

Wiedenbeck J., Cohan F. M. (2011) Origins of bacterial diversity through horizontal genetic transferand adaptation to new ecological niches. *FEMS Microbiol. Rev*., 35:957-976

Xie L., Rediske R. R., Hong Y., O'Keefe J., Gillett N. D., Dyble J., Steinman A. D. (2012) The role of environmental parameters in the structure of phytoplankton assemblages and cyanobacteria toxins in two hypereutrophic lakes. *Hydrobiologia*, 691:255-268

Zarraonaindia I., Smith D. P., Gilbert J. A. (2013) Beyond the genome: community-level analysis of the microbial world. *Biol Philos*, 28:261-282

Capítulo 1: Cianobactérias planctônicas de um reservatório tropical do sudeste do Brasil: estratégias adicionais para caracterizar uma rica comunidade de picocianobactérias

1. RESUMO

A análise taxonômica da comunidade de cianobactérias planctônicas foi realizada no reservatório tropical de Volta Grande, enquadrado como oligo-mesotrófico. A identificação das espécies de cianobactérias seguiu as normas do Código de Nomenclatura Botânica. A ordem Synechococcales apresentou 15 espécies, seguida de Chroococcales (8 espécies). A riqueza de populações picoplanctônicas foi uma característica marcante desse reservatório. Uma abordagem estatística, baseada em atributos morfológicos adicionas, foi adotada para alcançar uma delimitação mais robusta entre cinco espécies dos gêneros *Anathece* Komárek et Anagnostidis) Komárek, Kastovsky et Jezberová, *Cyanodictyon* Pascher e *Lemmermanniella* Geitler. A diversidade genética foi analisada através da técnica de metagenômica e o perfil taxonômico gerado apresentou uma alta diversidade intraespecífica em relação aos gêneros mais abundantes anotados. Estudos envolvendo taxonomia morfológica de picocianobactérias são menos frequentes que aqueles envolvendo espécies nano-microplanctônicas, uma vez que seu tamanho diminuto é frequentemente uma barreira para sua correta identificação, ou mesmo detecção. Dessa forma, este trabalho apresenta uma importante contribuição ao conhecimento desse grupo, apresentando o perfil taxonômico de picocianobactérias classificadas de acordo com critérios morfológicos e de acordo com o alinhamento e anotação das sequências do metagenoma. A complexidade morfológica observada pode ser relacionada à diversidade genética observada à nível intraespecífico, e o complexo perfil genético pode estar refletindo uma população cujos genomas de referência ainda são escassos. Finalmente, descrevemos também as espécies do gênero *Microcystis* observadas nesse sistema.

Planktonic cyanobacteria from a tropical reservoir of southeastern Brazil: additional strategies to characterize a picocyanobacteria rich community

1. ABSTRACT

A taxonomic analysis of planktonic cyanobacteria was performed in a tropical oligo-mesotrophic reservoir. The identification of the cyanobacteria species was done according to the Botanical Nomenclatural rules. The order Synechococcales presented 15 species, followed by Chroococcales (8 species). The richness of picoplanktonic components was a remarkable feature of this reservoir. A statistical approach, based on additional morphological attributes, was used to achieve a more robust delimitation among the five species belonging to the picoplanktonic genera *Anathece (*Komárek et Anagnostidis) Komárek Kastovsky et Jezberová, *Cyanodictyon* Pascher and *Lemmermanniella* Geitler. The genetic diversity was analyzed through metagenomics and pointed to a high intra-specific diversity within the most abundant genera annotated. Reports on picocyanobacteria taxonomy are rarer than on nanoplanktonic species, since their minute size is frequently a barrier for their correct identification, or even detection. Thus, this work represents an important contribution to increase the knowledge about this group, showing that phenotypical complexity could be related to intraspecific diversity. Furthermore, several species of the widely distributed *Microcystis* genus were described in this system.

Keywords: *Aphanocapsa*, *Anathece*, *Cyanodictyon*, *Lemmermanniella*, *Microcystis*, morphology, picocyanobacteria, tropical reservoir.

1. INTRODUCTION

The large diversity observed in cyanobacteria is directly related to their ability to adjust to the environment, through the adaptation and acclimatization of metabolic and phenotypic features (Komárek, 2006, 2011; Palinska et al., 2011). As a result of this adaptability skill, many cyanobacterial species show several morphotypes and ecotypes belonging to the same genotype (Komárek & Anagnostidis, 1988; Komárek et al., 2002; 2011), with distinctive morphologies as a result of differential gene expression (Castenholz & Norris, 2005). This means that morphological analyses incorporate the complex phenotypic responses of the genome to specific environmental conditions (Komárek & Anagnostidis, 1988; Komárek, 2006; 2011; Palinska et al., 2011).

The morphological complexity may lead to taxonomic confusion that represents one of the most serious problems in cyanobacterial studies (Komárek, 2006). The apparent simplicity of cells and colonies of these organisms can be misinterpreted. Indeed, cell shape and division patterns, along with their typical arrangement in the mucilaginous envelope, may lead to complex morphological analyses that hinder a precise identification, when researchers try to apply the International Code of Botanical Nomenclature. Thus, several authors have suggested recently the combined use of morphological, physiological, ecological and molecular characteristics to attain lower-rank levels of identification (Palinska et al., 2011; Strunecký et al., 2013; Komárek et al., 2014; Komárek, 2016). The modification or disappearance of several phenotypic traits in isolates, under culture conditions, is a drawback for the combined use of traditional morphological observations and more recent molecular approaches (Otsuka et al., 2000; Nguyen et al., 2012). As a result, several researchers are currently dealing with the phenotypic/genotypic paradigm in cyanobacteria studies (Palinska & Marquardt, 2008; Serizawa et al., 2008; Kumari et al., 2009; Yamamoto & Nakahara, 2009; Thomazeau et al., 2010; Strunecky et al., 2013; Dojani et al., 2013). Therefore, an accurate microscopy based taxonomic description, with the aid of statistical analysis, added with molecular information of the species in their natural environment, can provide a better overview of the community's structure, improving its understanding.

This study was performed in a large tropical oligo-mesotrophic reservoir, located in southeastern Brazil, in a transition zone between the Cerrado (Brazilian savannah) and

the Atlantic Forest biomes, both at present under severe anthropogenic impact. The coccoid cyanobacteria species were especially representative in the phytoplanktonic community and, interestingly, the richness and abundance of the small picoplanktonic cyanobacteria was a remarkable feature of this reservoir. Here, we present a morphology-based taxonomic study of these species. Because of the overlap of characters among some taxa, mainly observed in late development stages, a statistical approach based on categorization of distinctive morphological features was adopted to achieve a more robust delimitation among the picoplanktonic species *Anathece clathrata* (W.West et G.S.West) Komárek, Kaštovský et Jezberová, *Anathece minutissima* (W.West) Komárek, Kaštovský et Jezberová, *Cyanodictyon iac* Cronberg et Komárek, *Cyanodictyon planctonicum* Meyer and *Lemmermanniella pallida* (Lemmermann) Geitler. The metagenomic taxonomical profile of two samples were analyzed, in order to complement the morphological description, since metagenomics provides additional information from the whole community about instraspecific diversity (Oh et al., 2011, Bendall et al., 2016), but the database bias should always be considered, since the number of genomes sequenced is still increasing (Thomazeau et al., 2010; Luo et al., 2015). Our results provide important information about picocyanobacteria populations, especially because morphological, statistical and molecular techniques were applied. The oligo-mesotrophic feature of the Volta Grande reservoir is also remarkable, since most of the studies in tropical biome concern meso-eutrophic systems, with high levels of nutrients. The morphological description of cyanobacteria communities, particularly atypical communities, joins phenotypic traits to the molecular analysis aiming for precise community-population diversity through NGS ('*Next Generation Sequencing*").

1. MATERIAL AND METHODS

Study Area

Volta Grande Reservoir is located in the Rio Grande water basin, between the states of Minas Gerais and São Paulo, Brazil (Fig. 1), and has a flooded area of 222 km^2, a volume of 23 km^3 and up to 25 m depth (Braga & Gomiero, 1997). The local weather

presents two well defined seasons: rainy season, from October to March, and dry season, from April to September. It was initially classified as oligotrophic (Greco, 2002) but today, because of increased nutrient input, it shows oligo-mesotrophic traits (Lopes, 2013). The reservoir is used for hydropower generation.

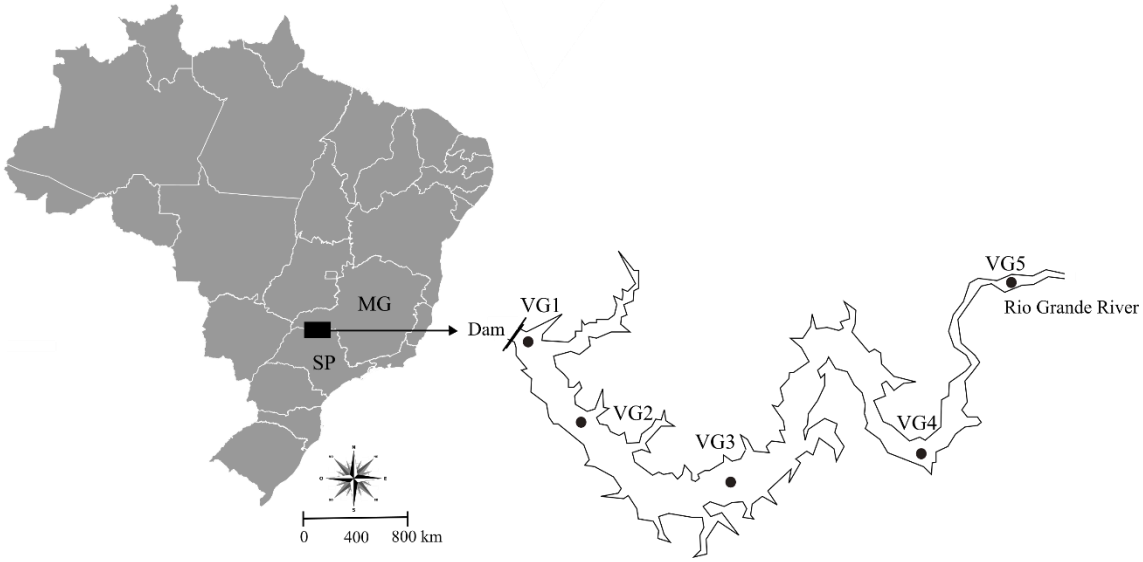


Fig. 1 - Volta Grande reservoir (S 20°01'76.8", W 48°19'11.5"), between Minas Gerais and São Paulo states (Brazil). The five sampling stations are indicated.

Sampling

Five sites (Fig. 1) were sampled monthly during 18 months (from January 2011 to June 2012) and distributed along the whole reservoir: VG1, located close to the dam (20°01’36.0” S and 48°12’59.8” W); VG2 located in the main body, close to the Água Comprida stream (20°07’45.2” S and 48°08’42.0” W); VG3, close to a recreational zone (Miguelopólis Beach) and São Miguel and Lageado streams (20°09’05.0” S and 48°03’09.6” W); VG4, between the main reservoir body and the Rio Grande river upstream (20°06’57.1 S and 47°55’04.6” W); VG5 in the lotic zone upstream (20°00’40” S and 47°52’19” W).

Phytoplankton samples were collected in the euphotic zone, with a 20 µm mesh size plankton net, towing by boat for 5 minutes. Integrated samples were also taken from the surface to the limit of the photic zone. All samples were fixed *in situ* with Lugol (4%) solution. Before the analysis, the integrated water samples were allowed to settle in

sedimentation chambers and the material, deposited at the bottom, was collected by help of a Pasteur pipet for further analysis under microscope. Nutrients concentrations (phosphorus, ammonium and nitrate) were measured by spectrophotometric methods on duplicate subsamples (Apha, 2005). In order to better characterize the ecological features of the system, we performed a linear regression analusis through generalized linear model (GLM), among environmental variables, considering spatial and seasonal variation. All analyses were performed in R language (R Core Team, 2015), using the Stats and Vegan package (Oksanen et al., 2015).

Taxonomic analysis

The taxonomic analyses were carried out under light microscopes (Olympus BX40 (at 400x magnification) and Zeiss Axioplan (400-1600x), Plan-Neofluar, NA:1.30). The organisms were measured and photographed with an image capture device Axiocam ERc 5s. For the identification of each taxon, specific literature was consulted, as cited in the description of each species. The classification system adopted was Hoffmann et al., (2005), with modifications, as proposed by Komárek, (2006). In order to visualize the mucilaginous envelope, the organisms were observed in phase contrast microscopy and some samples were stained with Indian ink. The cell dimensions were represented by their diameter (diam) or length and width (l/w), mean, standard error (SE), standard deviation (σ) and median. Furthermore, we added the number of individuals observed (N), number of measures taken (*n*), the Neighbor Cell Distance (NCD), which is the distance to the closest neighbor cell in same focus, length/width ratio (R) and respective medians (median NCD and median R). The most representative coccoid cyanobacteria taxa observed in pelagic (net samples and integrated samples) are presented in this work and an additional strategy to characterize the picocyanobacteria taxa is proposed, in order to improve the taxonomical assignments and morphological boundaries.

Picocyanobacteria Morphological Categorization

The remarkable morphological diversity of picocyanobacteria taxa in the entire study area and period required a special care concerning morphological delimitation. To

provide support to the identification and upgrade some attributes of cylindrical and rod-shaped picocyanobacteria species, we performed a statistical analysis, based on additional measurements of length and width (*n* was increased to ± 100) as well as their length/width ratio and the additional metric NCD. We defined five categorical variables:

1) presence or absence of cell rows or pseudo-filamentous rows;

2) colony size (less or greater than 100 µm);

3) colony form (irregular, lobular, elongated or rounded);

4) l/w ratio class (up to 1:2, between 1:2 and 1:3, above 1:3);

5) NCD class (up to 2 µm, between 2 µm and 3 µm, above 3 µm).

We performed a hierarchical clustering through average linkage (UPGMA, Euclidean distance) to check the consistency of the groups of individuals, initially defined through visual criteria. The average linkage method considers the distance among the cluster centroids, such that dissimilarities among groups should be greater than within groups (functions *hclust* and *vegdist* from Vegan R package) (Oksanen et al., 2015). Initially we defined groups of individuals according to visual and regular dimension attributes, which can be seen at the extremities of the cluster (Fig. 2a). According to the clustering output, we defined five candidate species (Fig. 2b), encompassing the sub-groups initially considered. In order to test the significance of the difference among these five groups, we performed an Analysis of Similarities (ANOSIM), which operates directly on a dissimilarity matrix (function "*anosim*" from Vegan R package). A linear generalized model (GLM) using Gaussian distribution was performed to compare the median of the length/width ratio and the NCD median according to the five candidate species, illustrated through boxplots (Fig. 2c). We adopted the median measure to achieve a better fit under the Gaussian model. The analyses were run using R language (R Core Team, 2015).

Metagenomic diversity profile

Two samples were chosen for metagenomic analyses, according to seasonality and total phytoplankton biomass. Because of the generally low cell density observed in this

reservoir (euphotic zone: min 22m, max 29.2m and mean 25.5m) the sample station with the overall highest phytoplankton biomass was chosen, looking for enough amount of material for DNA extraction and further sequencing process. The VG3 sampling station is located approximately in the middle of the reservoir, and one dry season (May/2011) and one rainy season (November/2011) samples were selected. The DNA was extracted according to Kurmayer et al., (2003) and the sequencing was performed under Nextera XT paired-end 2 x 250 bp, Illumina platform. The operational taxonomic unit strategy (OTU) was performed according to the QIIME pipeline (Caporaso et al. 2010), in which the raw reads (Illumina Hiseq paired ends) from the two metagenomic shotgun datasets (VG3_may, VG3_nov) were joined (multiple_join_paired_ends.py function) and the quality control step was obtained through split_libraries_fastq.py function using a phred quality threshold of 25 (datasets not multiplexed). The resulting output from split_libraries_fastq.py was screened for ribosomal sequences using Metaxa2 package (Bengtsson-Palme et al., 2015) and the sequences identified as bacteria ribosomal fragments were then passed on to pick closed reference OTUs pipeline and annotated according to SILVA ribosomal RNA gene database (Quast et al., 2013).

For BLAST/MEGAN approaches the paired reads were merged using Flash 1.2.11 (Magoč & Salzberg, 2011), allowing for both outie and innie orientation of merging fragments. Only the properly paired and combined reads were considered for BLAST alignment to NCBI nucleotide database. The BLAST output was analyzed in MEGAN Metagenomic Analyzer 5.10.7, which applies the lowest common ancestor (LCA) algorithm to assigns each BLAST hit to the lowest taxonomical node, according to user defined parameters (Huson et al., 2011) and alignment parameters, so the less specifically a read hits a taxa, the rootmost in the taxonomy it is placed (Huson et al., 2011; Huson & Mitra, 2012).

1. RESULTS AND DISCUSSION

Environmental Features

The summary of nutrient concentrations among all sampling area and period is presented in Table 1. The mean values were in general low, but total phosphorus and nitrate concentrations reached values as high as 56 $\mu g.L^{-1}$ and 204 $\mu g.L^{-1}$, respectively, confirming that the trophic status of Volta Grande Reservoir can today be classified as oligo-mesotrophic.

Tab. 1 - Nutrient concentrations ($\mu g.L^{-1}$) in five sampling station in Volta Grande reservoir (Brazil), from January 2011 to June 2012 (minimum recorded value, maximum, mean, standard error and median), TP: total phosphorus, SRP: soluble reactive phosphorus, NH_4: ammonium.

	TP	**SRP**	**NH4**	**Nitrite**	**Nitrate**
Min	7.30	0	0.10	1.50	88.90
Max	56.20	23.20	25.70	6.70	203.70
Mean	22.36	7.89	7.63	3.88	142.71
Std. error	1.12	0.61	0.47	0.14	3.11
Median	19.4	6.25	6.95	3.85	138.9

The linear regression of the environmental variables (dispersion parameter in accordance to Gaussian distribution) and sampling area/period pointed to some interaction between spatial and temporal distribution. Nitrate was higher (p= 0.001057) in samples VG4 and VG5, close to the lotic portion of the reservoir, and presented higher concentration in the dry season (p= 4.332e-09). The ammonium concentration was lower in sample VG3 (p=0.001959), especially in the dry season (p=0.004997). Dissolved phosphorus concentration was higher in the dry season (p=0.02543), while the euphotic zone depth (p=4.462e-11) and temperature (p=8.367e-12) were higher in the rainy season. The sampling sites close to the lotic portion of the reservoir showed higher levels of nutrients, while the sampling sites close to the dam showed higher euphotic zone depth. The dry season presented higher levels of nutrients.

Picocyanobacteria morphological delimitation

The samples that showed the highest biomass and richness of picocyanobacteria were combined and used for species identification, image capture and additional measurements of diameter, length/width ratio and NCD. The Neighbour Cell Distance (NCD) was defined as the distance between the two closest neighbour cells, at the same microscope focus. The aim of such additional attribute was to measure, numerically, the cell density within a colony, requiring a higher number of measurements to accomplish statistical criteria, indicated as the 'N' attribute in morphological metrics below. A cluster analysis was performed as an additional tool to delimitate some cypitical picocyanobacteria individuals. All the 31 morphologically discrete individuals presented large colonies and oval or rod shaped cells. Initially the individuals were classified in five groups, with a cophenetic correlation coefficient of 0.8855 (Fig. 2b), sustained by ANOSIM (p=0.001) permutation Similarity Analysis. The groups formed by clustering were significantly different according to the five categorical variables (cell rows, colony size, colony form, l/w ratio class and NCD class). The GLM analysis showed that the l/w ratio was significantly different among these groups (p =0.02, F value=2.6683). The NCD was also significantly different (p=0.002, F value=4.2362). The individuals grouped under the "Cplanc" group (*C. planctonicum*) showed larger R and NCD range of variation (Fig. 2c), and both metrics were important to distinguish between *Cyanodictyon* and *Anathece* species. Initial and late development stages of *C. planctonicum* were also distinguishable (Fig. 2b).

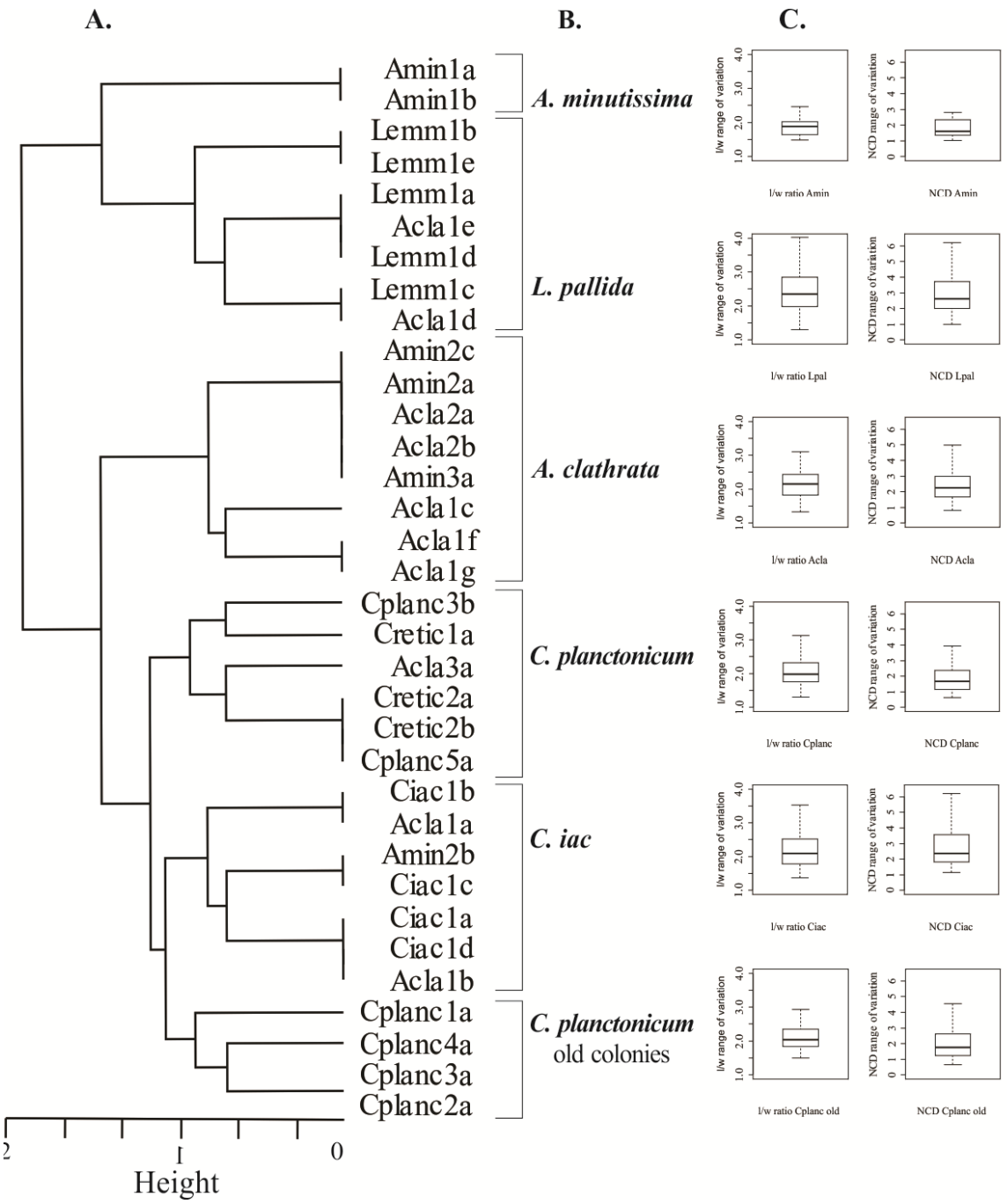


Fig. 2 - Cluster analysis (average linkagealgorithm, Euclidean similarity measure) among 31 morphologically discrete individuals from samples combined in order to increase cell density and reach a higher amount of measurements. A: Cluster showing the initial candidate taxa, grouped through visual criteria, B: Species identified through clustering appoach, C: Boxplot of each defined group showing the range of variation for R and NCD. Acla= *A. clathrata*; Amin= *A. minutissima*; Ciac= *C. iac*; Cplanc= *C. planctonicum*; Lemm= *L. pallida* (Cophenetic correlation coefficient = 0.8855).

Taxonomical analysis

Twenty three coccoid cyanobacteria taxa were recorded, all of them at the species level. The highest species richness was found in the Synechococcales order, with 15 species, while Chroococcales presented eight species. *Microcystis* Kützing et Lemmermann and *Aphanocapsa* Nägeli were the genera that presented the highest species richness, with five species each, followed by *Anathece* (Komárek et Anagnostidis) Komárek, Kastovsky et Jezberová and *Merismopedia* Meyen, with three species each.

A remarkable feature of the cyanobacteria community in this system is their small cell size. The community was predominantly composed by the Synechococcales order. This group of planktonic species, comprising cells of 0.5-2 µm in diameter, presents particularly difficult structure types, which differ only by cell density and by the cell organization within the colony. In addition, it is known that these traits may change or disappear under culture conditions (Sant'Anna et al., 2004), increasing the identification complexity. The statistical approach adopted here allowed a more accurate delimitation of five species of cylindrical and rod-shaped picoplanktonic cyanobacteria, which are described below.

Subclass Synechococcophycideae

Order Synechococcales

Family Synechococcaceae

Anathece (Komárek et Anagnostidis) Komárek, Kastovsky et Jezberová (2011)

The traditional subgenus *Anathece* was reviewed and established at the genus level by phylogenetic and morphological approaches (Komárek et al., 2011). The authors commented that *Anathece* and *Aphanothece* have different evolutionary lines, confirmed by phylogenetic analysis and the phenotypic characteristics. *Anathece* show small-celled, planktonic colonies with rod-like or oval cells, mostly (0.8)1-2(6) X 0.3-2 µm, usually less than 2 µm wide. Colonies are irregular, cells are irregularly distributed within the mucilage, without individual envelopes. *Anathece* species can be distinguished by their

colony morphology. In view of the present study, the length/width ratio seems to be an important character to be observed for the characterization of the species.

***Anathece clathrata* (W. West et G. S. West) komárek, kaštovský et jezberová**, *Eur. J. Phycol*.46(3): 315-326, 2011.

(Figs 3-8)

Basionym: *Aphanothece clathrata* W. West et G.S. West, Trans. Roy Irish Acad., 33(A): 111, 1906.

Dimensions: 1.1-3 x 0.6-1.1 µm l/w (mean 1.7 x 0.8 µm, SE 0.03 x 0.01, σ 0.3 x 0.1, median 1.7 x 0.8 µm, NCD: 0.8-8.2 µm, median NCD 2.4 µm, R 1:1.3-3.6 µm, median R 1:2.1 µm, N 8, n 123).

Colonies spherical (Fig. 3) to irregular (Fig. 8), more or less flat (Figs 4, 5), sometimes clathrate (Fig. 5), with more or less loosely aggregated cells. Mucilage homogeneous, diffluent. Cells always rod-shaped, homogeneous content, pale blue-green, without aerotopes.

This species differs from *Anathece minutissima* (W. West) Komárek, Kaštovský et Jezberová by the form of the colony, larger l/w ratio of the cells and more loosely (larger NCD) and regular arrangement of the cells in the colony (Fig. 2c).

Occurrence: planktonic in mesotrophic or eutrophic lakes and reservoirs. Records over all continents, except Africa (Guiry & Guiry 2015).

References: Komárková-Legnerová & Cronberg, (1994); Komárek & Anagnostidis, (1998); Joosten, (2006); McGregor et al., (2007); Komárek et al., (2011).

***Anathece minutissima* (W. West) Komárek, Kaštovský et Jezberová**, *Eur. J. Phycol.* 46(3): 315-326, 2011.

(Figs 9, 10)

Basionym: *Microcystis minutissima* W. West, Proc. Roy. Irish Acad., 31:35, 1912.

Dimensions: 1-2.2 x 0.6-1 µm l/w (mean 1.5 x 0.8 µm, SE 0.06 x 0.02, σ 0.3 x 0.1, median 1.5 x 0.8 µm, NCD: 1-2.8 µm, median NCD 1.6 µm, R 1:1.3-2.8 µm, median R 1:1.9 µm, N 6, n 21).

Colonies irregular (Fig. 9) and elongate (Fig. 10), with irregularly and more or less densely aggregated cells, showing regions with denser aggregates, sometimes with aspect of cloudy clusters. Mucilage wide, homogeneous, with diffluent margin. Cells

widely oval or shortly rod-shaped, homogeneous content, pale blue-green, without aerotopes.

The smaller NCD (median NCD 1.7 µm) and consequent more densely aggregate cells, along with the smaller R (median R 1:1.9 µm) and consequent more oval cells, separate *A. minutissima* from *A. clathrata.*

Occurrence: Freshwater, planktonic from oligotrophic to eutrophic lakes and reservoirs. Records over all continents, except North America region and Africa (Guiry & Guiry, 2015).

References: Komárek & Anagnostidis, (1998); Joosten, (2006); Komárek et al., (2011); Martins et al., (2012).

***Anathece smithii* (Komárková-Legnerová et Cronberg) Komárek, Kaštovský et Jezberová** *Eur. J. Phycol*. 46(3): 315-326, 2011.

(Figs 11, 12)

Basionym: *Aphanothece smithii* Komárková-Legnerová et Cronberg, Algolog. Stud. 72: 25, 1994.

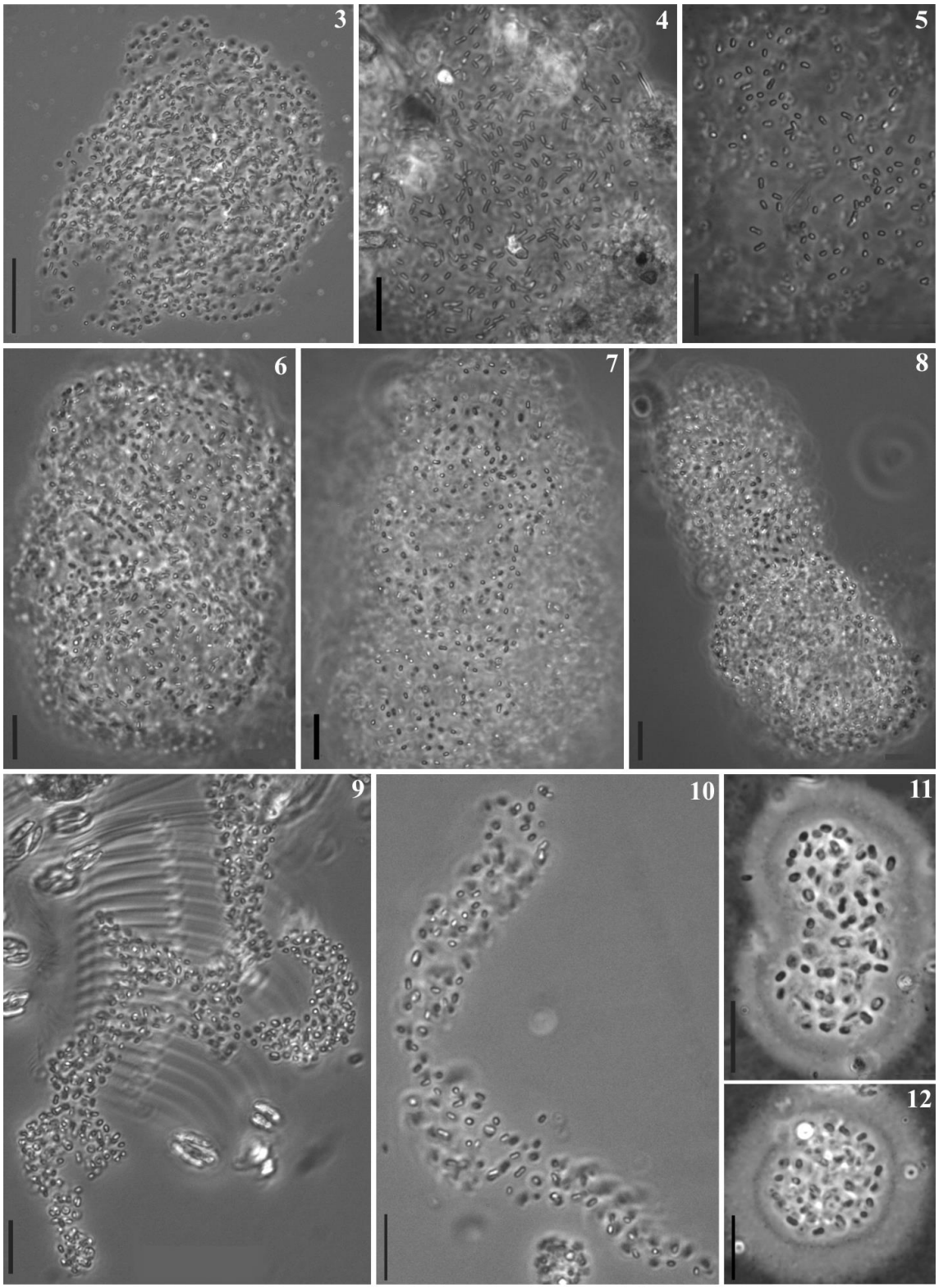
Dimensions: 1.2-3.7 x 0.9-1.5 µm l/w (mean 2.2 x 1.2 µm, SE 0.36 x 0.11, σ 0.9 x 0.3, median 2.1 x 1.1 µm, R 1:1-4.2 µm, median R 1:1.8 µm, N 2, n 8).

Colonies spherical to elongated, rarely irregular, cells loosely and more or less regularly distributed in the mucilage. Mucilage homogeneous, diffluent, wide. Cells oval, cell content homogeneous pale blue-green to grey, with small aerotopes.

The size and R of the cells, along with the dense and wide mucilage, were the most distinctive features.

Occurrence: planktonic in mesotrophic and slightly eutrophic, usually large water bodies (lakes and reservoirs). Europe, South America and Australia and New Zealand (Guiry & Guiry, 2015).

References: Komárková-Legnerová & Cronberg, (1994); Komárek & Anagnostidis, (1998); Tucci et al., (2006); Sant'Anna et al., (2007); Delazari-Barroso et al., (2007); Komárek et al., (2011); Martins et al., (2012).



Figs 3-12. (3) *Anathece chlatrata*; (4) *Anathece chlatrata*, chlatrate colony; (5) *Anathece chlatrata*, maximal length/width ratio; (6,7) *Anathece chlatrata*; (8) *Anathece chlatrata*, minimal length/width ratio; (9,10) *Anathece minutissima*; (11,12) *Anathece smithii.* Scale bar 10 µm.

***Cyanodictyon* Pascher (1914)**

This genus is characterized by irregularly reticulate to three-dimensional colonies. The cells form rows that interlace at varying levels, giving a three-dimensional aspect. The cells are more or less spherical to elongate. The main difference between *Cyanodictyon* and *Anathece* is the arrangement of the cells in the mucilage. Without the formations of the netlike colonies, *Cyanodictyon* could be mistaken for *Aphanothece* (Hickel, 1981; Joosten, 2006).

***Cyanodictyon iac* Cronberg et Komárek**, *Arch. Hydrobiol./Algolog. Stud.*75: 323-352, 1994.

(Figs 13-16)

Dimensions: 1-2.9 x 0.6-1.1 µm l/w (mean 1.6 x 0.8 µm, SE 0.07 x 0.02, σ 0.46 x 0.1, median 1.5 x 0.75 µm, NCD: 1.1-7.3 µm, median NCD 2.2 µm, R 1:1.4-3.5 µm, median R 1:2 µm, N 11, n 44).

Colonies spherical to irregularly spherical (Fig 16) or lobate (Figs 13-15), usually composed by several subcolonies (Figs 13-14, arrows). Mucilage homogeneous, diffluent. Cells oval to almost rod-shaped (Fig. 15), arranged in arcuate rows in the superficial layer of the colony (Figs 13, 14, arrows). Cell content homogeneous, pale blue-green.

The larger NCD (Fig. 2c) and consequently more loose arrangement of the cells within the colony, along with the presence of subcolonies (lobate colony), distinguishes this species from *Cyanodictyon planctonicum* Meyer.

Occurrence: planktonic in mesotrophic lakes and reservoirs. Europe and South America (Guiry & Guiry, 2015).

References: Cronberg & Komárek, (1994); Komárek & Anagnostidis, (1998); Sant'Anna et al., (2007).

***Cyanodictyon planctonicum* Meyer**, *Arch. Hydrobiol./Algolog. Stud.* 75: 183-188, 1994. **(Figs 17-24)**

Dimensions: 1.1-2.8 x 0.6-1.3 µm l/w (mean 1.6 x 0.8 µm, SE 0.03 x 0.01, σ 0.4 x 0.1, median 1.6 x 0.8 µm, NCD: 0.6-5.8 µm, median NCD 1.7 µm, R 1:1.3-3.2 µm, median R 1:2 µm, N 13, n 120).

Colonies more or less spherical to irregular, sometimes reticulate (Fig. 22), tridimensional (Figs 23, 24, arrows). Cells oval (Fig. 20) to rod-shaped (Figs 17, 18), straight or slightly curved (Fig. 17), arranged in pseudofilamentous rows (Figs 18-20, arrows). Cell content homogeneous, pale blue-green, without aerotopes.

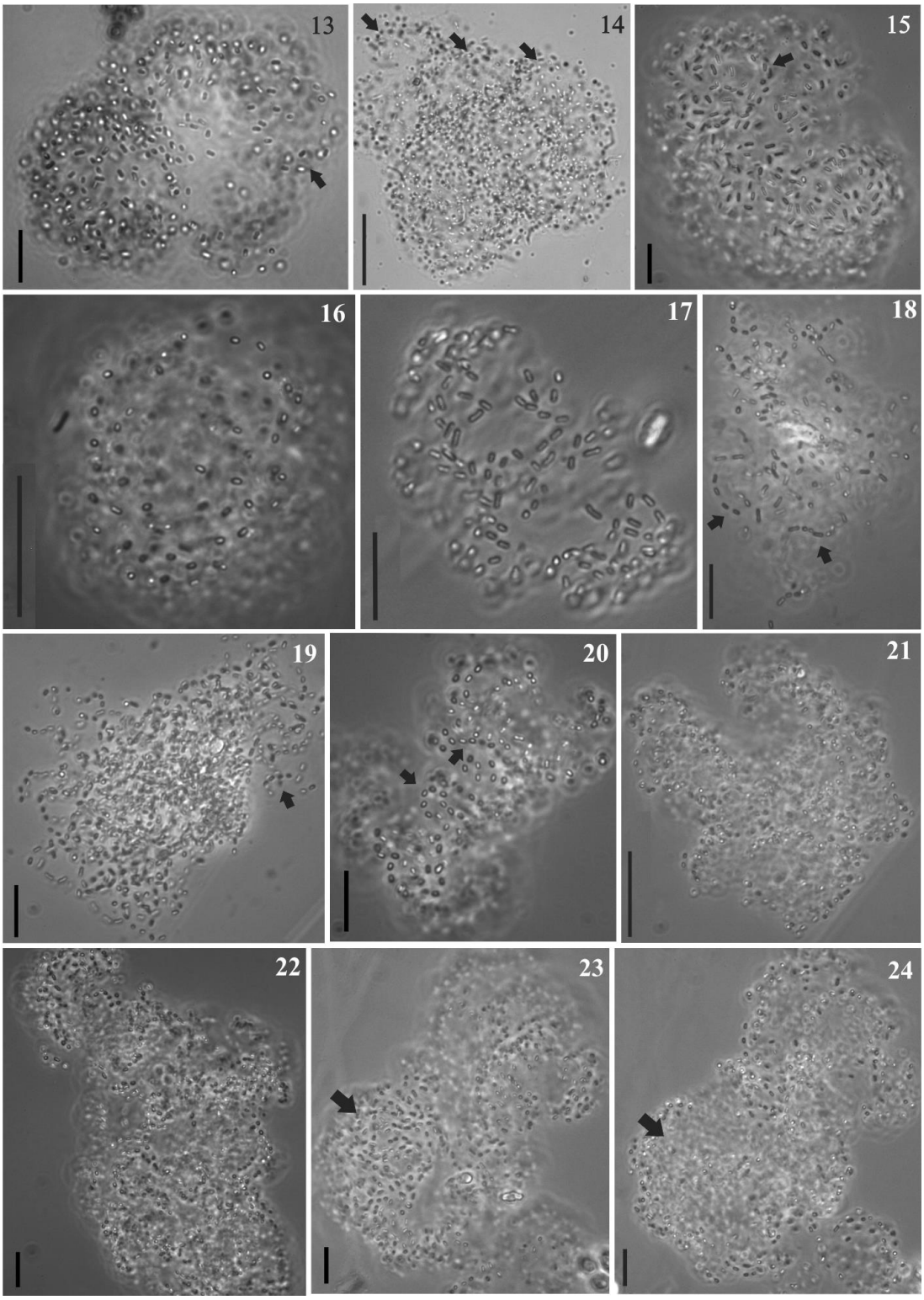
Depending on the developmental stage, the cells can be more (Figs 17, 18) or less (Figs 19-24) distant from each other (Fig. 2c, slightly higher upper limit of NCD metric), the cell density within the colony can vary widely as well. The netlike colonies distinguish this species from *Anathece clathrata.* The structureless and irregular colony and the more densely aggregated cells within the colony (smaller NCD) separate this species from *Cyanodictyon iac* (Fig. 2c).

We encountered some difficulty to set the boundaries between two very similar species of *Cyanodictyon*, with same ecological distribution. The morphological characteristics of *Cyanodictyon reticulatum* (Lemmermann) Geitler overlap several characteristics of *C. planctonicum* (Figs 18-20). According to Meyer, (1994), the difference between *C. planctonicum* and *C. reticulatum* is that the spherical cells of *C. reticulatum* form mucilaginous tubes (Fig. 20 arrow). According to Hickel, (1981), in *C. reticulatum* each row of cells is surrounded by thin mucilage, however in older colonies the mucilage may cover the entire colony, becoming similar to that found in *C. planctonicum* (Figs 18, 20). According to Komárková-Legnerová & Cronberg, (1994), the difference between these two species are the presence of gelatinous tubes in *C. reticulatum*, while in the early stages the mucilage surrounds the whole colony of *C. planctonicum*. In that paper the authors show the confusion between this two species and comment that *C. reticulatum* could be, in fact, the same species of *C. planctonicum* sensu Hickel when the cells are elongated. Joosten, (2006) cites *C. reticulatum* sensu Hickel as a synonym of *C. planctonicum* and comments that populations of *C. planctonicum* are initially misidentified as *C. reticulatum*, but the latter has spherical cells (Fig. 20). These descriptions seem to overlap in several characteristics, mainly with respect to large and old colonies. Since the specimens observed in our study have oval to rod-shape cells,

irregular, not elongated colonies, and do not present evident tubes of cells, as illustrated in the literature, we decided to identify our specimens as *C. planctonicum*.

Occurrence: freshwater, planktic in mesotrophic to eutrophic lakes and reservoirs. Europe, North America, South America, Australia and New Zealand (Guiry & Guiry, 2015).

References: Hickel, (1981); Komárková-Legnerová & Cronberg, (1994); Meyer, (1994); Komárek & Anagnostidis, (1998); Joosten, (2006); McGregor et al., (2007).



Figs 13-24. (13-15) *Cyanodictyon iac*, lobate colony, arrow indicates subcolonies; (16) *C. iac*, irregularly spherical colony; (17) *C. planctonicum*, young colony, cells loosely arranged; (18) *C. planctonicum*, arrow indicates pseudofilamentous rows, with apparent mucilaginous tube surrounding the row; (19) *C. planctonicum*, arrow indicates pseudofilamentous rows, old colony of *C. reticulatum* sensu Hickel (1981); (20) *C. planctonicum*, minimal length/width ratio; (21,22) *C. planctonicum*, old colonies, irregular and reticulate colonies; (23,24) *C. planctonicum*, old colonies, arrows indicates tridimensional aspect of the colony. Scale bar 10 µm.

***Lemmermanniella* Geitler (1942)**

Genus characterized by mucilaginous spheres or irregularly oval colonies, sometimes presenting subcolonies. The mucilage is homogeneous, diffluent, usually structureless. The cells are rod-shaped or cylindrical, several times longer than wide, solitary, more or less loosely distributed, placed in a peripheral layer, more or less tangentially arranged at the surface of the colony (Fig. 35, arrow). The cellular content is homogeneous, pale blue-green, without aerotopes. The mucilaginous compact colonies, the larger length of the cells, larger length/width ratio and cellular arrangement within the colony are the major characteristics that distinguish this genus from *Anathece.*

***Lemmermanniella flexa* Hindák** *Arch. Hydrobiol./Algolog. Stud.* 40: 393-401, 1985.
(Figs 35, 36)

Dimensions: 2.2-5.5 x 0.5-1 µm l/w (mean 3 x 0.7 µm, SE 0.3 x 0.04, σ 1.2 x 0.14, median 2.6 x 0.7 µm, NCD 1-6.1 µm, median NCD 3.2 µm, R 1:2.1-6.2 µm, median R 1:4 µm, N 3, n 13).

Colonies generally irregularly oval or elliptical. Mucilage homogeneous, conspicuous. Cells narrow, cylindrical, slightly bent with rounded ends, solitary, in two after division or in short pesudofilamentous rows, irregularly arranged in one layer, just beneath the surface of the colonial mucilage. Cell content homogeneous, pale blue-green or grey.

The dimensions are quite smaller than the original description. The usual slightly bent cell is an important characteristic.

Occurrence: planktonic in mesotrophic lakes and reservoirs. Europe and South America.

References: Komárek & Anagnostidis, (1998); Joosten, (2006).

***Lemmermanniella pallida* (Lemmermann) Geitler**, in *Engler & Prantl, Natürl. Planzenfam., Ed. 2,* 1b: 62, 1942.
(Figs 31-34)

Basionym: *Coelosphaerium pallidum* Lemmermann, Bot. Centralbl. 76: 154, 1898.

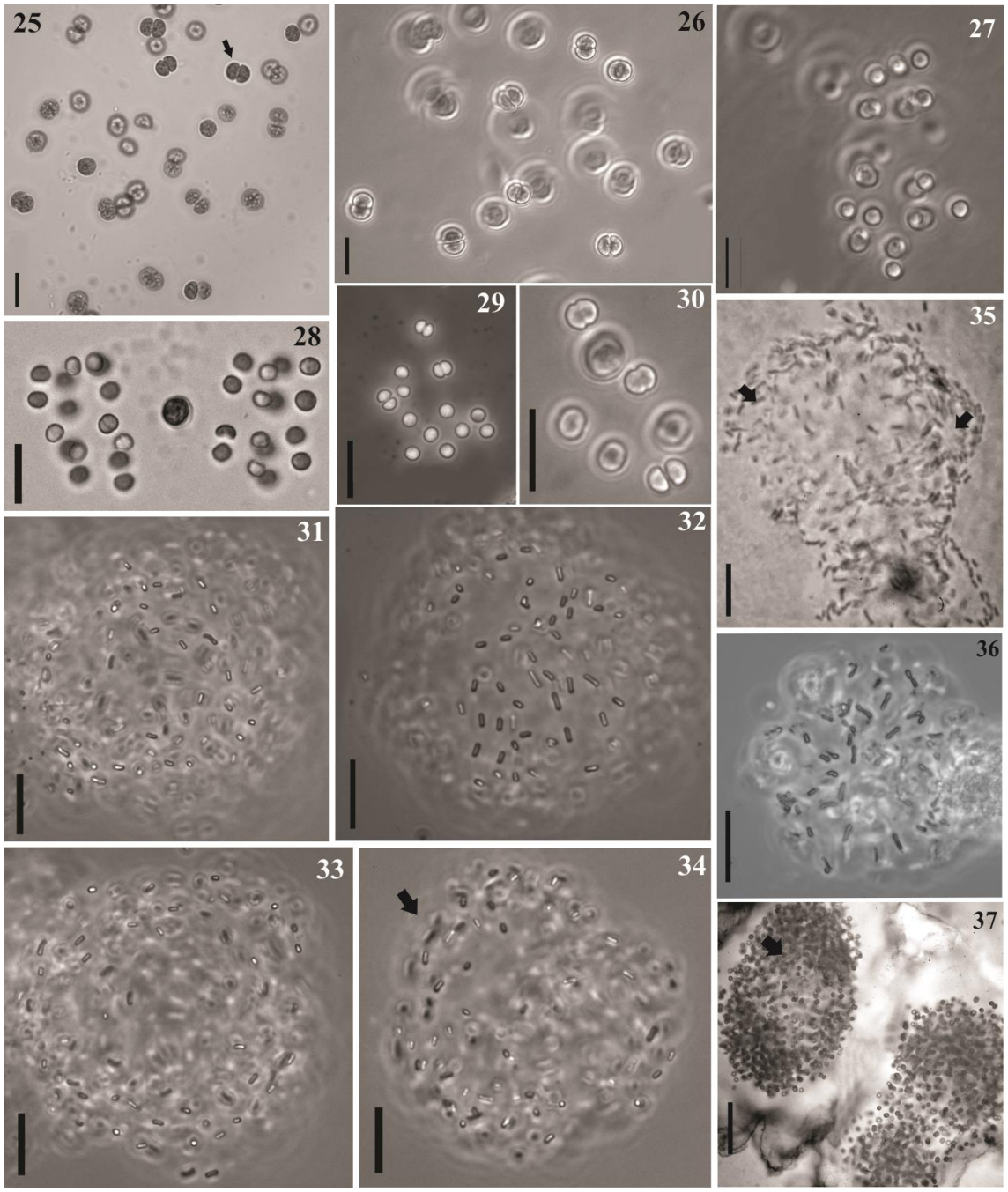
Dimensions: 1.4-3 x 0.6-0.8 µm l/w (mean 1.9 x 0.8 µm, SE 0.07 x 0.01, σ 0.4 x 0.05, median 1.8 x 0.8 µm, NCD: 1-7.2 µm, median NCD 2.6 µm, R 1:1.7-4 µm, median R 1:2.5 µm, N 8, n 36).

Colonies usually elliptical to oval. Mucilage homogeneous, conspicuous. Cells rod-shaped, usually straight, with rounded ends, irregularly and more or less densely distributed in a surface layer (Fig. 34, arrow). Cell content homogeneous, pale blue-green. This specimen present some similarities with *C. iac*, mainly in early stages, when both of them usually present round colonies, but the most common lobular aspect of *C. iac* colony and both R and NCD slightly larger distinguish *L. pallida* from *C. iac* (Fig. 2c).

The smaller length and more straight shape of the cells distinguish *L. pallida* from *L. flexa.* The occurrence of *L. pallida* is little known in Brazil, and it was only cited by Ferragut et al., (2005) for São Paulo state.

Occurrence: plankton or metaphyton, in large, mesotrophic to slightly eutrophic lakes and reservoirs. Records over all continents, except Africa (Guiry & Guiry, 2015).

References: Komárková-Legnerová & Cronberg, (1994); Komárek & Anagnostidis, (1998); Ferragut et al., (2005).



Figs 25-37. (25,26) *Limnococcus limneticus*, arrow indicates individual envelope detail; (27-30) *Chroococcus dspersus;* (31) *Lemmermanniella pallida*; (32) *Lemmermanniellapallida*, colony surface detail; (33) *Lemmermanniella pallida*; (34) *Lemmermanniella pallida*, arrow indicates cells arranged at the surface layer of the colony; (35,36) *Lemmermanniella flexa*; (37) *Sphaerocavum brasiliense*, arrow indicates a hollow colony detail. Scale bar 25-36: 10 µm; 37: 50 µm.

Family Merismopediaceae

***Aphanocapsa* Nägeli (1849)**

The generally numerous spherical cells, irregular, densely or sparsely, distributed within the colony, without aerotopes, are the genus major characteristics. As with several other taxa, the diameter of the cells is considered the distinctive character between the *Aphanocapsa* species. The habit and the arrangement of cells in the colony are also important features to enable the species identification. According to Komárek & Anagnostidis, (1998), the planktonic species are distinguished by cell arrangement and then by cell size. Among the species with densely arranged cells there are *A. incerta* (Lemmermann) Cronberg et Komárek and *A. holsatica* (Lemmermann) G. Cronberg et J. Komárek. Among the species with loosely arranged cells there are *A. delicatissima* West et G. S. West, *A. elachista* Westet G.S. West, *A. planctonica* (G.M. Smith) Komárek et Anagnostidis, *A. nubilum* Komárek et Kling and *A. conferta* (West et G.S. West) Komárková Legnerová et Cronberg. The authors define *A. elachista* as common in tropical countries, along with *A. koordersii* K.M. Strom. According to Sant'Anna et al., (2004), the species with confirmed tropical distribution and sparsely distributed cells are *A. delicatissima*, *A. elachista* and *A. koordersii*, separated from each other only by cell size.

***Aphanocapsa delicatissima* West et G. S. West**, *J. Linn. Soc., Bot.*40: 431, pl. 19: figs 2-3, 1912.

(Fig. 38)

Dimensions: 0.7-1.0 µm diam. (mean 0.9 µm, SE 0.02, σ 0.1, median 0.86 µm, N 7, n 26).

Colonies larger than 50 µm, elliptical to irregular, mucilage homogeneous, diffluent. Cells spherical, more or less densely and regularly distributed, with intercellular spaces not larger than two cells. Cell content homogeneous, pale blue-green.

This species can be recognized by the irregular form of the colony, cellular distribution within the colony and cell diameter. Young colonies are spherical to oval, but become lobate to irregular.

Occurrence: planktonic, probably cosmopolitan, recorded in several kinds and conditions of water bodies. Records over all continents, except Africa (Guiry & Guiry, 2015).

References: Komárková-Legnerová & Cronberg, (1994); Komárek & Anagnostidis, (1998); Werner, (2002); Sant'Anna et al., (2004); Tucci et al., (2006); McGregor et al., (2007); Domingues & Torgan, (2011); Nogueira et al., (2011); Martins et al., (2012).

***Aphanocapsa elachista* West et G. S. West**, *J. Linn. Soc., Bot.*30: 264-280, pls XIII-XVI, 1894.

(Figs 41,42)

Dimensions: 0.9-1.4 µm diam. (mean 1.1 µm, SE 0.04, σ 0.1, median 1.1 µm, N 5, n 14)

Colonies up to 50 µm, spherical to oval, loosely distributed, after division in two (Figs 41-42, arrows). Mucilage homogeneous, scarce, diffluent. Cells spherical, grey or pale blue-green content.

The cell diameter of the individuals observed is quite smaller than the original description (1.3-2 µm). Despite this, the colony morphology, cell arrangement and mainly the stability of the traits found in the samples analyzed seem to characterize a small population of *A. elachista*. The main difference between *A. elachista* and *A. koordersii* is the cell size, up to 2 µm in *A. elachista* and 2-3 µm in *A. koordersii.*

Occurrence: planktonic in mesotrophic to eutrophic waters, common in tropical countries. Records over all continents (Guiry & Guiry, 2015).

References: Komárek & Anagnostidis, (1998); Sant'Anna et al., (2004); Tucci et al., (2006); McGregor et al., (2007); Nogueira et al., (2011); Martins et al., (2012).

***Aphanocapsa incerta* (Lemmermann) Cronberg et Komárek**, *Arch. Hydrobiol./Algolog. Stud*.75: 323-352, 1994.

(Figs 39, 40)

Basionym: *Polycystis incerta* Lemmermann, Forschungsber. Boil. Stat. Plön 7: 132, 1899.

Dimensions: 0.8-1.4 µm diam. (mean 1.05 µm, SE 0.05, σ 0.17, median 0.1 µm, N 4, n 12).

Colonies rounded, rarely irregular, cellular distribution within the mucilage very dense and irregular. Mucilage homogeneous, diffluent. Cells spherical, blue green content. The highly dense and irregular arrangement of the cells is the distinctive characteristic.

Occurrence: commonly in plankton of mesotrophic to eutrophic water bodies, cosmopolitan. Records over all continents, except Africa (Guiry & Guiry, 2015).

References: Komárek & Anagnostidis, (1998); Sant'Anna et al., (2004); Tucci et al., (2006); McGregor et al., (2007); Domingues & Torgan, (2011); Martins et al., (2012).

***Aphanocapsa holsatica* (Lemmermann) Cronberg et Komárek** *Arch. Hydrobiol./Algolog. Stud*.75: 323-352, 1994.

(Figs 43, 44)

Basionym: *Microcystis holsatica* Lemmermann, Forschungsber. Boil. Stat. Plön 10: 150, 1903.

Dimensions: 1-1.2 µm diam. (mean 1.1 µm, SE 0.04, σ 0.1, median 1.1 µm, N 3, n 13).

Colonies irregular, lobate, elongate or clathrata, cells more or less densely aggregated. Mucilage homogeneous, diffluent, irregular outline. Cells spherical, grey or pale blue-green content.

The cell size of *A. holsatica* and *A. delicatissima* overlaps, but the colony morphology (*A. holsatica* elongate, usually lobate and clathrate) and the cell arrangement (*A. holsatica* dense and irregular) distinguish these two species.

Occurrence: plankton of mesotrophic to eutrophic freshwater bodies, cosmopolitan. Records over all continents, except Africa (Guiry & Guiry, 2015).

References: Komárek & Anagnostidis, (1998); Werner, (2002); Sant'Anna et al., (2004); McGregor et al., (2007); Martins et al., (2012).

***Aphanocapsa nubilum* Komárek et Kling**, *Arch. Hydrobiol./Algolog. Stud.* 61: 24-45, 1991.

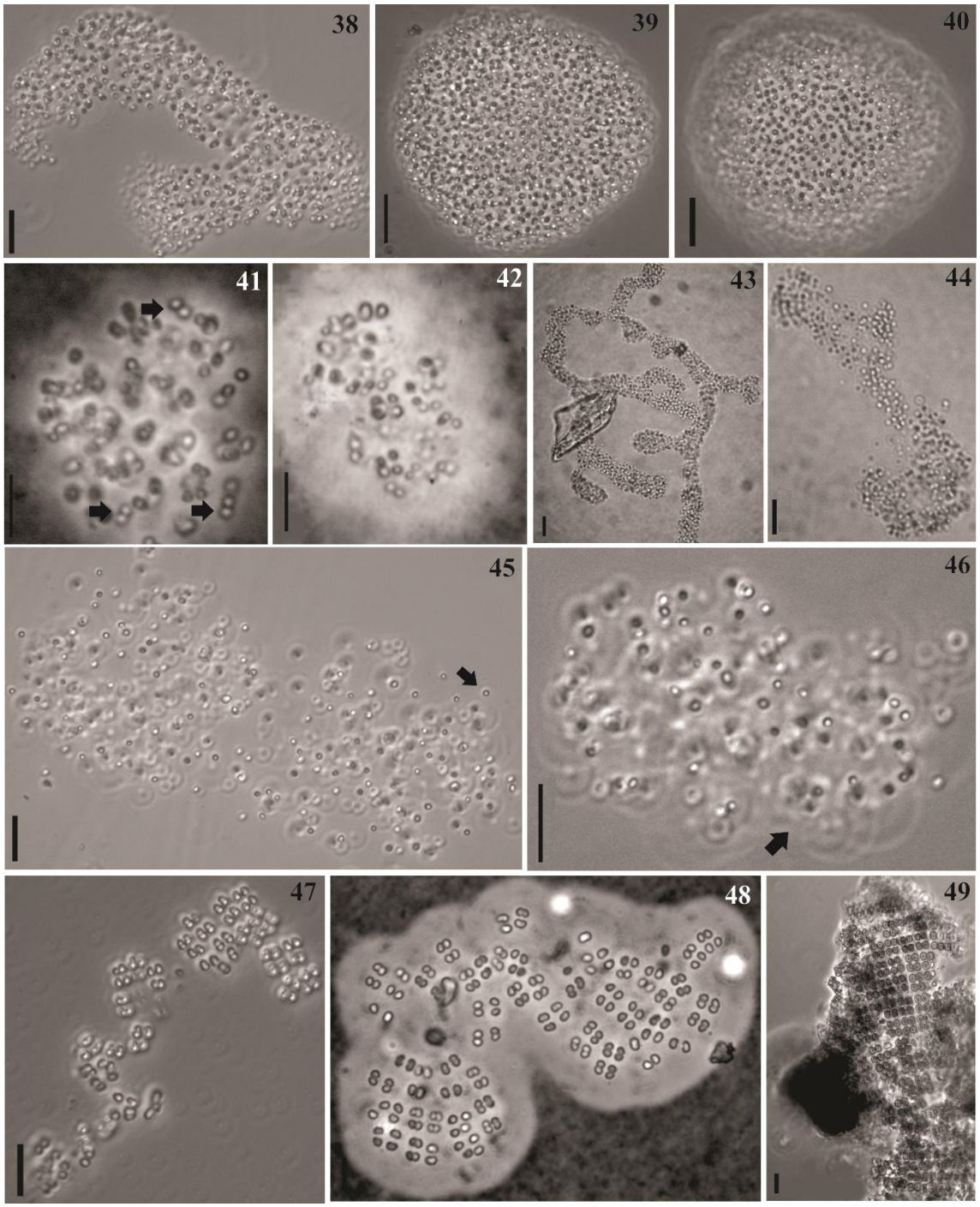
(Figs 45-46)

Dimensions: 1-2.5 µm diam. (mean 1.4µm, SE 0.1, σ 0.42, median 1.3µm, N 7, n 20).

Colonies irregularly spherical, cloud-shaped, mucilage conspicuous, following the outer margin of the cluster of cells, which can be seen as individual envelopes (Fig. 45 arrow). At the outer margin of the whole colony, the mucilage has a cloud-shape (Fig. 46 arrow), up to 3 µm wide around the border. Cells spherical, sparsely and more or less regularly distributed. Cell content homogeneous, pale blue-green.

Occurrence: planktonic in clear lakes and mesotrophic reservoirs, probably cosmopolitan. Europe, South America, Australia and New Zealand (Guiry & Guiry, 2015).

References: Komárek & Anagnostidis, (1998); Sant'Anna et al., (2007); Costa & Dantas, (2012).



Figs 38-49. (38) *Aphanocapsa delicatissima*; (39,40) *Aphanocapsa incerta*; (41,42) *Aphanocapsa elachista*; (43,44) *Aphanocapsa holsatica*; (45,46) *Aphanocapsa nubilum*; (47) *Merismopedia tenuissima*; (48) *Merismopedia punctata*; (49) *Merismopedia glauca*. Scale bar 10 µm.

***Limnococcus* (Komárek et Anagnostidis) Komárková, Jezberová, O. Komárek et Zapomelová (2010)** and ***Chroococcus* Nägeli (1849)**

Using phylogenetic analysis, Komárková et al., (2010) showed that *Chroococcus limneticus* Lemmermann does not belong to the real *Chroococcus* genus and should be described as a new genus named "*Limnococcus*". The difference is evident not only from the phylogenetic position of its 16S rRNA gene sequence, but also from its distinct morphology and ecology. The *Chroococcus* genus is characterized by cells or group of cells surrounded by mucilaginous envelopes that usually follow the cell form, sometimes lamellate, distinct or diffuse at the margin, while *Limnococcus* presents fine, homogeneous, and diffuse mucilage, in which the cells are irregularly arranged.

***Chroococcus dispersus* (Keissler) Lemmermann**, *Ark. Bot*. v. 2, n. 2, p. 102. 1904. **(Figs 27-30)**

Basionym: *Chroococcus minor* (Kützing) Nägeli var. *dispersus* Keissler, Verh. zool.-bot. Ges. Wien. 52: 311, 1902.

Dimensions: 2.5-3.9 μm diam. (mean 3.2 μm, SE 0.08, σ 0.4, median 3.3 μm, N 6, n 20).

Colony irregular, usually with groups of 4-8-16 cells, more or less distant and irregularly distributed. Colonial mucilage hyaline, diffuse, wide. Cells spherical, without individual mucilage, cell content homogeneous, pale blue-green, without aerotopes.

The smaller cells and the grouping are the main differences between *Chroococcus* (subg. *Limnococcus*) *dispersus* and *Limnococcus limneticus* (Lemmernann) Komárková, Jezberová, O. Komárek et Zapomelová.

Occurrence: planktonic, in clear lakes, oligotrophic or mesotrophic reservoirs. Records over all continents, except Africa (Guiry & Guiry, 2015).

References: Komárková-Legnerová & Cronberg, (1994); Komárek & Anagnostidis, (1998); Werner, (2002); Wehr & Sheath, (2003); Sant'Anna et al., (2004); Delazari-Barroso et al., (2007); Komárková et al., (2010); Martins et al., (2012).

***Limnococcus limneticus* (Lemmernann) Komárková, Jezberová, O. Komárek et Zapomelová**, *Hydrobiologia* 639: 69-83, 2010.

(Figs 25-26)

Basionym:*Chroococcus limneticus* Lemmermann, Bot. Zbl.76: 153, 1898.

Dimensions: 5-8.4 µm diam. (mean 5.9 µm, SE 0.16, σ 0.78, median 5.7 µm, N 7, n 23).

Colonies usually 8-16-32 celled, irregular, cells more or less distant from each other, sometimes with irregular groups. Mucilage hyaline, diffuse, wide. Cells spherical or widely oval before division, after division hemispherical, usually with individual mucilaginous envelope, cell content blue-green, homogenous to slightly granular, without aerotopes.

Limnococcus limneticus, along with *Chroococcus* (subg. *Limnococcus*) *dispersus* are widely distributed and showed an expressive number of individuals in the study area. Werner, (2002) also observed these two species together and observed that *L. limneticus* shows larger cells and do not form cell groups, like *C. dispersus. Chroococcus distans* (G.M. Smith) Komárková-Legnerová et Cronberg presents the cells even more distant from each other within the colony.

Occurrence: Freshwater, planktonic in oligotrophic or mesotrophic lakes and reservoirs. Records over all continents, except Africa (Guiry & Guiry, 2015).

References: Komárková-Legnerová & Cronberg, (1994); Werner, (2002); Wehr & Sheath, (2003); Komárková et al., (2010).

***Merismopedia glauca* (Ehrenberg) Kützing**, *Phycol. Germ.* p. 142, 1845.

(Fig. 49)

Basionym: *Gonium glaucum* Ehrenberg, Infunsionsp. 58, 1838.

Dimensions: 3-4.4 µm diam. (mean 3.9 µm, SE 0.15, σ 0.5, median 4.1µm, N 1, n 11).

Colonies flat, rectangular, up to 64-celled (rarely more), with more or less densely and regularly arranged cells. Colonial mucilage colourless, narrow. Cellular division in two perpendicular planes. Cells spherical to widely oval before division, hemispherical after division, with homogeneous or fine granular content.

Merismopedia glauca differs from *Merismopedia puncata* Meyen mainly by the arrangement of cells in the colony, more or less regularly and loosely disposed in *M. punctata*.

Occurrence: Freshwater, metaphytic, facultatively planktionic, frequent in shallow lakes, mesotrophic ponds and reservoirs. Cosmopolitan distribution. Records over all continents (Guiry & Guiry, 2015).

References: Komárková-Legnerová & Cronberg, (1994); Komárek & Anagnostidis, (1998); Sant'Anna et al., (2004); Tucci et al., (2006); McGregor et al., (2007); Martins et al., (2012).

***Merismopedia punctata* Meyen**, *Neues system der pflanzen-physiologie,* p. 440, plsX-XV, 1839.

(Fig. 48)

Dimensions: 2.4-3.5 µm diam. (mean 2.8 µm, SE 0.1 σ 0.3, median 2.7 µm, N 1, n 12).

Colonies flat, tabular, up to 64-celled (rarely more), usually in more or less regular rows with loosely arranged cells. Colonial mucilage wide, diffluent. Cells spherical to widely oval, cell content homogeneous, pale blue-green.

Besides the smaller cellular diameter, the cells arrangement is the main characteristic that distinguishes *M. punctata* from *M. glauca.*

Occurrence: planktonic and metaphytic in mesotrophic freshwater habitats. Records over all continents (Guiry & Guiry, 2015).

References: Komárek & Anagnostidis, (1998); McGregor et al., (2007).

***Merismopedia tenuissima* Lemmermann**, *Der grosse Waterneverstorfer Binnensee. Eine biologische Studie. Forschungsberichte aus der Biologischen Station zu Plön* 6: 154, 1898.

(Fig. 47)

Dimensions: 0.75-1.6 µm diam. (mean 1.14 µm, SE 0.1, σ 0.3, median 1.15µm, N 3, n 15).

Colonies flat, tabular, usually with subcolonies 8-16-32 celled. Cells in regular perpendicular rows, more or less densely aggregated, in groups of at least 4, spherical or oval (after division), with homogeneous, pale grey-blue or blue-green content.

M. tenuissima presents high morphological variability of the colony, but the minute size of the cells is a typical feature. This species showed a wide spatial and temporal distribution in the studied area.

Occurrence: planktonic and metaphytic in mesotrophic to moderately eutrophic, alkaline ponds, shallow lakes and reservoirs. Records over all continents, except Africa (Guiry & Guiry, 2015).

References: Komárková-Legnerová & Cronberg, (1994); Komárek & Anagnostidis, (1998); Joosten, (2006); Sant'Anna et al., (2007); McGregor et al., (2007).

Subclass Oscillatoriophycideae

Order Chroococcales

Family Microcystaceae:

***Microcystis* Kützing ex Lemmermann (1907)**

Microcystis is an important genus often found in the planktonic community, usually associated to bloom events. In the last years, the identification of the species belonging to this genus has been frequently discussed. Considering the inherent difficulties of their morphological characterization, the recent discoveries achieved through molecular approaches added new information, but also new uncertainties about the existence of morphospecies. However, Komárek & Komárková, (2002) emphasized the importance of knowing the *Microcystis* subgeneric diversity by traditional methods, because even when dealing with unidentifiable colonies, atypical stages or transient forms of *Microcystis*, which commonly occur in the plankton, several characteristic morphotypes can be observed. These morphotypes show distinct phenotypic and ecophysiological features, repeatedly occur in different regions and are stable in culture.

Therefore, every new contribution about the occurrence of *Microcystis* species in one environment is valuable. The classical methods, usually identifying the subgeneric forms based on their phenotypic traits, still have the same importance as the recent molecular approaches, since they can include information about particular adaptations of different morphospecies to the environment.

Here we describe five species of *Microcystis*, identified according to the International Code of Botanical Nomenclature.

***Microcystis aeruginosa* (Kützing) Kützing**, *Tab. Phycol*., 1:6, 1846.

(Figs 50, 51)

Basionym: *Micraloa aeruginosa* Kützing, Linnaea, 8:371, 1833.

Dimensions: 5-7 µm diam. (mean 5.4 µm, SE 0.17, σ 0.62, median 5 µm, N 9, n 45).

Colonies irregular in outline, lobate, usually shaped in subcolonies; mucilage difuse, forming a more or less wide margin around cells. Cells spherical, densely aggregated, with numerous aerotopes and pale blue green to dark green content (depending on the development stage).

The colonies found in Volta Grande reservoir showed a wide morphological variation, as expected and described for this species. Some specimens, in early developmental stages, were similar to *Radiocystis* Skuja. The feature that distinguishes these two genera is the cell division plane. When observing this feature, it becomes relatively easy to separate small colonies of *M. aeruginosa* from *Radiocystis*, since the cells of *M. aeruginosa* have three plans of division, while *Radiocystis* has one. Although apparently radiated cell lines are present, in *M. aeruginosa* we can observe cells in multiples plans. Other colonies were similar to *M. novacekii* (Komárek) Compère, however the mucilaginous envelope of *M. aeruginosa* presents no isolated cells in the periphery of the colony, as found in *M. novacekii*. The typical aspect of a *M. aeruginosa* colony is the irregular shape, densely aggregated cells, wide mucilaginous envelope, without delimited subcolonies or structured mucilage.

Occurrence: freshwater, planktonic in mesotrophic to eutrophic water bodies. Records over all continents (Guiry & Guiry, 2015).

References: Komárek&Anagnostidis, (1988); Hasler & Poulícková, (2002); Komárek & Komárková, (2002); Sant'Anna et al., (2004); Joosten, (2006); Delazari-Barroso et al., (2007); McGregor et al., (2007); Domingues & Torgan, (2011); Martins et al., (2012).

***Microcystis ichthyoblabe* Kützing** *Species algarum*. pp. [i]-vi, [1]-922. Lipsiae [Leipzig]: F.A. Brockhaus (1849).

(Fig. 61)

Dimensions: 2.5-3 µm diam. (mean 2.75 µm, SE 0.14, σ 0.29, median 2.75 µm, N 2, n 12).

The colonies are formed by subcolonies connected together, embedded in homogeneous, diffluent gelatinous envelopes, without holes, forming a margin around groups of cells, which can be seen as paths among the subcolonies. Mucilage diffuse, forming a narrow margin around the groups of cells. The cells are spherical, quite smaller than in other *Microcystis* species, densely and regularly aggregated, showing a flat aspect. Cell content pale blue green with aerotopes.

Occurrence: freshwater, planktonic in mesotrophic or slightly eutrophic, but not polluted lakes and reservoirs, rare in tropical countries. Europe, South America, Asia, Australia and New Zealand (Guiry & Guiry, 2015).

References: Komárková-Legnerová & Cronberg, (1994); Komárek & Anagnostidis, (1998); Komárek & Komárková, (2002).

***Microcystis novacekii* (Komárek) Compère**, Bull. Jard. Bot. Nat. Belg. 44: 19, 1974.

(Figs 52-55)

Basionym: *Diplocystis novacekii* Komárek, in Komárek et Ettl, Algologische Studien 206 p., 1958.

Dimensions: 5-6.3 µm diam. (mean 5.3 µm, SE 0.2, σ 0.5, median 5µm, N 7, n 30).

Colonies spherical to elliptical, compact, composed of usually spherical subcolonies. Mucilage wide and delimited with several solitary cells around the margin of the colony (Fig. 52). The cells are densely aggregated in the center of each subcolony (and therefore not forming holes) and spread out toward the colony margin. Cell content pale blue green to dark green, with aerotopes.

This species, along with *Microcystis protocystis* Crow, was very well represented in this plankton community. Some specimens had the colony and cell arrangement at the margin similar to those described by Hindák, (2006), with a few or none isolated cells

and several smaller subcolonies and more densely aggregated than the usually observed (Figs 54, 55).

Occurrence: freshwater, plankton in mesotrophic and eutrophic water bodies, common in the whole tropical zone. Europe, South America, Asia, Australia and New Zealand (Guiry & Guiry, 2015).

References: Komárek & Anagnostidis, (1998); Komárek & Komárková, (2002); Sant'Anna et al., (2004); Hindák, (2006); Joosten, (2006).

***Microcystis protocystis* Crow**, *The New Phytologist* 22(2): 59-68 (1923).

(Figs 56-58)

Dimensions: 3.7-7 µm diam. (mean 5.15 µm, SE 0.17, σ 0.83, median 5µm, N 11, n 40).

Colonies spherical to elliptical, mucilage diffuse forming a wide jagged margin. Cells sparsely distributed within the colony, with individual mucilaginous envelope (Fig. 56, arrow). Cell content pale blue green.

This is the *Microcystis* species more representative of the Volta Grande reservoir community. According to Sant'Anna et al., (2004), the sparse cell disposition gives *M. protocystis* the aspect of old or senescent colonies of different *Microcystis* species, like *M. aeruginosa.* For this reason, *M. protocystis* distribution should be wider than normally referred in the literature. Nguyen et al., (2012) observed that *M. protocystis* always showed a sparse arrangement of cells in a diffuse slime in different culture conditions.

Occurrence: freshwater, plankton in mesotrophic water bodies, probably one of the most common pantropical species. South America, Australia and New Zealand (Guiry & Guiry, 2015).

References: Komárek & Anagnostidis, (1998); Sant'Anna et al., (2004); Carvalho et al., (2008); Bittencourt-Oliveira et al., (2009); Domingues & Torgan, (2011); Martins et al., (2012); Nguyen et al., (2012); Rosini et al., (2013).

***Microcystis viridis* (A. Braun in Rabenhorst) Lemmermann,** *Abh. Nat. Ver. Bremen* 17: 342, 1903.

(Figs 59-60)

Basionym: *Polycystis viridis* A. Braun in Rabenhorst, Alg. Eur. 141-142: 1415. 1862.

Dimensions: 5-6.3 µm diam. (mean 5.3 µm, SE 0.2, σ 0.55, median 5µm, N 8, n 51).

This species is characterized by the typical packet-like subcolonies, more or less spherical or elongated. The mucilage surrounds the margin of the cell groups, which appear delimited by a slightly refractive outline (Fig. 59). The cells are densely aggregated.

Mucilage never forms a wide spherical margin around cell clusters and there are no solitary cells around subcolonies, what distinguishes this species from *M. novacekii* (sensu Hindák, 2006).

Occurrence: planktonic, slightly eutrophic lakes Europe, South America, Asia, Australia and New Zealand (Guiry & Guiry, 2015).

References: Komárek & Anagnostidis, (1998); Komárek & Komárková, (2002).

***Sphaerocavum* Azevedo et Sant'Anna (2003)**

This genus is characterized by the hollow colony, narrow mucilaginous envelope, diffluent and inconspicuous, cell division in two planes in successive generations (Azevedo & Sant'Anna, 2003). *Sphaerocavum* colonies, in young stages, appear like small hollow spheres, later hollow, spherical to irregular in shape, with holes in the colony surface, with cells content brownish or greenish, with aerotopes, more or less densely arranged, irregularly or in rows, always on the cell surface. Sant'Anna et al., (2004) highlights that the major difference between this genus and *Microcystis* is the hollow colony.

***Sphaerocavum brasiliense* Azevedo et Sant' Anna** *Algological Studies* 109: 79-92, 23 figs, 2003.

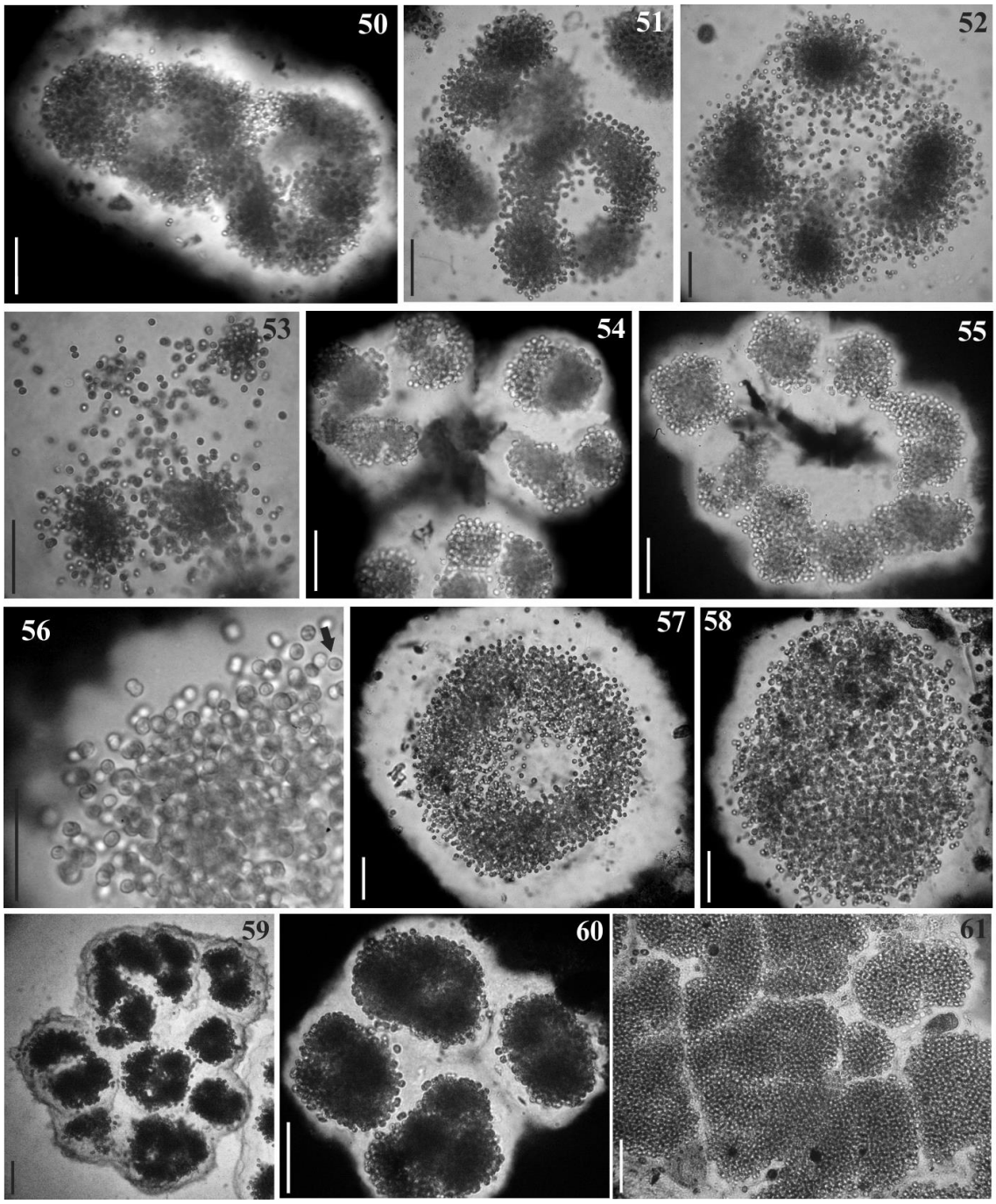
(Fig. 37)

Dimensions: 4.5-5.5 µm diam. (mean 5.2 µm, SE 0.12, σ 0.27, median 5µm, N 4, n 20).

The hollow colony is an evident feature of this species (Fig. 35 arrow). Several characters of the individuals observed do not match with the original description of this species. The population studied showed a wide mucilaginous envelope, while one of the typical characteristics of *Sphaerocavum,* even at genus level, is a narrow, inconspicuous mucilage. The size of the cells also was quite larger than in the original description, reaching 5.5 µm diam. This size variation has also been observed by other researchers. For example, Nogueira et al., (2011) observed a diameter between 3.2 to 5 µm. In this way, the population presented here is a morphotype of *S. brasiliensis*, with larger diameter and wide mucilaginous envelope. However, future molecular approaches are required to clarify these inconsistencies.

Occurrence: planktonic in mesotrophic to eutrophic water bodies. South America, Caribbean Islands, Australia and New Zealand (Guiry & Guiry, 2015).

References: Azevedo & Sant'Anna, (2003); Sant'Anna et al., (2004); Tucci et al., (2006), Nogueira et al., (2011); Rosini et al., (2013).



Figs 50-61. (50,51) *Microcystis aeruginosa*; (52) *Microcystis novacekii*; (53) *Microcystis novacekii*, young colony; (54,55) *Microcystis novacekii* sensu Hindák 2006; (56) *Microcystis protocystis*, arrow indicates individual envelope detail; (57,58) *Microcystis protocystis*; (59,60) *Microcystis viridis*; (61) *Microcystis ichtyoblabe*. Scale bar 50 µm.

Metagenomic Biodiversity Profile

The whole genome shotgun sequencing of samples VG3_May and VG3_Nov generated 551664 pairs of DNA fragments, according to Illumina paired ends sequencing strategy, with average length of 250bp. The pairs were maged using Flash 1.2.11 (Magoč & Salzberg, 2011), allowing for both outie and innie orientation of merging fragments, resulting in 450255 sequences with average length of 310bp. Both metagenomic datasets from the Volta Grande reservoir presented low sequencing coverage, i.e low average number of reads that align to known reference bases. The summary of metagenomic datasets and overall contribution of major taxa are presented in Table 2.

Tab. 2. Summary of metagenomic datasets and major taxa contribution. After pre-processing step, more than 80% of total sequences were merged. The relative contribution was computed according to the node summary (total number of assignments for phylum level or class level).

	VG3_May	**VG3_Nov**
Summary of metagenomic datasets		
Total_sequences post-processed	135069	416595
Flash merged read pairs	110504	339751
% merged	81,81	81,5
Megan assignments	18762	32857
% taxa contribution		
Actinobacteria (phylum)	10,6%	4,9%
Bacteroidetes/Chlorobi group	3,8%	5,5%
Chlamydiae/Verrucomicrobia group	0,2%	0,7%
Oscillatoriophycideae	53,5%	59,1%
Deinococcus-Thermus	0,4%	0,1%
environmental samples	0,9%	0,8%
Firmicutes	1,1%	1,1%
Planctomycetes	1,0%	2,0%
Proteobacteria (phylum)	20,4%	16,2%
Alphaproteobacteria	6,2%	4,0%
Betaproteobacteria	7,1%	6,5%
delta/epsilon subdivisions	0,7%	0,8%
Gammaproteobacteria	3,8%	2,6%

The Shannon-Weaver diversity index was slightly higher for sample VG3_Nov (3.2 for VG3_May and 3.4 for VG3_Nov), and the Simpson's reciprocal index was also higher for the sample VG_Nov (5.9 for VG3_May and 6.5 for sample VG3_Nov). Within Bacteria assignments, Cyanobacteria phylum presented highest contribution (Tab. 2), followed by Proteobacteria and Actinobacteria, according to the MEGAN assignments (BLASTnt alignment against NCBI nucleotide database). There was a higher contribution of Actinobacteria and Proteobacteria in sample VG3_May, while Cyanobacteria was

slightly higher in sample VG3_Nov, specially regarding *Microcystis* and *Synechococcus* abundances (Fig. 62).

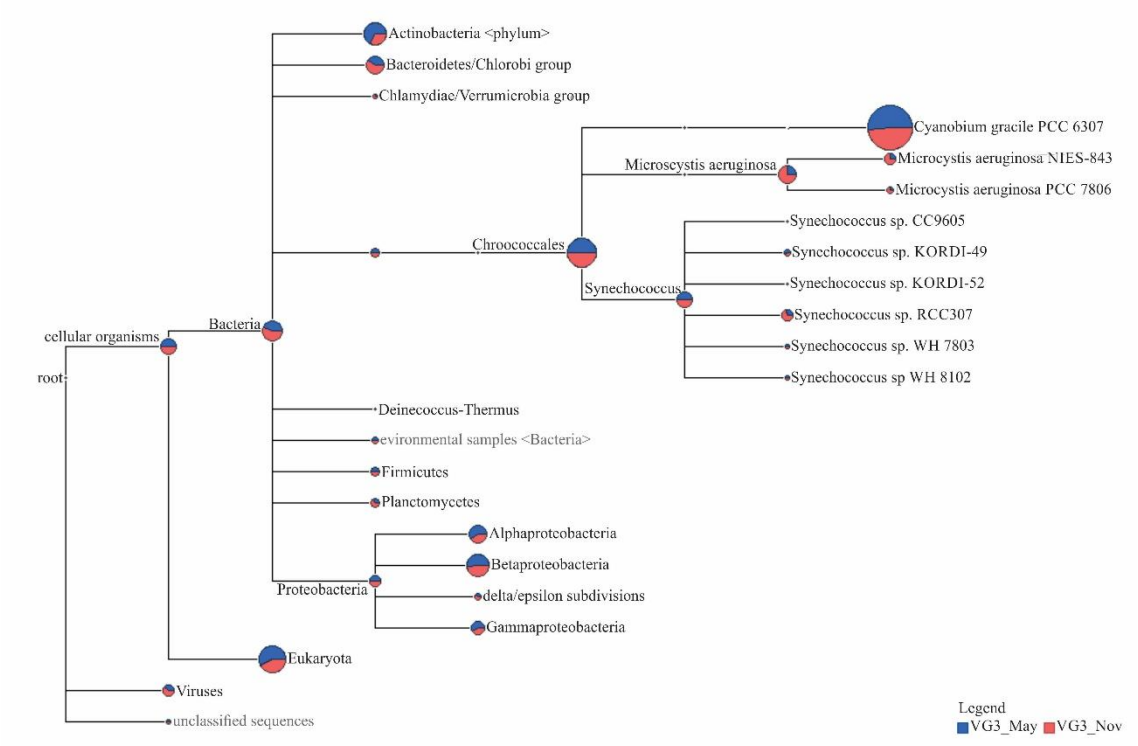


Fig. 62 - Phylogram showing the contribution of major taxa for samples VG3_May and VG3_Nov, according to Megan assignments. The samples were normalized to the sum of the smallest dataset.

Among the Cyanobacteria assignments, we could observe a predominantly pycoplanktonic community, with larger contribution of *Cyanobium gracile* and a diverse *Synechococcus* set of strains (Fig. 62). *Microcystis aeruginosa* was most abundant in the sample VG3_Nov, during the rainy season. This taxonomic profile is in accordance with our findings obtained by microscopy (see also Lopes, 2013). The great richness of picocyanobacteria morphotypes observed in lakes and the challenges concerning morphological and phylogenetic approaches hampers the accurate description of this important cyanobacterial group (Crosbie et al. 2003, Callieri et al. 2012). Since the morphological criteria usually generates unspecific annotations and the culture environment usually decharacterizes the colony morphology, we consider that *Cyanobium* and *Synechococcus* assignments were actually encompassing most of the morphotypes described under microscopy. In order to improve that taxonomic profile generated by MEGAN/NCBI approach, we performed a pick-closed OTUs pipeline after

the extraction of ribosomal sequences using Metaxa2, and according to SILVA database (for methods's details, see Chap. 4). A total of 63 operational taxonomic units (OTUs) were clustered according to RDA classifier (Caporaso et. al. 2010), VG3_May showing 40 OTUs and VG3_Nov showing 114. The biodiversity profile was in accordance to the diversity indexes generated by MEGAN, with higher diversity and evenness for sample VG3_Nov, as we can see in Figure 64. The lower Shannon and Simpson's reciprocal index for sample VG3_May seems to be related to the larger contribution of *Synechococcus* strains. Though with similar diversity pattern, the contribution of taxa was considerably distinct between the two methods.

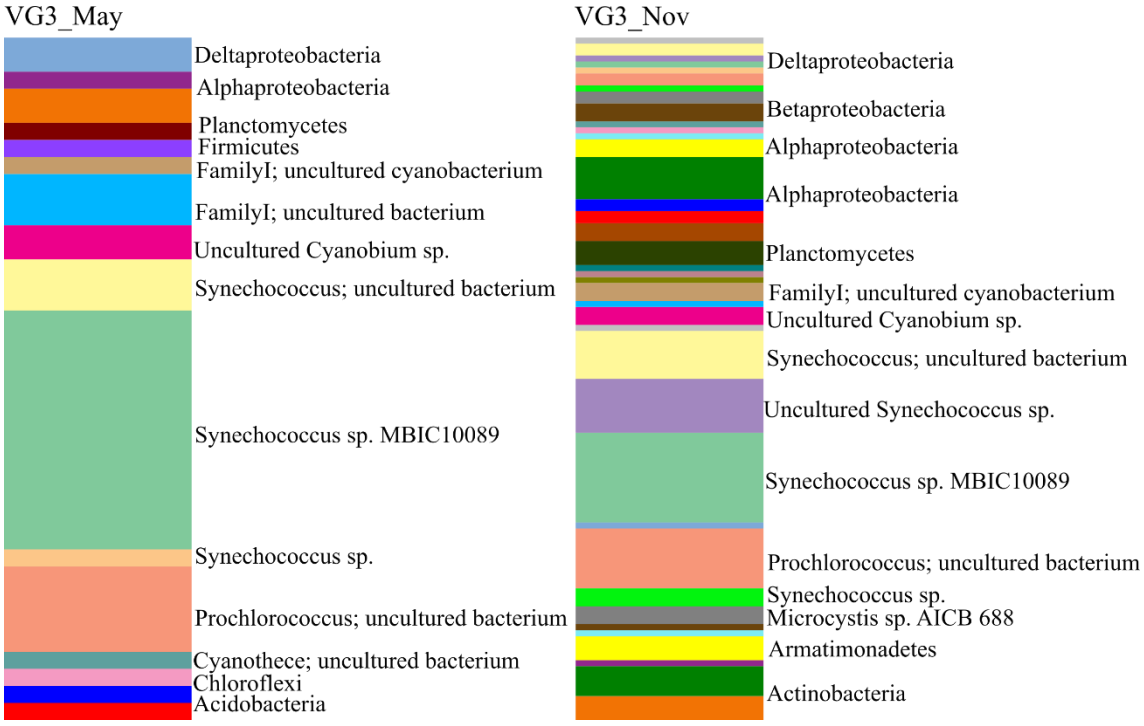


Fig. 63. Scattered plots showing the proportional taxa summary for ribosomal sequences extracted from metagenomic reads of the samples VG3_May and VG3_Nov using Metaxa2 and analyzed through QIIME pick OTUs pipeline, according to SILVA rRNA database.

The *Cyanobium* annotation was nearly absent according to SILVA rRNA database and cyanobacteria were represented mostly by the *Synechococcus* genus (Fig. 63). The large number of *Prochlorococcus* assignments could probably represent a diverse population yet uncharacterized and somewhat similar to *Prochlorococcus* ribosomal sequences. Such confusion between databases and methods is a result of a still far from

saturating natural diversity (Luo et al., 2015). Regardless of the low sequencing coverage and reference databases biases, the cyanobacteria and heterotrophic associated community showed distinct taxonomical structure between dry and rainy season, in accordance to the microscopy analysis and encompassing a particular group of cyanobacteria populations, which could be probably related to environmental conditions and seasonality.

Finally, we aligned the metagenomic fragments against the reference genome of *Cyanobium gracile* using NUCmer algorithm, MUMmer 3.23 (Kurtz et al., 2004), in order to explore the reasons behind the major contribution of this taxa according to MEGAN/LCA algorithm (Fig. 62). We could observe a very scarce coverage and similarity ranging from 80 to 100, mostly below 95% (Fig. 64), but the aligned reads were almost evenly spread over the whole *C. gracile* genome. The reason for the discrepancy between OTUs and MEGAN approaches may be related to the nature of each analysis, since ribosomal sequences (QIIME pipeline) are constrained to a more conserved genome region, while the shotgun metagenomics sequences (MEGAN/LCA algorithm) provide DNA fragments from any region, performing a genome-wide search and match. Future studies in this reservoir could reveal more details about what has been defined as *Cyanobium* by whole genome approach, and what has been called *Prochlorococcus* by ribosomal sequences strategy.

Fig. 64. Recruitment plot generated by MUMmer package showing the sequences from (a) VG3_May and (b) VG3_Nov aligned to *C. gracile* whole genome. The horizontal axis represents the genome-wide length of *C. gracile* genome while the vertical axis represents the metagenomic sequences similarity (%SIM) to *C. gracile* genome.

1. CONCLUSIONS

According to previous studies performed in the Volta Grande reservoir (Lopes, 2003), the low nutrient levels were probably the main factor that lead to the dominance of picocyanobacteria, which showed higher abundance during the rainy season. Lopes, (2013) used the molecular technique of DGGE (Denaturing Gradient Gel Electrophoresis) to study the cyanobacteria community. The author was able to detect the presence of *Synechococcus* and *Cyanobium*, showing a biodiversity profile comparable to the one we described, both by classical taxonomy and metagenomic approach.

To achieve an accurate taxonomic profile of a community dominated by picoplanktonic species, categorical and statistical criteria were necessary to define the

morphological boundaries between some species. Using a simple but robust statistical approach, we were able to describe morphologically five picocyanobacteria species *A. minutissima*, *A. clathrata*, *C. planctonicum*, *C.iac* and *L. pallida*. The large number of individuals observed (N) and measures taken (n) were essential to the robustness of the clustering, the permutation and variance analysis. The categorical variables, such as presence of cell rows, or colony morphology, along with NCD and l/w ratio measurements, revealed to be more helpful when put together than when considered separately. Therefore, the detailed observation of the complete set of attributes may be the best way to get closer to the true species richness estimation of a system and can, ultimately, provide the best phenotipical thresholds for population genetic diversity studies.

Despite the low sequencing coverage, the metagenomic strategy provided a coherent biodiversity profile, in pace with microscopy analysis and previous works performed in this reservoir, and showed that metagenomics can improve the range and the accuracy of both morphological-based and marker gene approaches. According to the brief taxonomical profile generated through MEGAN and OTU strategy we now know that highly diverse populations of picocyanobacteria are active in this reservoir, co-occurring with *Microcystis* population (Fig. 63) and heterotrophic bacteria associated. The microbial community usually display a large number of strain level annotations (Luo et a. 2015) as observed by the high intrapecific diversity of Synechococcus population. The strain-level profile can guide future investigations regarding the interaction between planktonic communities and environmental pressures. The remarkable picocyanobacteria population of the Volta Grande reservoir, detected through both morphological and molecular methods, can provide additional information and increase the knowledge on cyanobacteria and microbial communities of freshwater systems, since this reservoir, located within a chain of dams along the Rio Grande river, still presents meso-oligotrophic condition, contrasting with the vast majority of meso-eutrophic ecosystems in Brazilian territory.

1. ACKNOWLEDGEMENTS

This research was supported by a scholarship from CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) to M.L., and by funds provided by CEMIG (Companhia Eletrica de Minas Gerais) and FAPEMIG (Fundação de Apoio a Pesquisa de Minas Gerais) to A.G. We thank the staff of the Natural Sciences Museum (Zoobotanical Foundation of Rio Grande do Sul), mainly Dr. D.M. Talgatti, for valuable assistance in microscope photography and image editing.

1. REFERENCES

Apha - American Public Health Association (2005) Standard Methods for the Examination of Water and Wastewater. 21. ed. Washington dc.

Azevedo M.T.P., Sant'Anna C.L. (2003) *Sphaerocavum*, a new genus of planktic Cyanobacteria from continental water bodies in Brazil. *Algol. Stud.*, 109:79-92

Bendall M.L., Stevens S.L.R., Chan L.K., Malfatti S., Schwientek P., Tremblay J., Schackwitz W., Martin J., Pati A., Bushnell B., Froula J., Kang D., Tringe S.G., Bertilsson S., Moran M A., Shade A., Newton R. J., McMahon K.D., Malmstrom R.R. (2016) Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J*., 1-13

Bittencourt-Oliveira M.D.C., Costa Cunha M.C., Moura A.D.N. (2009) Genetic Polymorphism in Brazilian *Microcystis* spp. (Cyanobacteria) Toxic and Non-Toxic Through RFLP-PCR of the cpcBA-IGS. *Braz. Arch. Biol. Tech.*, 52: 901-909

Callieri C., Cronberg G., Stockner J.G. (2012) Freshwater picocyanobacteria: single cells, microcolonies and colonial forms. In Ecology of Cyanobacteria II, pp. 229-269, Springer Netherlands

Caporaso J.G., Kuczynski J., Stombaugh J., Bittinger K., Bushman F.D., Costello E. K., Huttley G.A. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat. Methods, 7(5): 335-336

Carvalho L.R., Pipole F., Werner V.R., Laughinghouse H.D., Camargo A.C.M., Rangel M., Konno K., Sant'anna, C.L. (2008) A toxic cyanobacterial bloom in an urban coastal lake, Rio Grande do Sul, Southern Brazil. *Braz. J. Microbiol.*, 39:761-769

Castenholz R.W., Norris T.B. (2005) Revisionary concepts of species in the Cyanobacteria and their applications. *Arch. Hydrobiol. Algol. Stud*., 117:53-69

Costa D.F., Dantas E.W. (2012) Diversity of phytoplankton community in different urban aquatic ecosystems in metropolitan João Pessoa, state of Paraíba, Brazil. *Acta Limnol. Bras*., 23(4):394-405

Cronberg G. & Komárek J. (1994) Planktic cyanoprokaryotes found in south Swedish lakes during the 12th International Symposium of Cyanophyte Research. *Arch. Hydrobiol./Algol. Stud.*, 75:323-352

Crosbie N.D., Pöckl M., Weisse T. (2003) Dispersal and phylogenetic diversity of nonmarine picocyanobacteria, inferred from 16S rRNA gene and cpcBA-intergenic spacer sequence analyses. *Appl. Envir. Microbiol*., 69(9):5716-5721

Delazari-Barroso A., Sant'anna C.L., Senna P.A.C. (2007) Phytoplankton from Duas Bocas Reservoir, Espírito Santo State, Brazil (except diatoms). *Hoehnea*, 34(2):211-229

Dojani S., Kauff F., Weber B., Büdel B. (2014) Genotypic and phenotypic diversity of cyanobacteria in biological soil crusts of the Succulent Karoo and Nama Karoo of Southern Africa. *Microb. Ecol.*, 67(2): 286-301

Domingues C.D., Torgan L.C. (2011) Fitoplâncton (exceto Chlorophyceae) de um lago artificial urbano no Sul do Brasil. *Braz. J. Bot.*, 34(3):463-480

Ferragut C., Lopes M.R.M., Bicudo D.C., Bicudo C.E.M., Vercellino I.S. (2005) Ficoflórula perifítica e

planctônica (exceto Bacillariophyceae) de um reservatório oligotrófico raso (Lago do IAG, São Paulo). *Hoehnea*, 32(2):137-184

Greco M.K.B. (2002) Balanço de massa de fósforo, evolução da eutrofização e o crescimento de macrófitas flutuantes no reservatório de Volta Grande (Minas Gerais/São Paulo). Tese de Doutoramento no Programa Ecologia, Conservação e Manejo da Vida Silvestre - ECMVS-UFMG. Belo Horizonte, março de 2002, 165 p.

Guiry M.D, Guiry G.M. (2015) AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. url: http://www.algaebase.org; searched on 27 July 2015.

Hasler P., Poulícková A. (2002) Planktic Cyanobacteria of the central and northern Moravia. *Czech Phycol.*, 2:25-32

Hickel B. (1981) *Cyanodictyon reticulatum* (Lemm.) Geitler (Cyanophyta) a rare planktonic blue-green alga refound in eutrophic lakes. *Algol. Stud.*, 27:111-118

Hindák F. (2006) Three planktonic cyanophytes producing water blooms in Western Slovakia. *Czech Phycol.*, 6:59-67

Hoffmann L., Komárek J., Kastovsky J. (2005) System of cyanoprokaryotes (cyanobacteria) – state in 2004. *Algol. Stud.* 117: 95-115.

Joosten A.M.T. (2006) Flora of the Blue-green Algae of the Netherlands I: The non-filamentous species of inland waters. KNNV Publishing, The Netherlands.

Komárek J. (2006) Cyanobacterial Taxonomy: Current Problems and Prospects for the Integration of Traditional and Molecular Approaches. *Algae*, 21

Komárek J. (2011) Introduction to the 18(th) IAC Symposium in Ceske Budejovice 2010, Czech Republic Some current problems of modern cyanobacterial taxonomy. *Fottea*, 11:1-7

Komárek J., Anagnostidis K. (1998) Cyanoprokaryota I. – In: Ettl H., Gärtner G., Heynig H., Mollenhauer D. (eds): Süßwasserflora von Mitteleuropa 19(1): 1-548, Gustav Fischer Verlag, Stuttgart-Jena

Komárek J., Kastovsky J., Jezberova J. (2011) Phylogenetic and taxonomic delimitation of the cyanobacterial genus *Aphanothece* and description of *Anathece* gen. nov. *Eur. J. Phycol.*, 46:315-326

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

Komárek J., Komarkova-Legnerova J., Sant'anna C.L., Azevedo M.T.D., Senna P.A.C. (2002) Two common *Microcystis* species (Chroococcales, Cyanobacteria) from tropical America, including *M. panniformis* sp nov. *Cryptogamie Algol.*, 23:159-177

Komárková J., Jezberova J., Komárek O., Zapomelova, E. (2010) Variability of *Chroococcus* (Cyanobacteria) morphospecies with regard to phylogenetic relationships. *Hydrobiologia*, 639:69-83

Komárková-Legnerová J., Cronberg G. (1994) Planktic blue-green algae from lakes in South Scania, Sweden. Part I. Chroococcales. *Algol. Stud.*, 72:13-51

Kumari N., Srivastava A.K., Bhargava P., Rai L.C. (2009) Molecular approaches towards assessment of cyanobacterial biodiversity. *Afr. J. Biotech.*, 8:4284-4298

Lopes A.M.M.B. (2013) Composição da comunidade de cianobactérias e outros grupos microbianos em dos reservatórios tropicais. Tese defendida junto ao Programa de Pós-Graduação em Ecologia, Conservação e Manejo da Vida Silvestre. 134 p.

Luo C., Knight R., Siljander H., Knip M., Xavier R.J., Gevers D. (2015) Constrains identifies microbial strains in metagenomic datasets. *Nat. Biotechnol.*, 33(10):1045-1052

Martins M.D., Branco L.H.Z., Werner V.R. (2012) Cyanobacteria from coastal lagoons of Southern Brazil: coccoid organisms. *Braz. J. Bot.*, 35(1):31-48

McGregor G., Fabbro L.D., Lobegeiger J.S. (2007) Freshwater planktic Chroococcales (Cyanoprokaryota) from North-Eastern Astralia: a morphological evaluation. *Nova Hedwigia*, 84:299-331

Meyer B. (1994) A new species of *Cyanodictyon* (Cyanophyceae, Chroorocalles) planktic in eutrophic lakes. *Algol. Stud.*, 75:183-188

Nguyen V.L.A., Tanabe Y., Matsuura H., Kaya K., Watanabe M.M. (2012) Morphological, biochemical and phylogenetic assessments of water-bloom-forming tropical morphospecies of *Microcystis* (Chroococcales, Cyanobacteria). *Phycol. Res.*, 60:208-222

Nogueira I.S., Gama Júnior W.A., D'alessandro E.B. (2011) Cianobactérias planctônicas de um lago artificial urbano na cidade de Goiânia, GO. *Braz. J.Bot.*, 34(4):575-592

Oh S., Caro-Quintero A., Tsementzi D., DeLeon-Rodriguez N., Luo C., Poretsky R., Konstantinidis K.T. (2011) Metagenomic Insights into Evolution, Function, and Complexity of the Planktonic Microbial Community of Lake Lanier, a Temperate Freshwater Ecosystem. *Appl. Env. Microbiol.*, 77(17):6000-6011

Oksanen F.J., Blanchet G., Kindt R., Legendre P., Minchin P.R., O'Hara R.B., Simpson G.L., Solymos P.M., Stevens H.H., Wagner H. (2015) Vegan: Community Ecology Package. R package version 2.3-0. http://CRAN.R-project.org/package=vegan

Otsuka S., Suda S., Li R., Matsumoto S., Watanabe M.M. (2000) Morphological variability of colonies of *Microcystis* morphospecies in culture. *J. Gen. Appl. Microbiol*., 46(1):39-50

Palinska K.A., Deventer B., Hariri K., Lotocka M. (2011) A taxonomic study on *Phormidium*-group (cyanobacteria) based on morphology, pigments, RAPD molecular markers and RFLP analysis of the 16S rRNA gene fragment. *Fottea*, 11:41-55

Palinska K.A., Marquardt J. (2008) Genotypic and phenotypic analysis of strains assigned to the widespread cyanobacterial morphospecies *Phormidium autumnale* (Oscillatoriales). *Arch. Microbiol.*, 189:325-335

R Core Team (2015) R: A Language and Environment for Statistical Computing. - R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Rosini E.F., Sant'anna C.L., Tucci A. (2013) Cyanobacteria from fishing ponds in the Metropolitan Region of São Paulo, Brazil. *Rodriguésia*, 64(2):399-417

Sant'anna C.L., Azevedo M.T.P., Senna P.A.C., Komárek J., Komárková J. (2004) Planktic Cyanobacteria from São Paulo State, Brazil: Chroococcales. *Braz. J. Bot.*, 27:213-227

Sant'anna C.L., Melcher S.S., Carvalho M.C., Gemelgo M.P., Azevedo M.T.P. (2007) Planktic Cyanobacteria from upper Tietê basin reservoirs, SP, Brazil. *Braz. J. Bot.*, 30:1-17

Serizawa H., Amemiya T., Enomoto T., Rossberg A.G., Itoh K. (2008) Mathematical modeling of colony

formation in algal blooms: phenotypic plasticity in cyanobacteria. *Ecol. Res.*, 23:841-850

Strunecký O., Komárek J., Johansen J., Lukesová A., Elster J. (2013) Molecular and morphological criteria for revision of the genus *Microcoleus* (Oscillatoriales, Cyanobacteria). *J. Phycol.*, 49(6):1167-1180

Thomazeau S., Houdan-Fourmont A., Coute A., Duval C., Couloux A., Rousseau F., Bernard C. (2010) The contribution of sub-Saharan African strains to the phylogeny of cyanobacteria: focusing on the nostocaceae (Nostocales, Cyanobacteria). *J. Phycol.*, 46:564-579

Tucci A., Sant'anna C.L., Gentil R.C., Azevedo M.T.P. (2006) Fitoplâncton do Lago das Garças, São Paulo, Brasil: um reservatório urbano eutrófico. *Hoehnea*, 33(2):147-175

Wehr J.D., Sheath R.G. (2003) Freshwater Algae of North America: Ecology and Classification. Academic Press, USA.

Werner V. R. (2002) Cyanophyceae/Cyanobacteria no sistema de lagoas e lagunas da planície costeira do estado do Rio Grande do Sul, Brasil – Tese defendida no Instituto de Biociências da Universidade Estadual Paulista, campus de Rio Claro. 402 p.

Yamamoto Y., Nakahara H. (2009) Seasonal variations in the morphology of bloom-forming cyanobacteria in a eutrophic pond. *Limnology*, 10:185-193

Capítulo 2: Desvendando a diversidade em um reservatório de água doce tropical: Cianobactérias e comunidade microbiana associada

2. RESUMO

Cianobactérias e comunidade microbiana associada de um reservatório de água doce tropical, impactado por florações de *Microcystis aeruginosa*, foram analisadas em escala temporal (quatro amostras ao longo de um ano seguindo o aumento progressivo de biomassa de cianobactérias no sistema), e espacial (amostragem em quatro locais diferentes, no período de floração). As variáveis ambientais apresentaram um gradiente que reflete as mudanças relacionadas ao evento de floração. Os perfis de biodiversidade e funcional foram analisados através da abordagem metagenômica, utilizando a técnica "*Whole Genome Shotgun"* (WGS). Foi observada a diminuição na diversidade à nível de classes, junto com a diminuição na equitabilidade, especialmente envolvendo Actinobacteria, Planctomicetia e Sphingobacteria ao longo do aumento de biomassa de cianobactérias. *Microcystis* apresentou um crescente número de cepas, variando de cinco para 15, acompanhando o aumento de biomassa. Apesar do perfil funcional ter sido estável entre as amostras, a distribuição de grupos ortólogos e módulos de funções relacionadas a processamento de informação ambiental, principalmente tranporte de membrana e transdução de sinal, foram mais representativos entre as anotações raras, ou seja, exclusivas a uma amostra. Uma vez que a comunidade não foi dominada por um único filo, as vias metabólicas observadas são provavelmente compartilhadas entre os membros da comunidade, e a sobreposição de atributos funcionais pode ser relacionada à complementaridade de funções entre populações. Nossos resultados demonstraram que o progressivo aumento de biomassa, associado a florações de cianobactérias, foi relacionado à dinâmica de nutrientes no sistema e que o perfil de biodiversidade da comunidade microbiana foi correlacionado a esse grandiente ambiental. A análise metagenômica permitiu a integração de dados ambientais à comunidade microbiana, uma vez que os fragmentos de DNA gerados pela técnica WGS representam potencialmente toda a comunidade local. O perfil funcional conservado entre todas as amostras sugere uma intricada rede de interações entre os membros da comunidade. Desta forma, futuras

investigações no reservatório de Marimbondo têm o potencial de desvendar interações entre ambiente, comunidade, populações e espécies.

Palavras-chave: Cianobactéria, metagenômica, comunidade microbiana de água doce, perfil taxonômico, perfil funcional, MEGAN, microdiversidade.

Unveiling diversity in a tropical freshwater reservoir: Cyanobacteria and associated microbial community.

2. ABSTRACT

The cyanobacteria and associated microbial community of a tropical water reservoir, impacted by Microcystis aeruginosa blooms, were analyzed over a temporal (four samples over one year following the progression of a bloom event), and spatial scale (four different sites in a period of high Microcystis biomass). The environmental variables showed a gradient ranging from clear water to turbid/high nutrients condition. The biodiversity and functional profile were analyzed through WGS metagenomic technique. A decrease in class diversity was observed, along with a decrease in evenness, mostly for the Actinobacteria, Planctomycetia and Sphingobacteria classes. The progressive increase in Cyanobacteria biomass probably promoted both the marked drop in richness and the decrease of contribution of remaining populations. Microcystis aeruginosa presented a crescent number of different strains, namely from five to fifteen, along with the increase in biomass. Despite the stable functional profile among samples, the distribution of modules and ortholog groups concerning functions related to environmental information processes, mostly membrane transport and signal transduction, were more representative within rare assignments, i.e., exclusive to one sample. Since the community was not dominated by a single phylum, the major pathways, modules and ortholog groups observed are likely shared between community members, and the overlap of functional assignments could be related to functional complementarity within populations. Our results showed that the progressive increase in cyanobacteria biomass, related to bloom establishment, was correlated to nutrient dynamics in the system, and the changes in microbial community profile were also correlated to such environmental gradient. The metagenomic approach allowed the integration of environmental and molecular data, since the reads yield by the WGS technique represent potentially the whole local community. The conserved functional profile suggests an intricate network of interactions among community members. Therefore, future investigations in the Marimbondo reservoir could yield additional information about the complex set of interactions observed between environment, community, populations and strains, providing more

accurate and fundamental knowledge, valuable for water management projects.

Keywords: Cyanobacteria, metagenomics, freshwater microbial community, taxonomical profile, functional profile, MEGAN, microdiversity.

2. INTRODUCTION

The most evident ecological impact of cyanobacteria is the potential for rapid increase in biomass and subsequent bloom settlement displayed by several different populations (Paerl & Paul, 2012; Paerl & Otten, 2013). Along with high growth rate, the ability for secondary metabolites production has recently been related to population structure (Hu et al., 2015). Cyanobacterial blooms usually contain one dominant population and a diverse mix of other photosynthetic and non-photosynthetic microorganisms engaged in a complex interactive network, from biodiversity dynamics to nutrient and energy cycling (Litchman et al., 2010; Zarraonaindia et al., 2013; Pimentel & Giani, 2013; Penn et al., 2014). *Microcystis* populations usually present a high level of genetic diversity, plasticity and microcystin producers and non-producers are normally found together (Wilson et al., 2005; Hu et al., 2015; Briand et al., 2015).

Metagenomics provides additional tools for understanding the biodiversity, ecology and metabolic features of a community, through direct access to the DNA content of the environment and, having enough coverage, it should also contain representatives from all predominant ecotypes (Rodriguez-Valera et al., 2009; Escalas et al., 2013). The impact of cyanobacteria in Brazilian water reservoirs has been largely explored (Sant'Anna et al., 2008; Piccin-Santos & Bittencourt-Oliveira, 2012; Wojciechowski & Padial, 2015; Palacio et al., 2015), but molecular approaches regarding the whole planktonic community are still rare. In this work we followed the progression of a bloom over a temporal and spatial scale in a tropical meso-eutrophic water reservoir. Through a metagenomic approach we sought for distinctness and changes in the taxonomical composition and functional contribution of cyanobacteria and associated microbial community along the bloom progression. Several works have suggested that companion heterotrophs may be essential for cyanobacteria biomass growth (Saxton et al., 2011; Shen et al., 2011; Steffen et al., 2012) and the local bacterial community may play an important role in the dominance of different phytoplankton species, since there is a strong correlation between population structure, environmental pressures and niche complementation (Bagatini et al., 2014; Cordero & Polz, 2014; Louati et al., 2015).

Freshwater ecosystems display rapid fluctuations in their physico-chemical and biological characteristics, an unlikely scenario to the differentiation of ecotypes, which

requires a stable niche (Humbert et al., 2013). This unstable environment ultimately prioritizes strain level diversity, genetic exchanges, metabolic trade-offs and social interactions (Moore et al., 1998; Coleman et al., 2006; Van Elsas & Bailey, 2002; Ulrich et al., 2005; Galhardo et al., 2007; Cordero et al., 2012). Previous studies on the functional profile of cyanobacteria communities have already tried to characterize and understand these processes, while discovering new genes, proteins and functions (Beck et al., 2012; Steffen et al., 2012; Knoop et al., 2013). There has also been an emphasis on the highly plastic feature of the *M. aeruginosa* genome (Kaneko et al., 2007) and the impact of nutrient dynamics on population stability through differential gene expression (Steffen et al., 2014). Therefore, we chose to explore the taxonomical and functional profile of planktonic cyanobacteria and the associated microbial community from a tropical meso-eutrophic freshwater ecosystem (Marimbondo reservoir, Brazil) and the various aspects of the interaction between the microbial community and the environmental gradient throughout the bloom progress over a sampling year. In spite of some constraints regarding sequencing attributes, we expected to provide additional information about complex microbial freshwater dynamics, through an exploratory metagenomic analysis, spanning from broad to fine scale.

2. MATERIAL AND METHODS

Sampling

Sampling was carried out in the pelagic zone of the Marimbondo reservoir, in Southeastern Brazil (Fig. 1). The reservoir receives the water from the Rio Grande river, has a flooded area of 438 km^2 and and is part of the Furnas Hydro Electrical Power System. The samples were taken using a Van Dorn sampler, over an annual period, covering the progress of a bloom event: November 2006 (T1), May 2007 (T3), August 2007 (ST4) and November 2007 (T5). On August 2007, three extra samples were taken in different sampling stations, referred as spatial sampling (S): ST4, S6, S7 and S8. Therefore, sample ST4 is the intersection between temporal and spatial scale. The spatial samples were taken during high cyanobacteria biomass conditions, foreknowing the metadata, such as microcystin production, aiming to sample relatively continuous environments. Due to the low sequencing coverage of reads, the sample T2 (February

2006) was excluded from this work.

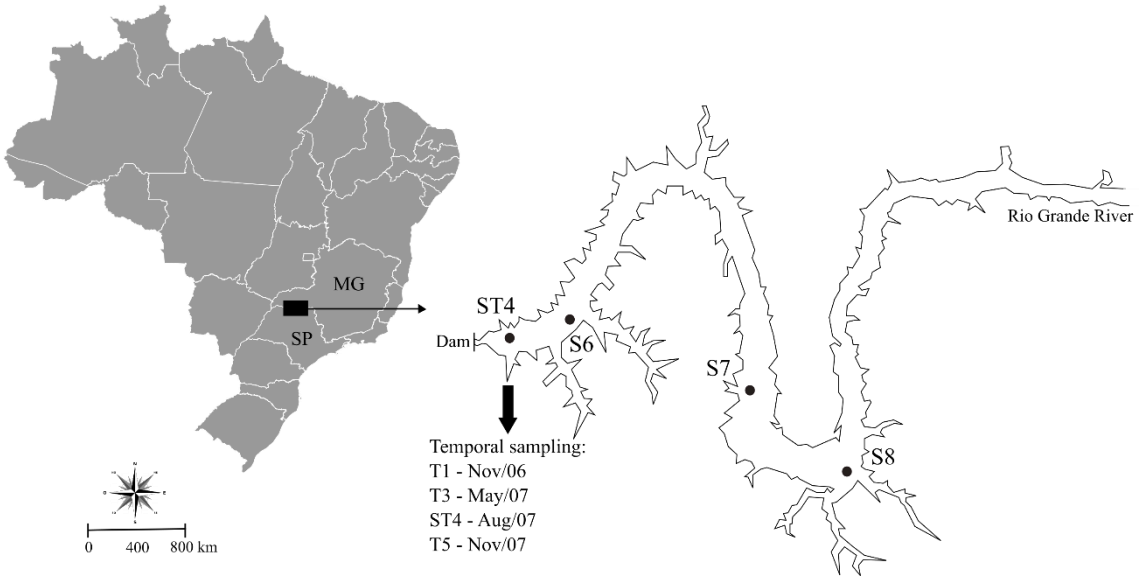


Fig. 1 - Marimbonfo reservoir, located between Minas Gerais and São Paulo states (Brazil), in the cities Icém (SP) and Fronteira (MG). The four spatial sampling stations are indicated. The temporal sampling was carried close to the dam, from November 2006 to November 2007.

Environmental variables

Cell counting and biomass estimation were carried out on an inverted microscope, using sedimentation chambers, and the phytoplankton, cyanobacteria and *Microcystis* biomass were expressed in $\mu m^3.ml^{-1}$. The physical-chemical variables analyzed were chlorophyll ($\mu g.L^{-1}$), organic suspended solids OSS ($mg.L^{-1}$), transparency (m), pH, dissolved oxygen DO ($mg.L^{-1}$), conductivity ($\mu S.cm^{-1}$), total phosphorus TP ($\mu g.L^{-1}$), orthophosphate ($\mu g.L^{-1}$), nitrate ($\mu g.L^{-1}$) total Kjeldahl nitrogen TKN ($mg.L^{-1}$). Nutrient analyses followed protocols described in APHA, (2005). Microcystin analyses ($\mu g.L^{-1}$) were performed by Enzyme-linked immuno-assay (ELISA), using a specific kit (Abraxis, ADDA Elisa) for the congener-independent detection of microcystins, following the manufacturer instructions. The complete set of environmental variables is presented in Supplementary Table S1.

Statistical analysis

Results obtained for the physical chemical variables were statistically analyzed through principal component analysis (PCA) using the “rda” function from Vegan package (Oksanen et al., 2015) in R programming language (R Core Team 2015). The "ordistep" function (Vegan package) and "corrplot" function (Corrplot package) guided the selection of variables to be included in the model in order to prevent multicollinearity and noise. The generalized linear model was constructed through the scores from the first axis, *Microcystis* biomass and microcystin concentration.

Sequencing and Pre-Processing of Sequence Datasets

The genomic DNA was extracted according to phenol-chlorophorm method (Kurmayer et al., 2003), then used for the library preparation and sequenced using a Roche 454 GS FLX platform. The raw sequences were cleaned with seq_crumbs software (http://bioinf.comav.upv.es/seq_crumbs/), trim_quality tool, under default parameters.

The metagenomic approach consists in the analysis of the genomic DNA from a whole community and can be performed in environmental single-gene surveys or random shotgun studies of all environmental genes (Fig. 2). The Whole Genome Shotgun (WGS) strategy is based on sequencing of the total DNA extracted from a sample, providing a genomic profile of all organisms within the community, constrained by sequencing technique and resulting sequencing coverage. The cleaning step, before the analysis of the metagenomics datasets, is therefore the crucial step towards reliable taxonomic and functional profiles (Gilbert & Dupont, 2011). The trim_quality tool removes regions of low quality (Phred score threshold 25) in the edges of the fragments using a sliding window (5bp).

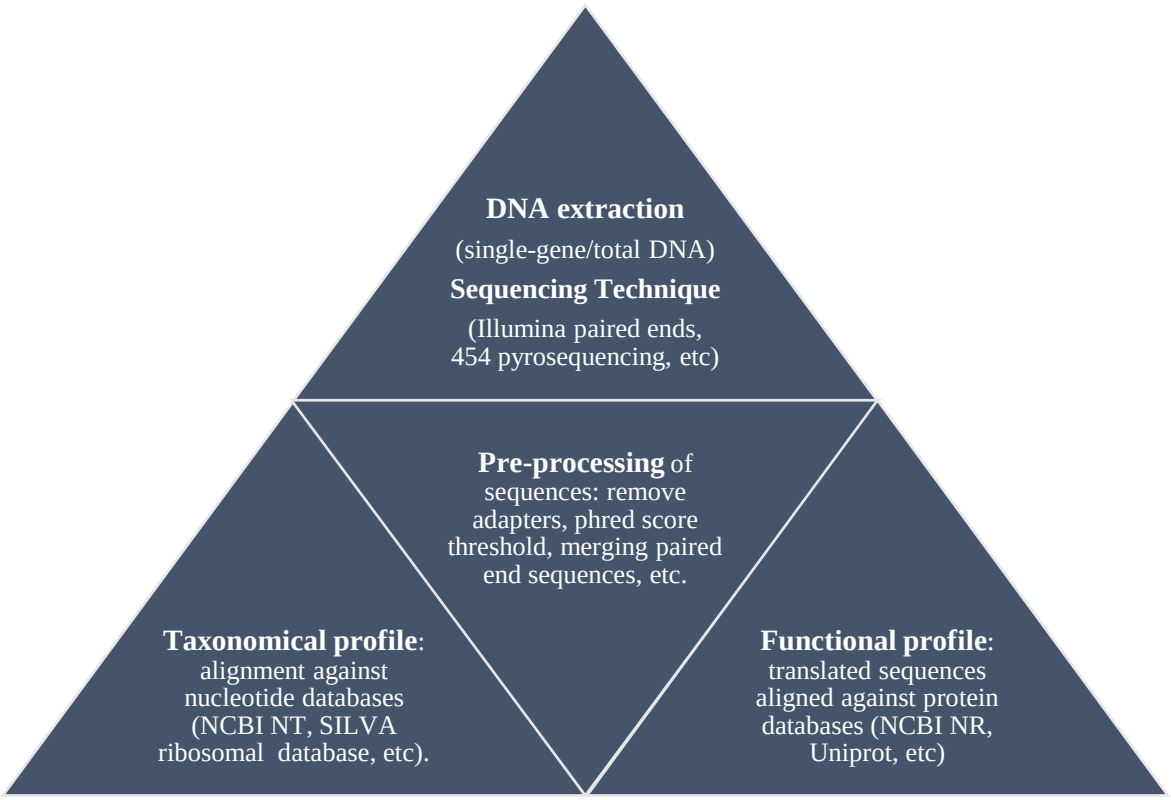


Fig. 2 - Fundamental steps constituting a metagenomics pipeline. The DNA extraction can be divided in target genes or total DNA and the extraction method should be carefully chosen for the type of material and sequencing requirements. The pre-processing of sequences is the crucial step, trimming adapters and low quality bases or stretches, which are a result of sequencing step and depends on sequencing platform adopted. The following steps are directly related to the study goals and targets. The genomic taxonomical profile classifies sequences according to metagenomics sequences (reads) similarity and alignment scores against nucleotide reference sequences and genomes. The functional profile aligns translated reads against protein databases, considering the six reading frames (sliding over each consecutive nucleotide: if a triplet, three successive nucleotides, constitutes an amino acid or a stop signaling, it is called a codon).

Biodiversity and functional profile using MEGAN

The cleaned sequences (Suppl. Table S2) were queried in the NCBI non-redundant protein sequences (nr) database using BLASTx and visualized through MEGAN Metagenomic Analyzer 5.10.7. The MEGAN procedure builds the taxonomic classification from the NCBI taxonomy database and applies the lowest common ancestor (LCA) algorithm (Fig. 3), which assigns each read to the lowest node that encompasses the set of taxa hits for that read, according to user defined parameters (Huson et al., 2011). To perform the functional analysis, MEGAN assigns each read to the functional role of the highest scoring gene in a BLASTx comparison against a protein database (Mitra et

al., 2011). The Marimbondo taxonomical and functional profile of the temporal and spatial sampling datasets were assigned under bit score threshold of 60, top percent of 10, min percent support of 0.5 and LCA percent of 70. The less specifically a read hits a taxa, the rootmost in the taxonomy it is placed (Huson et al., 2011; Huson & Mitra, 2012).

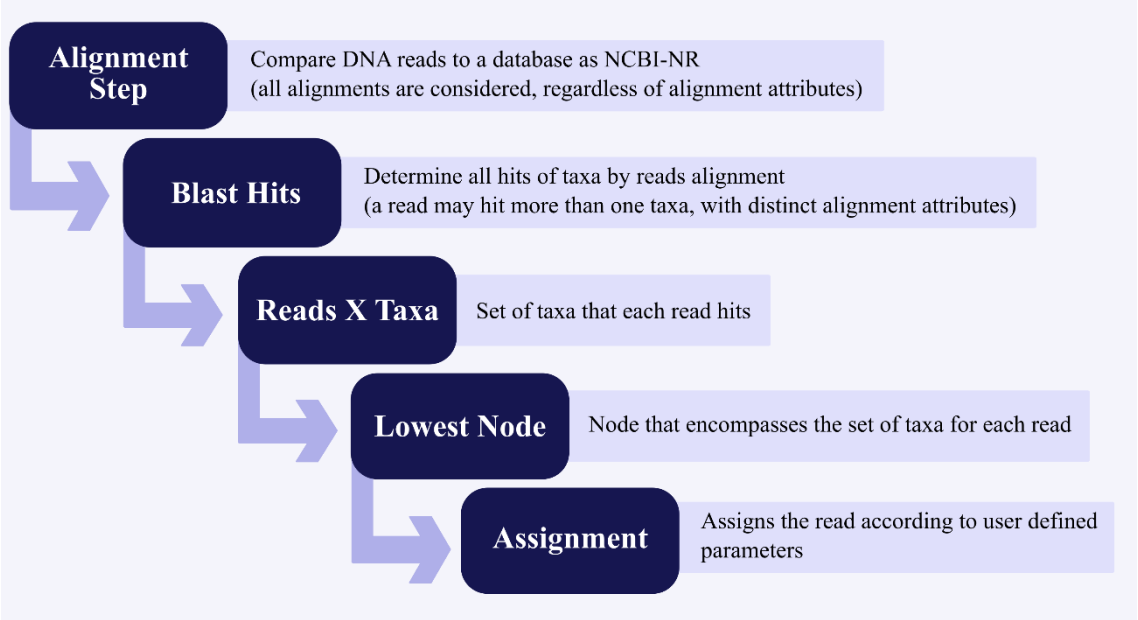


Fig. 3 - Lowest Common Ancestor (LCA) algorithm and taxonomic binning according to user defined parameters: minimum threshold for the bit score of hits, maximum threshold for the expected *E-value* of hits, maximum difference between the score of a hit and the highest hit score for that taxa, number of reads that must be assigned to each taxa and the percent of matches that the LCA of a read must cover. If a read doesn't have the alignment requirements as bit score, *E-value*, difference from the highest score and LCA percent of matches, it is placed one node back. If a taxa doesn't have the support requirements as number of reads assigned, it is not considered.

The samples were compared using square root normalization, as recommended by MEGAN manual, in order to minimize the size of sample bias. The "biome algorithm" was used to set the frequency distribution patterns for taxa, functions and pathways. The core biome assignments determine taxa and functions that appear in the majority of the samples. The shared biome determines the intersection of the taxonomic and functional content among the samples. The rare biome represents the taxa and functions unique to one sample.

The KEGG Analyzer tool was used to explore the functional content of all samples, regardless of sampling criteria. From each pathway, modules and ortholog

groups with higher contribution and frequency in all samples were described. All data are presented as relative contribution of the MEGAN assignments in order to avoid distortions relative to cell density and DNA concentration in each sample. The proportions were generated for each taxonomical rank or functional category considered, according to the comparison being performed, *i.e.* taxa assignments under cyanobacteria were divided by total cyanobacteria assignments and the sum of assignments for each functional category was divided by the total KEGG assignments. Assignments with number of hits above the average for the rank or functional category considered were named abundant. Modules with all ortholog groups present in the dataset were named complete, otherwise incomplete. Modules in which most ortholog groups were present in at least six samples were considered evenly distributed, otherwise fragmented. In order to explore the correlations between environmental variables and the class assignments generated by MEGAN, we performed a redundancy analysis (RDA).

MG-RAST Plataform

Additionally, the samples were uploaded and analyzed through MG-RAST platform, which offers automated quality control, annotation, comparative analysis and archiving services (Meyer et al., 2008). The metagenomes were stored under the following identifiers: 4636021.3 (T1), 4636023.3 (T3), 4636025.3 (ST4), 4636027.3 (T5), 4636015.3 (S6), 4636017.3 (S7) and 4636019.3 (S8). The data was compared to Kegg Orthology (KO) database using maximum e-value of 1e-5, minimum identity of 70%, minimum alignment length of 30 (measured in aa for protein and bp for RNA databases) and the data were normalized to values between 0 and 1.

2. RESULTS AND DISCUSSION

Environmental Features

The temporal and spatial samples showed notable patterns towards environmental variables, and an environmental gradient was evident on temporal scale. The first two axes of the temporal sampling PCA explained 88% of the variance (Tab. 1a, Fig. 4a), while the spatial sampling PCA explained 95.4% of the variance (Tab. 1b, Fig. 4b). Nitrate, chlorophyll (Chlor) and total phosphorus (TP) were the most important variables in the model ($p<0.01$), according to permutation tests (ordistep function, R Vegan package). In order to balance the variance explained among all seven samples, we also included organic suspended solids (OSS), inversely correlated to nitrate; transparency (Transp), inversely correlated to nitrate, chlor, TP and OSS; total Kjeldahl nitrogen (TKN), inversely correlated to conductivity, and Microcystin concentration (Micr), correlated to Chlor. According to cell counts under inverted microscope and biomass estimation, the temporal sampling showed an increase in biomass for total phytoplankton, total cyanobacteria and total *Microcystis* along with microcystin concentration (Fig. 5) from November 2006 (T1) to November 2007 (T5). The spatial sampling showed a lower microcystin concentration in the lentic sample ST4, near the dam, and in the nearest sample to the lotic region of the reservoir S8.

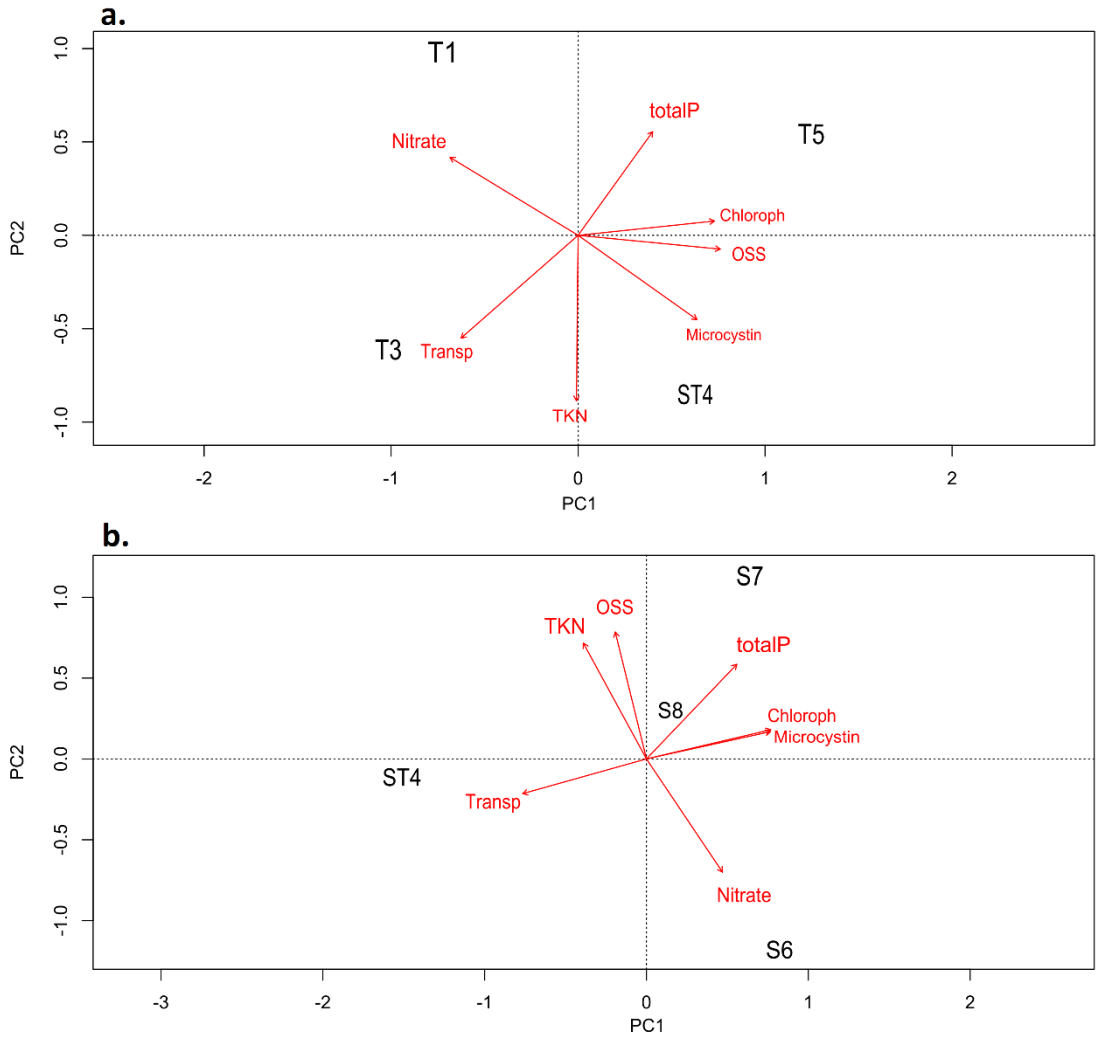


Fig. 4 - PCA ordination diagram of physical and chemical variables for temporal sampling dataset (a) and spatial sampling dataset (b). The environmental variables were logaritmized, scaled to unit variance (like correlations) and the ordination was displayed with symmetrical scaling between site and variables scores.

Tab. 1 - Summary of the Principal Component Analysis (PCA) of the temporal sampling dataset (a) and spatial sampling dataset (b) regarding the first (PC1) and second (PC2) components.

a. **Temporal sampling**	**PC1**	**PC2**
Eigenvalue	4.1046	2.0338
Proportion Explained	0.5864	0.2905
Cumulative Proportion	0.5864	0.8769
b. **Spatial sampling**	**PC1**	**PC2**
Eigenvalue	3.9431	2.735
Proportion Explained	0.5633	0.3907
Cumulative Proportion	0.5633	0.954

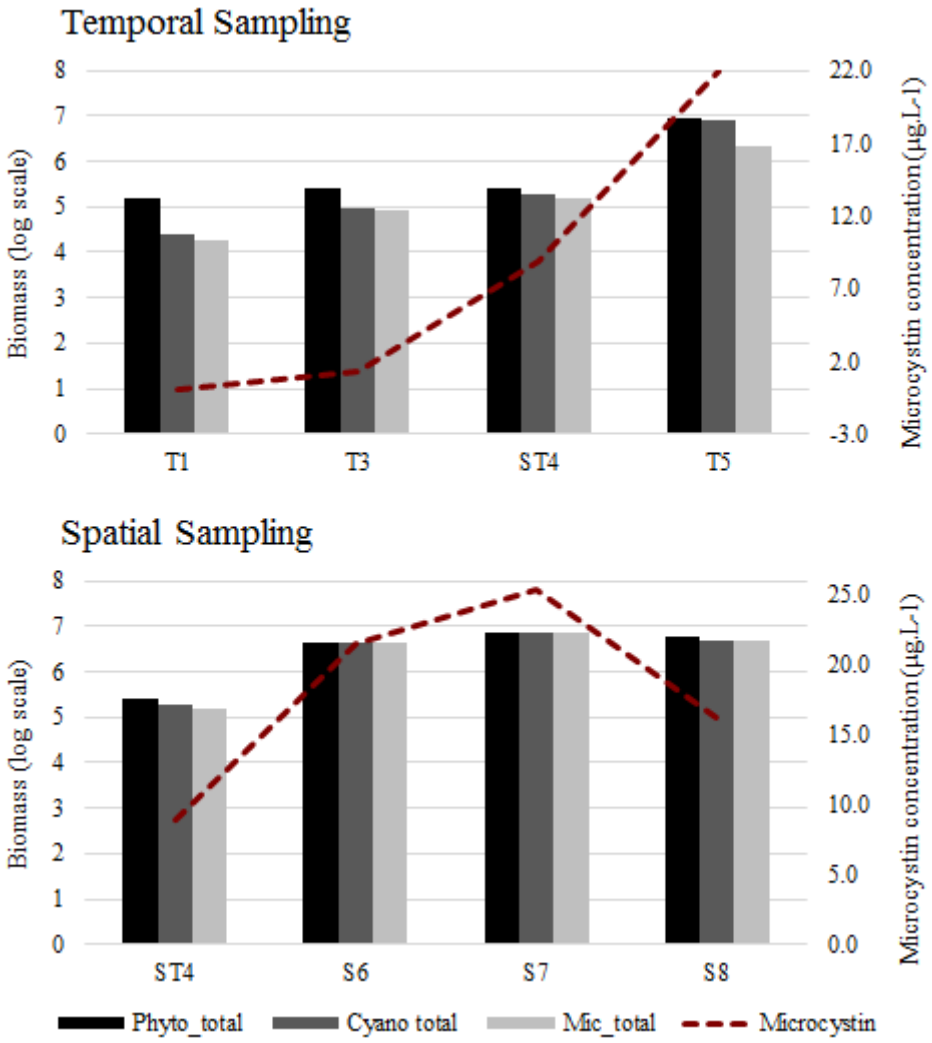


Fig. 5 - Biomass of total phytoplankton (Phyto_total), total cyanobacteria (Cyano total) and total *Microcystis* (Mic_total) estimated by counting under inverted microscope (presented in log scale) and microcystin concentration (dashed line, secondary y axis) for temporal (T1, T3, ST4 and T5) and spatial (ST4, S6, S7, S8) samplings.

Concerning the temporal sampling, the samples T1 and T3 were negatively correlated to the first axis, correlated to nitrate and transparency. Sample T5, with higher cyanobacteria biomass, was correlated to total phosphorus (TP) and Chlorophyll (Chloroph) (Fig. 4a). Sample T1 was positively correlated to nitrate, in opposite orientation to TKN. We observed a gradient from low biomass samples T1/T3 to the last temporal sample T5. The increase in Chlor and TP seems to be related to the increase in OSS, which increases turbidity (Haukka et al. 2006) and consequently decreases transparency (Fig. 4a). This same pattern of high nitrate under higher transparency/lower

production of Chlor and high TP/Chlor and TKN associated to high OSS was observed in a study using an ordination approach to evaluate water quality from 39 sites along the southern shore of Lake Ontario (Makarewicz & Lewis, 2015). The negative correlation between *Microcystis* biomass/microcystin concentration and nitrate has already been observed (Von Rückert & Giani, 2004; Xie et al., 2012; Steffen et al., 2014).

Concerning the spatial sampling, the sample ST4, in the lentic region, close to the dam, was negatively correlated to the first axis, in opposite orientation to TP, Cloroph and Microcystin concentration (Fig. 4b). The sample S6, in the same reservoir region of sample ST4, was correlated to nitrate, in opposite orientation to TKN and OSS. The samples S7 and S8, close to the lotic region presented higher levels of turbidity and nutrients (Fig. 4b). A turbidity and trophic status gradient seems to be related to the region of the reservoir, which receives suspended solids and nutrients from the river in the lotic portion, but the decrease in water flow towards the lentic region allows sedimentation and complexation of nutrients.

Microbial Community Profile

The biodiversity profile of 1.3 million reads was analyzed using MEGAN 5.10.7 (Suppl. Table S2). A high amount of sequences was not assigned to any taxa, possibly because these sequences are distantly related to any known sequence deposited in public databases up to now. The majority of the assigned reads were affiliated to the Bacteria domain (88% for both temporal and spatial samplings) and the remaining sequences were distributed between the Archaea and Eukarya domains. The variation of phyla contribution and Cyanobacteria orders, genera and *Microcystis* specific contribution in each dataset are presented in Table 2.

Tab. 2 - Relative contribution of taxa for temporal and spatial samplings. Only taxa with contribution => 0.02 are presented. The proportions correspond to the ratio of summarized assignments of phyla/total Bacteria, orders/total Cyanobacteria and for *Microcystis* strains the proportions were computed according to total strain-level assignments for each sample.

	Temporal Sampling				**Spatial Sampling**			
Phyla								
	T1	**T3**	**ST4**	**T5**	**ST4**	**S6**	**S7**	**S8**
Proteobacteria	0.33	0.30	0.43	0.31	0.41	0.42	0.46	0.32
Cyanobacteria	0.12	0.17	0.16	0.31	0.18	0.31	0.33	0.49
Bacteroidetes/Chlorobi	0.13	0.10	0.12	0.15	0.12	0.09	0.04	0.04
Actinobacteria	0.12	0.15	0.07	0.06	0.07	0.04	0.07	0.07
Planctomycetes	0.09	0.09	0.08	0.04	0.08	0.04	0.04	0.04
Verrucomicrobia	0.06	0.04	0.03	0.03	0.03	0.01	0.00	0.01
Order								
	T1	**T3**	**ST4**	**T5**	**ST4**	**S6**	**S7**	**S8**
Chroococcales	0.67	0.73	0.75	0.68	0.71	0.72	0.81	0.82
Oscillatoriales	0.11	0.07	0.07	0.13	0.08	0.08	0.03	0.02
Nostocales	0.06	0.04	0.05	0.05	0.08	0.07	0.05	0.03
Prochlorales	0.00	0.03	0.00	0.00	0.01	0.01	0.00	0.00
Genera								
	T1	**T3**	**ST4**	**T5**	**ST4**	**S6**	**S7**	**S8**
Microcystis	0.30	0.31	0.56	0.51	0.51	0.57	0.60	0.64
Pseudanabaena	0.08	0.04	0.05	0.10	0.04	0.04	0.00	0.00
Synechococcus	0.21	0.24	0.08	0.06	0.08	0.06	0.10	0.08
Cyanobium	0.06	0.07	0.02	0.02	0.02	0.01	0.03	0.02
Nostoc	0.00	0.00	0.03	0.01	0.04	0.04	0.03	0.01
Cyanothece	0.00	0.00	0.02	0.01	0.03	0.03	0.02	0.01
Anabaena	0.02	0.00	0.00	0.01	0.01	0.01	0.00	0.00
Microcystis								
	T1	**T3**	**ST4**	**T5**	**ST4**	**S6**	**S7**	**S8**
***Microcystis* genus**	0.04	0.03	0.05	0.04	0.04	0.05	0.05	0.06
***M. aeruginosa* specific level**	0.17	0.15	0.22	0.20	0.19	0.21	0.22	0.25
***M. aeruginosa* sub-specific level**	0.10	0.12	0.27	0.25	0.28	0.31	0.33	0.34
***M. aeruginosa* NIES-843**	0.27	0.23	0.18	0.14	0.14	0.16	0.15	0.16
***M. aeruginosa* PCC 9443**	0.27	0.20	0.10	0.12	0.08	0.10	0.10	0.10
***M. aeruginosa* PCC 9701**	0.00	0.15	0.08	0.07	0.06	0.06	0.06	0.07
***M. aeruginosa* PCC 9717**	0.15	0.15	0.09	0.07	0.07	0.07	0.07	0.08
***M. aeruginosa* PCC 9807**	0.00	0.15	0.08	0.07	0.06	0.07	0.07	0.07
***M. aeruginosa* PCC 9809**	0.15	0.00	0.08	0.07	0.06	0.07	0.07	0.08
***M. aeruginosa* PCC 9808**	0.00	0.13	0.09	0.07	0.07	0.08	0.08	0.08
***M. aeruginosa* TAIHU98**	0.15	0.00	0.06	0.07	0.05	0.06	0.05	0.06
***M. aeruginosa* PCC 9432**	0.00	0.00	0.00	0.06	0.07	0.04	0.04	0.00
***M. aeruginosa* PCC 9806**	0.00	0.00	0.08	0.06	0.06	0.07	0.07	0.07
***M. aeruginosa* DIANCHI905**	0.00	0.00	0.06	0.05	0.05	0.03	0.05	0.04
***M. aeruginosa* PCC 7806**	0.00	0.00	0.06	0.05	0.05	0.05	0.05	0.06
***M. aeruginosa* PCC 7941**	0.00	0.00	0.00	0.05	0.06	0.04	0.04	0.00
***M. aeruginosa* SPC777**	0.00	0.00	0.05	0.05	0.04	0.03	0.05	0.05
***Microcystis* sp. T1-4**	0.00	0.05	0.04	0.04	0.03	0.03	0.03	0.04

The pattern of frequency distribution of phyla assignments for each dataset (core assignments were common to all samples, while rare assignments were exclusive to one sample) pointed to a higher proportion of assignments shared among the samples. The core assignments covered 65% for temporal sampling and 40% for spatial sampling, the shared assignments among the samples covered 74% for temporal sampling and 45% for spatial sampling and the rare assignments, representing the taxa exclusive to one sample, covered 3% for temporal sampling and 5% for spatial sampling. For all samples, Proteobacteria, Cyanobacteria, Bacteroidetes, Actinobacteria and Planctomycetes were the most frequent and presented the higher relative contribution. Several authors have found a similar profile for freshwater microbial communities, however with distinct relative abundance and classes contribution (Bagatini et al., 2014; Eiler & Bertilsson, 2004; Shi et al., 2010; Oh et al., 2011; Steffen et al., 2012; Parfenova et al., 2013; Mou et al., 2013; Penn et al., 2014).

According the Shannon-Wiener and Simpson's reciprocal diversity indexes (Tab. 3), the samples with lower and growing cyanobacteria biomass (T1, T3, ST4, T5) presented higher diversity and the classes Actinobacteria and Planctomycetia were among the most representative taxa, since its contribution decreased from the low to the high biomass sample T5 (Fig. 6). The increase in cyanobacteria assignments seems to be related to the decrease in class diversity, which caused a more heterogeneous class contribution (Fig. 6, Tab. 3).

Tab. 3 - Diversity indexes for temporal and spatial datasets, according to the contribution of Bacteria reads assigned by MEGAN to the class level.

	Temporal sampling				**Spatial sampling**			
	T1	**T3**	**ST4**	**T5**	**ST4**	**S6**	**S7**	**S8**
Shannon-Weaver index	4.12	3.96	3.76	3.61	3.74	3.31	2.81	2.47
Simpson's reciproval index	12.51	10.60	9.22	7.12	9.15	6.46	5.11	3.40

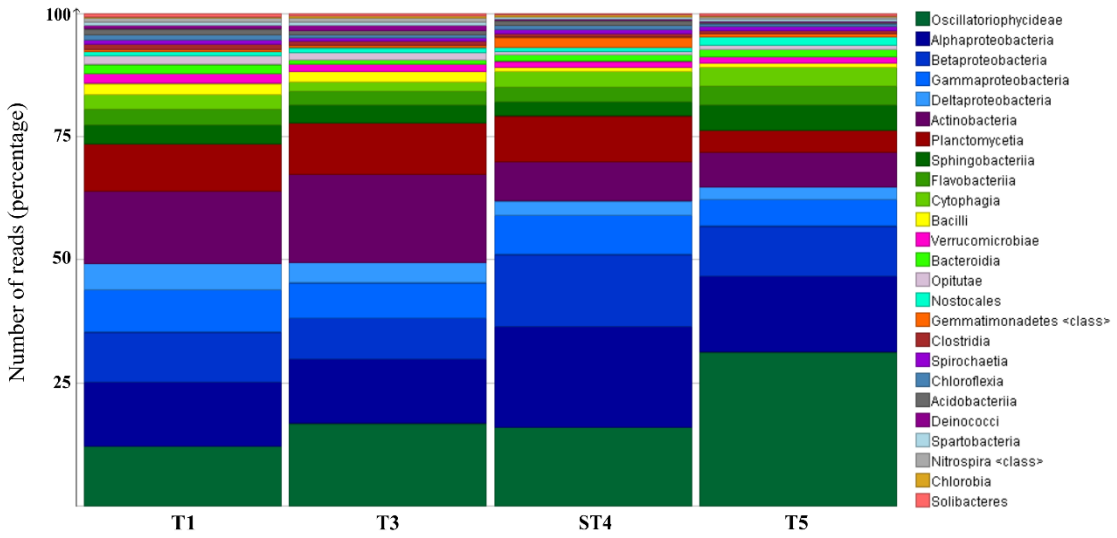


Fig. 6 Relative contribution of assignments for temporal samples at the class level, columns 100% stacked. Blue-green: Cyanobacteria; blue: Proteobacteria; green: Bacteroidetes; purple: Actinobacteria; red: Planktomycetia; Other colors: minor representative taxa.

The spatial dataset showed a stronger gradient of diversity from lentic to lotic region of the reservoir (Fig. 7). Bacteroidetes decreased in samples S7 ans S8. The Proteobacteria phylum was the main taxa for most of samples, and only in sample S8 did cyanobacteria reach 49% of total bacterial community, along with a marked drop in richness (Tab. 3). The lower richness and the absence of some major freshwater microbes, such as Gemmatimonadetes phylum, has already been observed in cyanobacteria impacted systems (Mou et al., 2013), as well as the association between proteobacteria and bloom progress (Worm & Sondergaard, 1998; Brunberg, 1999; Shi et al., 2010). Proteobacteria plays a potential crucial role in the aggregation of cells, formation of mucilage (Shen et al., 2011; Bagatini et al., 2014) and microcystin degradation (Maruyama et al., 2003; Mou et al., 2013; Zhu et al., 2014). Despite the overall stable contribution of the Proteobacteria phylum, we could observe some changes at class level assignments. Betaproteobacteria showed higher contribution in sample S7, while Alphaproteobacteria, Gammaproteobacteria and Deltaproteobacteria decreased in sample S8.

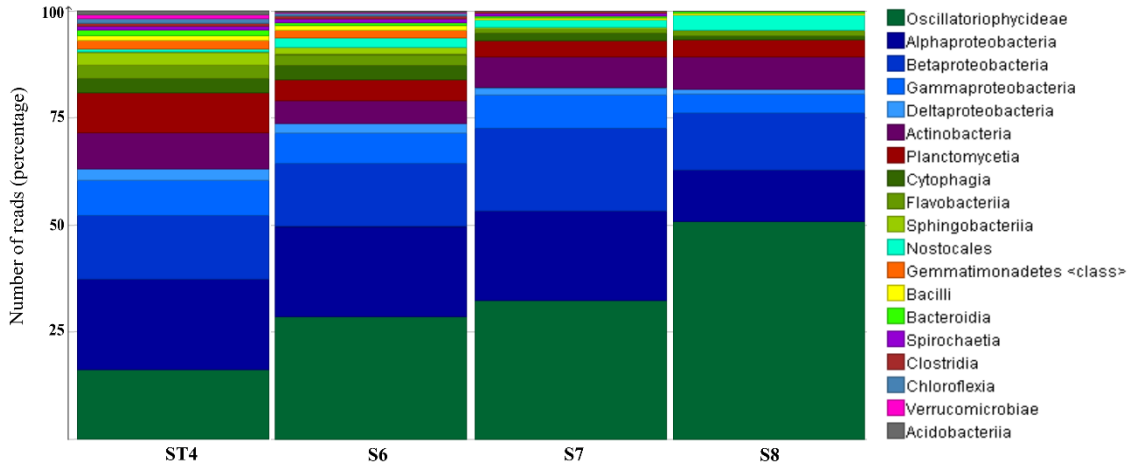


Fig. 7 Relative contribution of assignments for spatial samples at the class level, columns 100% stacked. Blue-green: Cyanobacteria; blue: Proteobacteria; green: Bacteroidetes; purple: Actinobacteria; red: Planktomycetia; Other colors: minor representative taxa.

The major taxonomic classes assigned by MEGAN correlated to environmental variables and temporal scale (Fig. 8), according to the redundancy analysis (RDA). The first two axes explained 89% of the variance (Tab. 4). The first axis clearly separated the first three temporal samples from the high cyanobacterial biomass samples. Chlorophyll and TKN presented the highest positive scores, while transparency the highest negative scores and only cyanobacteria presented strong positive scores on the first axis. The second axis separated most of the spatial samples (positive) from the temporal ones (negative), and OSS showed the highest positive score. Alphaproteobacteria and Betaproteobacteria showed a strong positive correlation with the second axis and OSS. Louati et al., (2015) found that Betaproteobacteria was more abundant in the cyanobacteria-associated bacterial community, instead of free living, and was not affected by bloom development. In the same study, the authors found higher proportions of Alphaproteobacteria under *Microcystis* bloom. The positive correlation between Betaproteobacteria and OSS could be reflecting the high efficiency of this class in degrading dissolved organic matter (Niemi et al., 2009) and its strong interaction with the cyanobacterial surrounding (Louati et al., 2015). The order Burkholderiales, known to be important in microcystin degradation (Mou et al., 2013), presented the highest contribution within the Betaproteobacteria class. Cyanobacteria (Oscillatoriophycideae and Nostocales) was positively correlated to the first axis and samples T5 and S8. Actinobacteria, Planktomycetia, Deltaproteobacteria and Bacteroidia correlated

positively to samples T1 and T3. The opposite orientation among Actinobacteria vector and those indicating higher levels of eutrophication reflects the inverse correlation between this class and cyanobacterial bloom conditions (Haukka et al., 2006; Debroas et al., 2009). Actinobacteria and Bacteroidia have been found among the most abundant taxa in several systems (Debroas et al., 2009; Humbert et al., 2009; Shen et al., 2011) and, according to Louati et al., (2015), Actinobacteria was found in the free living fraction of cyanobacteria cultures, but was absent in the cyanobacteria-associated fraction. The biodiversity profile seems to be conserved among freshwater systems and the differences could be interpreted as the outcome of the differing histories of these ecosystems and regional or continental selection pressures (Humbert et al., 2009), like an increased level of eutrophication (Haukka et al., 2006). The scores of the first axis correlated to *Microcystis* biomass (p=0.0149, F-value=13.25) and microcystin concentration (p=0.0207, F-value=11.091) (Dispersion parameter in agreement with Gaussian distribution).

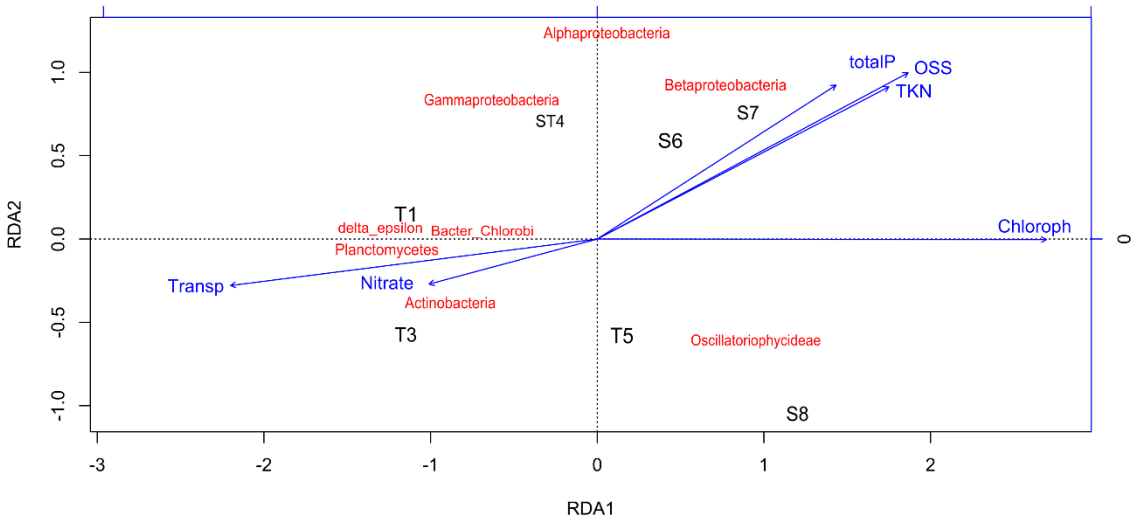


Fig. 8 - Ordination diagram of RDA between environmental variables and MEGAN taxonomic assignments under class level. The class assignments were transformed to relative contribution and environmental variables were logaritmized. Variables scaled to unit variance and ordination diagram displayed in symmetrically according to site and species scores.

Tab. 4 - Summary of the RDA analysis and scores for first and second axis.

	RDA1	RDA2
Eigenvalue	5.7099	2.3015
ProportionExplained	0.6344	0.2557
CumulativeProportion	0.6344	0.8901

Cyanobacteria Profile

Figure 9 and 10 represents the phylogram of the Bacteria super kingdom and Cyanobacteria phylum for temporal and spatial datasets, respectively. The order Chlorococcales was the most important and reached 68% of cyanobacteria assignments following the development of the bloom in the temporal sampling and 81% in spatial sampling (Tab. 5). Among the Oscillatoriales taxa, the most important genus was *Pseudanabaena*, which presented the highest contribution in sample T5, but it was not detected in spatial samples S7 and S8 (Tab. 5 and Fig. 10). According to cell counting and biomass estimation under inverted microscope, cyanobacteria presented the highest contribution in samples T5, S6, S7 and S8 (Fig. 5), in agreement with metagenomics assignments (Tab. 2), and the highest proportion of strain-level assignments from the most abundant species *M. aeruginosa* was found in spatial samples S6, S7 and S8 (Tab. 2). A similar pattern of order contribution within cyanobacteria was previously found by Steffen et al., (2012).

The amount of DNA extracted was not correlated to the number of total sequences after quality control step for each sample (Suppl. Table S1 and S2), while the total sequences for each sample were correlated to the number of assignments generated by MEGAN. According to the LCA algorithm, reads that do not have the requirements for assignment at some taxonomical node, either by alignment attributes or unknown/unspecific sequences, are moved to the previous one. The lower amount of reads assigned to specific and strain levels in the temporal samples T1 and T3 is related to the overall contribution off cyanobacteria community, instead of sequencing attributes. While the proportion and number of strain level assignments were higher in spatial dataset, sampled in a period of high cyanobacteria biomass, the proportion of other bacteria classes (Fig. 6) were higher in temporal dataset. The increase in cyanobacteria biomass followed by the increase in strain level assignments could be related to sequencing coverage, since the higher amount of biomass and DNA extracted, specially for samples S6, S7 and S8, allowed a more robust taxa classification, particularly

according to ‘min support’ LCA parameter, which sets the minimum number of reads presenting the same alignment parameters to assign a taxa to a node. Therefore, even if all *M. aeruginosa* strains detected in the study area were present in samples T1 and T3, the low abundance would hinder its placement according to LCA algorithm parameters, which would push the sequences back to generic level. In such scenario, we would expect a high amount of genus or specific level assignments, as observed by the higher proportion of *M. aeruginosa* specific level than sub-specific level in samples T1 and T3, while for all remaining samples the number of *M. aeruginosa* sub-specific assignments were higher (Tab. 2).

Sequences assigned as *M. aeruginosa* NIES-843, *M. aeruginosa* PCC 9443 and *M. aeruginosa* PCC 9717 were present in all samples but with higher proportion in sample T1, what could be a sequencing coverage bias or actually be representing the most abundant and frequent *M. aeruginosa* strains, which provided enough DNA sequences to be properly assigned even in the lowest cyanobacteria biomass samples. According to the frequency distribution of assignments through the samples for each dataset (Biome algorithm), cyanobacteria showed a slightly higher proportion of shared assignments for both samplings. The species assigned as common to samples T1, T3, ST4 and T5 were *Cyanobium gracile* PCC6307, *Microcystis aeruginosa* NIES-843, PCC9443 and PCC9717, *Synechococcus* sp. CB0101 and *Pseudanabaena biceps* PCC7429. The species assigned as part of the shared biome (assignments distributed in two or three samples) were *M. aeruginosa* PCC9701, PCC9807, PCC9808, PCC9809, TAIHU98 and *Microcystis* sp. T1-4. For spatial dataset the strains *M. aeruginosa* PCC9432 and PCC7941 were assigned as part of the shared biome, while the remaining strains were part of the core biome, present in all samples. The genetic distinctness of strains assigned as shared biome could be related to population dynamics, since the genetic plasticity associated to *M. aeruginosa* genome allow for genomic rearrangements, and the less frequent genes are recruited according to metabolic needs (Whitaker et al., 2005; Kaneko et al. 2007; Becker et al. 2012).

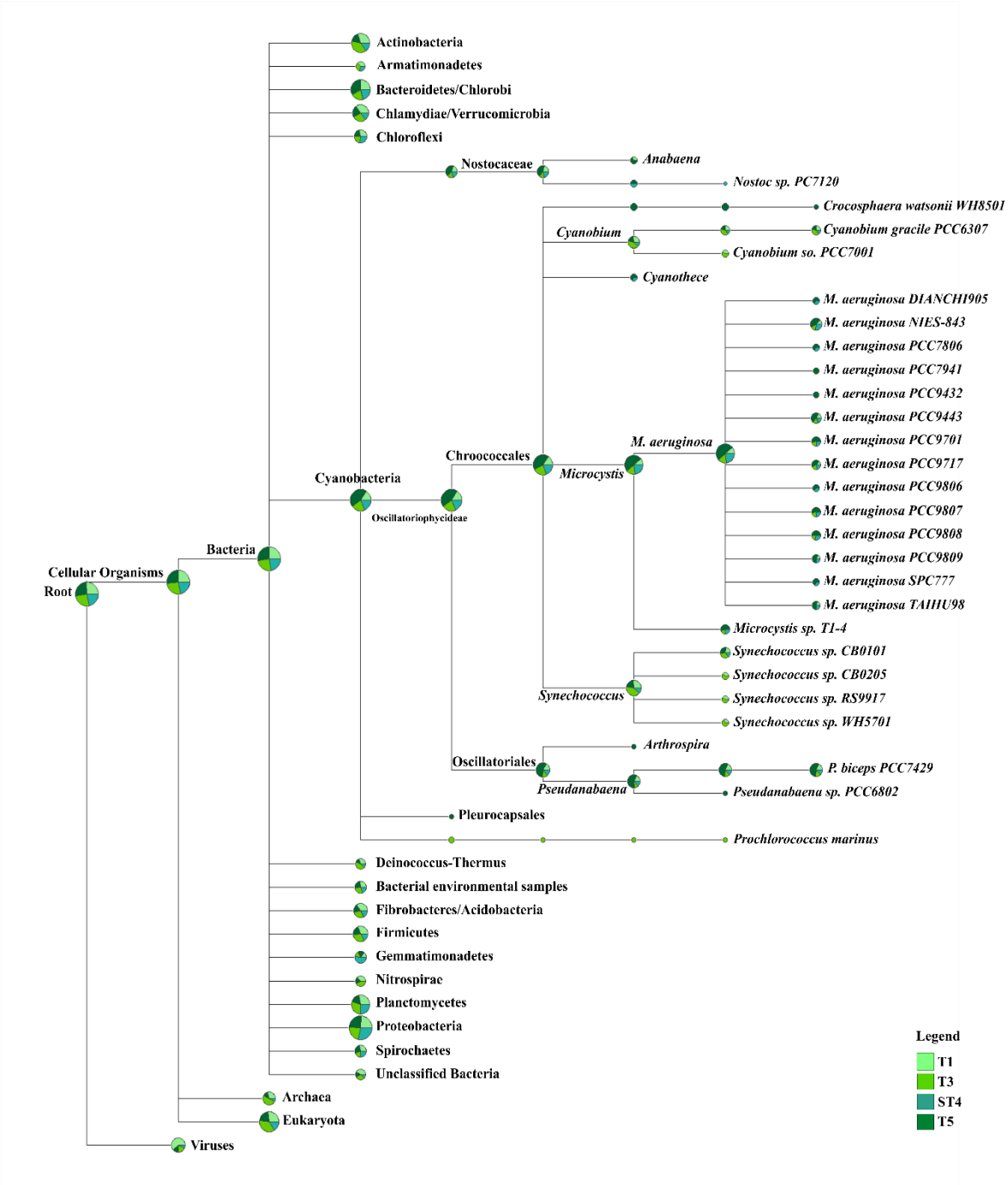


Fig. 9 - Phylogram of the temporal sampling, showing the Bacteria phyla and Cyanobacteria taxa assingned. Comparison among samples was performed using square root normalization. The pies represent summarized assignments (sum of all assignments below each level) and the phylogram layout is presented in logarithmic scale, to facilitate the visualization of differences between samples.

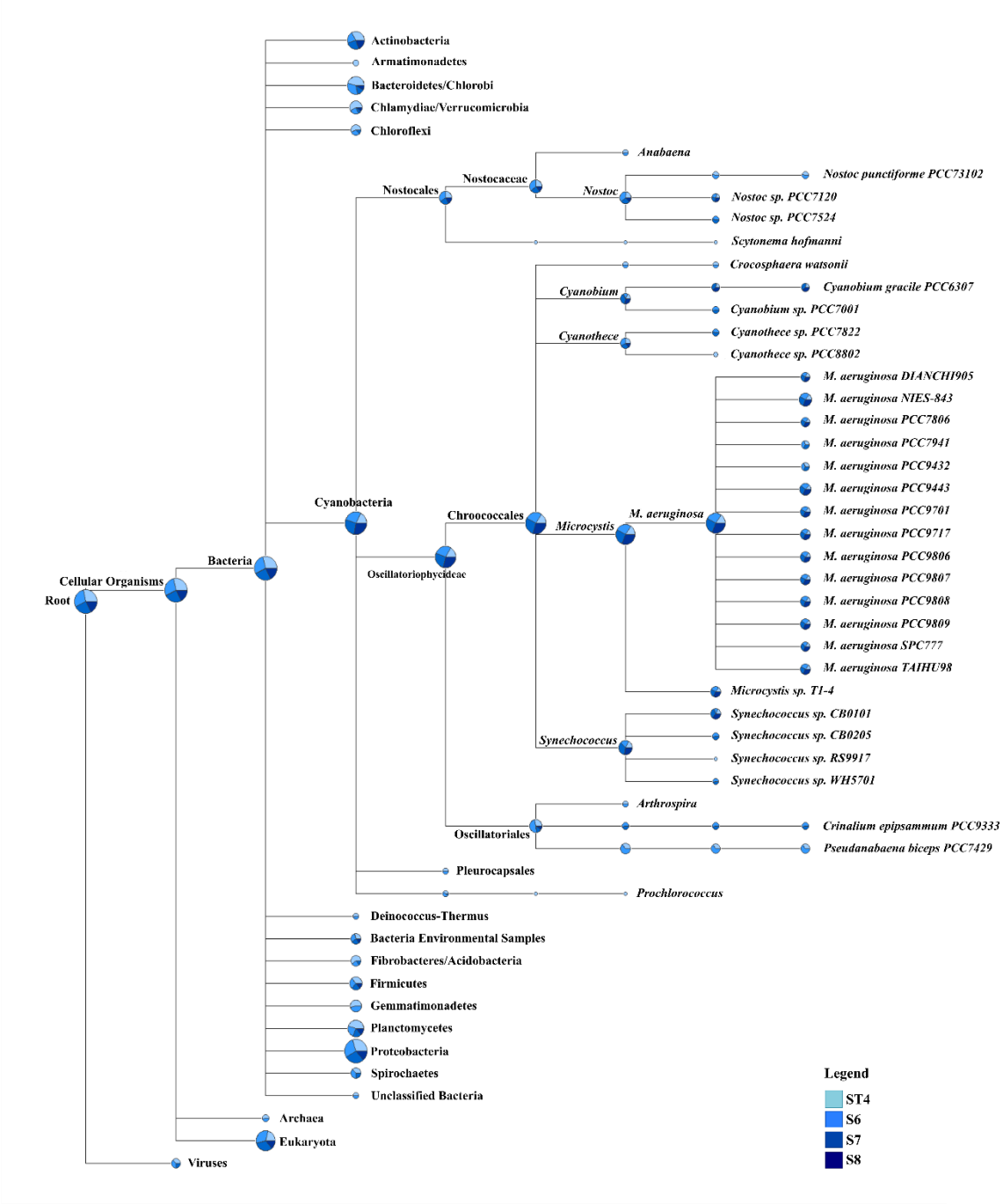


Fig. 10 - Phylogram of the spatial sampling, showing the Bacteria phyla and Cyanobacteria taxa assingned. Comparison among samples was performed using square root normalization. The pies represent summarized assignments (sum of all assignments below each level) and the phylogram layout is presented in logarithmic scale, to facilitate the visualization of differences between samples.

Functional Profile

The functional profile was build using all seven samples together, regardless of sampling design, in order to standardize the metagenomic datasets according to assignment counts, avoiding sequencing coverage bias by forcing each sample to be comparable among each other (squared root normalization method). The samples were highly similar according to functional profile generated by MEGAN/KEGG assignments. The complete set of most abundant and frequent functional modules and ortholog groups are presented in Supplementary Table S4 and S5. According to the overall contribution of functional categories, metabolism was the most highly represented category, followed by genetic information processing and environmental information processing (Tab. 5), a similar pattern found by Penn et al., (2014). The generally similar contribution of pathways among all samples suggests a relatively stable gene content pattern (Oh et al., 2011; Burke et al., 2011; Steffen et al., 2012), but we could also observe some punctual variations (Tab. 5). Nitrogen metabolism, glycerophospholipid metabolism, terpenoid backbone biosynthesis and bacterial secretion system slightly decreased in relation to the increasing biomass gradient from sample T1 to S8, while photosynthesis, antenna proteins, porphyrin and chlorophyll metabolism, RNA polymerase, nonribosomal peptide structures and carotenoid biosynthesis increased.

Tab. 5 - Relative contribution of functional pathways. The total counts were squared root normalized and the functional profile is presented according to the proportion of assignments under each functional category. The overall set of assignments (Total) shows the global functional profile of Marimbondo reservoir between November 2006 and November 2007 and along the lentic-lotic reservoir gradient.

PATHWAYS		**SAMPLES**						
	Total	**T1**	**T3**	**ST4**	**T5**	**S6**	**S7**	**S8**
Metabolism	74.4							
Amino Acid Metabolism	14.0							
Arginine and proline		13.8	14.2	15.6	15.8	15.6	15.6	16.7
Glycine. serine and threonine metabolism		13.3	12.4	12.1	13.0	12.4	12.2	12.2
Alanine. aspartte and glutamate metabolism		12.2	10.8	10.9	11.1	11.4	11.0	11.8
Cysteine and methionine metabolism		9.2	10.3	9.3	10.4	9.6	9.6	11.6
Phenylalanine. tyrosine and tryptophan biosynthesis		7.2	7.1	7.5	6.6	6.3	7.5	7.8
Lysine biosynthesis		5.2	5.2	5.5	4.8	4.8	4.8	5.4
Valine. leucine and isoleucine biosynthesis		4.5	4.3	4.2	4.3	4.5	4.7	4.9
Carbohydrate Metabolism	13.6							
Amino sugar and nucleotide sugar metabolism		11.4	11.1	11.3	11.4	10.8	10.4	12.4
Glycolysis / Gluconeogenesis		10.8	10.6	10.1	11.3	10.9	10.8	11.5
Pyruvate metabolism		10.7	10.0	11.0	10.8	11.1	11.4	11.5
Glyoxylate and dicarboxylate metabolism		10.1	9.1	9.6	9.4	11.0	12.0	10.5
Starch and sucrose metabolism		7.2	7.8	6.6	7.5	7.1	6.8	8.0

Citrate cycle (TCA cycle)		7.8	8.2	8.8	8.3	8.3	8.0	7.8
Butanoate metabolism		7.6	8.1	8.3	8.1	8.4	8.4	7.7
Energy Metabolism	12.1							
Oxidative phosphorylation		23.5	23.7	24.0	21,6	22,3	22,8	22,0
Nitrogen metabolism		20.8	19.6	18.3	18.7	19.0	18.3	16.6
Carbon fixation pathways in prokaryotes		16.6	16.0	17.9	16.3	16.3	16.4	14.3
Photosynthesis		7.2	8.4	7.2	10.6	9.4	10.2	12.5
Carbon fixation in photosynthetic organisms		8.6	8.3	9.1	8.4	8.8	8.4	9.2
Photosynthesis - antena proteins		2.0	2.4	2.2	3.5	3.2	3.3	4.9
Lipid Metabolism	4.2							
Fatty acid metabolism		20.3	20.9	23.4	21.6	23.4	22.9	18.8
Glycerolipid metabolism		11.5	12.9	11.5	15.2	15.1	14.9	18.8
Fatty acid biosynthesis		16.6	16.0	16.7	16.2	16.3	13.2	15.1
Glycerophospholipid metabolism		17.9	14.8	14.7	14.0	14.3	14.9	14.6
Nucleotide Metabolism	6							
Purine metabolism		56.6	57.1	56.2	56.3	56.6	57.4	57.1
Pyrimidine metabolism		43.4	42.9	43.8	43.7	43.4	42.6	42.9
Metabolism of Cofactors and Vitamins	8.7							
Porphyrin and chlorophyll metabolism		23.8	27.0	27.8	31.0	31.6	29.4	35.3
Ubiquinone and other terpenoid-quinone biosynthesis		9.5	11.0	10.1	9.7	10.6	11.9	10.4
Pantothenate and CoA biosynthesis		10.5	10.3	10.8	9.8	10.2	10.6	9.5
Nicotinate and nicotinamide metabolism		11.7	11.4	10.6	10.7	9,7	9,7	7,6
Riboflavin metabolism		5.6	4.1	4.9	4.6	5.3	5.2	4.0
Metabolism of Terpenoids and Polyketides	3.9							
Nonribosomal peptide structures		6.9	5.8	12.0	14.5	17.3	19.1	25.2
Terpenoid backbone biosynthesis		29.1	28.0	27.7	21.6	23.2	23.0	23.8
Carotenoid biosynthesis		11.6	14.7	12.0	16.7	13.6	14.9	14.9
Genetic Information Processing	13.1							
Transcription	0.9							
RNA polymerase		61.5	51.5	61.5	65.3	64.4	40.7	76.7
Translation	5.4							
Ribosome		43.7	43.5	42.8	44.8	44.6	39.6	42.8
Aminoacyl-tRNA biosynthesis		39.7	39.5	44.9	45.9	43.0	39.3	44.5
Folding, Sorting and Degradation	3.3							
RNA degradation		38.0	41.4	41.5	40.3	44.0	31.9	38.1
Protein export		26.3	25.1	23.8	25.6	27.7	23.1	25.0
Replication and Repair	3.4							
Homologous recombination		26.0	25.5	25.1	26.6	25.1	22.5	25.9
Mismatch repair		22.8	21.6	22.1	24.2	23.7	24.7	23.5
DNA replication		20.4	19.3	17.5	20.2	20.4	20.3	20.8
Environmental Information Processing	12.6							
Membrane Transport	6.5							
ABC transporters		76.0	79.3	77.3	81.4	77.4	79.7	80.4
Bacterial secretion system		21.7	19.0	19.6	16.7	20.3	18.5	17.9
Signal Transduction	5.8							
Two-component system		72.0	64.1	79.7	75.3	73.0	56.5	71.4

For the carbohydrate metabolism category, besides the overall homogeneous contribution of pathways, we could observe a heterogeneous contribution of individual modules among the samples. Amino sugar and nucleotide sugar metabolism, glycolysis/gluconeogenesis and glyoxylate and dicarboxylate metabolism presented modules and ortholog groups evenly distributed among all samples (Suppl. Table S4), while pyruvate metabolism and TCA cycle (Citrate cycle) presented fragmented modules and ortholog groups. We observed the limited distribution of two ortholog groups (K01959 and K01960) in samples T1 and T3, which could be part of module M00620, an incomplete reductive citrate cycle related to methanogenic bacteria. For starch and sucrose metabolism, within the *M. aeruginosa* related module M00565, the key enzymes for glycogen metabolism (EC 2.7.7.27, EC 2.4.1.21, EC 2.4.1.18 and EC 2.4.1.1) were evenly distributed throughout samples.

In the energy metabolism category, oxidative phosphorylation, nitrogen metabolism, carbon fixation pathways in prokaryotes and methane metabolism presented the highest number of assignments (Tab. 5). Proportion of photosynthesis and of photosynthesis-antenna protein pathways increased acordingly with the sample's cyanobacteria biomass (Fig. 5, Tab. 5). The most important modules observed were photosystem I (M00163), photosystem II (M00161), cytochrome b6f complex (M00162), related to Cyanobacteria, and F-type ATPase, prokaryotes and chloroplasts (M00157). Ortholog groups associated to Photosystem I and II were fragmented and distributed mainly over spatial samples (Suppl. Table S4). All antenna proteins were related to Cyanobacteria, mostly associated to phycobilisome and phycocyanin (Suppl. Table S5). The proportion of assignments under nitrogen metabolism slightly decreased over the period. Nitrate/nitrite transport system (M00438), NADH:quinone oxidoreductase, prokaryotes (M00144) and dissimilatory nitrate reduction (M00530) were incomplete but frequently distributed. This little decrease could be related to TKN increase, since *Microcystis* is known to show highest growth rates in response to ammonium, and lower growth rates in response to nitrate (Herrero et al., 2001; Chaffin & Bridgeman, 2014), but combined nitrogen sources can be taken up through permeases and metabolized to ammonium (Flores & Herrero, 2005).

Considering the metabolism of terpenoids and polyketides, two cyanobacteria associated modules (M00096 and M00364) were observed under the terpenoid backbone

biosynthesis pathway, and were complete and evenly distributed. All *M. aeruginosa* associated ortholog groups under the nonribosomal peptide structures pathway were seen evenly distributed, but with a substantial increase in proportion of assignments in the spatial samples (Fig. 11). The increasing microcystin synthetase protein assignments followed the chemical measurement of microcystin concentration. Remarkably, the observed decrease in microcystin concentration in sample S8, which presented the highest number of *mcy* gene assignments, shows that the abundance of genes is not always directly related to microcystin production. Besides this, the proximity of sample site S8 to the lotic region of the reservoir can be related to environmental conditions triggering (or not) the microcystin production. The relation between microcystin production and populational structure has been recently proposed (Hu et al., 2015), suggesting a complex ecological relationship between microcystin and *Microcystis* abundance and genotypes, not always following the abundance pattern (Wilson et al., 2005; Rinta-Kanto & Wilhelm, 2006; Xie et al., 2012; Pimentel & Giani, 2013; Penn et al., 2014; Briand et al., 2015).

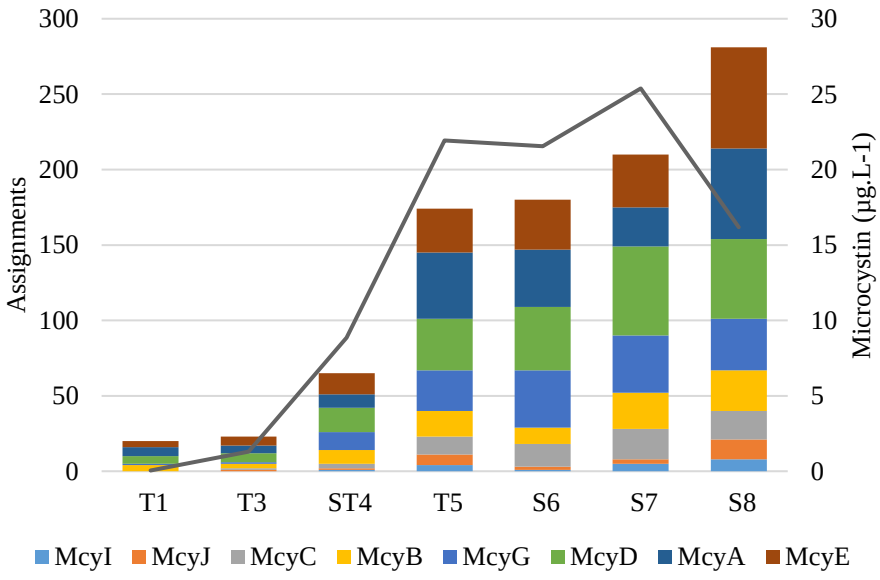


Fig. 11 - Microcystin synthetase gene assignments for each sample, showing a similar pattern as the chemical measurement of microcystin concentration.

Concerning the genetic information processing pathways, the ribosome module M00178 and aminoacyl-tRNA biosynthesis in prokaryotes (M00360), were observed as almost complete, and evenly distributed among all samples (Suppl. Table S4). We also detected the contribution of water-soluble domain of transmembrane ABC transporters,

usually found in cyanobacteria metabolic reconstructions (Tillet et al., 2000; Pope & Patel, 2008; Knoop et al., 2013). According to Penn et al., (2014), Proteobacteria presents a higher expression of phosphanate transport system, suggesting a likely role of the heterotrophic Proteobacteria in recycling reduced organic phosphorus. Urea and bicarbonate transport systems, related to Cyanobacteria, were found among the most important ABC transporter proteins. According to Penn et al., (2014), uptake of inorganic carbon is dominated by Cyanobacteria, corresponding to bicarbonate ABC transporters, while uptake of organic carbon reveals potential for niche overlap. Within the two-component system pathway, we observed modules related to nutrient response, circadian cycle and light response. Cyanobacteria are the only known prokaryotes with an endogenous circadian clock, shifting from respiration-dominated night metabolism, to biosynthesis and growth during the day (Knoop et al., 2013; Penn et al., 2014). The signal transduction category was the most important among the unique assignments and according to Steffen et al., (2012), it was among the category that showed the highest differences among the systems studied.

The nutrient dynamic of the system seems to be related to flexible genes, which presented a fragmented distribution throughout the samples and were mainly represented by pathways under environmental information processes (Tab. 5). According to Steffen et al., (2014), the changes they observed on gene expression of *M. aeruginosa* cultured populations are needed in order to maintain metabolic homeostasis, and the nutrient loadings into the system can affect not only specific genes involved in genome rearrangements but the whole community structure. The increase in energy metabolism, particularly photosynthesis, and nonribosomal peptide structure pathway assignments are directly related to the increase in cyanobacteria biomass. The highly similar contribution of central metabolism pathways and the somewhat fragmented contribution of genetic and environmental information processes pathways indicate a probable functional redundancy across taxa, in which any given function is potentially distributed across a variety of taxa (Burke et al., 2011; Steffen et al., 2014).

2. CONCLUSIONS

Cyanobacterial blooms usually contain one prominent bloom-forming primary producer and a diverse mix of other photosynthetic and non-photosynthetic microorganisms that contribute to nutrient and energy cycling (Litchman et al., 2010). Since microbial interactions could span all taxonomic ranks, from strain to superkingdom (Leibold et al., 2004; Zarraonaindia et al., 2013), functional stability could arise from pathway completion (Helling et al., 1987; Wintermute & Silver, 2010a,b), in which metabolic trade-offs could stabilize genes at medium frequencies owing to the partitioning of functional roles within the population (Cordero et al., 2012; Cordero & Polz, 2014). Despite the observed changes on biodiversity profile (Fig. 9-10), the similarity among the functional profiles of the temporal and spatial datasets (Tab. 5) indicates a functional redundancy across taxa and a conserved functional potential of the bloom-associated community (Oh et al., 2011; Burke et al., 2011; Steffen et al., 2012; Steffen et al., 2014; Penn et al., 2014). If different organisms are fulfilling the same roles, it is more likely that the communities differ in their taxonomic composition due to a gene centered dynamics, through high levels of genetic diversity being shaped by environmental pressures (Galhardo et al., 2007; Frangeul et al., 2008; Wiedenbeck & Cohan, 2011; Boon et al., 2014; Shapiro & Polz, 2014; Coleman et al., 2006; Van Elsas & Bailey, 2002; Ulrich et al., 2005). We were able to relate the physical-chemical variables to temporal and spatial components (Fig. 8), to biomass increase, microcystin production and to class level metagenomics assignments. The shift from a clear water state with high nitrate levels to turbid water, with increased levels of TP, TKN and Chlor seems to be related to the development of cyanobacteria biomass, shaping the remaining community (Fig. 6-7). Our results suggest a broad interactive network acting on Marimbondo reservoir, where the diversity at strain level can unveil the populational features subject to environmental conditions (Wiedenbeck & Cohan, 2011; Luo et al., 2015). Such diversity maintenance through recombination, rearrangements, single nucleotde variants or lateral gene transfer (Whitaker et al., 2005) might ensure that the overall population remains stable (Steffen et al., 2014; Biller et al., 2015). The reference database bias should therefore always be considered and the shift from taxonomical units to gene-centered analysis, or reverse ecology, may potentially improve the annotation accuracy (Shapiro & Polz, 2012), regardless of the whole genome thresholds.

The metagenomic biodiversity profile of the Marimbondo reservoir, under the progress and following the establishment of a *Microcystis* bloom event (Fig. 5), presented changes in taxa abundance (Fig. 6-7, Tab. 2, Suppl. Table S3), after the increase in cyanobacterial biomass (Fig. 5), especially Actinobacteria and Planctomycetia, whose contribution was higher under lower *Microcystis* biomass (Fig. 7, 10). Besides the decrease in contribution of some taxa, the decrease in diversity indexes was also observed (Tab. 3). Proteobacteria, mostly Alphaproteobacteria and Betaproteobacteria, seems to be related to Cyanobacteria bloom conditions, noticeable by the correlation between environmental variables and class assignments (Fig. 8). The functional profile presented an overall stable pathway contribution, in which central metabolism pathways were abundant and evenly distributed through the samples (Suppl. Table S4), while environmental information processing pathways presented a more fragmented distribution (Tab. 5). We conclude that *M. aeruginosa* microdiversity could be related to bloom establishment, and the associated heterotrophic microbial community changes their class contribution according to environmental conditions (Fig. 6 and Fig. 7). The similar functional profile points to a conserved potential, driven by adaptive strategies, and the entire community-environment perspective that resulted from this study was only possible through metagenomic approach.

2. ACKNOWLEDGEMENTS

This research was supported by a scholarship from CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) to M.L., and by funds provided by FAPEMIG (Fundação de Apoio a Pesquisa de Minas Gerais) to A.G. We thank the staff of the Phycology Laboratory (Dept. Botany/UFMG), for performing all laboratory analyses and the "Biodados" staff (Dept. of Biochemistry/UFMG), for assistance in bioinformatics. We also thank Msc. Cláudia Wermelinger for English review of the text.

2. REFERENCES

Ashby M. K., Houmard J. (2006) Cyanobacterial Two-Component Proteins: Structure, Diversity, Distribution, and Evolution. Microbiol. *Mol. Biol. Rev.*, 70(2):472-509

Apha - American Public Health Association (2005) Standard Methods for the Examination of Water and Wastewater. 21. ed. Washington dc.

Bagatini I. L., Eiler A., Bertilsson S., Klaveness D., Tessarolli L. P., Vieira A. A. H. (2014) Host-Specificity and Dynamics in Bacterial Communities Associated with Bloom-Forming Freshwater Phytoplankton. *PLoS ONE*, 9(1): e85950

Beck C., Knoop H., Axmann I. M., Steuer R. (2012) The diversity of cyanobacterial metabolism: genome analysis of multiple phototrophic microorganisms. *BMC Genomics*, 13:56

Biller S. J., Berube P. M., Lindell D., Chisholm S. W. (2015) *Prochlorococcus*: the structure and function of collective diversity. *Nat. Rev. Microbiol.,* 13:13-27

Boon E., Meehan C. J., Whidden C., Wong D. H. J., Langille M. G. I., Beiko R. G. (2014) Interactions in the microbiome: communities of organisms and communities of genes. *FEMS Microbiol Rev*, 38:90-118

Briand E., Bormans M., Gugger M., Dorrestein P. C., Gerwick W. H. (2015) Changes in secondary metabolic profiles of Microcystis aeruginosa strains in response to intraspecific interactions. *Envir. Microbiol*., doi:10.1111/1462-2920.12904

Brunberg A.K. (1999) Contribution of bacteria in the mucilage of *Microcystis* spp. (Cyanobacteria) to benthic and pelagic bacterial production in a hypereutrophic lake. *FEMS Microbiol. Ecol*., 29:13-22

Burke C., Steinberg P., Rusche D., Kjelleberg S., Thomas T. (2011) Bacterial community assembly based on functional genes rather than species. *PNAS*, 108(34):14288-14293

Chaffin J.D., Bridgeman T.B. (2014) Organic and inorganic nitrogen utilization by nitrogen-stressed cyanobacteria during bloom conditions. *J. Appl. Phycol.*, 26(1):299-309

Coleman M. L., Sullivan M. B., Martiny A. C., Steglich C., Barry K., DeLong E. F., Chisholm S. W. (2006) Genomic islands and the ecology and evolution of Prochlorococcus. *Science*, 311:1768–1770

Coleman M. L., Chisholm S. W. (2010) Ecosystem-specific selection pressures revealed through comparative population genomics. *PNAS*, 107(43):18634-18639

Cordero O. X., Ventouras L. A., DeLong E. F., Polz M. F. (2012) Public good dynamics drive evolution of iron aquisition strategies in natural bacterioplankton populations. *Proc. Natl. Acad. Sci. USA*, 109:20059–20064

Cordero O. X., Polz M. F. (2014) Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Rev. Microbiol.*, 12:263-273

Debroas D., Humbert J-F., Enault F., Bronner G., Faubladier M., Cornillot E. (2009) Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget – France). *Envir. Microbiol.*, 11(9):2412-2424

Desai M. M., Walczak A. M. (2015) Flexible gene pools - Rapid genetic exchange leads to mosaic genomes

in cyanobacterial populations. *Science*, 348:977-978

Eiler A., Bertilsson S. (2004) Composition of freshwater bacterial communities associated with cyanobacterial blooms in four Swedish lakes. *Envir. Microbiol.*, 6(12):1228-1243

Escalas A., Bouvier T., Mouchet M. A., Leprieur F., Bouvier C., Troussellier M., Mouillot D. (2013) A unifying quantitative framework for exploring the multiple facets of microbial biodiversity across diverse scales. *Envir. Microbiol.*, 15:2642-2657

Ettema T. J. G., Andersson S. G. E. (2009) The α-proteobacteria: the Darwin finches of the bacterial world. *Biol. Lett.*, 5:429-432

Flores E., Herrero A. (2005) Nitrogen assimilation and nitrogen control in cyanobacteria. *Bioch. Soc. Trans*., 33(1):164-167

Frangeul L., Quillardet, P., Castets A. M., Humbert J. F., Matthijs H. CP., Cortez D., Tolonen A., Zhang CC., Gribaldo S., Kehr J. C., Zilliges Y., Ziemert N., Becker S., Talla E., Latifi A., Billault A., Lepelletier A., Dittmann E., Bouchier C., Tandeau de Marsac N. (2008) Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics*, 9:274

Galhardo R. S., Hastings P. J., Rosenberg S. M. (2007) Mutation as a stress response and the regulation of evolvability. *Crit. Rev. Biochem. Mol. Biol*., 42:399–435

Ghai R., Mizuno C. M., Picazo A., Camacho A., Rodriguez-Valera F. (2014) Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing. *Molec. Ecol*., 23:6073-6090

Haukka K., Kolmonen E., Hyder R., Hietala J., Vakkilainen K., Kairesalo T., Haario H., Sivonen K. (2006) Effect of Nutrient Loading on Bacterioplankton Community Composition in Lake Mesocosms. *Microb. Ecol*., 51:137-146

Helling R.B., Vargas C.N., Adams J. (1987) Evolution of Escherichia coli during growth in a constant environment. *Genetics*, 116: 349-358

Herrero A., Muro-Pastor A.M., Flores E. (2001) Nitrogen control in cyanobacteria. *J. Bacteriol.*, 183(2): 411-425

Hu C., Rea C., Yu Z., Lee J. (2015) Relative importance of Microcystis abundance and diversity in determining microcystin dynamics in Lake Erie coastal wetland and downstream beach water. *J. Appl. Microbiol.*, 120:138-151

Humbert J-F., Dorigo U., Cecchi P., Le Berre B., Debroas D., Bouvy M. (2009) Comparison of the structure and composition of bacterial communities from temperate and tropical freshwater ecosystems. *Envir. Microbiol.*, 11(9):2339-2350

Humbert JF., Barbe V., Latifi A., Gugger M., Calteau A., Coursin T., Lajus A., Castelli V., Oztas S., Samson G., Longin C., Medigue C., Tandeau de Marsac N. (2013) A Tribute to Disorder in the Genome of the Bloom-Forming Freshwater Cyanobacterium *Microcystis aeruginosa*. *PLoS ONE*, 8(8): e70747

Huson D. H., Mitra S., Ruscheweyh H. J., Weber N., Schuster S.C. (2011) Integrative analysis of environmental sequences using MEGAN 4. *Genome Res.*, 21:1552-1560

Huson D. H., Mitra S. (2012) Introduction to the Analysis of Environmental Sequences: Metagenomics with MEGAN. *Evolut. Genom*., series Methods in Molecular Biology, 856:415-429

Kaneko T., Nakajima N., Okamoto S., Suzuki I., Tanabe Y., Tamaoki M., Nakamura Y., Kasai F., Watanabe A., Kawashima K., Kishida Y., Ono A., Shimizu Y., Takahashi C., Minami C., Fujishiro T., Kohara M., Katoh M., Nakazaki N., Nakayama S., Yamada M., Tabata S., Watanabe M. M. (2007) Complete Genomic Structure of the Bloom-forming Toxic Cyanobacterium Microcystis aeruginosa NIES-843. DNA Research, 14:247-256

Kashtan N., Roggensack S. E., Rodrigue S., Thompson J. W., Biller S. J., Coe A., Ding H., Marttinen P., Malmstrom R. R., Stocker R., Follows M. J., Stepanauskas R., Chisholm S. W. (2014) Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild *Prochlorococcus*. Science, 344:416-420

Knoop H., Gründel M., Zilliges Y., Lehmann R., Hoffmann S., Lockau W., Steuer R. (2013) Flux Balance Analysis of Cyanobacterial Metabolism: The Metabolic Network of *Synechocystis* sp. PCC 6803. *PLoS Comput Biol*, 9(6): e1003081

Kurmayer R., Christiansen G., Chorus, I. (2003) The abundance of microcystin-producing genotypes correlates positively with colony size in Microcystis sp. and determines its microcystin net production in Lake Wannsee. *Appl. Environ. Microbiol.*, 69:787-795

Leibold M. A., Holyoak M., Mouquet N., Amarasekare P., Chase J. M., Hoopes M. F., Holt R. D., Shurin J. B., Law R., Tilman D., Loreau M., Gonzalez A. (2004) The metacommunity concept: a frameworkfor multi-scale community ecology. *Ecol. Let.*, 7: 601-613

Litchman E., Tezanos Pinto P., Klausmeier C. A., Thomas M. K., Yoshiyama K. (2010) Linking traits to species diversity and community structure in phytoplankton. *Hydrobiologia*, 653:15-28

Louati I., Pascault N., Debroas D., Bernard C., Humbert, J. F., Leloup J. (2015) Structural Diversity of Bacterial Communities Associated with Bloom-Forming Freshwater Cyanobacteria Differs According to the Cyanobacterial Genus. *PLoS ONE*, 10(11): e0140614

Luo C., Knight R., Siljander H., Knip M., Xavier R.J., Gevers D. (2015) Constrains identifies microbial strains in metagenomic datasets. Nat. Biotechnol., 33(10):1045-1052

Makarewicz J. C., Lewis T. W. (2015) Exploring spatial trends and causes in Lake Ontario coastal chemistry: Nutrients and pigments. *J. Great Lakes Res*., 41(3):794-800

Maruyama T., Kato K., Yokoyama A., Tanaka T., Hiraishi A., Park H-D. (2003) Dynamics of Microcystin-Degrading Bacteria in Mucilage of *Microcystis*. *Microb. Ecol.*, 46:279-28

Meyer F., Paarmann D., D'Souza M., Olson R., Glass E.M., Kubal M., Paczian T., Rodriguez A., Stevens R., Wilke A., Wilkening J., Edwards R.A. (2008) The Metagenomics RAST server — A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9:386

Mitra S., Rupek P., Richter D. C., Urich T., Gilbert J. A., Meyer F., Wilke A., Huson D.H. (2011) Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics*, 12 (Suppl. 1): S21

Moore L. R., Rocap G., Chisholm S. W. (1998) Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature*, 393:464

Mou X., Lu X., Jacob J., Sun S., Heath R. (2013) Metagenomic Identification of Bacterioplankton Taxa and Pathways Involved in Microcystin Degradation in Lake Erie. *PLoS ONE*, 8(4):e61890

Niemi R. M., Heiskanen I., Heine R., Rapala J. (2009) Previously uncultured β-Proteobacteria dominate in

biologically active granular activated carbon (BAC) filters. *Wat. Res*., 43(20):5075-5086

Oh S., Caro-Quintero A., Tsementzi D., DeLeon-Rodriguez N., Luo C., Poretsky R., Konstantinidis K.T. (2011) Metagenomic Insights into Evolution, Function, and Complexity of the Planktonic Microbial Community of Lake Lanier, a Temperate Freshwater Ecosystem. *Appl. Environm. Microbiol.*, 77(17):6000-6011

Oksanen F.J., Blanchet G., Kindt R., Legendre P., Minchin P.R., O'Hara R.B., Simpson G.L., Solymos P.M., Stevens H.H., Wagner H. (2015) Vegan: Community Ecology Package. R package version 2.3-0. http://CRAN.R-project.org/package=vegan

Paerl H. W., Otten T. G. (2013) Harmful Cyanobacterial Blooms: Causes, Consequences, and Controls. *Environmental Microbiology*, DOI 10.1007/s00248-012-0159-y

Paerl H. W., Paul V. J. (2012) Climate change: Links to global expansion of harmful cyanobacteria. *Water Res*., 46(5):1349-1363

Palacio H. M., Ramírez J. J., Echenique R. O., Palacio J. A., Sant'anna C. L. (2015) Floristic composition of cyanobacteria in a neotropical, eutrophic reservoir. *Braz. J. Bot.*, 38(4):865-876

Parfenova V. V., Gladkikh A. S., Belykh O. I. (2013) Comparative analysis of biodiversity in the planktonic and biofilm bacterial communities in Lake Baikal. *Microbiology*, 82(1):91-101

Penn K., Wang J., Fernando S. C., Thompson J. R. (2014) Secondary metabolite gene expression and interplay of bacterial functions in a tropical freshwater cyanobacterial bloom. *The ISME Journal*, 8:1866–1878

Piccin-Santos V., Bittencourt-Oliveira M. C. (2012) Toxic Cyanobacteria in Four Brazilian Water Supply Reservoirs. *J. Envir. Protec*., 3:68-73

Pimentel J. S. M., Giani A. (2013) Estimating toxic cyanobacteria in a Brazilian reservoir by quantitative real-time PCR, based on the microcystin synthetase D gene. *J. Appl. Phycol.*, 25: 1545-1554

Pope P. B., Patel B. K. C. (2008) Metagenomic analysis of a freshwater toxic cyanobacteria bloom. *FEMS Microbiol. Ecol.*, 64:9-27

R Core Team (2015): R: A Language and Environment for Statistical Computing. - R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Rinta-Kanto J. M., Wilhelm S. W. (2006) Diversity of Microcystin-Producing Cyanobacteria in Spatially Isolated Regions of Lake Erie. *Appl. Envir. Microbiol*., 72(7):5083-5085

Rodriguez-Valera F., Martin-Cuadrado AB., Rodriguez-Brito B., Pašić L., Thingstad T. F., Rohwer F., Mira A. (2009) Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol*., 7:828-836

Rosen M. J., Davison M., Bhaya D., Fisher D. S. (2015) Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science*, 348: 1019-1023

Sant'anna C. L., Azevedo M. T. P., Werner V. R., Dogo C. R., Rios F. R., Carvalho L. R. (2008) Review of toxic species of Cyanobacteria in Brazil. *Algol. Stud.*, 126:251-265

Saxton M. A., Morrow E. A., Bourbonniere R. A., Wilhelm S. W. (2011) Glyphosate influence on phytoplankton community structure in Lake Erie. *J. Great Lakes Res.*, 37(4): 683-690

Shapiro B. J., Friedman J., Cordero O. X., Preheim S. P., Timberlake S. C., Szabó G., Polz M. F., Alm E. J. (2012) Population genomics of early events in the ecological differentiation of bacteria. *Science*, 336:48–51

Shapiro, B. J., Polz M. F. (2014) Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in Microbiology*, 22(5):235-247

Shen H., Niu Y., Xie P., Tao M., Yang X. (2011) Morphological and physiological changes in *Microcystis aeruginosa* as a result of interactions with heterotrophic bacteria. *Freshwater Biology*, 56: 1065–1080

Shi L., Cai Y., Wang X., Li P., Yu Y., Kong F. (2010) Community Structure of Bacteria Associated with *Microcystis* Colonies from Cyanobacterial Blooms. *J. Fresh. Ecol*., 25(2):193-203

Steffen M. M., Li Z., Effler T. C., Hauser L. J., Boyer G. L., Wilhelm S. W. (2012) Comparative Metagenomics of Toxic Freshwater Cyanobacteria Bloom Communities on Two Continents. *PLoS ONE*, 7(8): e44002

Steffen M. M., Dearth S. P., Dill B. D., Li Z., Larsen K. M., Campagna S. R., Wilhelm S. W. (2014) Nutrients drive transcriptional changes that maintain metabolic homeostasis but alter genome architecture in Microcystis. *The ISME Journal*, 8:2080-2092

Tillett D., Dittmann E., Erhard M., von Döhren H., Börner T., Neilan B. A. (2000) Structural organization of microcystin biosynthesis in *Microcystis aeruginosa* PCC7806: an integrated peptide–polyketide synthetase system. *Chem. Biol*., 7(10):753-764

Ulrich L. E., Koonin E. V., Zhulin I. B. (2005) One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol.*, 13:52–56

Van Elsas J. D., Bailey M. J. (2002) The ecology of transfer of mobile genetic elements. *FEMS Microbiol. Ecol.*, 42:187–197

Van Rossum T., Peabody M. A., Uyaguari-Diaz M. I., Cronin K. I., Chan M., Slobodan J. R., Nesbitt M. J., Suttle C. A., Hsiao W. W. L., Tang P. K. C., Prystajecky N. A., Brinkman F. S. L. (2015) Year-Long Metagenomic Study of River Microbiomes Across Land Use and Water Quality. *Front. Microbiol*. 6:1405

Von Rückert G., Giani A. (2004) Effect of nitrate and ammonium on the growth and protein concentration of Microcystis viridis Lemmermann (Cyanobacteria). *Revista Brasil. Bot.,* 27(2):325-331

Whitaker R. J., Grogan, D. W., Taylor J. W. (2005) Recombination Shapes the Natural Population Structure of the Hyperthermophilic Archaeon *Sulfolobus islandicus*. *Mol. Biol. Evol*., 22(12):2354-2361

Wiedenbeck J., Cohan F. M. (2011) Origins of bacterial diversity through horizontal genetic transferand adaptation to new ecological niches. *FEMS Microbiol. Rev*., 35:957-976

Wilson A. E., Sarnelle O., Neilan B. A., Salmon T. P., Gehringer M. M., Hay M. E. (2005) Genetic Variation of the Bloom-Forming Cyanobacterium Microcystis aeruginosa within and among Lakes: Implications for Harmful Algal Blooms. *Appl. Envir. Microbiol.*, 71(10):6126-6133

Wintermute E. H., Silver P. A. (2010) a. Emergent cooperation in microbial metabolism. *Mol. Syst. Biol*., 6:407

Wintermute E. H., Silver P. A. (2010) b. Dynamics in the mixed microbial concourse. *Genes Dev*., 24:2603-2614

Wojciechowski J., Padial A. A. (2015) Monitoring studies should consider temporal variability to reveal relations between cyanobacterial abundance and environmental variables. *An. Acad. Bras. Cienc*., http://dx.doi.org/10.1590/0001-3765201520140416

Worm J., Søndergaard M. (1998) Dynamics of heterotrophic bacteria attached to *Microcystis* spp. (Cyanobacteria). *Aq. Microb. Ecol.*, 14:19-28

Xie L., Rediske R. R., Hong Y., O'Keefe J., Gillett N. D., Dyble J., Steinman A. D. (2012) The role of environmental parameters in the structure of phytoplankton assemblages and cyanobacteria toxins in two hypereutrophic lakes. *Hydrobiologia*, 691:255-268

Zarraonaindia I., Smith D. P., Gilbert J. A. (2013) Beyond the genome: community-level analysis of the microbial world. *Biol Philos*, 28:261-282

Zhu L., Wu Y., Song L., Gan N. (2014) Ecological Dynamics of Toxic *Microcystis* spp. and Microcystin Degrading Bacteria in Dianchi Lake, China. *Appl. Envir. Microbiol*., 80(6):1874-1881

Capítulo 3: Explorando a variabilidade de uma população de *Microcystis* em processo de desenvolvimento de floração

3. RESUMO

O sucesso ecológico de cianobactérias tem sido relacionado à estrutura do seu genoma, altamente plástico e dinâmico, o qual permite que estratégias de diversificação desempenhem um papel fundamental na adequação de populações às condições ambientais. *Microcystis* é um dos mais representativos gêneros de cianobactérias produtoras de toxinas e capazes de desenvolver florações, sendo amplamente distribuído e não apresentando estrutura filogeográfica. Junto a isso, pouca evidência de diferenciação de ecótipos tem sido encontrada, mesmo com o amplo e crescente número de cepas sequenciadas. As mudanças derivadas do progresso do evento de floração foram exploradas, especialmente em relação à população de *Microcystis*, ao longo de um ciclo anual, em um reservatório de água meso-eutrófico, usando várias abordagens e estratégias metagenômicas. A amostragem foi realizada em zona pelágica, abrangendo as escalas temporal e espacial. O DNA foi sequenciado através da plataforma Roche 454 GS FLX e os conjuntos de sequências foram alinhados e classificados de acordo com BLASTx/MEGAN, recrutados contra genomas de referência usando o programa MUMmer, mapeados de acordo com BWA-MEM e variantes de base única foram detectados através do protocolo sugerido pelo programa GATK. A população de *Microcystis* apresentou uma estrutura altamente diversa, variando de cinco a 15 cepas diferentes ao longo do aumento em biomassa, quando um aumento na produção de microcistina foi também observado. Treze regiões hipervariáveis foram descritas, com atributos como baixa identidade e cobertura associados a funções relacionadas ao metabolismo compartilhado e genes flexíveis. O genoma flexível, ou conjunto de genes de frequência media-baixa, parece desempenhar um papel fundamental nessa população, e a variedade de anotações taxonômicas e funcionais observadas entre todas as cepas analisadas sugere que esta complexidade pode estar encobrindo uma estrutura populacional mais coesiva, a qual mantém uma alta dinâmica de diversidade entre os membros, centrada no conteúdo gênico e interação com variáveis ambientais.

Palavras-chave: Cianobactéria, diversidade genética, microdiversidade, detecção de variants populacional, análise de recrutamento, estrutura populacional, regiões hipervariáveis.

Exploring the variability within a *Microcystis* population under bloom development

3. ABSTRACT

The ecological success of cyanobacteria has been related to a highly plastic and dynamic genomic structure, which allows diversification strategies to play a major role in population fitness. *Microcystis* is one of the most representative genera of toxic bloom-forming cyanobacteria, being widely distributed and without phylogeographic structure. Also, little evidence of ecotype differentiation has been found, even with the large and increasing number of sequenced strains. We explored the changes arising from the bloom progress, especially within *Microcystis* population, over an annual cycle in a tropical meso-eutrophic water reservoir using several metagenomic strategies. Sampling was carried out in the pelagic zone, covering a temporal and spatial scale. The DNA was sequenced using a Roche 454 GS FLX platform. The samples were analyzed through BLASTx/MEGAN, recruited using MUMmer package, mapped by BWA-MEM and the variants called through GATK pipeline. The *Microcystis* population presented a highly diverse structure, ranging from five to fifteen different strains following the increase in biomass, when an increase in microcystin production was also observed. Thirteen hypervariable regions were described, showing low identity and low coverage attributes, associated to the flexible/shared genome functions found within these regions. Flexible genome, or medium-low frequency set of genes, seems to play a major role in this population, and the variety of taxonomical and functional annotations observed among all strains suggests that such complexity could be masking a very cohesive population structure, which structure would be centered on gene content and interactions with environmental variables, while maintaining a high level of diversity among its members.

Keywords: cyanobacteria, genetic diversity, microdiversity, population variant calling, recruitment, population structure, flexible gene pool, hypervariable regions

3. INTRODUCTION

Cyanobacteria are prokaryotic photosynthetic organisms that have been progressively increasing in importance around the world as a consequence of anthropic impacts in water bodies and watersheds (Sheath & Wehr, 2003; Whitton, 2012). Their long evolutionary time scale may have contributed to the high genomic diversity and adaptability observed among different cyanobacteria (Hess, 2011; Simm et al., 2015), buffering against extreme changes in the environment and increasing the chances of finding and fixing an adaptive genotype (Cordero & Polz, 2014). Through homologous recombination, rearrangements and lateral gene transfer, microdiversity (Kashtan et al., 2014) became an evolutionary strategy that allows this species to occupy changing environments, such as freshwater ecosystems (Castenholz & Norris, 2005; Wilson et al., 2005; Komárek et al. 2011; Komárková et al., 2011; Palinska et al., 2011; Humbert et al., 2013). Even among closely related strains or isolates a surprisingly divergent gene content is often found and the gene exchange dynamics could be an important way of interaction (Hess, 2011; Beck et al., 2012; Steffen et al., 2014).

Microcystis is one of the most representative genera of toxic bloom-forming cyanobacteria, widely distributed geographically (Kaneko et al., 2007) and without phylogeographic structure (Sabart et al., 2009; Van Gremberghe et al., 2011; Hu et al., 2015). Despite the large genotypic diversity that has been described (Xie et al., 2012), no clear evidence of intraspecific ecotype differentiation has been found for this species complex (Nguyen et al., 2012; Humbert et al., 2013). The 'pan-genome' analysis try to estimate the size of the gene repertoire accessible to any given taxa (Tettelin et al., 2008), considering the large number of genes with a fragmented or unique distribution within most microbial populations (Rodriguez-Valera et al., 2009). The difficulty in ranking different strains of *M. aeruginosa* to species level, the crescent number of different strains sequenced and the broad phenotypic diversity could be related to an open pan-genome, displayed by most microbial species, and to hypervariable genomic regions, which allow these organisms to respond to transient selective pressures in their local environment (Van Elsas & Bailey, 2002; Coleman et al., 2006; Pašić et al., 2009). This microdiversity could be the result of periodic selection events, resulting in groups of strains genetically clustered, linked by metabolic complementation, constituting what is been called genomic backbones (Shapiro & Polz, 2012; Rodriguez-Valera et al., 2009; Shapiro et al., 2012;

Kashtan et al., 2014). On the other hand, the diverse gene content characteristic of microdiverse populations could be related to a collective community genome, opened to the complementation among higher taxonomical ranks and continuously evolving through gene transfer, generating a cloud of distinct strains (Ehrlich et al., 2010) functionally linked even between different populations (Whitaker et al., 2005; Antonova & Hammer, 2011; Zhu et al., 2012; Steffen et al., 2014; Biller et al., 2015). Both concepts are very similar, differing on the range of genetic exchanges, the former considering populational-level interaction, while the later considering community-level interaction. Considering such microdiversity dynamics, cyanobacteria populations could be analysed as a mosaic genome, composed by strains with slightly distinct gene content, designed by the balance between environmental pressures and homologous recombination, keeping a set of gene variants able to spread and change its relative abundance and frequency according to local environmental conditions (Steffen et al., 2014; Rosen et al., 2015).

The metagenome encompasses all lineages within one sample and provides access to the functional gene composition of microbial communities (Rodriguez-Valera et al., 2009; Thomas et al., 2012). The whole genome shotgun technique (WGS) enables the search for differentiated genomic backbones and population features, thus allowing for an analysis that ranges from broad to fine scale. A frequently used strategy when looking for the genomic diversity within a population is genome recruitment, which is composed by reads mapped above 99% nucleotide identity which represents the preservation of genomic sequences in the natural environment (Pašić et al. 2009). The reads recruited usually present a major contribution of core or high-medium frequency genes (Mira et al., 2010; Knoop et al., 2013; Cordero & Polz, 2014), while hypervariable regions, usually observed through lower depth of coverage, are mostly composed of flexible or low-frequency genes (Tettelin et al., 2008; Mira et al., 2010; Shapiro et al., 2012; Cordero & Polz, 2014).

In this work, we explored the *Microcystis* population along the environmental gradient towards the bloom establishment in a tropical meso-eutrophic freshwater reservoir (Marimbondo reservoir, Brazil). Several metagenomic strategies were adopted in order to integrate the methods to the community complexity. The structural genome features that could be related to the environmental context were mostly observed through hypervariable regions and SNVs (single nucleotide variants) distribution among several *Microcystis* strains aligned over 99% of similarity and among the sequence-discrete

population, which aligned above 85% similarity and was associated to the shared genome. The description of such interactive-genome genes could guide further investigations regarding populations' diversity dynamics.

3. MATERIAL AND METHODS

Study Area

Sampling was carried out in the pelagic zone of Marimbondo reservoir, in Southeastern Brazil (Fig. 1), using a Van Dorn sampler. Samples covered a temporal (November 2006-T1, May 2007-T3, August 2007-ST4 and November 2007-T5), and spatial scale (August 2007-ST4, S6, S7 and S8).

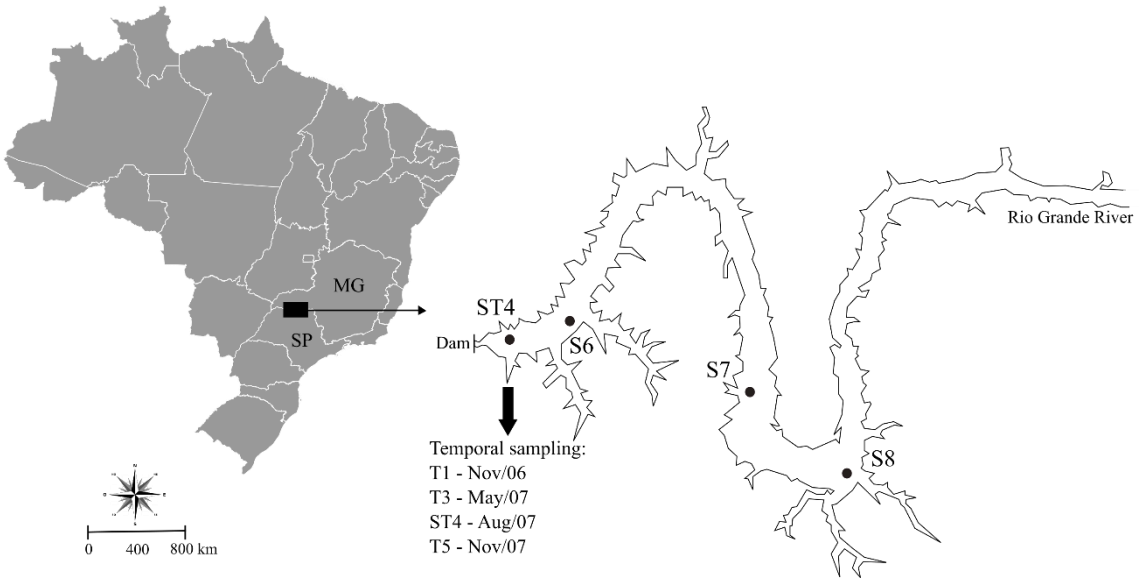


Fig. 1 - Marimbonfo reservoir, located between Minas Gerais and São Paulo states (Brazil), in the cities Icém (SP) and Fronteira (MG). The four spatial sampling stations are indicated. The temporal sampling was carried close to the dam, from November 2006 to November 2007.

Sequencing and pre-processing

The genomic DNA was extracted according to the phenol-chlorophorm method (Kurmayer et al., 2003), used for the library preparation and then sequenced using a Roche 454 GS FLX platform and Whole Genome Shotgun (WGS) strategy. The seven metagenomic samples were uploaded to MG-RAST platform, which offers automated quality control, annotation, comparative analysis and archiving services (Meyer et al., 2008).

The raw sequences were cleaned with seq_crumbs software (http://bioinf.comav.upv.es/seq_crumbs/), trim_quality tool, under default parameters. The cleaned sequences were searched against NCBI non-redundant protein sequences (NR) database using BLASTx and visualized through MEGAN Metagenomic Analyzer 5.1.5. MEGAN bases its taxonomic classification on the NCBI taxonomy, which is a hierarchically structured classification of all species that are represented at NCBI database (Huson et al. 2011). An initial attempt to look for variance patterns among the *Microcystis* strains assingned by BLASTx/MEGAN was performed through a recruitment approach, using MUMmer 3.23 (Kurtz et al., 2004). The original 454 datasets presented a generally insufficient depth and breadth of coverage, so the lack of robustness was a drawback that drove us into finding and adopting different strategies of research. In order to improve the coverage attribute, the seven post-processed samples were retrieved from the MG-RAST server and concatenated in one unique metagenomic dataset for CMIs proofing, variant calling and read mapping approaches (Fig. 2).

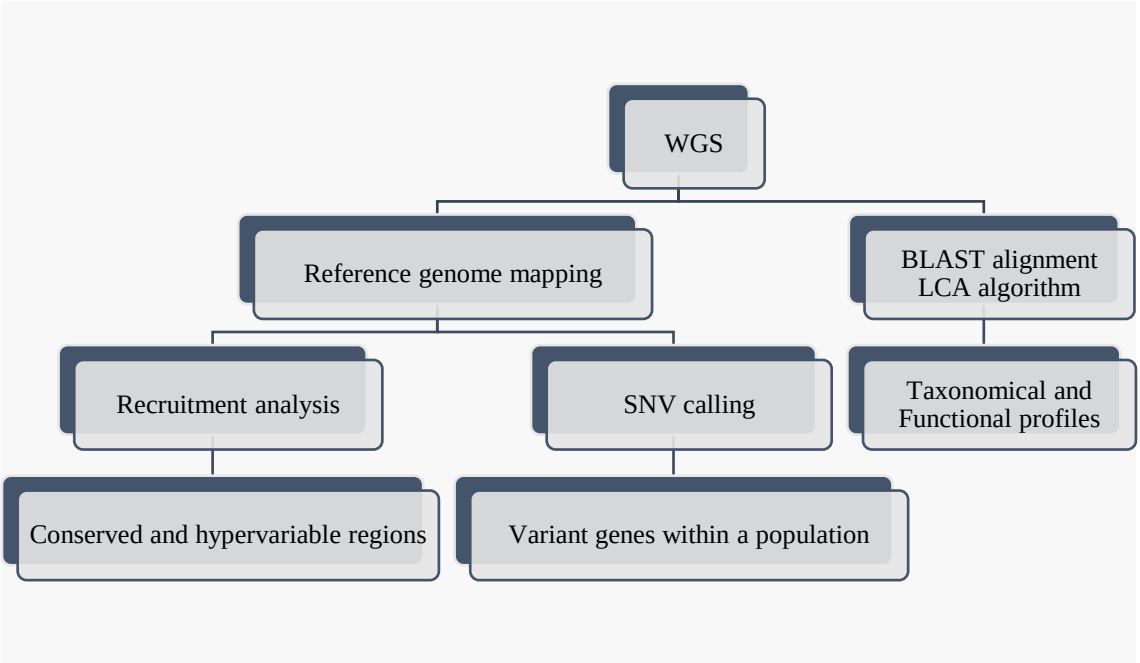


Fig. 2 - Different approaches adopted to the sequences generated through Whole Genome Shotgun (WGS) strategy in Marimbondo reservoir samples. The sequences were aligned to broad nucleotide or protein databases, composed of all available reference sequences, generating taxonomical or functional profiles of the whole community of reads from a sample. On the other hand, the metagenomics reads can be aligned to one reference genome of interest, usually the one with highest sequencing coverage (number of reads by genome position). The subsequent analysis can either describe the pattern of similarity from the whole community to the reference genome, namely the Recruitment analysis, or describe the genomic variability observed in environmental reads which properly aligned to the reference genome, through Single Nucleotide Variant (SNV) calling. Concerning the Recruitment analysis, the recruited reads represent the conserved genetic content of the reference genome within the sample, showing high similarity and sequencing coverage, while the hypervariable regions are characterized mainly by a low sequencing coverage, or even the complete absence of reads matching to some region. The SNV strategy applied to WGS datasets identifies the genes which present polimorfisms within a population of sequences mapped to the reference genome, i.e. allows to determine different genotypes within a population.

Hypervariable Regions

The first step was the alignment of all metagenome reads of the samples with higher *Microcystis* biomass, higher number of *M. aeruginosa* strains and recruited fragments to a custom database constructed by all available *Microcystis* sequences from NCBI nucleotide database through a BLASTn comparison, with a minimum *E-value* of 1e-5. A filter of 65% identity and minimum alignment length of 50bp was applied according to Steffen et al., (2012). All *Microcystis* BLAST hits from each dataset were then recruited to the reference genome using PROmer algorithm with default parameters. The PROmer algorithm was adopted because all matching and alignment routines are

performed on the six amino acid translation frames of the DNA input sequence. The show-coords and show-tiling utilities were used to locate and extract alignment information. Show-tiling determines the best mapped location of each query contig and help in the identification of syntenic regions and their contig's mapping to the reference (http://mummer.sourceforge.net/manual/).

Hypervariable regions were initially identified through coverage plot, generated by MUMmerplot utility, corresponding to regions larger than 10 kb, with low or none alignment, whose location were symmetric among all samples, regardless of other attributes. Such hypervariable regions were then called Candidate Metagenomic Islands (CMIs). The show-coords output, which presents the coordinates and alignment summary of all alignments present in the delta file, was used to locate the low % identity and/or low coverage alignments actually present in those regions. For each CMI, the genes annotated in that region were identified. Aditionally we looked for the Metagenomics Islands previously determined for *M. aeruginosa* according to Steffen et al., (2012). All *Microcystis* sequences and reference genome *M. aeruginosa* NIES-843 were retrieved from the NCBI nucleotide database and NCBI FTP site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Microcystis_aeruginosa_NIES_843_uid59101). The annotation of the reference genome was downloaded in ptt format and the database of Clusters of Orthologous Groups of proteins (COGs) was adopted in order to search for the function of genes within each CMI (Tatusov et al. 2000).

Read mapping and CMIs proofing

The unique metagenomic dataset, adopted in order to improve the coverage attribute, was mapped to fifteen reference genomes (*M. aeruginosa* NIES-843, PCC9443, PCC9809, PCC9717, PCC9807, PCC9432, PCC9806, PCC9808, PCC7941, PCC9701, NIES-44, DIANCHI905, SPC777, TAIHU98 and *M.* sp. T1-4) using Burrows-Wheeler Aligner-BWA-mem (Li & Durbin, 2009) and PROmer. A delta-filter of 97% of identity and one-to-one mapping of reference to query, allowing for 10 bp overlap, was applied to measure the breadth of coverage, while a filter of 99% identity and one-to-one mapping was applied to generate the set of metagenomic reads with unique and high identity alignment for each reference genome. The coverage summary was generated using

BEDtools Coveragebed function. The annotation of the BWA mapped reads according to generic feature format file-gff3 (downloaded from ftp://ftp.ensemblgenomes.org/pub/bacteria/release-30/gff3/bacteria_12_collection/) was performed using BEDtools insersectBed function. This function screens for overlaps between two sets of genomic features and allows the reporting of the original entries that overlapped, plus the number of base pairs of overlap between the two features. The output was handled to provide the coordinates of each read against the reference, the query identifier (Qid), the gene names, reference gene coordinates and size of overlap. The mapped reads were converted and sorted using Picard tools v. 1.138 (http://picard.sourceforge.net/) and the summary statistics and coverage were obtained using SAMtools and Genome Analysis Tool Kit v.3.4 (GATK).

Single Nucleotide Variant calling

The Single Nucleotide Variant calling (SNV) is na importante tool to the evaluation of intraspecific diversity, which has been suggested to comparatve genomic analysis of microbial community, specially to environemntal metagenomic samples, since the punctual genetic variations at nucleotide level can be detected even among reads from different populations (Nielsen et al., 2011; Olson et al., 2015), pottentially pointing to shared genes working on populational and community functional dynamics. Initially the base qualities were empirically recalibrated and indel realignment was performed for each dataset using the GATK tools. The variant calling of the fourteen realigned datasets (except NIES-44) were performed using Haplotypecaller tool using the following parameters: ploidy 1, minimum phred-scaled confidence threshold in which variants should be emitted from 10 and called from 20. To calculate the empirical allele frequency we extracted the AD (allele depth) and DP metrics from the format field of the VCF output. The allele depth (AD) calculation as performed by HaplotypeCaller consider only reads that statistically favor one allele over the other and the sum of AD may be different than the individual sample depth, especially when there are many non-informative reads. The depth of coverage (DP) from format field only consider the filtered reads. In this way, the variant allele frequency was taken as the ratio of AD and DP metrics from format field. The variants were then annotated using SnpEff tool (Cingolani et al., 2012a), focusing only on coding regions, and the output was handled using SnpSift tool

(Cingolani et al., 2012b). The variants are classified according to protein transcription impact, which could be synonymous (the change on a nucleotide does not changes the amino-acid and has no impact in protein transcription) or non-synonymous (the change on a nucleotide changes the amino-acid, potentially impacting protein transcription) (Fig. 3). Among the non-synonymous variants, we focused on high impact (disruptive) variants, according to SnpEff categories, which probably cause protein truncation, loss of function or trigger nonsense-mediated decay, working as indicators of rearrangement and genomic exchanges (Cingolani et al., 2012a; Rosenthal et al., 2015). All tables were then uploaded to a custom MySQL database in order to perform the subsequent analysis and functional annotation according to Gene Ontology database – GO (Ashburner et al., 2000, Gene Ontology Consortium, 2015).

Genomic DNA	TTC	TTT	ATC	TCC	TGC
mRNA	AAG	AAA	UAG	AGG	ACG
Amino acids	Lys	Lys	Stop codon	Arg	Thr
Variant effect	-	Synonymous	Non-Synonymous		

Fig. 3 - Effects of a single nucleotide variant on transcription (mRNA) and translation (amino acids). The variant effect can be synonymous (blue), meaning that there is no subsequent change in protein synthesis, or non-synonymous, meaning that the changing in the mRNA transcribed modifies the codon, translating different amino acids or even disrupting the amino acid sequence that would constitute a protein.

3. RESULTS

Microcystis Biodiversity Profile

The samples were uploaded to the MG-RAST platform and can be found under the accession numbers 4636015.3 to 4636028.3. The ‘best hit’ for the taxonomical classification according to M5RNA database, which integrates SILVA, Greengenes and RDP databases, pointed to a *Microcystis* population composed by 29 species and 45 *M. aeruginosa* strains. Since the taxonomical profile presented a pattern distinct from BLAST and read mapping approaches, along with a very low depth of coverage, the MG-RAST taxonomical profile for *Microcystis* population was not considered in this work.

According to the biodiversity profile generated by the BLASTx/MEGAN approach, *Microcystis* sequences ranged from 12% of total cyanobacteria assignments in sample T1 to 68% in sample T5, following the increase in *Microcystis* biomass (see Chap. 2 for more details), and from 56% in sample ST4 to 61% in sample S6 for spatial sampling. Along with the increasing abundance of *Microcystis* population in the high biomass samples, the number of distinct strains also increased (Fig. 4), what could be a result of the higher sequencing coverage (1.44 for sample T1 and 6.21 for sample T5 for the most abundant strain NIES-843, according to BWA alignment and SAMtools depth function), allowing to the LCA algorithm to place the reads in more specific nodes. The reference strains *M. aeruginosa* NIES-843, PCC9443 and PCC9717 were present in all samples. These strains presented the highest contribution in temporal sampling, showing a drecreasing amount of assignments along with the increase in *Microcystis* biomass and a similar contribution among spatial samples (Fig. 4). The strains PCC7941 and PCC9432 were present only in samples with higher biomass (T5, S6, S7 and S8). The summary of sequencing attributes is presented in the Supplementary Table S1.

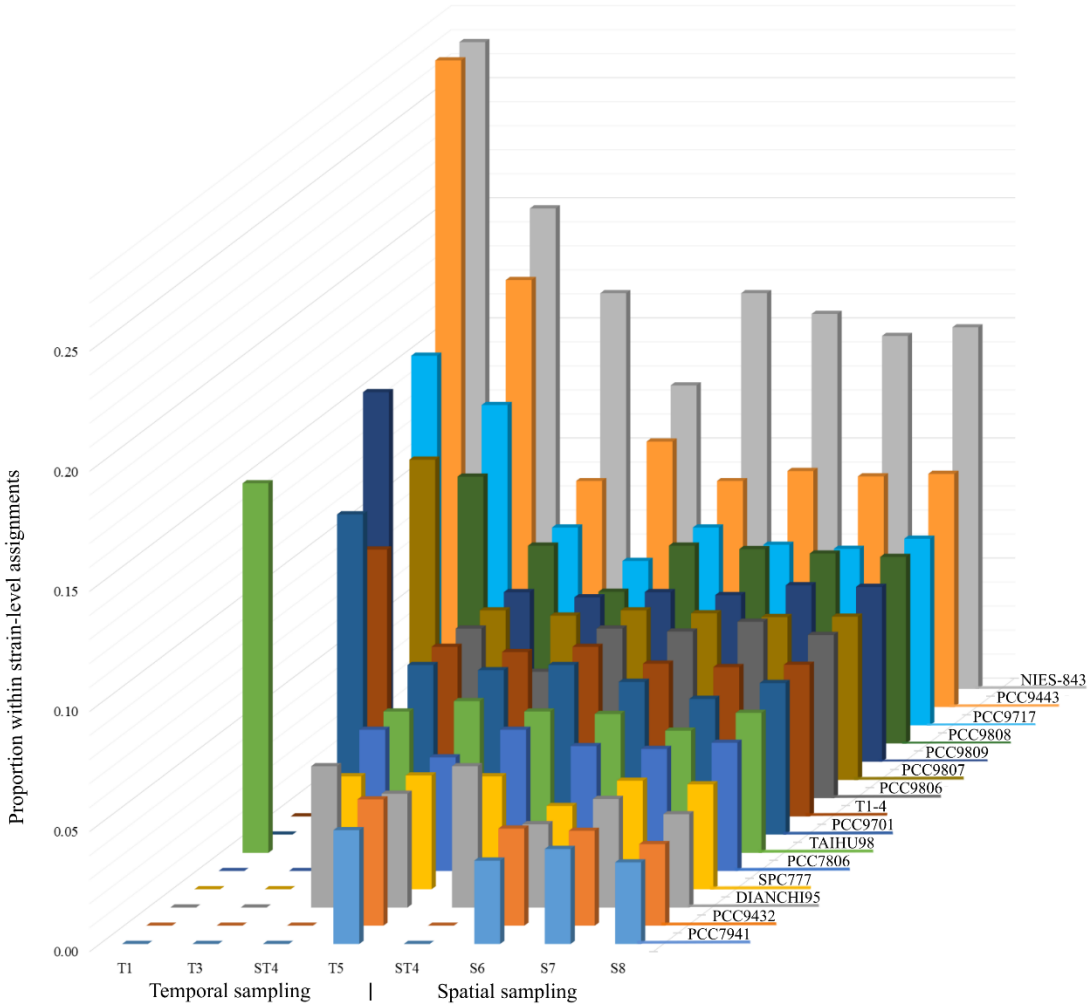


Fig. 4 - Microdiversity of *M. aeruginosa* population in the temporal (T1, T3, ST4, T5) and spatial (ST4, S6, S7, S8) samples, according to the proportion of strain-level assignments for each sample.

Recruitment Analysis

The initial step in the search for hypervariable regions was the visual comparison of coverage plots generated with MUMmerplot, using the output of PROmer and Delta-Filter of each sample as input (Fig. 5). The summary of recruitment is presented in table 1, showing the low sequencing coverage of each sample. Thirteen low depth coverage regions, including six metagenomic islands previously described by Steffen et al., (2012), were analyzed in terms of gene content and alignment attributes (Tab. 2). All regions showed lower average of alignment identity (91.8%), similarity (93.5%) and coverage (61%) of query to reference genome. The CMIs of each sample showed some differences in position and length. The general coordinates, length, gene content and alignment

attributes of all CMIs and SMIs, along with the flanking regions (regions before and after each CMI, with the same length) are presented in Supplementary table S3. The major COG families among CMIs were Replication, recombination and repair (L), Function unknown (S), General function prediction only (R), Defense mechanisms (V), Posttranslational modification, protein turnover, chaperones (O), Cell wall/membrane/envelope biogenesis (M) and several Hypothetical proteins. The complete list of coordenates, genes and COG annotations of the thirteen CMIs are presented in Supplementary table S4. As expected for the *Microcystis aeruginosa* genome, a large number of genes encoding hypothetical proteins, transposases, predicted general function proteins and function unknown were found over the whole genome (Kaneko et al., 2007; Steffen et al., 2012; Frangeul et al., 2008). The shared genome or low-frequency genes encompasses an exchangeble community gene pool, typically encoding several transposases and integrases (L), related to DNA turnover (Coleman et al., 2006; Cordero & Polz, 2014; Frangeul et al., 2008). The restriction modification systems, which behave as mobile genetic elements and may induce genome variability (Kobayashi 2004), belong to defense mechanisms COG family (V) and were found in four CMIs (Fig. red circles and blue triangles), namely IS element (insertion element) and MITE (Miniature Inverted-repeat Transposable Elements).

Tab. 1 - Summary of recruitment analysis of the individual metagenomics datasets to the most abundant *Microcystis* strain *M. aeruginosa* NIES-843. From total number of post-processed sequences, the proportion of *Microcystis* sequences pointed by BLASTn and alignment filters are presented. The *Microcystis* sequences were then recruited to *M. aeruginosa* NIES-843 genome. The depth of coverage of the recruited sequences ranged from 1.3 to 12X.

	RL1	**RL3**	**RL4**	**RL5**	**RL6**	**RL7**	**RL8**
Total seqs	156648	191694	133480	272381	207192	273083	131760
% Microcystis seqs	3.69	2.28	10.13	21.70	20.40	11.33	43.85
% NIES-843 recruitment	35.44	40.11	35.24	26.81	29.68	31.32	27.30
% NIES-843 coverage	1.31	0.92	3.57	5.82	6.05	3.55	11.97

Tab. 2 - Position, length and gene content of the Candidate Metagenomic Islands (CMIs) of *M. aeruginosa* NIES-843 for the Marimbondo Reservoir. The genes comprised between the initial and final genes represent the hypervariable regions according to local *Microcystis* populations.

	Start	**End**	**Length**	**Initial Gene**	**Final Gene**
CMI-1	35073	53810	18737	MAE00490	MAE00720
CMI-2	119506	138612	19106	MAE01450	MAE01700
CMI-3	419030	432679	13649	MAE04720	MAE04840
CMI-4	440125	463705	23580	MAE04950	MAE05210
CMI-5	1125258	1145978	20720	MAE12730	MAE12890
CMI-6	1872886	1885373	12487	MAE20990	MAE21110
CMI-7	2230040	2253990	23950	MAE24680	MAE24880
CMI-8	2669299	2682840	13541	MAE29240	MAE29440
CMI-9	2813778	2829267	15489	MAE30870	MAE31100
CMI-10	3615690	3629905	14215	MAE39310	MAE39510
CMI-11	3760041	3776694	16653	MAE40940	MAE41100
CMI-12	3813793	3828466	14673	MAE41460	MAE41610
CMI-13	4673904	4684868	10964	MAE50850	MAE50990

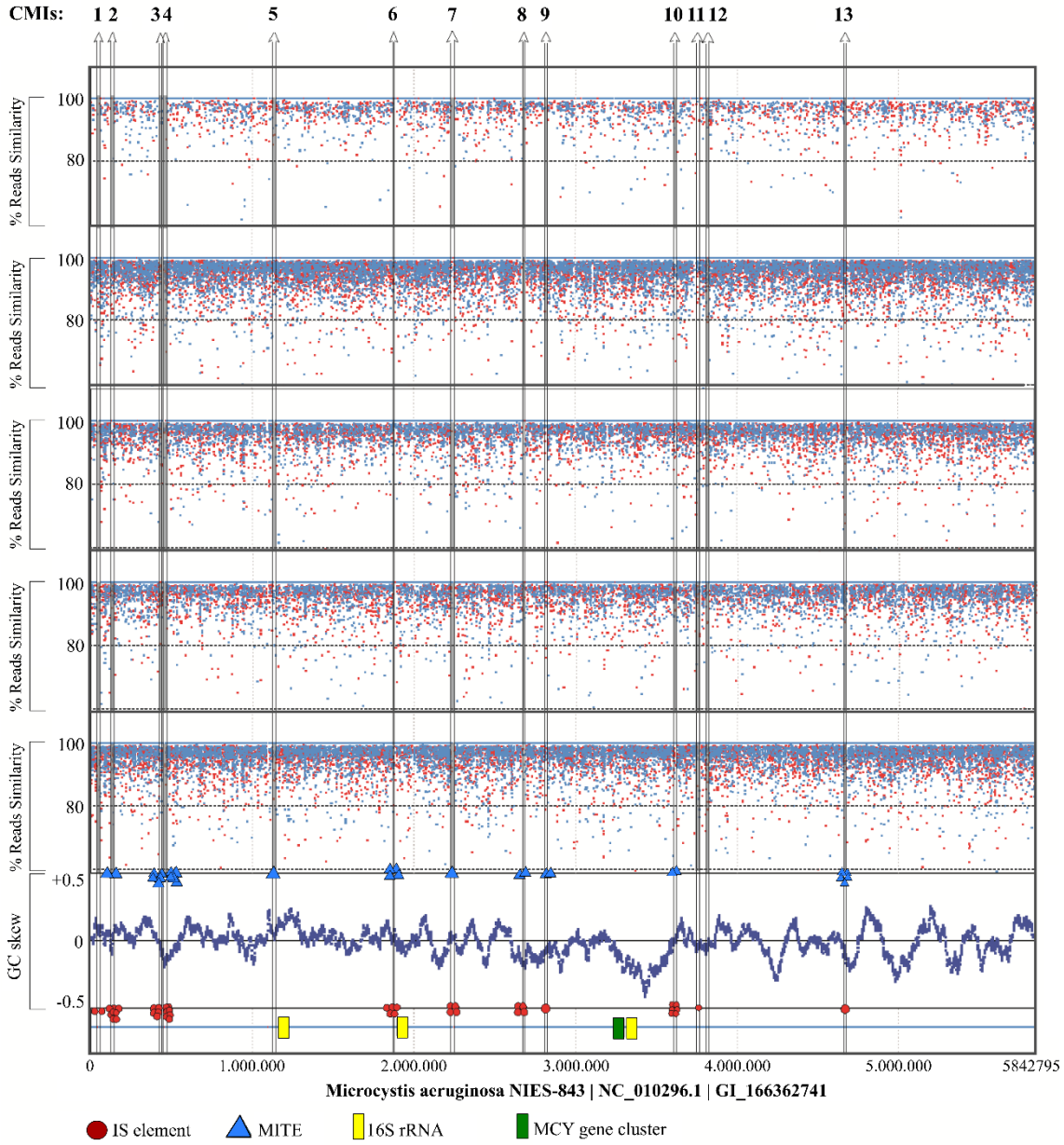


Fig. 5 - Coverage plot (% similarity of metagenomics reads along the complete reference genome) showing the recruited fragments and low similarity (below 95%) alignments to reference genome *M. aeruginosa* NIES-843 and the Candidate Metagenomic Islands (CMIs) for the five individual samples under high *Microcystis* biomass. The vertical arrows point to the symmetric position of the thirteen CMIs, greyish fill indicates the presence of the CMI in each sample, while blank fill indicates the absence of the CMI in each sample. The placement of mobile elements near to CMIs are indicated by red circles (insertion elements) and blue triangles (Miniature Inverted-repeat Transposable Elements). 16S rRNA gene clusters are represented by yellow rectangles and the microcystin (*MCY*) gene cluster is represented by green rectangle.

Since the individual samples presented low coverage against the dominant strain NIES-843 (6-10X), the seven post-processed samples were retrieved from MG-RAST server and concatenated in one unique metagenomic dataset. This unique dataset presented enough coverage (22-25X, Suppl. Table S2) and was used to check each hypervariable region. We could confirm the alignment attributes of all regions, which were then called Candidate Metagenomic Islands (CMIs), since further analyses are needed to delimitate their boundaries (Tab. 2 and Fig. 6). The coordinates, length, gene content and alignment attributes of the thirteen CMIs are presented in Supplementary Table S3. Due to the increased reads abundance and consequent sequencing coverage of the unique dataset, the coverage plot presented a large amount of reads aligned from 85% to 100% of similarity (Fig. 6), encompassing genera and species levels. With the increased depth of coverage of the unique dataset, some low coverage regions observed in individual samples (Fig. 5) were not detected in the unique coverage plot, but CMIs 3, 4, 7, 8, 9, 10, 11 and 12 could be clearly visualized.

These hypervariable regions represent reference genes absent or under variation within the population, while the recruited reads are part of the sequence-discrete population, which is composed by reads mapped above 99% nucleotide identity to NIES-843, i.e. the population of high similarity reads which can be individually annotated as genes of different species or strains according to current nucleotide and protein databases, not necessarly belonging to the reference genome population, and closely related populations whose reads mapped below 90% identity (Bendall et al., 2016). These two structural features (recruited reads and hypervariable regons) could be seen as guidelines to map the diversification dynamics, since the reference-based hypervariable genes described (Tab. 2) could indicate potential targets for populational linkage, *i.e*. the genes and respective functional roles which group a set of distinct populations through metabolic exchanges or complementarity, in order to improve the ecological fitness and reach the community stability. Such diverse functional annotation, tracked within hypervariable regions in a mostly recruited reference genome, could be related to the mosaic nature of *M. aeruginosa* populations (Rosen et al. 2015) and the alternate genomic structures (Steffen et al., 2014), meaning that a population alone is not able to deal with environmental fluctuations, requiring a set of distinct sub-specific members which present differential gene content and whose relative abundance in the sample could be related to population or community demands. The reference-based sequence-discrete population could set the gene pool for both population/community and flexible/core

genome compartments (Beck et al., 2012; Bendall et al. 2016), according to identity cutoffs and reference genomes availability.

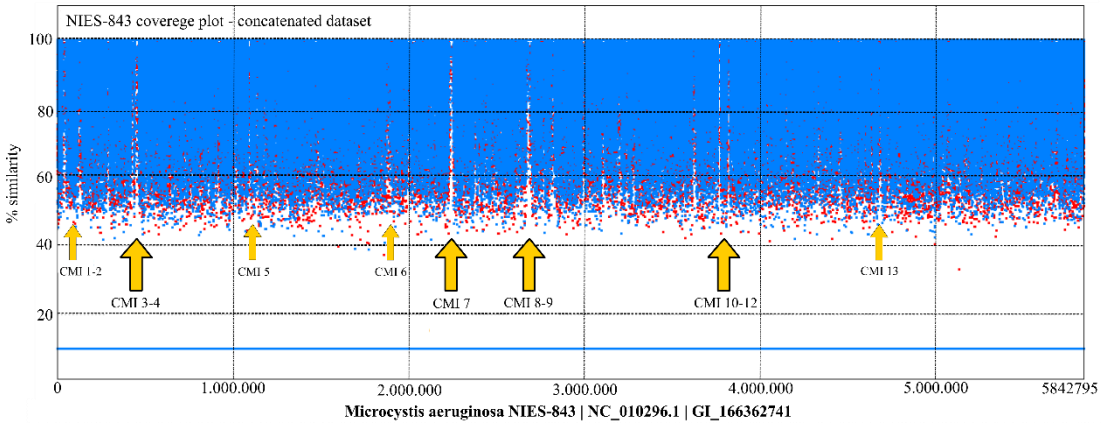


Fig. 6 Coverage plot (% similarity of metagenomics reads along the complete reference genome) showing the recruited fragments from concatenated metagenomic dataset to the reference genome *M. aeruginosa* NIES-843, showing the Candidate Metagenomic Islands (CMIs) previously described individually (CMI 1-13) and highlighting the CMIs which presented low coverage above 80% of similarity (larger arrows).

Among the reads recruited under 99% identity (highlighted in Suppl. Table S4), we could observe some differential alignments, presenting non-synonymous variants in eight of the thirteen CMIs (Tab. 3). The recruited reads (showing similarity to reference genome above 99%) located within the boundaries of hypervariable regions are part of the *M. aeruginosa* NIES-843 sequence-discrete population and can be indicators of target genes under diversity dynamics driven by local environmental pressures, reinforced by the presence of moderate impact non-synonymous variants (Cordero & Polz, 2014). The functional annotation among the recruited reads within CMIs encompassed mostly hypothetical/uncharacterized proteins and transposases, which are directly related to diversity dynamics, being regulated by environmental conditions and driven genome rearrangements (Steffen et al., 2014). CMI-6 originally comprised the genes from MAE_20990 to MAE_21100, most of them transposases and hypothetical proteins, but when checking this region in the concatenated dataset, we could observe an extension of that region with lower depth of coverage than the average (Suppl. Table S4). Interestingly, this extended region mapped against ribosomal genes in samples T5 and S6, along with two ABC transporters, a Penicillin-binding protein and a glutathione dehydrogenase in samples T5 to S8, all of them presenting moderate impact variants. The lower depth of

this region, associated to moderate impact variant occurrence, functions annotated and the GC skew observed (Fig. 5, just after CMI-6), indicate that the variant genes observed in this extended region could be under medium-frequency dynamics (Cordero & Polz, 2014), which are not located in hypervariable regions, but take part of the metabolic trade-offs, i.e. the environmental pressures may favour some functional traits, changing the frequency of shared genes related to these metabolic requirements, even among genes working on central metabolism.

Tab. 3 - Differential gene alignments, showing some low coverage and non-synonymous variant impact among the metagenomic reads recruited under 99% identity to *M. aeruginosa* NIES-843 reference genome, in eight of the thirteen CMIs. avgDP: average read depth after quality filtering step; avgAD: average allelic depths for the environmental alleles constituting a poliformism; Freq: ratio of polimorfisms and read depth (aflter filtering steps). Freq=1 indicates that all reads aligned presented the poliformic allele.

Samples	Gene	Protein	Impact	avgDP	avgAD	Freq
	CMI-1					
T5	MAE_00520	hypothetical protein	-	-	-	-
	MAE_00680	hypothetical protein	mod	2.4	2.4	1
	MAE_00590	peptidase PatA-like protein	mod	3.8	3.8	1
	MAE_00640	peptidase PatA-like protein	mod	3.9	3.9	1
	MAE_00710	aldo/keto reductase	mod	14.2	13	0.9
	CMI-2					
S6	MAE_01440	hypothetical protein	-	-	-	-
S7	MAE_01520	hypothetical protein	-	-	-	-
	CMI-3					
S7	MAE_04740	hypothetical protein	mod	9.7	9.7	1
	CMI-4					
S6	MAE_05070	transposase	-	-	-	-
S8	MAE_04970	hypothetical protein	mod	11.8	11.8	1
	CMI-5					
T5	MAE_ 12730	ASPIC/UnbV protein	mod	2.3	2.3	1
	MAE_12840	beta ketoacyl-acyl carrier	mod	4.3	4.1	0.9
	MAE_12800	hypothetical protein	-	-	-	-
	MAE_12890	pirin-like protein	mod	7.2	6.5	0.9
	MAE_12810	hypothetical protein	mod	2	2	1
	MAE_12880	hypothetical protein	mod	3	3	1
	CMI-7					
T5	MAE_24880	methyltransferase	mod	12.4	11.2	0.9
	CMI-11					
T5	MAE_41100	*UreE*	-	-	-	-
	CMI-12					
T5/S7/S8	MAE_41600	peptidase	mod	7	6.8	1

Read Mapping and Variant Calling

The depth of coverage for individual samples to each reference genome presented a similar distribution (Fig. 7), with mean depth of 2X for temporal sampling and 3X for spatial sampling. The samples T5 and S8 showed the highest number of reads aligned by position for each *M. aeruginosa* strain mapped (Fig. 7). The individual samples presented a generaly low sequencing coverage, hampering the mapping interpretation, so the concatenated dataset, composed of all individual metagenomics datasets, was adopted in order to improve the coverage attribute for subsequent read mapping analysis. The summary of alignments of the concatenated samples against the fifteen complete reference genomes of *M. aeruginosa* is presented in Supplementary Table S2. Twenty two to 25% of the reads of the concatenated dataset mapped against all reference genomes, presenting 18 to 22 average depth of coverage and 0.77 to 0.84 breadth of coverage (Fig. 8 dashed line). The breadth of coverage represents the proportion of bases of the whole reference genome that was covered by the metagenomics reads. Two to 3% of the reads aligned uniquely with 99% identity to each reference genome.

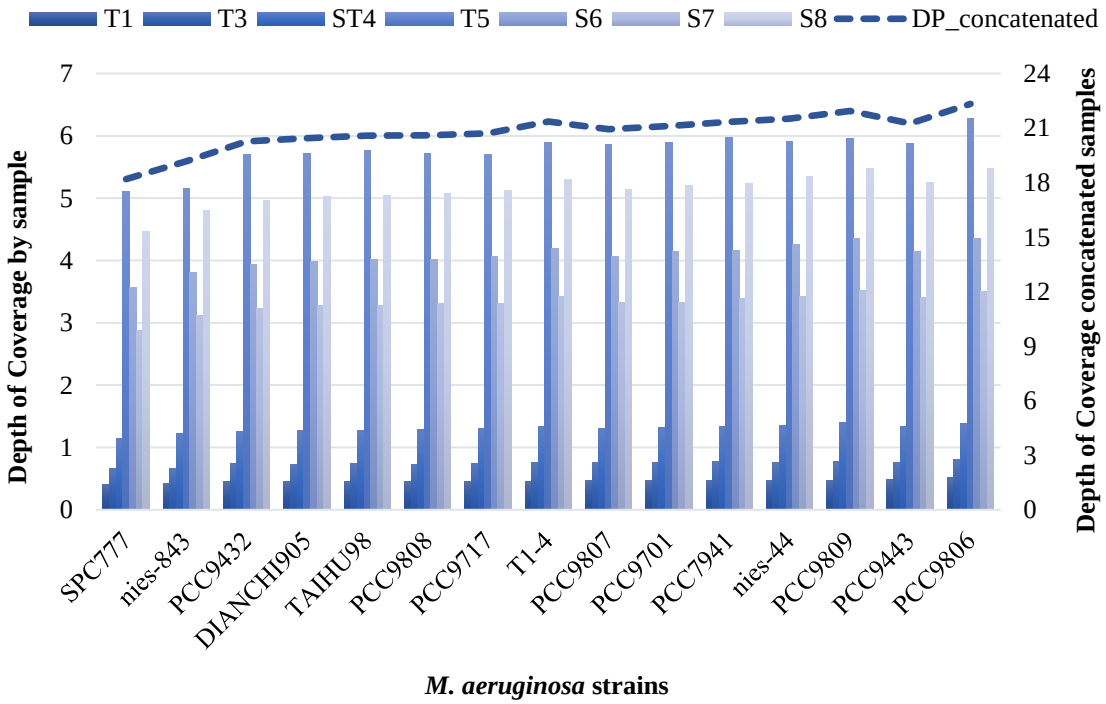


Fig. 7 - Depth of coverage (DP) for individual metagenomic datasets mapped to 15 *M. aeruginosa* complete reference genomes. Bars represent individual DP, dashed line represent the overall DP for the concatenated dataset aligned to each reference genome.

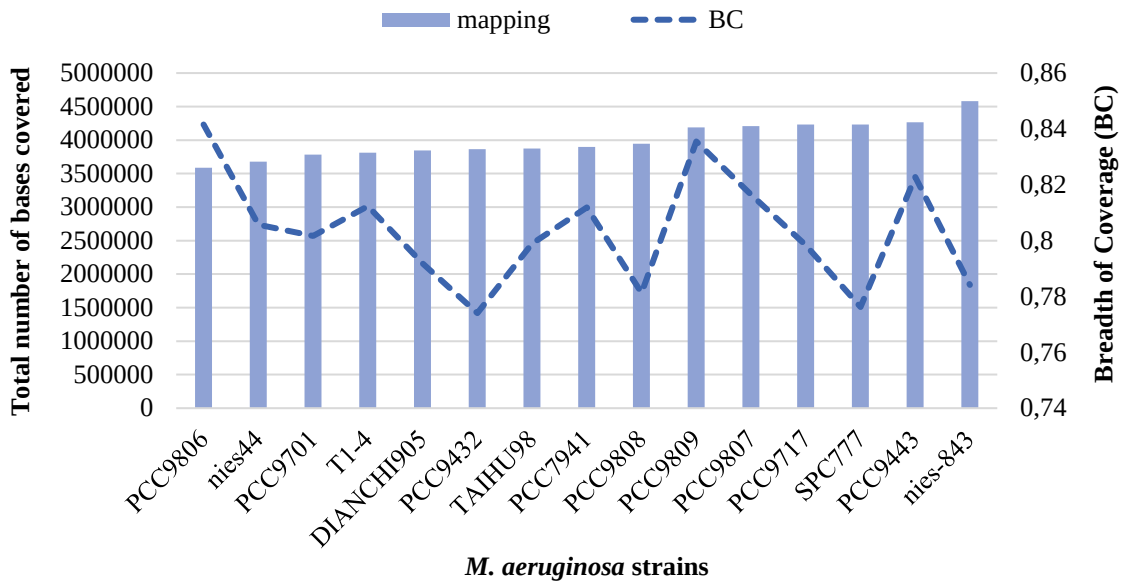


Fig. 8 - Total number of bases from concatenated metagenomics datasets mapped to 15 *M. aeruginosa* complete reference genomes at mininum coverage depth of 5 reads (bars) and breadth of coverage (dashed line) of reads mapped under BWA-mem default parameters to each *M. aeruginosa* strain. The breadth of coverage was measured using SAMtools mpileup function by dividing the number of total mtagenomic bases mapped at minimum 5 DP by total length of each reference genome.

The strains PCC9809, PCC7941 and PCC9806 presented the highest depth of coverage. According to the total number of metagenomic bases mapped at minimum 5X DP to each reference genome (Fig. 8) the proportion of reference genome covered was higher for strains PCC9809 and PCC9806, and lower for strains PCC9432, PCC9808 and SPC777. In fact, the strains PCC7941 and PCC9432 were assigned only in samples T5, S6, S7 and S8 according to LCA algorithm, but were mapped for all samples. The reads mapped sparsely over the whole reference genomes with a heterogeneous depth of coverage (Suppl. Table S5). The contigs with average depth of coverage distribution above the 75 percentile and the respective genes, among these contigs, also with average depth of coverage distribution above 75%, were analyzed in more detail and are presented in Supplementary Table S6. Twenty to 35% of the most abundant genes were uncharacterized proteins and 3-7.5% were transposases. NIES44 presented the highest amount of mobile element proteins. DIANCHI905 and NIES843 had seven microcystin genes among the most abundant set of genes (Tab. 4). The highest depth of coverage was seen in *mcyD*, followed by *mcyE* and then *mcyG*. The lack of *mcy* genes in NIES44, TAIHU98 and PCC9806 are in accordance to previous studies (Okano et al., 2015; Hu et al., 2015). Four cyanopeptolin synthetase genes (*mcnB*, *mcnC*, *mcnE*, *mcnF*) were among

the most abundant genes in nine genomes, mostly DIANCHI905, PCC9808 and NIES843. DIANCHI905 and TAIHU98 were the only strains which presented aeruginosin synthetase genes among its major set of genes (Tab. 4).

Tab. 4 - Number of metagenomic reads mapped to non-ribosomal peptide synthetases genes (rows) within the 75 percentile set of the most abundant contigs and genes for all *M. aeruginosa* reference genomes (columns). Reference genomes absent in this table presented no mapping or mapping below required alignment thresholds.

	DIANCHI 905	**PCC 7941**	**PCC 9809**	**PCC 9808**	**PCC 9807**	**PCC 9717**	**PCC 9701**	**PCC 9443**	**PCC 9432**	**TAIHU 98**	**NIES 843**
Microcystin biosynthesis											
mcyA	405	406	374	387	420	108	0	235	0	0	389
mcyB	293	297	165	168	292	0	0	302	0	0	283
mcyC	195	200	0	0	199	194	0	196	0	0	202
mcyD	586	585	586	586	586	0	0	585	0	0	586
mcyE	497	496	498	497	496	0	0	496	0	0	496
mcyG	420	418	418	422	421	0	0	420	0	0	418
mcyH	95	0	107	106	0	0	0	0	0	0	107
Cyanopeptolin biosynthesis											
mcnB	123	0	0	129	0	0	0	0	0	0	121
mcnC	815	838	0	836	0	236	0	0	728	0	847
mcnE	242	235	261	239	268	139	241	0	257	0	234
mcnF	101	102	103	103	0	0	0	0	103	0	102
Aeruginosin biosynthesis											
aerA	318	0	0	0	0	0	0	0	0	317	0
aerB	0	0	0	0	0	0	0	0	0	208	0
aerJ	105	0	0	0	0	0	0	0	0	203	0
aerK	0	0	0	0	0	0	0	0	0	86	0
aerN	129	0	0	0	0	0	0	0	0	0	0
Unclassified non-ribosomal peptide synthetases											
mipB	0	118	0	0	0	0	0	0	0	0	0
MICAH_2230003	0	0	320	0	0	0	0	0	0	0	0
MICAH_2220002	0	0	240	0	0	0	0	0	0	0	0
MICAH_2500004	0	0	234	0	0	0	0	0	0	0	0
MICAB_5570005	0	0	0	0	0	318	0	0	0	0	0
MICAB_5570012	0	0	0	0	0	240	0	0	0	0	0
MICAB_2620006	0	0	0	0	0	146	0	0	0	0	0
MICAK_2550003	0	0	0	0	0	0	235	0	0	0	0
MAE_56590	0	0	0	0	0	0	0	0	0	0	321
MAE_56520	0	0	0	0	0	0	0	0	0	0	240

We compared the gene content with the Minimal Gene Set (Gil et al., 2004) to evaluate the consistency among the fifteen reference genomes. From 205 genes composing the Minimal Gene Set, 174 genes presented a heterogeneous distribution among the fifteen strains (Suppl. Table S7). *M. aeruginosa* strains DIANCHI905 and TAIHU98 presented the highest number of genes mapped. Through GATK variant call pipeline, we tracked for polimofisms among such housekeeping gene set and 61 genes presented high impact variants, with distinct occurrence according to each *M. aeruginosa* strain (Suppl. Table S8). The high (disruptive) impact in protein probably causes protein truncation, loss of function or triggers nonsense-mediated decay. In this way, high impact variants can be indicators of rearrangement and genomic exchanges (Cingolani et al., 2012a; Rosenthal et al., 2015). Two genes, *dnaG* and *dnaK*, part of the most abundant and frequent set of genes, presented high impact variants for all fourteen genomes analyzed through the GATK variant call pipeline, with variant allele frequency of 1.0 for most strains (Suppl. Tab. S8). The Thioredoxin gene *trxA*, part of the nucleic acids metabolism and essential for photosynthetic growth, presented high impact variants in twelve of fourteen genomes (Tab. 5), most of them with variant allele frequency of 1.0 and presented large codon deletions on six genomes. Other five genes showed high impact variants frequently distributed (*metK*, *pheT*, *polA*, *pyrG*, *rpoC*). The allele depth, depth of coverage of filtered reads and variant allele frequency of 61 genes in each strain are presented in Supplementary Table 8.

The occurrence (1 or 0) of some housekeeping and related genes among the fifteen genomes is presented in Table 6. Protein names and associated gene ontology are presented in Supplementary Table S7. A mosaic distribution of genes can be seen among the strains, mainly for genes with contribution above 75%, as observed by Steffen et al., (2014) for *M. aeruginosa* populations. We highlight the more frequent distribution of the ATP-dependent zinc metalloprotease *FtsH*, an integral membrane component, and the cell division protein *FtsZ*, responsible for protein polymerization. The Phycocyanin alpha and beta subunit were also frequently distributed, along with the Protease 4 *sppA*. Among the major genes whose contribution was below 75%, we observed a broad distribution of bicarbonate transport ATP-binding and Nitrate transport ATP-binding genes, related to nitrate transmembrane transporter activity.

Tab. 5 - Summary of total variants detected for *trx*A gene among fourteen *M. aeruginosa* genomes (rows). Low=low impact variant, Mod=moderate impact variant, High=high impact variant, S_cod=synonymous coding effect, NS_cod=non-synonymous coding effect, FS=frame shift effect, Fq=variant allele frequency calculated by AD/DP ratio, Ref=Reference DNA stretch, Alt=alternative/query variant. Average mapping quality ranged from 150.7-278.

		Low	**Mod**	**High**			
	Total	**S_cod**	**NS_cod**	**FS**	**Fq**	**Deletions**	**Effect**
Nies843	43	21	18	4	1.0		
PCC9432	24	8	15	0	1.0	HE972530/3694	Stop_Lost+Codon_Deletion
PCC9717	25	14	10	0	1.0	HE972704/9355	Stop_Lost+Codon_Deletion
PCC9701	39	21	15	2	1.0	CAIQ01000111/3803	Frame_Shift+Stop_Lost
PCC9809	29	14	13	2	1.0		
PCC9807	26	16	9		1.0	HE973360/553837	Stop_Lost+Codon_Deletion
PCC9806	23	9	12	1	1.0	HE973251/25412	Frame_Shift+Stop_Lost
PCC7941	37	20	16	0	1.0	HE973150/37801	Stop_Lost+Codon_Deletion
TAIHU98	20	8	8	3	1.0	contig2/345187	Stop_Lost+Codon_Deletion
DIANCHI905	24	11	10	3	1.0		

Tab. 6 - Occurrence (1 or 0) of some housekeeping genes (rows) in each *M. aeruginosa* reference genome (columns) observed among the 75 percentile depth of coverage mapping features.

Genes with depth of coverage above 75 percentile															
	DIANCHI 905	**PCC 7941**	**PCC 9809**	**PCC 9808**	**PCC 9807**	**PCC 9806**	**SPC 777**	**T1-4**	**PCC 9717**	**PCC 9701**	**PCC 9443**	**PCC 9432**	**NIES 44**	**TAIHU 98**	**NIES 843**
ftsH	1	1	1	1	1	1	0	1	1	1	1	0	0	1	1
ftsI	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0
ftsX	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
ftsY	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
glnA	1	0	0	0	0	1	0	0	1	1	1	1	0	1	1
glnH	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
glnN	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
glnP	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
gatA	0	0	1	0	1	1	1	0	0	0	1	0	0	0	1
gatB	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
gltX	1	0	1	0	0	0	0	0	1	0	0	1	0	1	1
gltS	1	1	1	0	0	1	0	1	1	1	0	1	0	1	0
gltA	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0
gltB	0	0	0	1	0	1	0	1	1	1	1	1	0	1	1
gltD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
gyrB	0	1	0	1	0	0	0	0	0	1	0	0	0	1	1
gyrA	0	0	0	0	0	1	0	0	1	0	0	0	0	1	1
pgi	0	0	0	0	1	1	0	1	0	0	0	0	0	1	1
recA	0	0	0	1	0	1	1	0	0	0	0	0	0	0	1
recF	1	0	0	1	0	1	0	1	0	1	1	0	0	0	1
recQ	1	1	1	1	0	1	0	0	1	0	0	0	0	1	1
recN	0	1	1	0	1	0	0	0	1	1	1	0	0	0	1
recG	0	0	1	0	0	0	0	1	1	0	0	0	0	1	1
recJ	0	0	1	1	1	0	0	1	1	1	1	0	1	1	1
cmpC	1	0	0	1	1	1	0	0	0	0	0	0	0	1	0
nrtC	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
CphA	0	1	1	1	0	0	0	0	0	1	0	0	0	1	1
Genes with depth of coverage below 75 percentile															
ftsZ	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
tpiA	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
nrtD	1	1	1	1	1	1	0	1	1	0	1	1	0	1	0
nrtA	0	1	1	1	0	1	1	1	1	1	1	1	0	1	0
nrtC	0	1	1	1	1	1	0	1	1	1	1	1	0	0	0
nrtB	0	1	1	1	1	0	0	1	1	1	1	1	0	0	0
cmpA	1	1	1	1	1	1	0	1	1	0	1	1	0	1	0
cmpB	0	1	1	1	1	1	1	1	1	0	1	1	0	0	0
cmpC	1	1	1	1	1	1	0	1	1	1	1	1	0	1	0
cmpD	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0
ptsJ	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
sppA	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0
spsCA	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
spsA	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

3. DISCUSSION

Freshwater ecosystems are very dynamic and the selective pressures imposed on an unstable niche requires a proportionally dynamic genetic diversification, through increasing mutation rates, homologous recombination, rearrangements and lateral gene transfers (Galhardo et al., 2007; Frangeul et al., 2008; Humbert et al., 2013). The ecological strategy based on high genomic plasticity enables the microbial population to maintain an adaptive potential, where population structure changes according to environmental conditions in order to improve ecological fitness (Coleman et al., 2006; Kashtan et al., 2014; Steffen et al. 2014). The natural functions and ecological roles played by cyanobacteria secondary metabolites are still not well understood, but they are probably related to growth conditions (Briand et al., 2012; 2015). Along with the increase in microdiversity, we observed the presence of a non-microcystin producer PCC9432, exclusively in samples T5 to S8, during periods of high *Microcystis* biomass (Fig. 4). *M. aeruginosa* populations present a highly plastic genomic feature (Kaneko et al., 2007), with genes scattered throughout the genome and subjected to rearrangements according to regulator proteins and environmental conditions (Steffen et al., 2014). This mosaic structure has recently been described as random sets of variant alleles and genes whose dynamics depend on environmental pressures (Rosen et al., 2015; Desai & Walczak, 2015). Few studies have considered the actual structure of natural populations, which would enhance the screening of entire communities and allow multiple strategies, scales and connections (Cordero & Polz, 2014; Bendall et al., 2016).

Otsuka et al., (2001) unified five *Microcystis* morphospecies into a single species under the Rules of the Bacteriological Code, and NIES-843 was proposed as the type strain of *Microcystis* genus. The recruitment analysis showed a wide sequence-discrete population (Oh et al., 2011; Bendall et al., 2016), with a wide range of similarity to NIES843 genome (Fig. 3). The reads also mapped to all fifteen reference genomes, from 0.4 to 0.7 breadth of coverage and 21 to 24 depth of coverage (Fig. 8 and Suppl. Table S2). This scattered mapping (Suppl. Table S5) could be, in fact, showing an alternate genomic structure drawn by environmental constraints (Steffen et al., 2014). This mosaic-like genomic structure comprises both the core and flexible genome (Beck et al., 2012) and the strategy based on genetic diversity maintenance could be mixed-up with taxonomical diversity. The CMIs pointed to the most likely genomic regions under variation within the NIES-843' sequence-

discrete population (Fig. 5 and 6), along with the functional profile (Suppl. Table S6), mostly concerning flexible or low-frequency genes, encompassing transposases, integrases associated to DNA turnover, hypothetical and unknown proteins (Coleman et al., 2006; Cordero & Polz, 2014; Frangeul et al., 2008). Therefore, the presence of mobile elements in almost all CMIs (Fig. 5) is a feature of hypervariable regions, and its scattered distribution throughout the *M. aeruginosa* genome indicates a genome-wide variability/adaptability potential (Kashtan et al., 2014). The presence of non-synonymous coding variants in most of the genes described as >99% identity within nine of the thirteen CMIs indicates that, as expected for hypervariable regions, there is an underlying variation in course, through random mixture of alleles (Rosen et al., 2015) or selective gene sweeps (Shapiro et al., 2012), coordinated by environmental-guided regulation and covering the genome-wide structure (Steffen et al., 2014; Bendall et al., 2016). According to Pašić et al., (2009), metagenomic islands constitute a feature that is conserved within species regardless of their geographic origin. On the other hand, gene-content variation might be an evolutionary response to local, rapidly varying biotic interactions (Cordero & Polz, 2014). CMIs 2, 3, 6, 7, 9 and 11 correspond to Steffen's metagenomics islands for *M. aeruginosa* NIES843 in a bloom-impacted lake (Steffen et al., 2012). Our findings suggest that the number and distribution of hypervariable regions is an environmental induced feature, concerning the adaptation dynamics of each population.

The mapping against the fifteen reference genomes reinforces the lack of phylogeographic structure observed for *Microcystis* population (Sabart et al., 2009, Hu et al., 2015). We could observe a somewhat distinct gene content for each strain, but the higher depth of coverage for transposases and uncharacterized proteins points to a flexible, medium-low frequency set of genes playing the major roles in this system (Suppementary Table S6). The search for the Minimal Gene Set (Gil et al., 2004) and some housekeeping genes pointed to a mosaic-like gene occurrence within strains (Tab. 6), and part of them were not among the most abundant genes. In fact, the flexible genome functions presented the higher abundance in this system. Humbert et al., (2013) found a heterogeneous distribution for some of these genes among the twelve strains described in their work. Despite the wide SNV distribution and impact found within the minimal gene set, none of them was mapped in a CMI (Suppl. Table S4), showing that diversity dynamics can be observed even among conserved genes (Fraser et al., 2009). Several high impact variants were observed within this gene set (Suppl. Table S7). Two genes, *dnaG* and *dnaK*, part of the most abundant and

frequent set of genes, presented high impact variants in all strains analyzed through SNV calling. *DnaG* is a DNA primase, essential for the replication machinery, while *dnaK* is part of the chaperone system, involved in protein folding and translocation of proteins through biological membranes (Gil et al., 2004). The cyanobacterial m-type *Trx* gene (*TrxA*) receives electrons from the photosynthetic chain via ferredoxin and ferredoxin-dependent thioredoxin reductase and is tightly linked to photosynthetic activity in response to light (Gil et al., 2004; Alexova et al., 2011). The *TrxA* active site usually presents a high degree of sequence identity (Florencio et al. 2006), but for this population, a high impact and high amount of variants were found in that gene. Another important set of variants (six frameshifts, 14 non-synonymous and 60 synonymous coding) was found for the RNA polymerase subunit beta *rpoC2*, that encodes components of the basic transcription machinery. The gamma subunit *rpoC1* is considered a cyanobacterial discriminatory marker (Yoshida et al.. 2008). The generally higher proportion of synonymous over non-synonymous variants suggests that most genetic variations within this sequence-discrete population might be neutral, thus allowing many highly similar genotypes to coexist without outcompeting each other (Bendall et al., 2016), but the broad distribution of high impact variants observed in this population could also show key genes whose activity may be linked to bloom development.

3. CONCLUSIONS

The microdiversity observed in this study could be considered a taxonomical bias, masking microbial interactions that spans all taxonomic ranks while trying to fit genomic information in coherent artificial units (Konopka, 2009; Zarraonaindia et al., 2013; Boon et al., 2014). The progressive increase in microdiversity that was observed together with the increase in biomass and microcystin concentration (see Chap. 2), opened questions and promoted doubts about the actual population we were looking at. In fact, a high phenotypical variability used to be found in classical taxonomic analysis of water bodies with presence of *Microcystis* populations, from meso-oligotrophic (see Chap. 1) to hyper-eutrophic environmental conditions. The recruitment analysis showed a sequence-discrete population, within a wide range of similarity (Fig. 6), and deeper investigations could unveil a gene-centered dynamic at strain level. Compare metagenomics data to reference genomes and indentify SNV patterns (Suppl. Table S5 and S7) can lead to an accurate differentiation of co-occurring strains (Luo et al., 2015). The CMIs described are a local population feature,

concerning the environmental variables at sampling time (Fig. 5). By changing the environmental conditions, we could expect distinct hypervariable locations. The better these regions are mapped the more information about genome-wide and gene-centered genome rearrangements and recombinations should be available, so that genome-linkage patterns could arise from Metagenomics Islands discovering and description.

Several alternative answers could be proposed to try to explain the "appearance" and "disappearance" of some strains (Fig. 4), but our mapping against all fifteen *M. aeruginosa* strains (Fig. 7), along with the hypervariable regions and high-moderate impact variants described (Suppl. Table S8), reinforced the mosaic/cloud gene distribution (Steffen et al., 2014; Roosen et al. 2015) and community of genes (Boon et al., 2014) theories proposed in recent publications. Our data represents a current challenge among microbial community next-generation sequencing studies: the multiple and puzzling mapping against potentially the whole genome present in a sample, either accurately defined as strains or ecotypes, or as intermediary, complementary, transient and dynamic pieces of a much more parsimonious population structure, centered on gene content and biochemical interactions. The low coverage of our dataset required the adoption of multiple strategies, in order to improve the accuracy of our findings, based on complementary results. All CMIs and genes highlighted in this study are probably related to *Microcystis* population dynamics and dominance, but the low coverage attribute of Marimbondo metagenomics dataset do not allow reliable statements. Despite such technique barrier, we believe that this work may have provided valuable clues about potential candidates to further investigations.

3. ACKNOWLEDGEMENTS

This research was supported by a scholarship from CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) to M.L., and by funds provided by FAPEMIG (Fundação de Apoio a Pesquisa de Minas Gerais) to A.G. We thank the staff of the Phycology Laboratory (Dept. Botany/UFMG), for performing all laboratory analyses and to "Biodados" staff (Dept. Biochemistry and Imunology/UFMG), for assistance in bioinformatics. We also thank Msc. Cláudia Wermelinger for English review of the text.

3. REFERENCES

Alexova R., Haynes P.A., Ferrari B.C., Neilan B.A. (2011) Comparative protein expression in different strains of the bloom-forming cyanobacterium *Microcystis aeruginosa*. *Mol. Cell. Proteomics*, 10(9):M110-003749

Antonova E.S., Hammer B.J. (2011) Quorum-sensing autoinducer molecules produced by members of a multispecies biofilm promote horizontal gene transfer to Vibrio cholera. *FEMS Microbiol. Let*., 322(1):68-76

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., Sherlock G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.*, 25(1):25-9

Beck C., Knoop H., Axmann I. M., Steuer R. (2012) The diversity of cyanobacterial metabolism: genome analysis of multiple phototrophic microorganisms. *BMC Genomics*, 13:56

Bendall M.L., Stevens S.L.R., Chan L.K., Malfatti S., Schwientek P., Tremblay J., Schackwitz W., Martin J., Pati A., Bushnell B., Froula J., Kang D., Tringe S.G., Bertilsson S., Moran M.A., Shade A., Newton R.J., McMahon K.D., Malmstrom R.R. (2016) Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.*, 1–13

Biller S.J., Berube P.M., Lindell D., Chisholm S.W. (2015) *Prochlorococcus*: the structure and function of collective diversity. *Nat. Rev. Microbiol.*, 13:13-27

Boon E., Meehan C.J., Whidden C., Wong D.H.J., Langille M.G.I., Beiko R.G. (2014) Interactions in the microbiome: communities of organisms and communities of genes. *FEMS Microbiol Rev*, 38:90-118

Briand E., Bormans M., Quiblier C., Salençon M.J., Humbert J.F. (2012) Evidence of the Cost of the Production of Microcystins by *Microcystis aeruginosa* under Differing Light and Nitrate Environmental Conditions. *PLoS ONE*, 7(1):e29981

Briand E., Bormans M., Gugger M., Dorrestein P.C., Gerwick W.H. (2015) Changes in secondary metabolic profiles of Microcystis aeruginosa strains in response to intraspecific interactions. *Envir. Microbiol*., doi:10.1111/1462-2920.12904

Castenholz R. W., Norris T. B. (2005) Revisionary concepts of species in the Cyanobacteria and their applications. *Arch. Hydrobiol. Algol. Stud.*, 117:53-69

Cingolani P., Platts A., Wang le L., Coon M., Nguyen T., Wang L., Land S.J., Lu X., Ruden D.M. (2012)a A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, 6(2):80-92

Cingolani P., Patel V. M., Coon M., Nguyen T., Land S. J., Ruden D. M., Lu X. (2012)b Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. In:

Toxicogenomics in non-mammalian species, 3: 35. Frontiers E-books

Coleman M.L., Sullivan M.B., Martiny A.C., Steglich C., Barry K., DeLong E.F., Chisholm S.W. (2006) Genomic islands and the ecology and evolution of Prochlorococcus. *Science*, 311:1768–1770

Cordero O.X., Ventouras L.A., DeLong E.F., Polz M.F. (2012) Public good dynamics drive evolution of iron aquisition strategies in natural bacterioplankton populations. *Proc. Natl. Acad. Sci. USA*, 109:20059–20064

Cordero O.X., Polz M.F. (2014) Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Rev. Microbiol.*, 12:263-273

Desai M.M., Walczak A.M. (2015) Flexible gene pools - Rapid genetic exchange leads to mosaic genomes in cyanobacterial populations. *Science*, 348:977-978

Ehrlich G.D., Ahmed A., Earl J., Hiller N.L., Costerton J.W., Stoodley P., Post J.C., DeMeo P., Hu F.Z. (2010) The distributed genome hypothesis as a rubric for understanding evolution in situ during chronic bacterial biofilm infectious processes. *FEMS Imm. Medical Microbiol.* 59(3):269-279

Florencio F.J., Pérez-Pérez M.E., López-Maury L. Mata-Cabana A., Lindahl M. (2006) The diversity and complexity of the cyanobacterial thioredoxin systems. *Photosynth. Res*., 89:157-171

Frangeul L., Quillardet, P., Castets A.M., Humbert J.F., Matthijs H.C.P., Cortez D., Tolonen A., Zhang C.C., Gribaldo S., Kehr J.C., Zilliges Y., Ziemert N., Becker S., Talla E., Latifi A., Billault A., Lepelletier A., Dittmann E., Bouchier C., Tandeau de Marsac N. (2008) Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics*, 9:274

Fraser C., Alm E.J., Polz M.F., Spratt B.G., Hanage W.P. (2009) The bacterial species challenge: making sense of genetic and ecological diversity. *Science*, 323(5915):741-746

Galhardo R.S., Hastings P.J., Rosenberg S.M. (2007) Mutation as a stress response and the regulation of evolvability. *Crit. Rev. Biochem. Mol. Biol*., 42:399–435

Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucl. Acids Res.*, 43(Database issue):D1049–D1056

Gil R., Silva F.J., Peretó J., Moya A. (2004) Determination of the Core of a Minimal Bacterial Gene Set. *Microbiol. Mol. Biol. Rev*., 68(3):518-537

Hess W.R. (2011) Cyanobacterial genomics for ecology and biotechnology. *Curr. Opin. Microbiol.*, 14:608-614

Hu C., Rea C., Yu Z., Lee J. (2015) Relative importance of Microcystis abundance and diversity in determining microcystin dynamics in Lake Erie coastal wetland and downstream beach water. *J. Appl. Microbiol.*, 120:138-151

Humbert J.F., Barbe V., Latifi A., Gugger M., Calteau A., Coursin T., Lajus A., Castelli V., Oztas S., Samson G., Longin C., Medigue C., Tandeau de Marsac N. (2013) A Tribute to Disorder in the Genome of the Bloom-Forming Freshwater Cyanobacterium *Microcystis aeruginosa*. *PLoS ONE*, 8(8): e70747

Huson D.H., Mitra S., Ruscheweyh H.J., Weber N., Schuster S.C. (2011) Integrative analysis of environmental sequences using MEGAN 4. *Genome Res.*, 21:1552-1560

Kaneko T., Nakajima N., Okamoto S., Suzuki I., Tanabe Y., Tamaoki M., Nakamura Y., Kasai F., Watanabe A., Kawashima K., Kishida Y., Ono A., Shimizu Y., Takahashi C., Minami C., Fujishiro T., Kohara M., Katoh M., Nakazaki N., Nakayama S., Yamada M., Tabata S., Watanabe M. M. (2007) Complete Genomic Structure of the Bloom-forming Toxic Cyanobacterium *Microcystis aeruginosa* NIES-843. *DNA Research*, 14:247-256

Kashtan N., Roggensack S.E., Rodrigue S., Thompson J.W., Biller S.J., Coe A., Ding H., Marttinen P., Malmstrom R.R., Stocker R., Follows M.J., Stepanauskas R., Chisholm S.W. (2014) Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild *Prochlorococcus*. Science, 344:416-420

Knoop H., Gründel M., Zilliges Y., Lehmann R., Hoffmann S., Lockau W., Steuer R. (2013) Flux Balance Analysis of Cyanobacterial Metabolism: The Metabolic Network of *Synechocystis* sp. PCC 6803. *PLoS Comput Biol*, 9(6): e1003081

Komárek J., Kastovsky J., Jezberova J. (2011) Phylogenetic and taxonomic delimitation of the cyanobacterial genus *Aphanothece* and description of *Anathece* gen. nov. *Eur. J. Phycol.*, 46:315-326

Komárková J., Jezberova J., Komárek O., Zapomelova, E. (2010) Variability of *Chroococcus* (Cyanobacteria) morphospecies with regard to phylogenetic relationships. *Hydrobiologia*, 639:69-83

Konopka A. (2009) What is microbial community ecology? *ISME J.*, 3: 1223–1230.

Kurmayer R., Christiansen G., Chorus, I. (2003) The abundance of microcystin-producing genotypes correlates positively with colony size in Microcystis sp. and determines its microcystin net production in Lake Wannsee. *Appl. Environ. Microbiol.*, 69:787-795

Kurtz S., Phillippy A., Delcher A.L., Smoot M., Shumway M., Antonescu C., Salzberg S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, 5(2):R12

Li H., Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754-1760

Luo C., Knight R., Siljander H., Knip M., Xavier R.J., Gevers D. (2015) Constrains identifies microbial strains in metagenomic datasets. Nat. Biotechnol., 33(10):1045-1052

Meyer F., Paarmann D., D'Souza M., Olson R., Glass E.M., Kubal M., Paczian T., Rodriguez A., Stevens R., Wilke A., Wilkening J., Edwards R.A. (2008) The Metagenomics RAST server — A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9:386

Mira A., Martín-Cuadrado A.B., D'Auria G., Rodríguez-Valera F. (2010) The bacterial pan-genome: a new paradigm in microbiology. *Int. Microbiol*., 13: 45-57

Nguyen V.L.A., Tanabe Y., Matsuura H., Kaya K., Watanabe M.M. (2012) Morphological, biochemical and phylogenetic assessments of water-bloom-forming tropical morphospecies of *Microcystis* (Chroococcales, Cyanobacteria). *Phycol. Res*., 60: 208-222

Oh S., Caro-Quintero A., Tsementzi D., DeLeon-Rodriguez N., Luo C., Poretsky R., Konstantinidis K.T. (2011) Metagenomic Insights into Evolution, Function, and Complexity of the Planktonic Microbial Community of Lake Lanier, a Temperate Freshwater Ecosystem. *Appl. Environm. Microbiol.*, 77(17):6000-6011

Okano K., Miyata N., Ozaki Y. (2015) Whole Genome Sequence of the Non-Microcystin-Producing *Microcystis aeruginosa* Strain NIES-44. *Genome Announc.*, 3(2):e00135-15

Otsuka S., Suda S., Shibata S., Oyaizu H., Matsumoto S., Watanabe M.M. (2001) A proposal for the unification of five species of the cyanobacterial genus *Microcystis* Kützing ex Lemmermann 1907 under the rules of the Bacteriological Code. *Int. J. System. Evol. Microbiol.*, 51(3):873-879

Palinska, K. A., Deventer, B., Hariri, K. & Lotocka, M. (2011): A taxonomic study on Phormidium-group (cyanobacteria) based on morphology, pigments, RAPD molecular markers and RFLP analysis of the 16S rRNA gene fragment. - Fottea 11: 41-55.

Pašić L, Rodriguez-Mueller B., Martin-Cuadrado A., Mira A., Rohwer F., Rodriguez-Valera, F. (2009) Metagenomic islands of hyperhalophiles: the case of *Salinibacter ruber*. *BMC Genomics*, 10:570

Rodriguez-Valera F., Martin-Cuadrado A.B., Rodriguez-Brito B., Pašić L., Thingstad T.F., Rohwer F., Mira A. (2009) Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol*., 7:828-836

Rosen M.J., Davison M., Bhaya D., Fisher D.S. (2015) Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science*, 348: 1019-1023

Rosenthal E., Blue E., Jarvik G.P. (2015) Next-generation gene discovery for variants of large impact on lipid traits. *Curr. Opinion Lipidol.*, 26(2):114-119

Sabart M., Pobel D., Latour D., Robin J., Salençon M.J., Humbert J.F. (2009) Spatiotemporal changes in the genetic diversity in French bloom-forming populations of the toxic cyanobacterium, *Microcystis aeruginosa*. *Env, Microbiol. Rep*., 1(4):263-272

Shapiro B.J., Friedman J., Cordero O.X., Preheim S.P., Timberlake S.C., Szabó G., Polz M.F., Alm E.J. (2012) Population genomics of early events in the ecological differentiation of bacteria. *Science*, 336:48–51

Shapiro, B.J., Polz M.F. (2014) Ordering microbial diversity into ecologically and genetically cohesive units.

Trends in Microbiology, 22(5):235-247

Sheath R. G., Wehr J. D. (2003) Freshwater Algae of North America. Elsevier Science, San Diego, California. 918 p.

Simm S., Keller M., Selymesi M., Schleiff E. (2015) The composition of the global and feature specific cyanobacterial core-genomes. *Front. Microbiol*., 6:219

Steffen M.M., Li Z., Effler T. C., Hauser L.J., Boyer G.L., Wilhelm S.W. (2012) Comparative Metagenomics of Toxic Freshwater Cyanobacteria Bloom Communities on Two Continents. *PLoS ONE*, 7(8): e44002

Steffen M.M., Dearth S.P., Dill B.D., Li Z., Larsen K.M., Campagna S.R., Wilhelm S.W. (2014) Nutrients drive transcriptional changes that maintain metabolic homeostasis but alter genome architecture in Microcystis. *ISME J.*, 8:2080-2092

Tatusov R.L., Galperin M.Y., Natale D.A., Koonin E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic acids research*, 28(1):33-36.

Tettelin H., Riley D., Cattuto C., Medini D. (2008) Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol*., 11(5):472-7

Thomas T., Gilbert J., Meyer F. (2012) Metagenomics - a guide from sampling to data Analysis. *Microbial Informatics and Experimentation*, 2:3

Van Elsas J.D., Bailey M.J. (2002) The ecology of transfer of mobile genetic elements. *FEMS Microbiol. Ecol.*, 42:187–197

Van Gremberghe I., Leliaert F., Mergeay J., Vanormelingen P., Van der Gucht K., Debeer A.E., Lacerot G., Meester L.D., Vyverman W. (2011) Lack of Phylogeographic Structure in the Freshwater Cyanobacterium *Microcystis aeruginosa* Suggests Global Dispersal. *PLoS ONE*, 6(5):e19561.

Whitaker R.J., Grogan, D.W., Taylor J.W. (2005) Recombination Shapes the Natural Population Structure of the Hyperthermophilic Archaeon *Sulfolobus islandicus*. Mol. Biol. Evol., 22(12):2354-2361

Whitton B. A. (2012) Ecology of Cyanobacteria II: Their Diversity in Space and Time. Springer Science & Business Media, 760 p.

Wilson A. E., Sarnelle O., Neilan B. A., Salmon T. P., Gehringer M. M., Hay M. E. (2005) Genetic Variation of the Bloom-Forming Cyanobacterium Microcystis aeruginosa within and among Lakes: Implications for Harmful Algal Blooms. *Appl. Envir. Microbiol.*, 71(10):6126-6133

Xie L., Rediske R.R., Hong Y., O'Keefe J., Gillett N.D., Dyble J., Steinman A.D. (2012) The role of environmental parameters in the structure of phytoplankton assemblages and cyanobacteria toxins in two hypereutrophic lakes. *Hydrobiologia*, 691:255-268

Yoshida M., Yoshida T., Satomi M., Takashima Y., Hosoda N., Hiroishi S. (2008) Intra-specific phenotypic and genotypic variation in toxic cyanobacterial *Microcystis* strains. *J. Appl. Microbiol*., 105:407-415

Zarraonaindia I., Smith D.P., Gilbert J.A. (2013) Beyond the genome: community-level analysis of the microbial world. *Biol Philos*, 28:261-282

Zhu M., Xu Y., Li R. (2012) Genetic Diversity of Bloom-Forming *Microcystis* (Cyanobacteria) Populations in a Hyper-Eutrophic Pond in Central China. *Curr. Microbiol*., 65(3):219-224

Capítulo 4: A variabilidade subjacente em uma população de cianobactérias dominante: mudando a pergunta para ‘quem mais’?

4. RESUMO

As técnicas metagenômicas alcançam uma variedade de escalas analíticas, desde perfis taxonômicos amplos até detecção de variantes de base única. O foco na estrutura da população, ou mesmo entre populações tem sido sugerido, uma vez que ele possibilita uma visualização mais realista. A técnica de sequencimanto de fragmentos randômicos globais (*Whole Genome Shotgun* - WGS) podem revelar um perfil mais complexo, porém preciso da estrutura das populações. Foram analisadas sequências metagenômicas de um reservatório de água urbano hipereutrófico (Lagoa da Pampulha, Brasil), impactado por constantes florações de cianobactérias, com o objetivo de explorar a estrutura da população dominante, pertencente à ordem Nostocales. Pertencem a essa ordem dois gêneros altamente diversos, *Cylindrospermopsis* e *Raphidiopsis*, os quais compartilham mais de 99% de similaridade, mas apresentam vários genes exclusivos e diferem até mesmo em clusteres inteiros. Tendo sido encontrado um amplo conjunto de anotações taxonômicas distintas, de acordo com diferentes métodos e parâmetros de alinhamento, foram realizados o mapeamento e a detecção de variantes de base única. *R.brookii* D9 e *C. raciborskii* T3 parecem ser as melhores candidatas compondo o genoma central dessa população, mas a falta do genoma completo de *C. racibroskii* T3 e o alinhamento com cepas sul-americanas indica que talvez estejamos lidando com um ecótipo, não produtor de cilindrospermopsina, estreitamente relacionado a *R. brookii* D9 e *C. racibroskii* T3. A ampla similaridade das sequências metagenômicas com múltiplos genomas de referência da ordem Nostocales, e a diversidade observada na anotação funcional sugerem que a sobreposição entre anotações taxonômicas pode estar refletindo a variabilidade fundamental de populações microbianas em ambiente natural, cuja estrutura é determinada principalmente pelas pressões ambientais locais. Mais estudos são necessários para revelar não apenas uma possível população de *Cylindrospermopsis* sul-americana, mas também uma rede complexa de interações, com um genoma central amplamente conservado entre membros da ordem Nostocales, e uma dinâmica de mosaico entre cepas.

Palavras-chave: Nostocales, *Cylindrospermopsis*, *Raphidiopsis*, metagenômica, estrutura populacional, diversidade genética, análise de variantes de base única.

The underlying variability within a dominant cyanobacteria population: changing the question to ‘who else’?

4. ABSTRACT

Metagenomic strategy reaches a wide variety of analytical scales, from broad taxonomic profiles to fine scale variant calling. The focus on population structure, or even within populations, has been suggested, since it provides a more realistic overview. The Whole Genome Shotgun technique reveal a very complex but accurate population structure profile. The metagenomic sequences from an urban hypereutrophic water reservoir (Pampulha lake, Brazil), impacted by constant cyanobacteria blooms, were analyzed in order to explore the structure of the dominant population, which belongs to Nostocales order. This order encompassed two highly microdiverse genera, *Cylindrospermopsis* and *Raphidiopsis*, which share nucleotide similarity above 99%, but have several unique genes and even entirely missing clusters among their strains. Because we found a broad set of distinct taxonomic annotations, according to different methods and alignment thresholds, a fine scale read mapping and variant calling was performed. *R.brookii* D9 and *C. raciborskii* T3 appeared to be the best candidates to compose the backbone genome of this population, but the lack of the complete genome sequence for *C. racibroskii* T3 and the alignment with south-american partial sequences indicate that maybe we were dealing with an ecotype, non-CYN producer, closely related to *R. brookii* D9. The sequence-wide similarity of the metagenomic reads against multiple reference genomes under the Nostocales order and the diverse gene annotation observed suggest that the overlap between taxonomical assignments could be reflecting the underlying variability of wild microbial populations, whose structure is determined mostly by local environmental pressures. More studies are necessary to reveal not only a potential south-american *Cylindrospermopsis* population, but also a complex interaction network within the Nostocales order, with a broad and highly conserved backbone genome, and a mosaic dynamics between strains.

Keywords: Nostocales, *Cylindrospermopsis*, *Raphidiopsis*, metagenomics, population structure, genetic diversity, variant calling

4. INTRODUCTION

Populations are usually defined as coexisting individuals of the same species, which means that they share a common gene pool, but for bacteria, instead, populations can be operationally recognized as groups of coexisting individuals that are highly clustered at the genotypic and phenotypic levels (Cordero & Polz, 2014; Coleman & Chisholm, 2010). This means that, even within the same species/population, a large number of distinct strains can be identified, each one working on its own ecological niche (Wiedenbeck & Cohan, 2011) but sharing a backbone genome (Kashtan et al., 2014). A backbone genome could be seen as the unifying genetic potential of a population, whereby the population dynamics can alternate the genomic structures (Steffen et al., 2014). Cyanobacteria coexisting strains interacting within populations have bem reported (Moore et al., 1998; Coleman et al., 2006; Piccini et al., 2011; Shapiro et al., 2012; Shapiro & Polz, 2012; Kashtan et al., 2014; López-Pérez et al., 2013; Humbert et al., 2013). Beck et al., (2012) observed that from 21238 distinct CLOGs (Clusters of orthologous groups), identified across all cyanobacterial strains studied, 65% consisted of single genes that have no likely ortholog in any other considered strain and only about 3% of CLOGs were assigned to all strains. This indicate that the pan-genome of cyanobacteria is not closed and, in fact, is increasing according to the number of strains considered (Simm et al., 2015). When the pan-genome concept of the whole community is considered, instead of a specific taxonomical rank, it rather closely relates to the '*community of genes'* concept, introduced by Boon et al., (2014), in which the interactions among distinct population members are gene-centered and directly associated to habitat-specific demands. These two concepts are further closely related to the metacommunity theory of microbial systems, defined by Leibold et al., (2004), in which local communities potentially interact through multiple species.

Microbial community ecology has been focused in terms of entities and their respective relative abundance, but taxonomic and phylogenetic approaches alone may be insufficient in providing the claimed answers, so larger taxonomic groups may be of interest because they share one or more important attributes (Boon et al., 2014), expanding the perception of the community's potential, rather than the individual description. The toxigenic genus *Cylindrospermopsis* is highly successful in freshwater environments and has been reported to be distributed worldwide (Padisák, 1997; Moustaka-Gouni et al., 2009; Stucken et al., 2010) with an important impact in South America (Bittencourt-Oliveira et al., 2011a; b; Bittencourt-Oliveira et al., 2012a; b; Piccini et al., 2011; 2013; Sukenik et al., 2012; Hoff-

Risseti et al., 2013; Cirés et al., 2014; Moura et al., 2015; Elias et al., 2015). *C. raciborskii* strains can produce the hepatotoxin cylindrospermopsin (CYN), a tricyclic alkaloid inhibitor of protein synthesis (Falconer & Humpage, 2006) or neurotoxins associated with paralytic shellfish poisoning (PSP), specifically the tetrahydropurine saxitoxin (STX) (Lagos et al., 1999; Kellmann et al., 2008). The closely related genus *Raphidiopsis* is reported to produce CYN and/or deoxycylindrospermopsin (doCYN), anatoxin-a or PSP toxins (Li et al., 2001; Namikoshi et al., 2003; Yunes et al., 2009). There is some evidence about a biogeographic pattern concerning toxin production in certain Nostocales (Cirés et al., 2014), such as *C. raciborskii*, with CYN-producing strains occurring in Australia and Asia but non-CYN-producing strains occurring in Africa, Europe, and North America (Haande et al., 2008) and PSP toxin-producing strains occurring in Brazil and Uruguay (Piccini et al 2011, Lagos et al. 1999).

In this study, a hypereutrophic urban reservoir was analyzed through Whole-Genome Shotgun in order to explore the structure of the dominant cyanobacteria population. The Pampulha lake presents persistent cyanobacterial blooms and a seasonal dominance shift, in which the Nostocales order is prevalent most of the year, but by the end of the dry – start of the rainy season, the contribution of Chroococcales order increases (Figueredo & Giani, 2011; Lopes, 2013; Lopes et al., *in review*). *Microcystis aeruginosa* and *Cylindrospermopsis raciborskii* are amidst the most widely distributed toxic cyanobacteria in Brazil (Tucci & Sant'Anna, 2003; Costa et al., 2006; Crossetti & Bicudo, 2008; Sant'Anna et al., 2008; Mello et al., 2012; Moura et al., 2015) and the coexistence and alternation of these species have already been observed around the world (Moustaka-Gouni et al., 2009; Xie et al., 2012; Marinho et al., 2013; Lopes, 2013; Antunes et al., 2015). Besides the ability to produce several types of cyanotoxins, the Nostocales order, particularly members of *Aphanizomenon* and *Cylindrospermopsis*, has showed an apparent ability to disperse from tropical to subtropical and temperate latitudes, a phenomenon that has been associated with global warming (Cirés et al., 2014; Sukenik et al., 2012; Bouvy et al., 2003; Simm et al., 2015).

The study of interactions among wild populations and closely related strains can reveal genomic features acting throughout niche complementation and ecological dynamics (Cordero & Polz, 2014; Coleman et al., 2006). Through metagenomics, the analysis of the genomic DNA from a whole community is made possible (Gilbert & Dupont, 2011; Wiedenbeck & Cohan, 2011; Oh et al., 2011), potentially uncovering dependencies among community members (Hug et al., 2012; Steffen et al., 2012; Penn et al., 2014). Ultimately,

metagenomic datasets enhance genome-wide investigation (Escalas et al., 2013; Bendall et al., 2016) and improve the taxonomical overview by allowing different strategies and scales of analysis (Suenaga, 2012). Accessing environmental sequences from virtually any organism or any genomic region (Caro-Quintero & Konstantinidis, 2012; Shapiro & Polz, 2014) allows us to seek for the inter/intraspecific variability which, in fact, constitutes an underlying ecological feature (Fraser et al., 2009).

4. MATERIAL AND METHODS

Study Area

The Pampulha lake is a hypereutrophic artificial shallow lake, with 2.4 km^2 and average depth of 5m (maximum depth 16m), located in the city of Belo Horizonte, Minas Gerais state, Brazil (19°55'09 "S, 43° 56 '47" W). It was constructed in 1936 in order to prevent flooding, to contribute to water supply and for recreation, but around 1970 the lake was already impacted by wastewater input, leading to a eutrophication process and cyanobacterial blooms (Giani, 1994; Goodwin, 1997; Figueredo & Giani, 2001; Lopes, 2013). The Pampulha lake was sampled according to regional seasonality, characterized by a hot rainy season, from October to March, with higher concentration of nitrite, and a colder, dry season, from April to September, with higher concentration of total phosphorus and ammonia (Costa, 2015). Sampling was carried out near the central pelagic region in January 2011 (P1M) and April 2012 (P3M) using a Van Dorn bottle.

Sequencing and Pre-Processing of Sequence Datasets

The genomic DNA was extracted according to Kurmayer et al., (2003) and the sequencing was performed under Nextera XT paired-end 2 x 250 bp, Illumina platform. For the Genome Analysis Toolkit - GATK (McKenna et al., 2010) and read mapping, the datasets were trimmed using Trimmomatic 0.32 (Bolger et al., 2014), paired end mode and the following parameters: leading low quality 15, trailing low quality 20, 4-base wide sliding window, cutting when the average quality per base drops below 30 and minimum length 50bp. The paired outputs were then merged using Flash 1.2.11 (Magoč & Salzberg, 2011),

allowing for both outie and innie orientation of merging fragments and extending the maximum overlap to 300 bp, since a large proportion of the sequences presented almost full overlapping. Only the properly paired and combined reads were considered. The increased accuracy resulting from choosing just high quality and properly paired reads allows for a more confident mapping and variant calling.

MG-RAST Plataform

The merged sequences from metagenomic samples P1M and P3M were uploaded into the Metagenomics RAST server (MG-RAST), a public resource for the automatic phylogenetic and functional analysis of metagenomes (Meyer et al., 2008). The sequences were processed according to standard pipeline and a cutoff of 90-97% of identity and a 30 bp alignment length was applied to the subsequent analysis in order to avoid false positives.

Read Mapping

The P1M and P3M samples were aligned to seven complete reference genomes (*Anabaena cylindrica*, *A. variabilis*, *Nostoc punctiforme*, *N.* sp. PCC7524, *Trichormus azollae*, *C. raciborskii* CS505 and *R. brookii* D9) and 12 partial sequences (*A.* sp 90 anabaenopeptin biosynthesis gene cluster, *A. siamensis*, *Planktothrix agardhii* contig 5, *R. curvata* 16S sequence, *C. raciborskii* T3 and AWT205 toxin cluster, *C. raciborskii* cena302, 303, 305, MVCC19, MVCC14 and HM9917) using MUMmer 3.23 (Kurtz et al., 2004) and Burrows-Wheeler Aligner-bwa mem (Li & Durbin, 2009). MUMmer is one of the most commonly used whole genome alignment method (Olson et al., 2015) and the PROmer algorithm was used for complete genome alignment, with default parameters. NUCmer was used for partial sequences, with default parameters. Delta-filter was used to select alignments under identity cutoffs (97% and 99%) and specific mapping parameters (uniqueness or coverage depth). The show-coords output presents the coordinates and alignment information of all alignments present in the delta file and was used to locate and extract alignment summaries. Recruitment plots were generated by MUMmerplot utility. The alignment depth of coverage (aDP), concerning all reads aligned to some region, was measured using bwa alignments and the GATK DepthOfCoverage algorithm, while the breadth of coverage was measured using MUMmer delta-filter with one-to-one mapping of

reference to query with alignment overlap tolerance of 10 bp and identity cutoffs of 97% for complete genomes and 99% for partial sequences.

Single Nucleotide Variant Calling (SNV calling)

The two trimmed and merged shotgun datasets (P1M and P3M) were mapped against reference sequences of the two more abundant nostocales species pointed by MG-RAST: *Cylindrospermopsis raciborskii* CS505 and *Raphidiopsis brookii* D9. The sequences were mapped using bwa mem, following a standard pipeline informed by the GATK best-practices version 3.5 (DePristo et al., 2011; Van der Auwera et al., 2013). The mapped reads were converted and sorted using Picard tools v. 1.138 (http://picard.sourceforge.net/) and the summary statistics and coverage were obtained using SAMtools and Genome Analysis Tool Kit v.3.4. In order to avoid false-positive SNV calls and improve the variant calling accuracy, the base qualities were empirically recalibrated and indel realignment was performed for each dataset, mostly because we are dealing with a populational SNV calling and the sequence coverage is low to moderate (Nielsen et al., 2011). The resulting variants of an initial variant call followed by variant filtration were used as known sites for the final variant call, since currently there is no reference sites available. The following parameters were adopted to recalibration variant filtering: Variant Confidence/Quality by Depth (QD) >10, Mapping Quality (MQ) > 40.0, Phred-scaled Probability that a Polymorphism Exists (QUAL) > 800, Approximate Read Depth (aDP) > 30 and Genotype Quality (GQ) > 30.

The variant calling of the two realigned samples was made with the Haplotypecaller tool using the following parameters: ploidy 1, minimum phred-scaled confidence threshold at which variants should be emitted from 10 and called from 50. The two realigned samples were called together, according to multi-sample procedure on GATK pipeline, which includes a linkage-based analysis, improving the accuracy of SNV calls (Nielsen et al., 2011; Nielsen et al., 2012). The final variants were filtered using VariantFiltration and SelectVariants tools with the following parameters: QD >10, MQ > 40.0, QUAL > 1000, aDP > 30 and GQ > 30. For some analysis the 'concordance' filter was used to consider only the variants present in both datasets. The high quality variants were then annotated using SnpEff tool (Cingolani et al., 2012a) according to the generic feature format file (gff3) for each reference genome, available in http://bacteria.ensembl.org/ (GCA_000175835.1 and GCA_000175855.1), focusing only on coding regions, and the output was handled using

SnpSift tool (Cingolani et al. 2012b). In order to calculate the empirical allele frequency we extracted the Allelic Depths (AD) and the Depth of Coverage (DP) parameters from format field of VCF outputs. The AD outputs the depth of coverage of reference and alternate allele by sample, while DP outputs the filtered depth of coverage for each sample. The frequency of alternate alleles was calculated as the ratio alternate AD/DP. The VCF outputs were converted to table and loaded into a custom MySQL database in order to perform the functional annotation according to Gene Ontology database – GO (Ashburner et al. 2000; Gene Ontology Consortium, 2015).

4. RESULTS AND DISCUSSION

MG-RAST Profile

The samples can be found under the accession numbers 4635711.3 (P1M) and 4635712.3 (P3M). After the quality control step, only 0.5% of the sequences of each sample were identified as ribosomal sequences, according to M5RNA database, which integrates SILVA, Greengenes and RDP databases. The SILVA SSU database showed the most robust RNA annotation. For P1M sample, SILVA SSU accounted for 6068 hits, while RDP 1882 and Greengenes 1831. For P3M sample, SILVA SSU accounted for 2462 hits, while RDP 1186 and Greengenes 1162. The metagenome summary is presented in Table 1. The functional assignments according to hierarchical classification and a 97% identity cutoff showed a distinct metabolic profile between the samples. Sample P1M presented higher assignments under regulation and cell signaling, stress response, membrane transport and clustering based subsystems, while sample P3M presented higher assignments under cell division and cell cycle, nitrogen metabolism, RNA metabolism and amino acids and derivatives.

Tab. 1 - Metagenome summary of preprocessing, gene calling and RNA identification according to MG-RAST pipeline.

Sequences atributes	**P1M**	**P3M**
num seqs	3742137	3089456
average length	279	258
failed QC	456948	326179
passed QC	3285189	2763277
ribosomal RNA genes	18973	14250
seqs with predicted proteins known function	1695369	1378887
seqs with predicted proteins unknown function	1350169	1079266
no rRNA genes or predicted proteins	220678	290874
predicted protein coding regions	1799842	1405708
assigned annotation	458731	332019
no significant similarities to the protein database	1341111	1073689
features assigned to functional categories	335454	227548

The taxonomical profile according to best hit classification was generated by comparing the dataset to the M5RNA database using a maximum e-value of 1e-5, a minimum identity of 90%, and a minimum alignment length of 30 bp. Bacteria domain accounted for 56% of hits for sample P1M and 55% for sample P3M. According to Bacteria phyla distribution, Cyanobacteria contributed to the larger portion of the bacterial community, followed by Proteobacteria and Actinobacteria (Fig. 1), as observed by other studies in freshwater systems (Bagatini et al., 2014; Eiler & Bertilsson, 2004; Shi et al., 2010; Steffen et al., 2012; Penn et al., 2014).

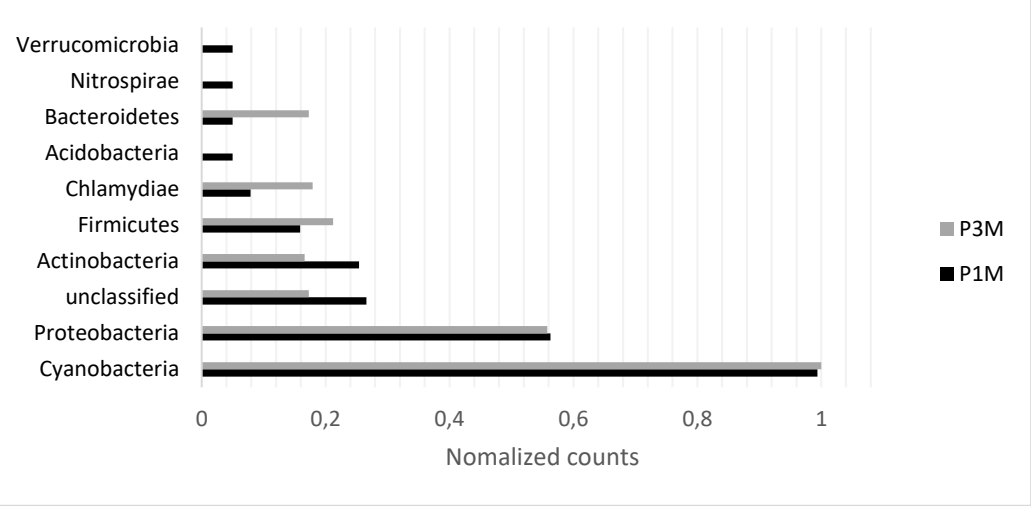


Fig. 1 - Phyla distribution in the Bacteria domain. Data normalized according to MG-RAST pipeline.

In the Cyanobacteria phylum, the Chroococcales order presented higher contribution in sample P3M, while Nostocales order presented higher contribution in sample P1M (Fig. 2 A). Concerning the most abundant genera, *Raphidiopsis* and *Cylindrospermopsis* presented the highest contribution in both samples (Fig. 2 B). In the Chroococcales order, *Microcystis* presented the highest contribution in sample P3M (Fig. 2 C).

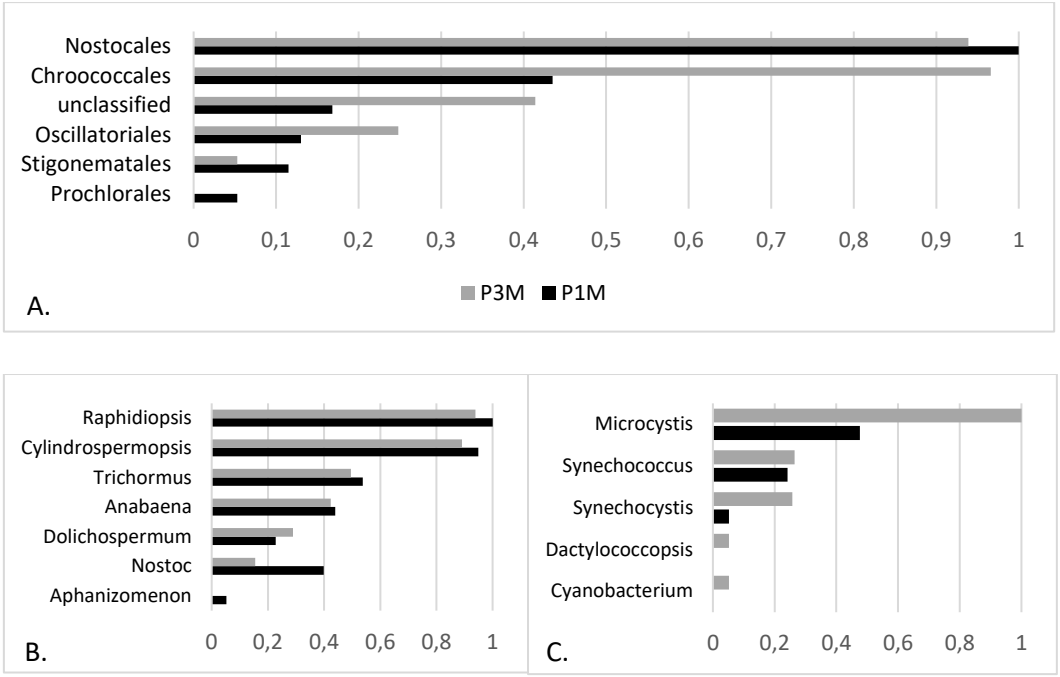


Fig. 2 - (A) Order and genera distribution in the Cyanobacteria phylum. (B) Nostocales genera. (C) Chroococcales genera. Data normalized according to MG-RAST pipeline.

This seasonal interplay between *Microcystis* and Nostocales, specially *Cylindrospermopsis*, has already been studied (Moustaka-Gouni et al., 2009; Wu et al., 2009; Xie et al., 2012; Marinho et al., 2013; Rzymski et al., 2014; Lopes et al., *in review*). The *Raphidiopsis* genus was present in both samples, but with much higher contribution in sample P1M, annotated as the Brazilian strain *R. brookii* D9. *Cylindrospermopsis* genus presented high contribution in both samples and the Australian strain *C. raciborskii* CS-505 presented the highest number of hits (Tab. 2). Regarding strain level annotation in the Nostocaceae family (Tab. 2), we could observe that even the highest identity was below 99%. Two strains which presented lower abundance, *Anabaena siamensis* TISTR 8012 and the Brazilian strain *C. raciborskii* T3 (Zagatto, 1998) presented the highest average alignment identity (>98%), followed by *R. brookii* D9, above 97%. The highest abundance

of *C. raciborskii* CS-505 was reached under 93% average identity. *R. brookii* D9 and *C. raciborskii* T3 were both isolated from Billings reservoir (São Paulo, Brazil). According to Boon et al., (2014), uniform taxonomic or phylogenetic thresholds may fail to adequately delineate ecologically cohesive units, especially in microorganisms whose genomes can change rapidly, and so the large contribution of unclassified sequences at the strain level and the low average identity, even among the most abundant strains, points to a complex population structure, where the individual annotation could be blurred by variation in the population gene content.

Tab. 2 - Strain level annotation for Nostocaceae family according to MG-RAST showing the average identity (avgID), average alignment length (avgLEN), total number of hits (#hits) and total number of reads aligned (abundance).

	Nostocaceae strains	**avgID**	**avgLEN**	**#hits**	**abundance**
P1M	*Raphidiopsis brookii* D9	97.80	131.69	2912	514020
P1M	*R. brookii* D9 Unclassified	97.72	131.08	3007	499859
P1M	*Cylindrospermopsis raciborskii* CS-505	92.81	126.44	2524	276781
P1M	*C. raciborskii* CS-505 Unclassified	93.20	126.05	2583	273686
P3M	*R. brookii* D9	97.61	127.12	2764	225171
P3M	*R. brookii* D9 Unclassified.	97.59	126.43	2819	220120
P3M	*C. raciborskii* CS-505	92.94	121.83	2452	130114
P3M	*C. raciborskii* CS-505 Unclassified	93.26	121.97	2484	127215
P1M	*C. raciborskii* T3	98.95	125.03	31	5184
P3M	*C. raciborskii* T3	98.85	116.31	30	2593
P1M	*Anabaena azollae*	94.83	149.93	5	1336
P3M	*A. azollae*	93.14	139.01	5	764
P1M	*A. cylindrica*	93.30	130.87	3	625
P3M	*A.cylindrica*	92.68	122.70	3	266
P1M	*Nostoc commune*	90.55	143.36	3	206
P1M	*C. raciborskii* AWT205	89.14	77.81	6	106
P3M	*Nostoc commune*	88.54	118.34	2	69
P1M	*A.siamensis* TISTR 8012	98.34	99.65	2	51

Read Mapping

When comparing a metagenomic dataset against completely sequenced reference genomes, differences related to recent community evolution, evidences of public goods sharing (McInerney et al., 2011) or gene gain/losses can arise. Thus, the application of the metacommunity theory in a set of metagenomic reads mapped against closely related sequenced genomes can expand the practical application of the metagenome in relation to the habitat and to the partitioning of genes among organisms (Boon et al., 2014; Shapiro &

Polz, 2014; Bendall et al., 2016). It has been reported that *C. raciborskii* can also coexist with morphotypes assigned to the closely related genus *Raphidiopsis*, showing >99% of nucleotide similarity, but with several distinctive features in terms of gene content (Stucken et al., 2010). In order to explore the population variation related to *Cylindrospermopsis* and *Raphidiopsis* genera, we performed the read mapping of the two metagenomic datasets against the reference genomes *C. raciborskii* CS-505 and *R. brookii* D9 and the alignment summary is presented in Table 3. The sample P1M, with higher contribution of Nostocales, presented the largest percentage of mapping for both strains, greatest depth and mean coverage. The mean depth, mean coverage depth and mean coverage breadth by contig and by genome is presented in Supplementary Table S1.

Tab. 3 - Alignment and coverage summary of the two metagenomic samples P1M and P3M, according to BWA-MEM aligner and BEDtools Coveragebed function. avgDepthCov: average depth of coverage; avgBreadthCov: average breadth of coverage; %Mapped: percent of total reads mapped.

	avgDepthCov	**avgBreadthCov**	**%Mapped**
P1M_CS505	275.5	0.8	33.8%
P3M_CS505	115.7	0.8	21.2%
P1M_D9	319.8	1.0	36.6%
P3M_D9	134.0	1.0	23.2%

From 91 CS505 contigs, 53 presented breadth of coverage above 0.8 and seven presented less than 0.1, for both samples. The contig CS505_3, which encompasses the cylindrospermopsin biosynthesis cluster (*cyr*), presented 0.5 breadth of coverage and average depth of coverage 177.6 for sample P1M and 71.6 for sample P3M. The entire cylindrospermopsin cluster was absent, but the flanking regions were covered (Fig. 3), which agrees with the biogeographical patterns proposed in several studies (Piccini et al., 2011; Haande et al., 2008; Saker et al., 2003; Lagos et al., 1999; Pomati et al., 2004).

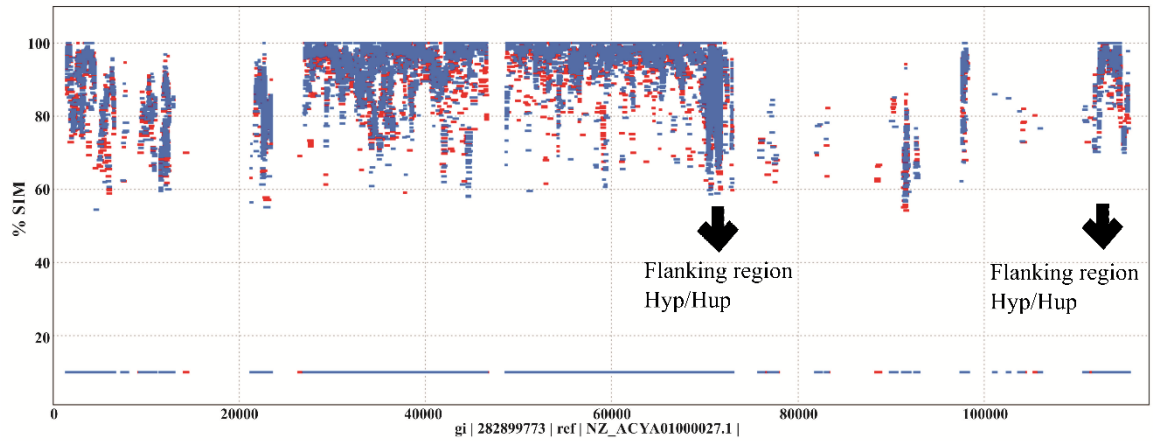


Fig. 3 - Recruitment plot of the P1M dataset to CS505_3 contig, showing the low breadth of coverage and the absence of cylindrospermopsin gene cluster and high depth of coverage of genes from the flanking regions *Hyp*/*Hup*. Y axis represent the percent of similarity of reads to the contig (X axis).

The nitrogen cluster (*nif*) presented in contig CS505_63 was covered, but some other genes from this contig were missing between CRC_02781 and CRC_02785 genes (Fig. 4), most of them hypothetical proteins, some transposases and CRISPR.

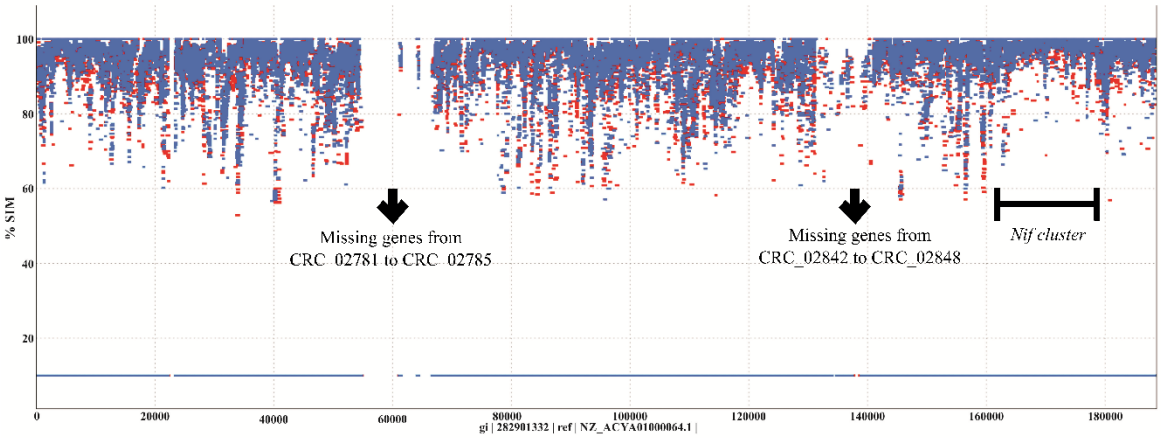


Fig. 4 - Recruitment plot of P1M dataset to CS505_63 contig, showing the alignment of the reads to the nitrogen fixation gene cluster, but with some missing genes in other regions of the contig. Y axis represent the percent of similarity of reads to the contig (X axis).

Similarly, the CS505_48 contig, which encompasses the glycolipid layer formation cluster, was mostly covered but some hypothetical genes and transposases were missing in regions “A” and “C”, intercalated by region “B” (Fig. 5), where the core gene CRC_02015 Photosystem II reaction center protein *PsbA*/*D1* was highly covered. The glycolipid layer formation cluster presented covered genes, from CRC_02047 to CRC_02052, encompassing

the genes *hgl*E,G,D,C,A and *het*M. We could observe a large number of low similarity reads aligned to this region (Fig. 5), as well as to other regions of the contig. Such population of reads aligned with low similarity could be related to populational genetic diversity, encompassing individuals with high similarity to *C. raciborskii* CS505 genes (above 99% similarity) as well individuals from other strains (above 95% similarity) or even other species (below 95% similarity).

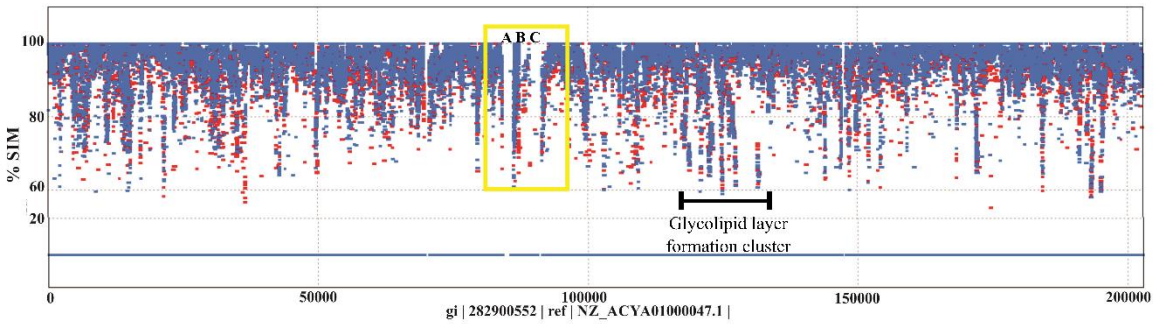


Fig. 5 - Recruitment plot of P1M dataset to CS505_48 contig, highlighting some missing genes and a low similarity population of reads aligned to the glycolipid layer formation cluster. A: absent transposases CRC_02011 to CRC_02014. B: CRC_02015 Photosystem II reaction center subunit IX. C: absent hypothetical protein and transposases CRC_02016 to CRC_02020. Y axis represent the percent of similarity of reads to the contig (X axis).

Low similarity alignments could also be seen in contig CS505_13, in addition to an overall high coverage (Fig. 6). The genes present in those regions showed a high depth of coverage, mostly encoding hypothetical proteins, transposases, predicted general function proteins and unknown function genes, which have been found over the whole genome of cyanobacteria (Kaneko et al., 2007; Steffen et al., 2012; Frangeul et al., 2008) and are framed as part of the shared genome.

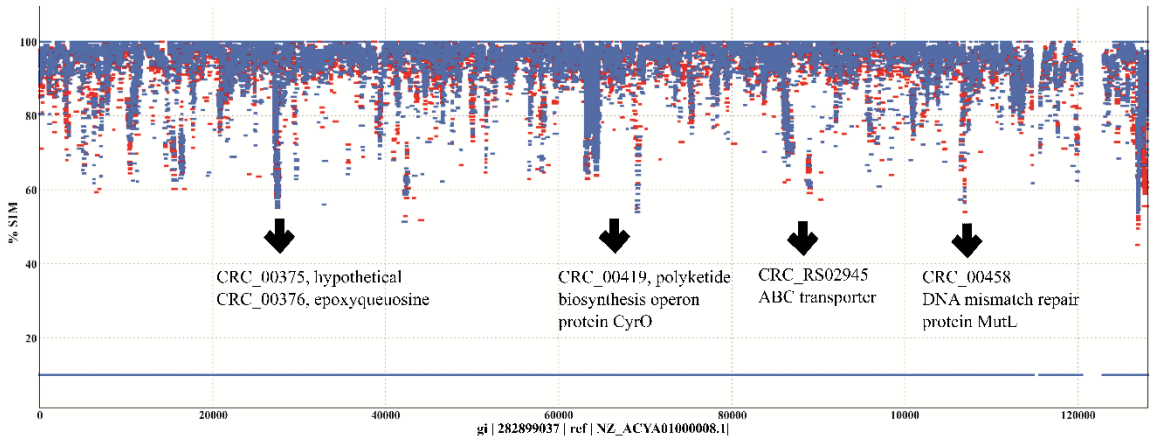


Fig. 6 - Recruitment plot of P1M dataset to CS505_13 contig, highlighting highly covered genes with a low similarity population of reads also aligned. Y axis represent the percent of similarity of reads to the contig (X axis).

Both datasets mapped against all *R. brookii* D9 contigs, and all of them showed over 0.8 breadth of coverage and presented a very homogeneous read distribution (Suppl. Table S1). The D9_5 contig, encompassing the saxitoxin cluster, along with amino acid metabolism and ion metabolism pathways, presented a broad and deep coverage, but was not among the contigs with the deepest coverage (Fig. 7). The saxitoxin cluster in the P1M sample presented a mean depth of coverage of 112.4, compared to the 311.2 mean for the entire contig. Two *R. brookii* D9 unique genes were not covered in sample P1M, CRD_02148 and CRD_02149, located between the *sxtM* and *sxtL* genes (Fig. 8). These genes are involved in oxidoreductase activity, iron ion binding and sulfotransferase activity.

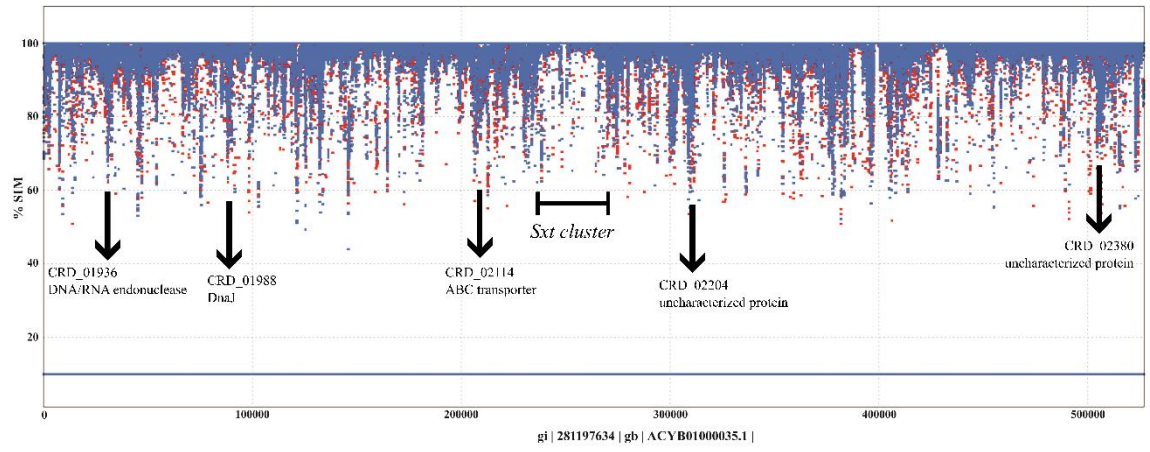


Fig. 7 - Recruitment plot of P1M dataset against D9_5 contig, highlighting the lower coverage under the saxitoxin gene cluster. Y axis represent the percent of similarity of reads to the contig (X axis).

Despite the overall lower depth and the prevalence of Chroococcales order, sample P3M showed a similar mapping profile. Gene CRD_02142 (*sxtQ*) presented lower depth in sample P3M and a breadth of coverage of 0.6 (Fig. 8). The *sxtQ* gene differs between *R. brookii* D9 and *C. raciborskii* T3 by only one amino acid, so we extracted and blasted the sequences against NCBI nucleotide database and all sequences presented 100% identity with T3 *sxtQ*.

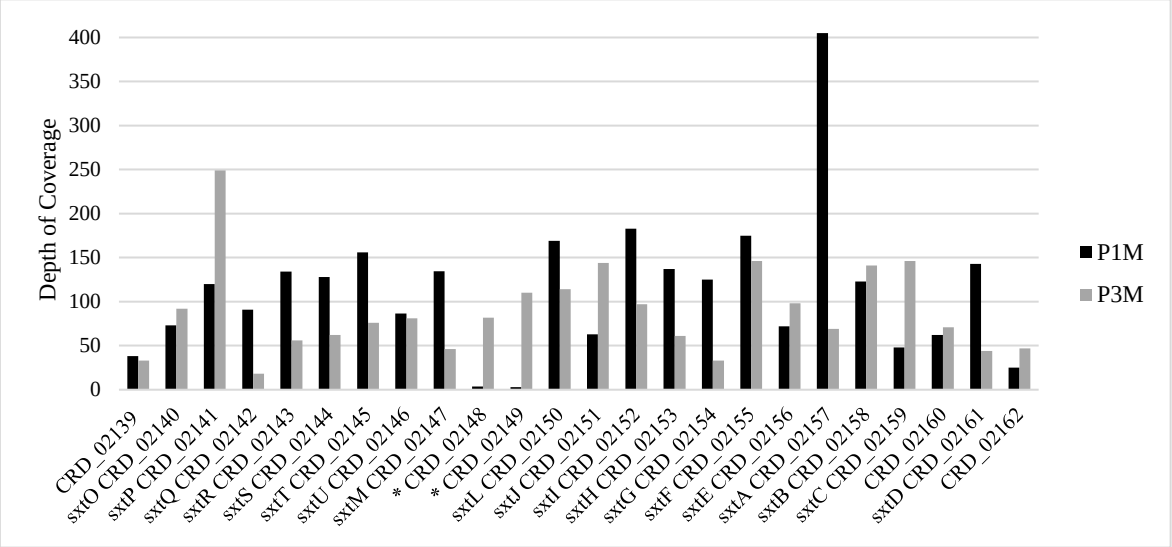


Fig. 8 - Depth of coverage of samples P1M (more Nostocales) and P3M (more Chroococcales) for several saxitoxin gene clusters of *R. brookii* D9, highliting the differences among the samples.

Among the lower coverage contigs, we observed some regions with an important decrease in coverage. For contig D9_1144, we observed a long region of low coverage encompassing unique *R. brookii* D9 genes from CRD_00152 to CRD_00182 in sample P1M (Fig. 9) and from CRD_00147 to CRD_00182 in sample P3M. These genes are involved in secondary metabolite biosynthesis, cell wall/membrane biogenesis, and carbohydrate transport and metabolism, and they seem to be related to medium-frequency genes, whose dynamics depend on functional roles within the population (Cordero & Polz, 2014).

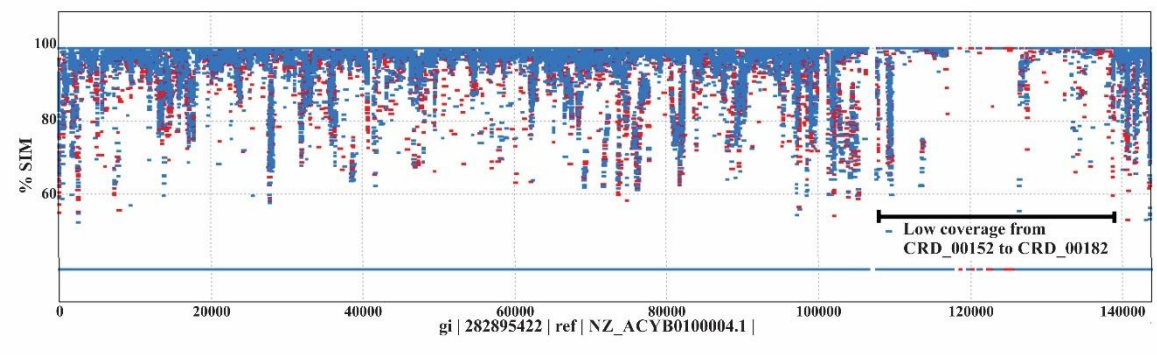


Fig. 9 - Recruitment plot of P1M dataset against D9_1144 contig, highlighting a large lower coverage region. Y axis represent the percent of similarity of reads to the contig (X axis).

The gene cluster (CRD_00336 to CRD_00339), related to nitrogen control and consisting of three nitrile hydrases (NhlE, NthA, NthB) and one associated cobalamin synthesis protein (Stucken 2010), was missing in contig D9_1260 in sample P3M, and presented lower coverage for sample P1M (Fig. 10).

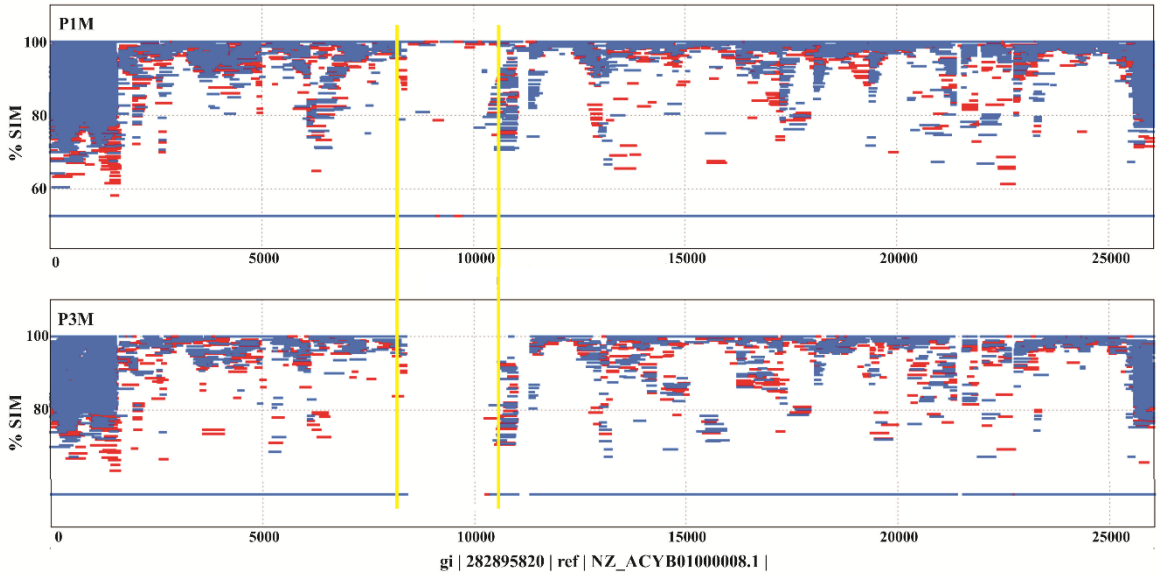


Fig. 10 - Recruitment plot of P1M and P3M datasets against D9_1260 contig, highlighting a low coverage region. Y axis represent the percent of similarity of reads to the contig (X axis).

The MG-RAST pipeline pointed to *C. raciborskii* CS505 and *R. brookii* D9 as the major strains that made up the population, however, the read mapping and alignment of low coverage regions pointed to an otherwise complex taxonomical structure. Because of the lack of CS505 *cyr* gene cluster, the missing unique genes on D9 *sxt* cluster and the complete alignment of the metagenomic reads to the T3 *sxt* cluster, which presents 7 additional genes

(Stucken et al., 2010), we also performed a recruitment analysis for the *C. raciborskii* T3 strain (Fig. 11).

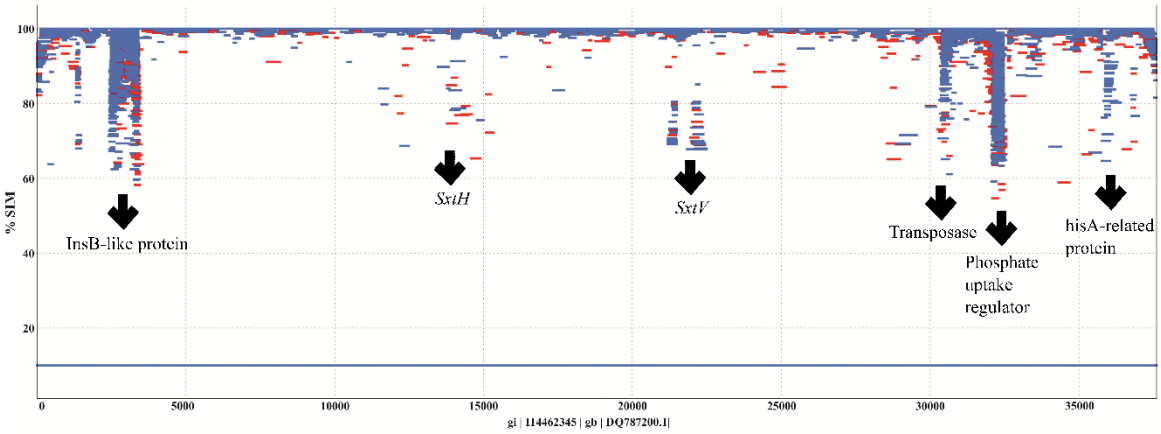


Fig. 11 - Recruitment plot of P1M dataset against T3 sxt cluster, highlighting some deep covered genes above 70% identity. Y axis represent the percent of similarity of reads to the contig (X axis).

C. raciborskii T3 presents a PSP toxin biosynthesis gene cluster (*sxt*) and the saxitoxin biosynthetic pathway encompasses 26 proteins (Kellmann et al., 2008). The alignment of both samples against the complete toxin cluster of *C. raciborskii* T3 generated 100% mapping among all genes, mostly with 100% breadth of coverage. The average depth of coverage in sample P1M was 114.4, while in sample P3M it was 47.8 (Fig. 11).

The large number of different taxa could be interpreted as insufficient nucleotide databases or high population variability. The mapping of several annotated taxa from the Nostocaceae family, which alignments were filtered above 97% of identity and one-to-one mapping, pointed to a genus level diversity (Tab. 4) of partial sequences.

Tab. 4 - Depth and breadth of coverage concerning reads aligned above 97% of similarity to Nostocaceae species. avgDP=average number of reads aligned (SAMtools); AvgBreadthCov=average breadth of coverage measured for sequences above 97% of identity, concerning the whole reference sequence.

		avgDP	AvgBreadthCov
A. cylindrica PCC7122	P1M	40.8	0.036
	P3M	20.1	0.033
A. variabilis PCC29413	P1M	32.5	0.027
	P3M	17.5	0.025
A. siamensis	P1M	103.4	1.000
	P3M	61.3	0.436
Nostoc sp. PCC7524	P1M	33.3	0.027
	P3M	18.25	0.026
N. punctiforme	P1M	29.1	0.018
	P3M	16	0.017
N. azollae	P1M	45.8	0.040
	P3M	21.9	0.037
Raphidiopsis curvata	P1M	323.7	1.000
	P3M	221.7	1.000
C. raciborskii T3	P1M	123.3	1.000
	P3M	50.3	1.000

The *Anabaena siamensis* species was pointed as abundant according to the MG-RAST pipeline, and even though it does not have a complete sequenced genome yet, the aligned sequences showed 100% breadth of coverage in sample P1M and 43% in sample P3M. The contigs of *R. brookii* D9 showed higher depth and breadth of coverage than the CS505 contigs. The 16S complete sequence of *R. curvata* showed 100% coverage and high depth of coverage in both samples, but no alignment for *cyr* genes. We also checked for alignments against all *Cylindrospermopsis* sequences from Brazil and Uruguay (Tab. 5).

Tab. 5 - Brazilian, Uruguayan and Australian partial *Cylindrospermopsis raciborskii* sequences aligned to P1M and P3M metagenomic datasets.

	cyr cluster	**sxt cluster/genes**	**16S**	**other**
C. raciborskii T3	-	ABI75088 - ABI75121	JQ707295.1	-
C. raciborskii CP-2	-	-	HM991731.1	-
C. raciborskii cena302	-	JX175236.1 (sxtA)	JQ707291.1	-
	-	JX175234.1 (sxtB)	-	-
	-	JX175232.1 (sxtI)	-	-
C. raciborskii cena303	-	JX175237.1 (sxtA)	JQ707292.1	-
	-	JX175235.1 (sxtB)	-	-
	-	JX175233.1 (sxtI)	-	-
C. raciborskii cena305	-	KC894587.1 (sxtA)	JQ707293.1	-
	-	KC894588.1 (sxtB)	-	-
	-	KC894589.1 (sxtI)	-	-
C. raciborskii cena306	-	-	JQ707294.1	-
C. raciborskii CYP011K	-	-	JQ707296.1	-
C. raciborskii MVCC14	-	JX105890.1 (sxtM)	HQ112347.1	HQ112345.1 (nifH)
	-	JX105887.1 (sxtF)	-	-
C. raciborskii MVCC19	-	-	HQ112348	HQ112346 (nifH)
C. raciborskii AWT205	CyrO	-	-	ABX60157 (transp)

The saxitoxin cluster of *C. raciborskii* T3 was 100% covered, with 100% breadth of coverage and higher depth in sample P1M. We also found 100% identity and breadth of coverage for all *sxt* genes and 16S ribosomal RNA gene of CENA302, CENA303, CENA305 and CENA306 (only 16S sequence available), while no reads aligned against *cyr* genes. According to Hoff-Risseti et al., (2013), the 16S rRNA sequences of the CENA302, CENA305 and T3 strains, isolated from the Billings reservoir, showed similarity above 99% with that of the *R. brookii* D9, found in the same reservoir. In the same study, the authors found a distinct cluster for *Raphidiopsis* genera, however, the Brazilian *R. brookii* D9 strain clustered together with T3 and CENA302, 303 and 305 strains. The Uruguaian strains MVCC14 and MVCC19 presented full alignment for the partial *sxtM* gene and the *sxtF* complete cds, the 16S ribosomal sequences, *nifH* and a tranposase, but no alignment against the recently described *cyr* genes in the Uruguaian saxitoxin producers (Piccini et al., 2013). The relatedness of the 16S ribosomal gene sequences of Uruguaian and Brazilian strains has already been observed (Piccini et al., 2013), as well as the detection of PSP. The c*yrA* sequence of the T3 strain presented no alignments, as also observed for *cyr* genes from the Australian strains *C. raciborskii* CYP011K and CS505, but all 16S ribosomal RNA sequences mapped above the alignment cutoffs. In contrast, the c*yrO* gene from the type strain *C. raciborskii* AWT205 presented alignment over 97% of identity but breadth of coverage of 0.6. The *cyrO* gene was also found in the CS509 strain, a non-CYN-producer

strain, between the *hyp* genes, and this putative *cyr* regulatory protein may have an alternative function, or at least may be not essential for CYN-production (Sinha et al., 2014).

Phenotypic, physiological, and genetic variability regarding *Cylindrospermopsis* and *Raphidiopsis* populations have been well documented through a vast bibliography (Gugger et al., 2005; Kellmann et al., 2008; Moustaka-Gouni et al., 2009; Stucken et al., 2010; Bittencourt-Oliveira et al., 2011b; 2012a; Riccini et al., 2011; 2013; Elias et al., 2015). The presence of different genotypes in the same study area has also been observed and it has been proposed that different *C. raciborskii* ecotypes might undergo different fitness and physiological adaptations, guided by selective pressures (Piccini et al., 2011; Hoff-Risseti et al., 2013). Moustaka-Gouni et al., (2009) observed structurally distinct reproductive stages in the developmental cycle in planktonic Nostocales and suggested that even the distinction between *Cylindrospermopsis* and *Raphidiopsis* should be questioned. Our study shows that, in a same population, a vast diversity of alignments that are both high in identity and in breadth of coverage can be found. That may result in an inflated taxa profile, since the annotation will depend on sequencing and interpretation methods. In spite of the high depth observed in some regions/contigs, none of the complete genomes recruited showed full breadth of coverage. *R.brookii* D9 and *C. raciborskii* T3 seem to be the best candidates to compose the backbone genome (Kashtan et al., 2014) of this population, since they presented the most robust alignment identity and coverage.

The broad alignment overlap among the strains could be reflecting an unknown population structure, in which a totally new strain may be discovered, or an interacting network of gene dynamics driven by selective pressures could be in action (Whitaker et al., 2005; Antonova & Hammer, 2011; Zhu et al., 2012; Biller et al., 2015). The latter hypothesis means that *Cylindrospermopsis* and *Raphidiopsis* strains might have been coevolving in obligate interspecies, whose interactions are mutually beneficial and highly specific (Boon et al., 2014).

Population SNV Calling

Since *C. raciborskii* and *R. brookii* D9 were annotated as the dominant taxa in this lake, the variant calling for the two complete genomes was performed in order to look for punctual differences between these strains and the actual local nostocaceae population. SNV

calling aims to determine in which positions there are variants or in which positions at least one of the bases differs from a reference sequence (Nielsen et al., 2011) and it is essential for comparative genomics, as it yields insights into nucleotide level organismal differences and sequence-discrete populations (Olson et al., 2015; Bendall et al. 2016). For SNV discovery, using raw reads can provide greater resolution than using a genome assembly, but for reference-based SNV discovery, reads can fail to align to regions of high divergence, so assembly and merging paired-end methods allow alignments onto more divergent sequences, even from different bacterial species within the same genus (Olson et al., 2015). According to Cingolani et al., (2012a), the variants can be categorized in low, modifier, moderate and high impact. The high (disruptive) impact in the protein probably causes protein truncation, loss of function or triggers nonsense mediated decay. Therefore, high impact variants can be indicators of rearrangement and genomic exchanges (Rosenthal et al., 2015). In our study, the P1M (rainy period) and P3M (dry period) samples presented a highly similar variant profile, according to both reference genomes analyzed (*C. raciborskii* CS505 and *R. brookii* D9). The total number of variants was higher for CS505, which could be a result of the lower alignment/coverage rate for CS505 strain. Besides, the overall percentage of high impact variants was slightly higher for the D9 genome (Fig 12 and Tab. 6), mostly regarding the D9_5 contig. The higher proportion of 'modifier' (usually non-coding variants, unspecific predictions), 'silent' and 'intergenic' variants (downstream and upstream region) in both datasets could be reflecting a sequence-discrete population, composed of co-occurring genotypes under neutral genetic variation, allowing many highly similar genotypes to coexist without outcompeting each other (Bendall et al., 2016). The contigs that presented the higher number of variants are presented in Table 6.

Tab. 6 - Variant annotation summary for *C. raciborskii* and *R. brookii* D9. Data extracted from SnpEff summary, showing the total number of variants detected, the ratio of transitions and transversions (Ts/Tv), variant effect and average variant rate.

	Cylindrospermopsis	***Raphidiopsis***
Total variants	125551	71148
Ts/Tv ratio	2.626	2.5992
High impact	0.18%	0.25%
Low impact	6.04%	4.97%
Moderate impact	2.67%	2.67%
Modifier impact	91.1%	92.1%
Missense functional class	30.2%	34.5%
Nonsense functional class	0.135%	0.21%
Silent functional class	69.66%	65.3%
Downstream region	44.8%	44.95%
Exon region	8.9%	7.9%
Intergenic region	1.2%	1.6%
Transcript region	0.005%	0.007%
Upstream region	45.1%	45.56%
Average Variants rate	29	44

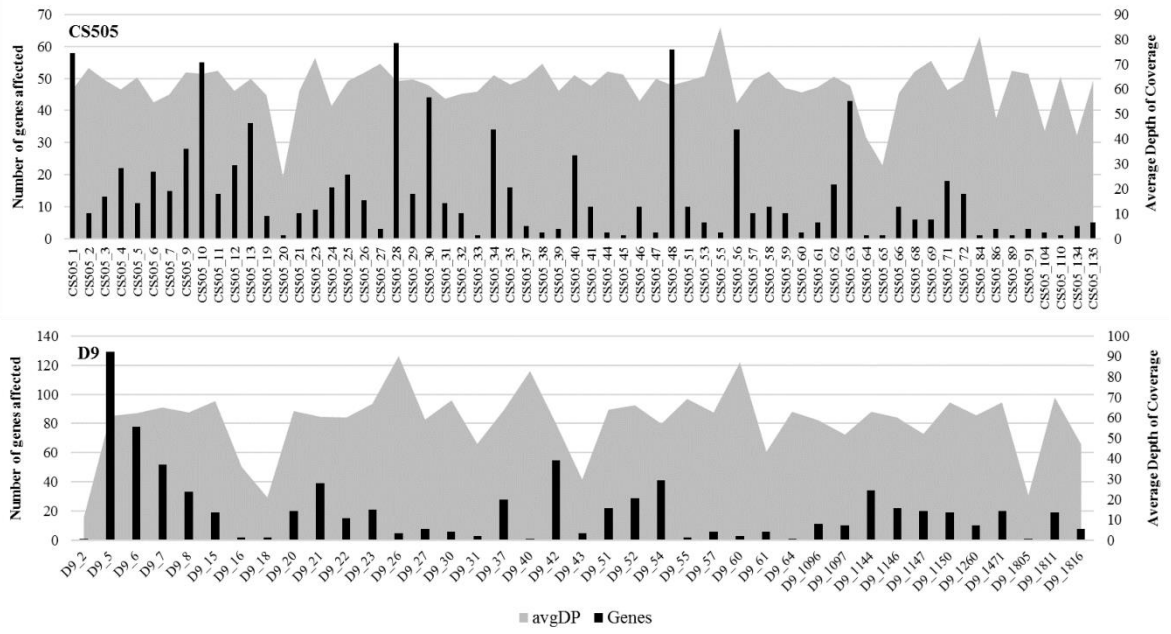


Fig. 12 - Overall distribution of high impact variants in all CS505 and D9 contigs, in sample P1M (same distribution but higher coverage than P3M). Bars (left axis) represent number of genes affected for each contig. Gray area (right axis) illustrates the coverage, according to the depth of respective variant coding regions.

Tab. 7 - Reference contigs with larger amount of SNVs, total number of variants detected, variant rate (average number of base pairs between two variants), number of high impact variants and breadth of coverage for each sample. The variants summary was extracted from 'concordance' file between P1M and P3M samples.

Strains	**Total variants**	**Variant rate**	**High impact**	**Breadth of coverage**	
	P1M/P3M	P1M/P3M	P1M/P3M	P1M	P3M
D9_5	11723	44	277	1.0	1.0
D9_6	5826	46	156	1.0	0.9
D9_42	5525	37	128	1.0	1.0
D9_54	3218	39	115	1.0	1.0
D9_21	4177	30	113	1.0	1.0
D9_7	4842	49	107	1.0	1.0
D9_1144	2967	48	88	1.0	0.9
CS505_28	8449	30	160	0.8	0.8
CS505_48	8060	25	140	1.0	1.0
CS505_1	8058	26	129	0.9	0.9
CS505_10	6955	23	131	1.0	1.0
CS505_63	6944	27	106	0.9	0.9
CS505_13	5151	24	94	1.0	1.0
CS505_30	4983	31	90	0.8	0.8
CS505_56	4939	37	81	0.7	0.7
CS505_4	4604	23	85	1.0	0.9

In the coding regions of CS505, the SNVs were observed in more detail for the contigs with higher amount of variants and higher variant rate (Tab. 7). The functional category of high impact variants in coding regions of selected contigs were annotated according to the Gene Ontology database. The total set of genes and functions are presented in Supplementary Table S3 and S4. The uncharacterized proteins and unknown functions were prevalent in both samples. Among the most frequent functions annotated for genes under high impact variants for CS505 contigs 1, 10, 13, 28, 48 and 63, are: ATP binding (the cell's source of energy and phosphate), integral component of membrane (which discerns the appropriate spatial relationship between a membrane and a gene product), hydrolase activity, transferase activity (both enzyme functions that catalyse multi-stage reactions), plasma membrane (phospholipid bilayer and associated proteins), DNA binding (gene product interacting selectively and non-covalently with DNA), catalytic activity (enzyme activity), metal ion binding (interacting selectively and non-covalently with any metal ion), membrane (lipid bilayer along with all the proteins and protein complexes embedded in it and attached to it.), cobalamin metabolism (vitamin B12), intracellular (the living contents of a cell), magnesium ion binding (interacting selectively and non-covalently with magnesium ions), oxidoreductase activity (catalysis of an oxidation-reduction reaction) and ATPase activity (ATP hydrolase activity). According to Sinha et al. (2014), the CS505 genome encodes a

large number of ATP Binding Cassette (ABC) transport-related genes, mainly responsible for the transport of glycerophospholipids and showing a strain-specific nature, suggesting a specific ecological adaptation concerning membrane structure and permeability. We found ATP binding and integral component of the membrane as the major functions associated with high impact variants, which can be related to the specific nature of membrane structure, transporters and transposases, owing to the highly plastic nature of the *C. raciborskii* genomes (Sinha et al., 2014).

The most frequent functions of the *R. brookii* D9 genes under high impact variants were ATP binding, DNA binding, metal ion binding, oxidoreductase activity, cytoplasm, hydrolase activity, integral component of membrane, intracellular, transferase activity, membrane, methyltransferase activity (catalysis of the transfer of a methyl group to an acceptor molecule), iron-sulfur cluster binding, ATPase activity, magnesium ion binding and nucleic acid binding. Besides the distinct toxin profile, the more prominent differences between CS-505 and D9 at the genomic level are the absence of nitrogenase and heterocyst development gene clusters in D9, which, on that basis, presents more genes involved in amino acid transport and metabolism (Stucken et al., 2010). We could observe a similar functional profile between the genes under high impact variants for the two strains, encompassing central metabolism functions involved in energy, DNA processes, ion interactions and membrane. However, a large number of strain related functions was also observed (Suppl. Table S3 and S4). That could be related to pathway complementation, in which the population members interchange functional capabilities (Helling et al., 1987; Wintermute & Silver, 2010) through adaptive genomic regions (Shapiro et al., 2012; Shapiro & Polz, 2014; Rosen et al., 2015; Bendall et al., 2016).

Among the genes with larger number of high impact variants (Tab. 8) we observed some central metabolism genes for contig CS505_1, like the sequence-specific DNA binding CRC_00508 and the DNA polymerase III subunit beta dnaN, responsible for most of the replicative synthesis in bacteria. For contig CS505_3, which encompasses the cylindrospermopsin biosynthesis cluster (cyr), the flanking genes hypB, A, E, D, F presented high coverage (aDP, Tab. 8), but some of these genes presented high impact variants, with alternate allele frequency of 1. The presence of such non-synonymous variants on these genes could be related to the large number of low similarity alignments observed in recruitment plot (Fig. 3). The contigs CS505_48 ans CS505_63 also presented some central metabolism genes with high impact variants, among them a Leucine-tRNA ligase, a CHAD

protein (conserved histidine α-helical domain), predicted to either chelate metals or serve as phosphoacceptors acting in the interface between nucleotide and polyphosphate metabolism (Lakshminarayan & Aravind, 2002), a ribosomal RNA small subunit methyltransferase and the Peptidoglycan-binding *LysM*. The *nifB* and *nifS* genes, part of the nitrogenase operon, were among the high impact genes, presenting, in GATK terms, frameshift plus stop-gained and stop-lost effects, respectively. The nitrogen metabolism and phosphorus metabolism genes were conserved among three *Cylindrospermopsis* strains CS-505, CS-506, and CS-509 (Sinha et al., 2014), but we found non synonymous coding variation for all nitrogenase operon genes and for nitrite/nitrate uptake and reduction operon genes, including frameshift, codon change plus codon deletion/insertion and frameshift plus stop-gained for *nrtA* gene, a twin-arginine translocation protein, part of the nitrite/nitrate uptake and reduction operon, and codon change plus codon insertion/deletion and frameshift for *narB* gene. Similarly, we found non synonymous coding variation for all *pho* operon genes, including high impact frameshift effect for *phoB*, a multi-component transcriptional regulator protein, and the phosphate-binding protein CRC_01734.

Tab. 8 - *C. raciborskii* CS-505 genes mapped with larger number of high impact variants from selected contigs showing the approximate depth of coverage (aDP), the variant effect detected (Effect), the allele depth (AD), the filtered depth of coverage (DP) and the alternate allele frequency (Freq). Variant calling data from 'concordance' function between the samples P1M and P3M and allele frequency data from the deepest high impact variant for each gene. FS=frameshift, UN=uncharacterized protein.

Contig	Gene	Gene Name	aDP	Effect	AD	DP	Freq
CS505_1	CRC_00508	DNA-binding	246	FS	46	46	1
CS505_1	CRC_00540	ATPase	644	FS/Stop-gained	20	20	1
CS505_1	CRC_00689	UN	540	FS	64	64	1
CS505_1	CRC_00696	dnaN	264	FS	51	51	1
CS505_3	CRC_01700	hypB	282	FS	59	59	1
CS505_3	CRC_01703	hypD	268	FS	54	54	1
CS505_3	CRC_01720	hypF	574	FS	46	46	1
CS505_10	CRC_00042	narB	438	FS	39	39	1
CS505_10	CRC_00046	nrtA	297	FS	76	76	1
CS505_10	CRC_00098	UN	216	FS	52	54	0.9
CS505_13	CRC_00375	hypothetical	172	Syn/Non-Syn	68	68	1
CS505_13	CRC_00376	queG	251	Syn/Non-Syn	57	57	1
CS505_13	CRC_00419	cyrO	1639	Syn/Non-Syn	236	289	0.8
CS505_13	CRC_00436	ABC transporter	170	Syn/Non-Syn	54	54	1
CS505_13	CRC_00458	mutL	353	Syn/Non-Syn	72	72	1
CS505_25	CRC_00858	ntcA	159	FS	53	53	1
CS505_28	CRC_01116	phoB	839	FS	41	41	1
CS505_28	CRC_01144	ABC transporter-like	278	FS	33	33	1
CS505_28	CRC_01043	Rho	348	FS	20	23	0.8
CS505_40	CRC_01734	Phosphate-binding	270	FS	64	64	1
CS505_48	CRC_02036	leuS	584	FS	57	57	1
CS505_48	CRC_01988	CHAD	242	FS	21	21	1
CS505_48	CRC_02047	hglE	989	FS	43	43	1
CS505_48	CRC_02048	hglG	326	Syn/Non-Syn	47	47	1
CS505_48	CRC_02049	hglD	312	Syn/Non-Syn	50	50	1
CS505_48	CRC_02050	hglC	677	FS	43	43	1
CS505_48	CRC_02051	hglA	294	Syn/Non-Syn	49	49	1
CS505_48	CRC_02052	hetM	370	Syn/Non-Syn	39	39	1
CS505_48	CRC_02113	UN	218	FS	57	57	1
CS505_63	CRC_02750	UN	385	FS	50	50	1
CS505_63	CRC_02819	metallophosphoesterase	346	FS	50	50	1
CS505_63	CRC_02763	Arsenical pump membrane	251	FS	40	46	0.8
CS505_63	CRC_02767	RNA methyltransferase	313	FS	51	51	1
CS505_63	CRC_02831	UN	76	FS	24	24	1
CS505_63	CRC_02832	lysM	346	FS	51	53	0.9
CS505_63	CRC_02891	nifB	326	FS/Stop-gained	45	45	1
CS505_63	CRC_02889	nifS	345	Start-lost	59	59	1

Among the *R. brookii* D9 contigs with higher number of variants (Tab. 9) we highlight the CRD_02147 *sxtM* gene, probably involved in the export of PSP toxins (Kellmann et al., 2008), which presented a high impact variant through stop-gained effect due to a transversion. Some of the genes which presented a high depth of coverage and a broad read alignment below 95% similarity (Fig. 7) carry high impact variants. There were

two genes that aligned to the missing genes regions of contig D9_1144 (Fig. 9) and presented frameshift variants, but due to low DP and low quality they were filtered off from SNV call. The unique genes Putative Nitrile hydratase activator (NhlE), Putative Nitrile hydratase alpha subunit (NthA), Putative Nitrile hydratase beta subunit (NthB) and Cobalamin synthesis protein/P47K were missing in contig D9_1260 in sample P3M, and presented lower coverage for sample P1M (Fig. 10), but we found some frameshift variants just before this region, encompassing genes CRD_00332 to CRD_00335 (Tab. 9), three of them uncharacterized proteins which could be related to the subsequent missing genes.

Tab. 9 - *R. brookii* D9 genes mapped with larger number of high impact variants from selected contigs showing the approximate depth of coverage (aDP), the variant effect detected (Effect), the allele depth (AD), the filtered depth of coverage (DP) and the alternate allele frequency (Freq). Variant calling data from 'concordance' function between the samples P1M and P3M and allele frequency data from the deepest high impact variant for each gene. FS=frameshift, UN=uncharacterized protein.

Contig	Gene	Gene Name	aDP	Effect	AD	DP	Freq
D9_5	CRD_01936	DNA/RNA endonuclease	246	FS	46	46	1
D9_5	CRD_01988	Heat shock DnaJ-like	644	FS/Stop-gained	20	20	1
D9_5	CRD_02114	ABC transporter-like	540	FS	64	64	1
D9_5	CRD_02147	sxtM					
D9_5	CRD_02204	UN	264	FS	51	51	1
D9_5	CRD_02380	UN	282	FS	59	59	1
D9_6	CRD_02594	Twin-arginine translocation	268	FS	54	54	1
D9_6	CRD_02485	UN	574	FS	46	46	1
D9_6	CRD_02671	Peptidase C14	438	FS	39	39	1
D9_42	CRD_01534	UN	297	FS	76	76	1
D9_42	CRD_01593	UN	216	FS	52	54	0.9
D9_42	CRD_01475	plqA	172	Syn/Non-Syn	68	68	1
D9_54	CRD_01849	UN	251	Syn/Non-Syn	57	57	1
D9_54	CRD_01778	UN	1639	Syn/Non-Syn	236	289	0.8
D9_54	CRD_01814	Dioxygenase protein	170	Syn/Non-Syn	54	54	1
D9_7	CRD_02900	serine/threonine phosphatase	353	Syn/Non-Syn	72	72	1
D9_7	CRD_02925	cell envelope-related	159	FS	53	53	1
D9_7	CRD_02788	UN	839	FS	41	41	1
D9_7	CRD_02934	DevB	278	FS	33	33	1
D9_1144	CRD_00156	Nucleotidyl transferase	18	FS	-	-	-
D9_1144	CRD_00170	Phosphoheptose isomerase	44	FS	-	-	-
D9_1260	CRD_00332	3'-5' exonuclease	140	FS	43	43	1
D9_1260	CRD_00333	UN	321	FS	47	56	0.8
D9_1260	CRD_00334	UN	82	FS	60	61	0.9
D9_1260	CRD_00335	UN	164	FS	47	47	1

To complement the population variant analysis, we also performed a variant call for *C. raciborskii* T3 toxin biosynthesis gene cluster. After the realignment, filtration and recalibration steps, following the GATK pipeline, 70 remaining variants were distributed throughout 24 of the 34 genes, with a ts/tv ratio of 1.9 (lower than for CS505 and D9, see Tab. 5). The alr5035-like protein (ABI75088.1) presented the highest number of variants (Tab. 7) and the InsB-like protein, SxtH, SxtV, ORF6 transposase, OmpR family protein and HisA-related protein, which presented a high depth of coverage and a large number of low similarity alignments (Fig. 11) also presented single nucleotide polimorfisms (Tab. 10).

Tab. 10 - Summary of variant call for T3 toxin cluster showing the allele and alignment attributes for the variants with highest QUAL. Ref: reference allele, Alt: alternate allele, aDP: approximate depth of coverage, P1M_freq/ P3M_freq: alternate allele frequency considering filtered reads for each sample.

protein_id	protein	Ref	Alt	aDP	P1M_freq	P3M_freq
ABI75088.1	alr5035-like protein	A	C	144	1	1
ABI75089.1	sxtD	A	T	144	1	1
ABI75090.1	InsB-like protein	CATTG	C	1241	0.4	0.5
ABI75093.1	sxtB	T	C	104	0.9	0.9
ABI75094.1	sxtA	T	C	107	0.9	0.9
ABI75097.1	sxtG	C	T	76	0.9	0.9
ABI75098.1	sxtH	C	A	103	0.8	1
ABI75099.1	sxtI	T	C	91	0.9	1
ABI75100.1	sxtJ	G	A	69	0.9	1
ABI75102.1	sxtL	A	G	140	0.9	0.9
ABI75103.1	sxtM	A	C	86	0.8	1
ABI75106.1	sxtW	C	T	100	0.9	1
ABI75107.1	sxtV	T	G	108	0.9	1
ABI75110.1	sxtS	G	A	93	0.9	1
ABI75111.1	cyanophage S-PM2 protein	A	G	98	0.9	1
ABI75113.1	sxtQ	C	G	95	0.9	1
ABI75114.1	sxtP	C	T	78	0.9	0.9
ABI75115.1	sxtO	A	C	78	0.6	0.6
ABI75116.1	ORF6 transposase	C	T	551	0.8	0.8
ABI75117.1	sxtY	C	T	190	0.8	0.9
ABI75118.1	sxtZ	C	T	182	0.9	1
ABI75119.1	OmpR family	T	C	146	1	1
ABI75120.1	ORF8 HisA-related protein	G	A	173	1	1
ABI75121.1	ORF9 unknown	T	C	140	0.9	1

Interestingly, there were 3 discrepant variants between the samples, in contrast to the exact same variant profile for both samples against CS505 and D9. In sample P1M, we found an extra variant for ABI75119.1, a two-component OmpR family protein, with a transition in the position 669, below average mapping quality (Phred-scaled probability that a REF/ALT polymorphism exists) and depth quality (QD). Sample P3M presented an extra

variant in the aforementioned protein, though in position 204, and a deletion in the transposase InsB-like protein, also below average mapping quality and QD. Such variations could be part of the population dynamics of this strain, but could be indicating, instead, a distinct population of a saxitoxin producer species.

4. CONCLUSIONS

This work presented the strain level analysis of a complex microbial community through shotgun metagenomic strategies, which enabled exploration in different scales, from broad taxonomical profile to fine scale functional analysis of the strain variants (Bendall et al., 2016). We could observe a highly diverse sequence-discrete population (Oh et al., 2011; Bendall et al., 2016), recruiting several nostocaceae members simultaneously (see Suppl. Table S2). This set of mapped genes across different genera/species/strains indicates that such population presents a wide backbone genome (Kashtan et al., 2014), and the strain level of interaction probably plays the major role, regardless of taxonomical ranks (Luo et al., 2015). The definition of bacterial populations as ecologically distinct gene-flow units is an important step towards the better understanding of the microbial genotypic variation (Cordero & Polz, 2014; Shapiro & Polz, 2014), inasmuch as homologous recombination and/or selective genome-wide/gene sweeps constantly shape the population structure according to environmental pressures (Shapiro et al., 2012; Rosen et al., 2015; Bendall et al., 2016). In our study, the overall community profile pointed to a distinct taxonomical (Fig. 1 and 2) and functional profile between samples (Suppl. Fig. 1) and guided us into looking for the major cyanobacteria population acting in this system. The OTU strategy resulted in closely clustered OTUs that presented divergent taxonomical annotation (Fig. 4, 5 and 6). This generated a high level of uncertainty and, along with the read mapping against several population member candidates (Suppl. Table S2), it became clear that, depending on the method and scale, the taxonomical annotation can be either inflated (Tab. 2) or unspecific.

Through recruitment and SNV calling strategies, which allows the observation of intra-population diversity (Bendall et al., 2016), we could see a large number of high-impact variants throughout the CS505 and D9 contigs analyzed (Fig. 16). In spite of the conserved nature of nitrogen and phosphorus metabolism genes related to *Cylindrospermopsis* strains CS-505, CS-506, and CS-509 (Sinha et al., 2014), we found non-synonymous coding

variation for all nitrogenase operon genes and for nitrite/nitrate uptake and reduction operon genes (Fig. 17, CS505_63), probably responsible for the lower coverage observed for *Nif* cluster (Fig. 8). Similarly, we found non-synonymous coding variation for all *pho* operon genes. Among D9's high impact variants, the CRD_02147 *sxtM* gene (Fig. 18, D9_5 first half) and *devB* gene, part of the glycolipid layer formation cluster, reinforced that a sequence-discrete population (Bendall et al., 2016) is being reflected on these genes, under medium-frequency variation dynamics (Cordero & Polz, 2014), since a great part of the high impact genes detected are related to central metabolism and important accessory clusters. The general mapping and variant pattern was very similar in both rainy and dry season samples, even under larger *Microcystis* contribution. The larger proportion of synonymous variants, mostly "Modifier" (usually non-coding variants or variants affecting non-coding genes) and "Silent" (Tab. 5) suggests that most of the variation within this nostocaceae sequence-discrete population might be neutral, allowing many highly similar genotypes to coexist (Bendall et al., 20016). The lower Ti/Tv ratio found for *C. raciborskii* T3 strain (1.9, while 2.6 for CS505 and D9) could represent either a different genotype from CS505 and D9, or a low quality call for the T3 reference sequence (complete saxitoxin cluster sequence), generating false positive variants (Liu et al., 2012). In fact, the higher overall coverage presented by D9 relatively to CS505, the absence of *Cyr* cluster in CS505 recruitment, the lack of two exclusive D9 genes from *sxt* cluster (Fig. 12), the high impact variants on *sxtM* and *devB* D9 genes and the complete mapping of T3 *sxt* cluster (Fig. 15) seems to indicate that a yet unknown population of *Cylindrospermopsis* could be playing a major role in the Pampulha lake population dynamics. We can speculate, according to the mapping against south-american partial sequences, that we might be dealing with an ecotype, non-CYN producer, closely related to *R. brookii* D9, however, carrying the nitrogenase and nitrite/nitrate uptake operons under variation, as observed by the low coverage and high impact variants found for *nifB* and *nifS*, and the non-synonymous variants for the *nrtA* and *narB* genes.

More studies and a deeper understanding of this population, particularly regarding the discussed genes, are necessary and carry the potential to unveil not only a potential south-american *Cylindrospermopsis* population, but a complex interaction network, with a broad, highly conserved backbone genome within the Nostocales order, and a fine scale mosaic dynamics between strains (Steffen et al., 2014; Desai & Walczak, 2015). Some theories are quite interesting regarding the lack of the CYN cluster, presence of the SXT cluster or even

the presence of both, as described by Piccini et al., (2013). The Black Queen Hypothesis (Morris et al., 2012) proposes that a member of a population can lose one or more genes by requesting the product from another producing member, creating a dependency on other community members regarding function and thus demanding a certain degree of community stability. Furthermore, the appearance of an adaptive mutant or a recombinant by periodic selection events could select those ecotypes best adapted to a given environment (Cohan, 2005; Shapiro et al., 2012). In addition, the 'Kill the Winner' model states that those members that are more successful competitors, in terms of resource exploitation and growth rate, are subject to increased predation pressure by highly host-specific phage (Winter et al., 2010; Cordero & Polz, 2014), generating a top-down selective maintenance of intra-specific diversity.

In short, we started from a broad overview of the community and ended with a complex population in which several strains, but none exactly, could be assigned. In broad scale, the samples presented distinct taxonomical and functional profiles, but when looking at the nostocaceae intraspecific scale, differences almost disappear between the samples. The *Cylindrospermopsis/Raphidiopsis* population from the Pampulha lake seems to be stable, since its variant pattern revealed to be very similar between samples. The small differences observed between the samples on the T3 variant call may be explained by the lower depth of coverage in sample P3M, when *Microcystis* population presented higher contribution. The alignments to *Anabaena* and *Nostoc* genera point to a mosaic-like population structure, where a broad backbone genome is at work and encompasses different species, thus suggesting that, after much concern in trying to understand *who* is there, there is still much to be done to identify who *else*.

4. ACKNOWLEDGEMENTS

This research was supported by a scholarship from CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) to M.L and by funds provided by FAPEMIG (Fundação de Apoio a Pesquisa de Minas Gerais) and by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) to A.G. We thank the staff of the Phycology Laboratory (Dept. Botany/UFMG), for performing all laboratory analyses and providing the entire dataset of this work. We thank the ShapiroLab team, under supervision of Dr. Jesse B. Shapiro, where M.L stayed for a 4 months period, with financial support of the "Science Without Border" Program, and learned several bioinformatic techniques.

4. REFERENCES

Antonova E.S., Hammer B.J. (2011) Quorum-sensing autoinducer molecules produced by members of a multispecies biofilm promote horizontal gene transfer to Vibrio cholera. *FEMS Microbiol. Let*., 322(1):68-76

Antunes J.T., Leão P.N., Vasconcelos V.M. (2015) *Cylindrospermopsis raciborskii*: review of the distribution, phylogeography, and ecophysiology of a global invasive species. *Front. Microbiol.*, 6

Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., Sherlock G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.*, 25(1):25-9

Bagatini I.L., Eiler A., Bertilsson S., Klaveness D., Tessarolli L.P., Vieira A.A.H. (2014) Host-Specificity and Dynamics in Bacterial Communities Associated with Bloom-Forming Freshwater Phytoplankton. *PLoS ONE*, 9(1): e85950

Beck C., Knoop H., Axmann I.M., Steuer R. (2012) The diversity of cyanobacterial metabolism: genome analysis of multiple phototrophic microorganisms. *BMC Genomics*, 13:56

Bendall M.L., Stevens S.L.R., Chan L.K., Malfatti S., Schwientek P., Tremblay J., Schackwitz W., Martin J., Pati A., Bushnell B., Froula J., Kang D., Tringe S. G., Bertilsson S., Moran M. A., Shade A., Newton R.J., McMahon K.D., Malmstrom R.R. (2016) Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J.*, 1–13

Bengtsson-Palme J., Hartmann M., Eriksson K.M., Pal C., Thorell K., Larsson D.G.J., Nilsson R.H. (2015) Metaxa2: Improved Identification and Taxonomic Classification of Small and Large Subunit rRNA in Metagenomic Data. *Mol. Ecol*., 15(6):1403-1414

Biller S.J., Berube P.M., Lindell D., Chisholm S.W. (2015) *Prochlorococcus*: the structure and function of collective diversity. *Nat. Rev. Microbiol.*, 13:13-27

Bittencourt-Oliveira M.C., Piccin-Santos V., Kujbida P., Moura A.N. (2011)a Cylindrospermopsin in Water Supply Reservoirs in Brazil Determined by Immunochemical and Molecular Methods. *J. Water Resource and Protection*, 3:349-355

Bittencourt-Oliveira M.C., Moura A.N., Hereman T.C., Dantas E.W. (2011)b Increase in Straight and Coiled *Cylindrospermopsis raciborskii* (Cyanobacteria) Populations under Conditions of Thermal De-Stratification in a Shallow Tropical Reservoir. *J. Water Resource and Protection*, 3:245-252

Bittencourt-Oliveira M.C., Buch B., Hereman T.C., Arruda-Neto J.D.D.T., Moura A.N., Zocchi S.S. (2012)a Effects of light intensity and temperature on *Cylindrospermopsis raciborskii* (Cyanobacteria) with straight and coiled trichomes: growth rate and morphology. *Brazil. J. Biol.*, 72(2):343-351

Bittencourt-Oliveira M.C., Dias S.N., Moura A.N., Cordeiro-Araújo M.K., Dantas E.W. (2012)b Seasonal dynamics of cyanobacteria in a eutrophic reservoir (Arcoverde) in a semi-arid region of Brazil. *Braz. J. Biol.*, 72(3):533-544

Bolger A.M., Lohse M., Usadel B. (2014) Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170

Boon E., Meehan C.J., Whidden C., Wong D.H.J., Langille M.G.I., Beiko R.G. (2014) Interactions in the

microbiome: communities of organisms and communities of genes. *FEMS Microbiol Rev*, 38:90-118

Bouvy M., Nascimento S.M., Molica R.J.R., Ferreira A., Huszar V., Azevedo S.M.F.O. (2003) Limnological features in Tapacura reservoir (northeast Brazil) during a severe drought. *Hydrobiologia*, 493:115-130

Caporaso J.G., Kuczynski J., Stombaugh J., Bittinger K., Bushman F.D., Costello E. K., Huttley G.A. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, 7(5): 335-336

Caro-Quintero A., Konstantinidis K.T. (2012) Bacterial species may exist, metagenomics reveal. *Env. Microbiol.*, 14(2):347-355

Cingolani P., Platts A., Wang le L., Coon M., Nguyen T., Wang L., Land S.J., Lu X., Ruden D.M. (2012)a A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, 6(2):80-92

Cingolani P., Patel V.M., Coon M., Nguyen T., Land S.J., Ruden D.M., Lu X. (2012)b Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. In: *Toxicogenomics in non-mammalian species*, 3: 35. Frontiers E-books

Cirés S., Wörmer L., Ballot A., Agha R., Wiedner C., Velázquez D., Quesada, A. (2014) Phylogeography of cylindrospermopsin and paralytic shellfish toxin-producing Nostocales cyanobacteria from Mediterranean Europe (Spain). *Appl. Env. Microbiol.*, 80(4):1359-1370

Cohan F.M. (2005) Periodic selection and ecological diversity in bacteria. In Selective sweep (pp. 78-93). Springer US.

Coleman M.L., Sullivan M.B., Martiny A.C., Steglich C., Barry K., DeLong E.F., Chisholm S.W. (2006) Genomic islands and the ecology and evolution of Prochlorococcus. *Science*, 311:1768–1770

Coleman M.L., Chisholm S.W. (2010) Ecosystem-specific selection pressures revealed through comparative population genomics. *PNAS*, 107(43):18634-18639

Cordero O.X., Polz M.F. (2014) Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Rev. Microbiol.*, 12:263-273

Costa I.A.S., Azevedo S.M.F., Senna P.A.C., Bernardo R.R., Costa S.M., Chellappa N.T. (2006) Occurrence of toxin-producing cyanobacteria blooms in a Brazilian semiarid reservoir. *Braz. J. Biol.*, 66(1B):211-219

Crossetti L.O., Bicudo C.E.M. (2008) Adaptations in Phytoplankton Life Strategies to Composed Change in a Shallow Urban Tropical Eutrophic Reservoir, Garças Reservoir, over 8 Years. *Hydrobiologia*, 614:91-105

DePristo M., Banks E., Poplin R., Garimella K., Maguire J., Hartl C., Philippakis A., del Angel G., Rivas M.A., Hanna M., McKenna A., Fennell T., Kernytsky A., Sivachenko A., Cibulskis K., Gabriel S., Altshuler D., Daly M. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Gen.*, 43:491-498

Desai M.M., Walczak A.M. (2015) Flexible gene pools - Rapid genetic exchange leads to mosaic genomes in cyanobacterial populations. *Science*, 348:977-978

Eiler A., Bertilsson S. (2004) Composition of freshwater bacterial communities associated with cyanobacterial blooms in four Swedish lakes. *Envir. Microbiol.*, 6(12):1228-1243

Elias L.M., Silva-Stenico M.E., Alvarenga D.O., Rigonato J., Fiore M.F., de Lira S.P. (2015). Molecular and Chemical Analyses of Cyanobacterial Blooms in Tropical Lagoons from Southeast Brazil. *J. Water Resource and Protection*, 7(1):50

Escalas A., Bouvier T., Mouchet M. A., Leprieur F., Bouvier C., Troussellier M., Mouillot D. (2013) A unifying quantitative framework for exploring the multiple facets of microbial biodiversity across diverse scales. *Envir. Microbiol.*, 15:2642-2657

Falconer I.R., Humpage A.R. (2006) Cyanobacterial (blue-green algal) toxins in water supplies: Cylindrospermopsins. *Environ. Toxicol.*, 21:299-304

Figueredo C.C., Giani A. (2001) Seasonal variation in the diversity and species richness of phytoplankton in a tropical eutrophic reservoir. *Hydrobiologia,* 445(1):165-174

Frangeul L., Quillardet, P., Castets A. M., Humbert J. F., Matthijs H. CP., Cortez D., Tolonen A., Zhang CC., Gribaldo S., Kehr J. C., Zilliges Y., Ziemert N., Becker S., Talla E., Latifi A., Billault A., Lepelletier A., Dittmann E., Bouchier C., Tandeau de Marsac N. (2008) Highly plastic genome of *Microcystis aeruginosa* PCC 7806, a ubiquitous toxic freshwater cyanobacterium. *BMC Genomics*, 9:274

Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucl. Acids Res.*, 43(Database issue):D1049–D1056

Gilbert J.A., Dupont C.L. (2011) Microbial Metagenomics: Beyond the Genome. *Annu. Rev. Mar. Sci.*, 3:347-71

Gugger M., Molica R., Le Berre B., Dufour P., Bernard C., Humbert J.F. (2005) Genetic diversity of *Cylindrospermopsis* strains (Cyanobacteria) isolated from four continents. *Appl. Envi. Microbiol.*, 71(2):1097-1100

Haande S., Rohrlack T., Ballot A., Roberg K., Skulberg R., Beck M., Wiedner C. (2008) Genetic characterisation of *Cylindrospermopsis raciborskii* (Nostocales, Cyanobacteria) isolates from Africa and Europe. *Harmful Algae*, 7:692-701

Helling R.B., Vargas C.N., Adams J. (1987) Evolution of Escherichia coli during growth in a constant environment. *Genetics,* 116: 349-358

Hoff-Risseti C., Dörr F.A., Schaker P.D.C., Pinto E., Werner V.R., Fiore M.F. (2013) Cylindrospermopsin and Saxitoxin Synthetase Genes in *Cylindrospermopsis raciborskii* Strains from Brazilian Freshwater. *PLoS ONE*, 8(8): e74238

Hug L.A., Beiko R.G., Rowe A.R., Richardson R.E., Edwards E.A. (2012) Comparative metagenomics of three Dehalococcoides containing enrichment cultures: the role of the non-dechlorinating community. *BMC Genomics*, 13:327

Hughes J.B., Hellmann J.J., Ricketts T.H., Bohannan B.J. (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl. Env. Microbiol.*, 67(10):4399-4406

Humbert J.F., Barbe V., Latifi A., Gugger M., Calteau A., Coursin T., Lajus A., Castelli V., Oztas S., Samson G., Longin C., Medigue C., Tandeau de Marsac N. (2013) A Tribute to Disorder in the Genome of the Bloom-Forming Freshwater Cyanobacterium *Microcystis aeruginosa*. *PLoS ONE*, 8(8): e70747

Kaneko T., Nakajima N., Okamoto S., Suzuki I., Tanabe Y., Tamaoki M., Nakamura Y., Kasai F., Watanabe A., Kawashima K., Kishida Y., Ono A., Shimizu Y., Takahashi C., Minami C., Fujishiro T., Kohara M., Katoh M., Nakazaki N., Nakayama S., Yamada M., Tabata S., Watanabe M. M. (2007) Complete Genomic Structure of the Bloom-forming Toxic Cyanobacterium *Microcystis aeruginosa* NIES-843. *DNA Research*, 14:247-256

Kashtan N., Roggensack S.E., Rodrigue S., Thompson J.W., Biller S.J., Coe A., Ding H., Marttinen P., Malmstrom R.R., Stocker R., Follows M.J., Stepanauskas R., Chisholm S.W. (2014) Single-Cell Genomics

Reveals Hundreds of Coexisting Subpopulations in Wild *Prochlorococcus*. Science, 344:416-420

Kellmann R., Mihali T.K., Jeon Y.J., Pickford R., Pomati F., Neilan B.A. (2008) Biosynthetic intermediate analysis and functional homology reveal a saxitoxin gene cluster in cyanobacteria. *Appl. Envir. Microbiol.*, 74(13):4044-4053

Kurmayer R., Christiansen G., Chorus, I. (2003) The abundance of microcystin-producing genotypes correlates positively with colony size in Microcystis sp. and determines its microcystin net production in Lake Wannsee. *Appl. Environ. Microbiol.*, 69:787-795

Kurtz S., Phillippy A., Delcher A.L., Smoot M., Shumway M., Antonescu C., Salzberg S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, 5(2):R12

Lagos N., Onodera H., Zagatto P.A., Andrinolo D., Azevedo S., Oshima Y. (1999) The first evidence of paralytic shellfish toxins in the freshwater cyanobacterium *Cylindrospermopsis raciborskii*, isolated fro m Brazil. *Toxicon.*, 37:1359-1373

Leibold M.A., Holyoak M., Mouquet N., Amarasekare P., Chase J.M., Hoopes M.F., Holt R.D., Shurin J.B., Law R., Tilman D., Loreau M., Gonzalez A. (2004) The metacommunity concept: a frameworkfor multi-scale community ecology. *Ecol. Let*., 7: 601-613

Li H., Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754-1760

Li H., Xing P., Chen M., Bian Y., Wu Q. L. (2011) Short-term bacterial community composition dynamics in response to accumulation and breakdown of *Microcystis* blooms. *Water Res.*, 45(4):1702-1710

Li R., Carmichael W.W., Brittain S., Eaglesham G.K., Shaw G.R., Liu Y., Watanabe M.M. (2001) First report of the cyanotoxins cylindrospermopsin and deoxycylindrospermopsin from *Raphidiopsis curvata* (Cyanobacteria). *J. phycol.*, 37:1121-1126

Liu Q., Guo Y., Li J., Long J., Zhang B., Shyr Y. (2012) Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. *BMC genomics*, 13(8):1

Lopes A.M.M.B. (2013) Composição da comunidade de cianobactérias e outros grupos microbianos em dois reservatórios tropicais. PhD thesis, Universidade Federal de Minas Gerais, 133 p.

Lopes A.M.M.B., Figueredo C.C., Giani A. Species seasonal changes during persistent cyanobacterial blooms: A molecular, morphological and ecological characterization. *Hydrobiologia* (in review)

López-Pérez M., Gonzaga A., Rodriguez-Valera F. (2013) Genomic Diversity of "Deep Ecotype" *Alteromonas macleodii* Isolates: Evidence for Pan-Mediterranean Clonal Frames. *Genome Biol. Evol.*, 5(6):1220-1232

Luo C., Knight R., Siljander H., Knip M., Xavier R.J., Gevers D. (2015) Constrains identifies microbial strains in metagenomic datasets. Nat. Biotechnol., 33(10):1045-1052

Magoč T., Salzberg S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957-2963

Marinho M.M., Souza M.B.G., Lürling M. (2013) Light and phosphate competition between *Cylindrospermopsis raciborskii* and *Microcystis aeruginosa* is strain dependent. *Microb. Ecol.*, 66(3):479-488

McInerney J.O., Pisani D., Bapteste E., O'Connell M.J. (2011) The Public Goods Hypothesis for the evolution of life on Earth. *Biol. Direct.*, 6:41

McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., Altshuler D., Gabriel S., Daly M., DePristo M.A. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20:1297-1303

McMurdie P.J., Holmes S. (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE.*, 8(4):e61217

Mello M.M.E., Soares M.C.S., Roland F., Lürling M. (2012) Growth inhibition and colony formation in the cyanobacterium *Microcystis aeruginosa* induced by the cyanobacterium *Cylindrospermopsis raciborskii*. *J. Plank. Res.*, fbs056

Meyer F., Paarmann D., D'Souza M., Olson R., Glass E.M., Kubal M., Paczian T., Rodriguez A., Stevens R., Wilke A., Wilkening J., Edwards R.A. (2008) The Metagenomics RAST server — A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9:386

Moore L.R., Rocap G., Chisholm S.W. (1998) Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature*, 393:464

Morris J. J., Lenski R. E., Zinser E. R. (2012) The Black Queen hypothesis: evolution of dependencies through adaptive gene loss. *Am.Soc. Microbiol*., 3(2):e00036-12

Moura A.D.N., Bittencourt-Oliveira M.D.C., Chia M.A., Severiano J.S. (2015) Co-occurrence of *Cylindrospermopsis raciborskii* (Woloszynska) Seenaya & Subba Raju and *Microcystis panniformis* Komárek et al. in Mundaú reservoir, a semiarid Brazilian ecosystem. *Acta Limnol. Brasil.*, 27(3):322-329

Moustaka-Gouni M., Kormas K.A., Vardaka E., Katsiapi M., Gkelis S. (2009) *Raphidiopsis mediterranea* Skuja represents non-heterocytous life-cycle stages of *Cylindrospermopsis raciborskii* (Woloszynska) Seenayya et Subba Raju in Lake Kastoria (Greece), its type locality: Evidence by morphological and phylogenetic analysis. *Harmful Algae*, 8:864-872

Namikoshi M., Murakami T., Watanabe M.F., Oda T., Yamada J., Tsujimura S., Nagai H., Oishi S. (2003) Simultaneous production of homoanatoxin-a, anatoxin-a, and a new non-toxic 4-hydroxyhomoanatoxin-a by the cyanobacterium *Raphidiopsis mediterranea* Skuja. *Toxicon.*, 42:533-538

Nielsen R., Paul J.S., Albrechtsen A., Song Y.S. (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Gen.*, 12:443-451

Nielsen R., Korneliussen T., Albrechtsen A., Li Y., Wang J. (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PloS one*, 7(7):e37558

Oh S., Caro-Quintero A., Tsementzi D., DeLeon-Rodriguez N., Luo C., Poretsky R., Konstantinidis K.T. (2011) Metagenomic Insights into Evolution, Function, and Complexity of the Planktonic Microbial Community of Lake Lanier, a Temperate Freshwater Ecosystem. *Appl. Environm. Microbiol.*, 77(17):6000-6011

Olson N.D., Lund S.P., Colman R.E., Foster J.T., Sahl J.W., Schupp J.M., Keim P., Morrow J.B., Salit M.L., Zook J.M. (2015) Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front. Genet*., 6:235

Padisák J. (1997) *Cylindrospermopsis raciborskii* (Woloszynska) Seenayya et Subba Raju, an expanding, highly adaptive cyanobacterium: worldwide distribution and review of its ecology. *Arch. Hydrobiol.*, 107(4):563-593

Padmavathi P., Veeraiah K. (2009) Studies on the influence of *Microcystis aeruginosa* on the ecology and fish

production of carp culture ponds. *Afr. J. Biotechnol.*, 8(9):1911-1918

Penn K., Wang J., Fernando S.C., Thompson J.R. (2014) Secondary metabolite gene expression and interplay of bacterial functions in a tropical freshwater cyanobacterial bloom. *ISME J.*, 8:1866–1878

Piccini C., Aubriot L., Fabre A., Amaral V., González-Piana M., Giani A., Figueredo C.C., Vidal L., Kruk C., Bonilla S. (2011) Genetic and eco-physiological differences of South American *Cylindrospermopsis raciborskii* isolates support the hypothesis of multiple ecotypes. *Harmful Algae*, 10:644-653

Piccini C., Aubriot L., D'Alessandro B., Martigani F., Bonilla S. (2013) Revealing Toxin Signatures in Cyanobacteria: Report Genes Involved in Cylindrospermopsin Synthesis from Saxitoxin-Producing *Cylindrospermopsis raciborskii*. *Adv. Microbiol.*, 3:289-296

Pomati F., Moffitt M.C., Cavaliere R., Neilan B.A. (2004) Evidence for differences in the metabolism of saxitoxin and C1+ 2 toxins in the freshwater cyanobacterium *Cylindrospermopsis raciborskii* T3. *Biochimica et Biophysica Acta*, 1674(1):60-67

Quast C., Pruesse E., Yilmaz P., Gerken J., Schweer T., Yarza P., Peplies J., Glöckner F.O. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucl. Acids Res.*, 41 (D1):D590-D596

Rosen M.J., Davison M., Bhaya D., Fisher D.S. (2015) Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science*, 348: 1019-1023

Rosenthal E., Blue E., Jarvik G.P. (2015) Next-generation gene discovery for variants of large impact on lipid traits. *Curr. Opinion Lipidol.*, 26(2):114-119

Rzymski P., Poniedziałek B., Kokociński M., Jurczak T., Lipski D., Wiktorowicz K. (2014) Interspecific allelopathy in cyanobacteria: Cylindrospermopsin and *Cylindrospermopsis raciborskii* effect on the growth and metabolism of *Microcystis aeruginosa*. *Harmful Algae*, 35:1-8

Saker M.L., Nogueira I.C., Vasconcelos V.M., Neilan B.A., Eaglesham G.K., Pereira P. (2003) First report and toxicological assessment of the cyanobacterium *Cylindrospermopsis raciborskii* from Portuguese freshwaters. *Ecotox. Envir. Safety*, 55(2):243-250

Sant'anna C.L., Azevedo M.T.P., Werner V.R., Dogo C.R., Rios F.R., Carvalho L.R. (2008) Review of toxic species of Cyanobacteria in Brazil. *Algol. Stud.*, 126:251-265

Shapiro B.J., Friedman J., Cordero O.X., Preheim S.P., Timberlake S.C., Szabó G., Polz M.F., Alm E.J. (2012) Population genomics of early events in the ecological differentiation of bacteria. *Science*, 336:48–51

Shapiro, B.J., Polz M.F. (2014) Ordering microbial diversity into ecologically and genetically cohesive units. *Trends in Microbiology*, 22(5):235-247

Shi L., Cai Y., Wang X., Li P., Yu Y., Kong F. (2010) Community Structure of Bacteria Associated with *Microcystis* Colonies from Cyanobacterial Blooms. *J. Fresh. Ecol*., 25(2):193-203

Simm S., Keller M., Selymesi M., Schleiff E. (2015) The composition of the global and feature specific cyanobacterial core-genomes. *Front. Microbiol*., 6:219

Sinha R., Pearson L.A., Davis T.W., Muenchhoff J., Pratama R., Jex A., Burford M.A., Neilan B.A. (2014) Comparative genomics of *Cylindrospermopsis raciborskii* strains with differential toxicities. *BMC Genomics*, 15:83

Steffen M.M., Li Z., Effler T.C., Hauser L.J., Boyer G.L., Wilhelm S.W. (2012) Comparative Metagenomics of Toxic Freshwater Cyanobacteria Bloom Communities on Two Continents. *PLoS ONE*, 7(8): e44002

Steffen M.M., Dearth S.P., Dill B.D., Li Z., Larsen K.M., Campagna S.R., Wilhelm S.W. (2014) Nutrients drive transcriptional changes that maintain metabolic homeostasis but alter genome architecture in Microcystis. *The ISME Journal*, 8:2080-2092

Stucken K., John U., Cembella A., Murillo A.A., Soto-Liebe K., Fuentes-Valdés J.J., Friedel M., Plominsky A.M., Vásquez M., Glöckner G. (2010) The Smallest Known Genomes of Multicellular and Toxic Cyanobacteria: Comparison, Minimal Gene Sets for Linked Traits and the Evolutionary Implications. *PLoS ONE*, 5(2):e9235

Suenaga H. (2012) Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities. *Env. Microbiol.*, 14(1):13-22

Sukenik A., Hadas O., Kaplan A., Quesada A. (2012) Invasion of Nostocales (cyanobacteria) to subtropical and temperate freshwater lakes–physiological, regional, and global driving forces. *Front. Microbiol.*, 3:86

Tucci A., Sant'Anna C.L. (2003) *Cylindrospermopsis raciborskii* (Woloszynska) Seenayya & Subba Raju (Cyanobacteria): weekly variation and relation with environmental factors in an eutrophic lake, São Paulo, SP, Brazil. *Rev. Bras. Bot.*, 26(1):97-112

Van der Auwera G.A., Carneiro M., Hartl C., Poplin R., del Angel G., Levy-Moonshine A., Jordan T., Shakir K., Roazen D., Thibault J., Banks E., Garimella K., Altshuler D., Gabriel S., DePristo M. (2013) From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr. Protoc. Bioinformatics*, 43:11.10.1-11.10.33

Whitaker R.J., Grogan, D.W., Taylor J.W. (2005) Recombination Shapes the Natural Population Structure of the Hyperthermophilic Archaeon Sulfolobus islandicus. Mol. Biol. Evol., 22(12):2354-2361

Wiedenbeck J., Cohan F.M. (2011) Origins of bacterial diversity through horizontal genetic transferand adaptation to new ecological niches. *FEMS Microbiol. Rev*., 35:957-976

Winter C., Bouvier T., Weinbauer M.G., Thingstad T.F. (2010) Trade-offs between competition and defense specialists among unicellular planktonic organisms: the "killing the winner" hypothesis revisited. *Microbiol. Mol. Biol. Rev*., 74:42-57

Wintermute E.H., Silver P.A. (2010) Dynamics in the mixed microbial concourse. *Genes Dev*., 24:2603-2614

Wu Z., Shi J., Li R. (2009) Comparative studies on photosynthesis and phosphate metabolism of *Cylindrospermopsis raciborskii* with *Microcystis aeruginosa* and *Aphanizomenon flos-aquae*. *Harmful Algae*, 8(6):910-915

Xie L., Rediske R.R., Hong Y., O'Keefe J., Gillett N.D., Dyble J., Steinman A.D. (2012) The role of environmental parameters in the structure of phytoplankton assemblages and cyanobacteria toxins in two hypereutrophic lakes. *Hydrobiologia*, 691:255-268

Yunes J.S., De La Rocha S., Giroldo D., Silveira S.B., Comin R., Bicho M.S., Melcher S.S., Sant'Anna C.L., Vieira A.A.H. (2009) Release of carbohydrates and proteins by a subtropical strain of Raphidiopsis brookii (cyanobacteria) able to produce saxitoxin at three nitrate concentrations. *J. Phycol.*, 45:585-591

Zagatto P.A. (1998) Toxicity of the algae Cylindrospermopsis raciborskii isolated from Billings reservoir, Taquacetuba branch Technical Report, September 1998 Environmental Agency of São Paulo State (CETESB), São Paulo, SP, pp 23.

Zhu M., Xu Y., Li R. (2012) Genetic Diversity of Bloom-Forming Microcystis (Cyanobacteria) Populations in a Hyper-Eutrophic Pond in Central China. *Curr. Microbiol*., 65(3):219-224

CONCLUSÃO GERAL

As avaliações populacionais da diversidade genética da comunidade planctônica dos reservatórios de Volta Grande, Marimbondo e Pampulha revelaram uma alta diversidade genética de cianobactérias, observada através de diversas abordagens e escalas. Através da abordagem metagenômica, especialmente o sequenciamento randômico de amostras ambientais (*Whole-Genome Shotgun* - WGS), foi possível integrar ecologia e genômica, alcançando desde uma escala ampla de análise até a escala fina, a nível de nucleotídeos. Mesmo apresentando uma baixa cobertura de fragmentos de DNA, uma alta diversidade intraespecífica foi observada no metagenoma do reservatório de Volta Grande. Da mesma forma, apesar da baixa cobertura de fragmentos de DNA das amostras individuais do reservatório de Marimbondo, observamos a correlação do gradiente ambiental ao perfil taxonômico molecular. Apesar das mudanças na contribuição das diversas classes da comunidade como um todo, observamos um perfil funcional estável ao longo do período de amostragem. Análises mais específicas e dados mais robustos são necessários para alcançar a verdadeira dinâmica funcional no reservatório de Marimbondo. A intrincada rede de alinhamentos observada no metagenoma do reservatório da Pampulha não apenas sugere a presença de uma população adaptada às condições ambientais locais, como corrobora a alta diversidade intraespecífica observada em todos os conjuntos de dados analisados, enfatizando o potencial da análise de comunidades sob a perspectiva WGS em reservatórios de água doce.

Com um mesmo conjunto de dados, múltiplas técnicas e escalas de análise podem ser adotadas quando trabalhamos com amostras de WGS. Através dos quatro capítulos, apresentamos uma notável microdiversidade em relação às populações dominantes de três reservatórios. Além disso, várias evidências de uma constante manutenção de diversidade à nível de cepas foram apresentadas, apontando regiões hipervariáveis, passíveis de rastreamento entre as diversas cepas de uma mesma população, e que, junto aos SNVs, podem revelar conexões dentro da estrutura genômica (*genome linkage*) relacionadas à dinâmica de rearranjo e recombinação, abrindo caminho para o verdadeiro mecanismo de resposta das populações planctônicas às pressões ambientais, principalmente relacionadas ao impacto antrópico.

Concluímos, por fim, que perfis metagenômicos são o primeiro passo para alcançar correlações rigorosas entre a comunidade planctônica e as variáveis ambientais. Através

desses perfis, é possível planejar vias de análise compatíveis com a robustez do conjunto de dados e diretamente relacionadas à hipótese de trabalho. A mudança de perpectiva que vem acontecendo nos últimos anos ressalta o conteúdo gênico como o protagonista de toda a dinâmica de populações e comunidades, e diversas teorias têm tentado explicar a multiplicidade de cepas e a estabilidade funcional de comunidades. Nossos dados mostraram que microdiversidade é um atributo de comunidades de cianobactérias, que essas diversas cepas se sobrepõem em um mosaico de genes que se variam de acordo com as pressões impostas pelo ambiente, constantemente sob processo adaptativo através de variações à nível de nucleotídeos, genes, clusteres de genes ou regiões genômicas amplas. A adoção de modelos preditivos e monitoramento de sistemas aquáticos utilizando a abordagem metagenômica constituiria um grande passo em direção ao equilíbrio de ecossistemas e sustentabilidade, permitindo um nível de controle e precisão jamais alcançado até então.